大数据智能下数据脱敏的思考

王红凯[1],龚小刚[1],叶卫[1],陈超[1],马新强[2,3*],姚进强[4],刘勇[2*]

1. 国网浙江省电力有限公司信息通信分公司,杭州 310027
2. 浙江大学智能系统与控制研究所,杭州 310027
3. 重庆文理学院人工智能学院,重庆 402160
4. 浙江省交通集团检测科技有限公司,杭州 310030

摘要　随着人工智能技术在大数据与高性能计算的推动下飞速发展,涌现出大量创新性的方法,对现有的数据安全与脱敏方法带来了诸多挑战。从当今大数据与智能技术发展的现状入手,讨论了数据脱敏技术的内涵、工作流程、系统机制、典型脱敏案例,展望了未来数据脱敏技术的发展趋势,从技术、需求和法治、管理机制方面提出了一些数据脱敏的建议。

关键词　大数据;人工智能;数据脱敏;信息安全

随着大数据时代的到来,海量数据中蕴藏的巨大价值得以挖掘,同时也带来了隐私信息与关键性敏感数据保护方面的困难。如何在实现数据高效共享应用的同时,保护敏感信息不被泄露,已成为数据安全智能开发的关键环节。目前,在大数据背景下,面对复杂多源海量异构环境,要知道敏感信息在哪,并且知道哪些数据参考或应用了这些敏感数据是非常困难的。如何解决这些必须解决的问题,更值得深度思考。

当前国内外数据安全解决方案提供商对脱敏技术研究主要集中在以下两点[1-4]:一是敏感数据智能准确识别技术;二是数据脱敏算法与规则的优化。目前国内外还没有一家数据安全厂家能提供成熟完善的脱敏解决方案。Gartner 认为,数据脱敏应成为相关企业在软件开发、数据分析和培训时的强制选项。目前,国外数据脱敏的主要实践者包括 IBM、ORACLE 和 Informatica,它们凭借在传统数据库行业较早的进入时间、较深厚的实践经验和

收稿日期:2019-10-13;修回日期:2019-12-30

基金项目:浙江大学工业控制技术国家重点实验室开放课题(ICT170330,ICT1800413,ICT1900358);广东省重点领域研发计划项目(2019B010120001);重庆市发改委重大产业技术研发项目(2018148208);浙江省重点研发计划项目(2019C01004)

作者简介:王红凯,高级工程师,研究方向为网络与信息安全,电子信箱:whkzju@163.com;马新强(通信作者),教授,研究方向为大数据智能化与信息安全,电子信箱:xinqma@zju.edu.cn;刘勇(共同通信作者),教授,研究方向为大数据、人工智能与信息安全,电子信箱:yongliu@iipc.zju.edu.cn

引用格式:王红凯, 龚小刚, 叶卫, 等. 大数据智能下数据脱敏的思考[J]. 科技导报, 2020, 38(3): 115-122; doi: 10.3981/j.issn.1000-7857.2020.03.008

技术积累,占据了多数市场份额。目前国内数据脱敏的研究和应用尚处在起步阶段,通信运营商、能源(电力)、银行、大型国企根据自身需求制订了一些数据脱敏解决方案[5-7],但目前多以静态脱敏为主,设计流程不够灵活多变,工具能力有限,专用性较强,配置规则不够简单易用,维护相对困难,不能满足数据交互流量的不断增长和复杂多变的安全处理需求。对于动态脱敏,中国的数据安全厂家也在积极探索中,在能源和银行业也有实际应用案例[8-9],但目前还无法以统一的商用模式应用在各行各业,当前的动态脱敏还是具有较强的行业定制化特性。

传统的数据脱敏系统往往需要较多的人工干预,在新任务、新环境下用户的初始配置工作较大[7,10-12],并且受限于传统数据脱敏技术[13]和系统内置算法的繁杂程度[14-15],用户往往需要一定的业务和技术基础,这会导致传统数据脱敏系统的入门成本过高。尤其是在大数据智能化时代,亟需将数据安全脱敏技术与人工智能的自主学习和强大的数据分析能力相结合[16-19],在不需过多人工干预的情况下,显著加强数据脱敏系统的可靠性和易用性,在不降低脱敏安全性的基础上,实现易学习、免配置、自动脱敏和自适应脱敏算法等功能,以大幅度降低企业部署数据脱敏系统的成本。

1 相关概念及内涵

在讨论大数据时代的敏感数据智能识别及分析过程中,应明确以下几个相关概念及关系。

1.1 数据脱敏

数据脱敏就是在保存数据原始特征的同时改变其数值,从而保护敏感数据免于未经授权而被访问,同时又可以进行相关的数据处理,可以在保留数据意义和有效性的同时保持数据的安全性,并遵从数据隐私规范。借助数据脱敏,信息依旧可以被使用,并与业务相关联,不会违反相关规定,而且也避免了数据泄露的风险。首先就是如何识别敏感数据,确定敏感数据的定义,以及研究敏感数据的依赖。应用程序是十分复杂并且完整的。知道敏感信息在哪,并且知道哪些数据参考或应用了这些敏感数据是非常困难的。

静态脱敏一般用于非生产环境。在不能将敏感数据存储于非生产环境的场合中,通过脱敏程序转换生产数据,使数据内容与数据间的关联能满足测试与开发问题排查的需要,同时进行数据分析、数据挖掘等分析活动。而动态脱敏通常用于生产环境,在敏感数据被赋权个体访问时才对其进行脱敏,并能够根据策略执行相应的脱敏方法。静态脱敏与动态脱敏的区别在于是否在使用敏感数据时才进行实时的脱敏。

据统计,当前全球数据脱敏主要投入实际应用的技术有三大类:数据失真技术、数据加密技术和匿名化限制发布技术(包括:k-匿名、L-多样性、数据抑制、数据扰动和差分隐私)[3,8]。

1.2 传统数据脱敏系统的工作流程

传统数据脱敏系统的工作流程,一般如图1所示。

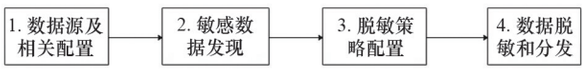


图1 传统数据脱敏系统工作流程

其中对于动态脱敏系统,其数据源相关配置一般由管理员用户预先配置完成,无须普通用户涉及;而静态脱敏系统,则须根据每次任务的特点和要求进行按需配置,此过程因受业务需求所限,一般来说难以实现自动化或免配置。

而敏感数据的自动识别一般分为两部分内容:(1)敏感数据的识别;(2)敏感数据在不同表格、不同字段之间的关联关系的识别。

传统数据脱敏系统的敏感数据发现和关联关系识别,一般都是通过人工配置和正则表达式匹配来实现的,如图2所示。

而其识别的准确程度主要取决于正则表达式的规则设置是否精准合理。但总体来说,传统的正

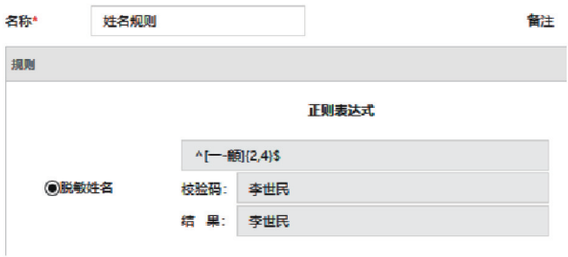


图2　规则式脱敏系统示意

则表达式匹配敏感数据的精准度和性能往往不可兼得:正则匹配规则设置越简单则系统性能越好,但简单规则会导致识别率相对较差;而设置相对复杂的正则规则虽然可以一定程度上提升识别率,但往往对系统性能影响较大,而且对于人工配置的经验和技术要求较高,一般业务人员难以胜任。

对此,亟需通过将人工智能技术应用于敏感数据和关联关系识别阶段,可有效解决传统正则方式性能与准确性不可兼得的痛点,也可省去专业人员对于正则识别规则的维护,真正达到免配置和自动功能的实现,为用户带来价值提升。

2　大数据时代下的数据脱敏系统机制

数据脱敏系统设计提供多种预置算法库,包括映射、随机、散列、和加密共四大类,每一大类中都包含若干算法。系统可以根据用户的业务需求,灵活搭配,使针对敏感数据的脱敏能够满足数据原始属性、可重复性、可逆性、关联性、可追踪性以及准确性的要求。

一般对应的数据脱敏系统原型工作流程逻辑,如图3所示,系统实施步骤可以分解为:(1) 收集用户当前业务系统需求进行数据模型分析,建立系统原型及使用环境;(2) 对用户当前业务系统数据进行智能数据分析,通过机器自学习识别出当前业务系统敏感数据,进行自动聚类,提出敏感数据分类分级规则方案;(3) 对智能发现的敏感数据进行关联关系分析以及数据变形规则分析,同时对系统进行分析;(4) 根据第3步的分析,对用户权限、脱敏规则、脱敏数据表、脱敏函数、脱敏数据关联关系及脱敏流程进行配置,导入到系统原型;(5) 对敏感数据进行数据的变形工作,包括脱敏函数库的自

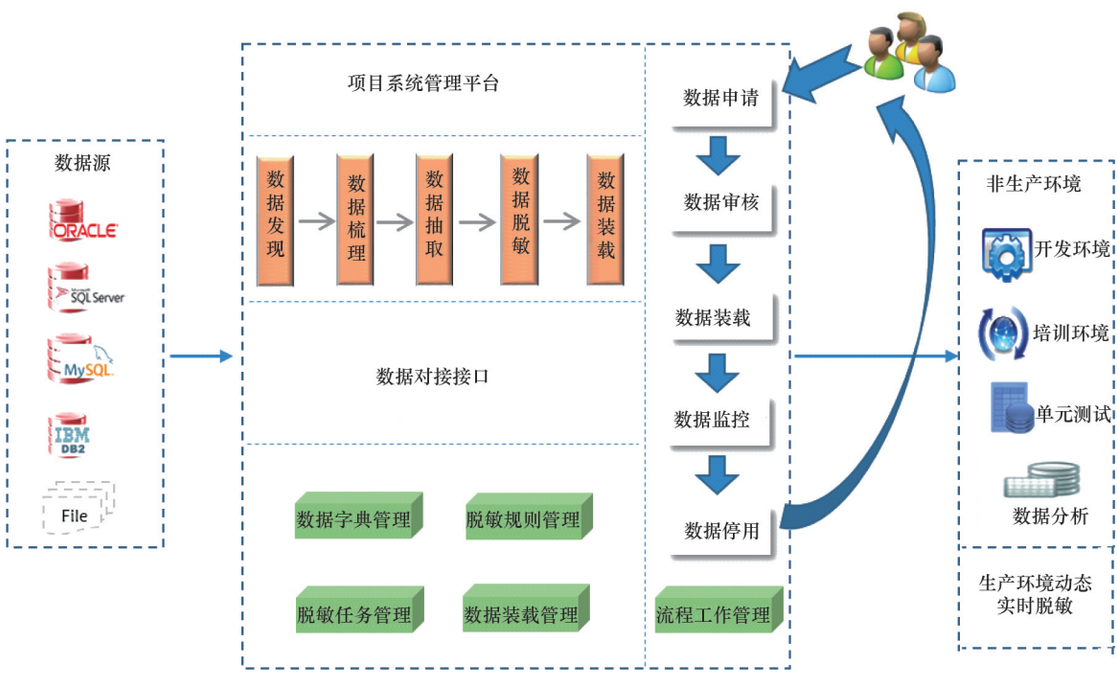


图3　数据脱敏系统工作逻辑流程图

定义、扩展及丰富;(6) 进行脱敏数据的分发,包括数据加载到其他库、数据加载到本地库、数据在线脱敏使用;(7) 通过脱敏后的结果对系统原型支撑技术进行验证。

3 大数据智能化背景下的数据脱敏技术分析

3.1 基于人工智能的敏感数据自动分类和识别

敏感数据识别是实现数据脱敏的重要前提条件,例如采用模式识别的方法实现脱敏信息的自动识别[9,20]。针对不同种类的数据,其敏感特征的检测方法会有所差异。通过训练集获得特征数据后,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,用于敏感数据定级。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数据的特征值进行匹配,当匹配命中时系统自动记录当前敏感数据的敏感级值。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数据进行补充。对部分业务系统样本数据和元数据进行分类训练研究,最后分类建立敏感数据集,从而实现敏感数据准确识别。

3.2 基于机器学习的数据关联关系识别和保持

对于结构化数据,特别是不同数据元素之间关系非常复杂的数据集,往往存在同一数据表中某字段与另外字段有对应关系。而一般来说,数据脱敏前后这种对应关系不应破坏,否则将导致该字段的使用价值不复存在。一般来说,对于需要进行数据统计、需要参考量的情况下,对数据的关联性要求较高。应用机器学习算法对大规模数据之间的关联关系识别和保持进行研究[5,17]。

3.3 基于用户使用模型学习的智能自适应脱敏算法

目前,传统的数据脱敏系统往往需要用户针对所有已识别的敏感数据,逐个进行脱敏策略的配置[20-23]。该方式除了对于运维人员的技术和业务理解程度有一定要求之外,在面临海量数据时,也存在较大的人工操作工作量。若拟在上述人工智能敏感数据自动发现和数据库关联关系识别技术的基础上,需研究基于用户使用模型智能自适应脱敏算法。在应对静态脱敏场景时,通过用户对于脱敏结果的特性进行限定后,由系统自动挑选合适算法进行脱敏,免除了用户逐个字段进行策略配置的问题;而在应对动态脱敏场景时,通过一段时间对于用户配置策略的学习,实现敏感数据的自动策略配置和免配置的功能。从而显著提升了脱敏系统的易用性和可用性,降低了企业相关系统部署和应用成本。

4 典型案例实验分析

一般的脱敏算法主要包含:特征算法和通用算法。以一个典型的运营商应用业务场景中的数据脱敏为例(表1),采用智能自适应脱敏算法在海量运营商用户中随机抽取36万个用户的静态属性表、套餐开通数据表、Top10APP使用数据表,月流量语音使用数据表总共31列的属性中进行智能自适应脱敏。其中ref_id属性已被提前加密,唯一性保证连接4个表。

表1 待脱敏数据

数据表类	属性名字
用户静态属性表	'ref_id','age','gender','province_code', 'city_code','terminal_brand','product_id', 'product_name','join_date', 'work_area','house_area'
套餐开通数据表	'ref_id','pay_first_tag', 'pay_first_fee', 'pay_first_time', 'count_pay_time','count_pay_fee', 'state_code', 'state_change_time'
用户使用行为表	'ref_id','fee','balance','caller_duration', 'called_duration', 'charge_money_total','novoice_days', 'noflow_days', 'arpu','mou','dou','dou_free','is_acct', 'beyond_package'
用户偏好表	'ref_id','top_10_app'

以用户静态属性表为例介绍实验过程和实验结果。如图4(a)所示,是原始为脱敏的数据视图,如图4(b)所示为经过第1轮的迭代脱敏发现,有部

ref_id	age	gender	province_code	city_code	work_area	house_area	terminal_brand	product_id	product_name	join_date
298068	29.0	男	34	354	320812	320804.0	华为	90348805.0	国宝卡	20121209
471392	33.0	女	38	480	350802	350802.0	苹果	90063345.0	腾讯大王卡	20190607
282044	46.0	女	34	343	321112	321102.0	苹果	90363912.0	S10新星粉卡	19991117
271656	31.0	女	34	330	429006	429006	维沃	90063345.0	腾讯大王卡	20171102
266580	51.0	男	34	330	320106	320282	小米	90063345.0	腾讯大王卡	19990201
534260	NULL	男	51	510	440111	440111.0	维沃	90063345.0	腾讯大王卡	20180410
109162532	26.0	女	51	530	440605	440117.0	苹果	90063345.0	腾讯大王卡	20170717
220948	NULL	男	84	841	610112	610112.0	苹果	90350506.0	地王卡	20140407

(a) 脱敏前

ref_id	age	gender	province_code	city_code	work_area	house_area	terminal_brand	product_id	product_name	join_date
*	[28, 32]	男	**	3**	320812	320804	华为	90348805	*	*
*	[33, 37]	女	**	4**	350802	350802	苹果	90063345	*	*
*	[43, 47]	女	**	3**	321112	321102	苹果	90363912	*	*
*	[28, 32]	女	**	3**	429006	429006	维沃	90063345	*	*
*	[48, 52]	男	**	3**	320106	320282	小米	90063345	*	*
*	[23, 27]	男	**	5**	440111	440111	维沃	90063345	*	*
*	[23, 27]	女	**	5**	440605	440117	苹果	90063345	*	*
*	[23, 27]	男	**	8**	610112	610112	苹果	90350506	*	*
*	[23, 27]	男	**	5**	441302	441302	维沃	90155946	*	*

(b) 第1轮算法迭代脱敏后

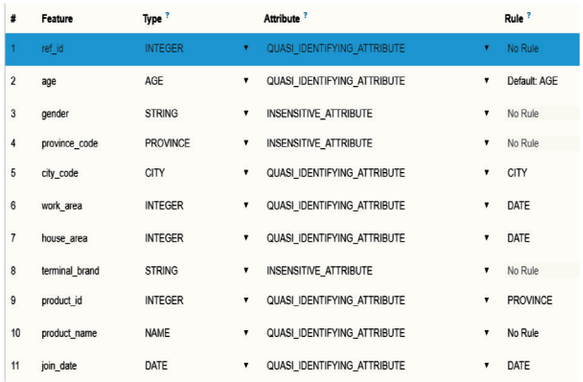
ref_id	age	gender	province_code	city_code	work_area	house_area	terminal_brand	product_id	product_name	join_date
*	*	男	34	3**	3*****	3*****	华为	903488**	国宝卡	********
*	*	女	38	4**	3*****	3*****	苹果	900633**	腾讯大王卡	********
*	*	女	34	*	*	*	苹果	*	*	*
*	*	女	34	3**	4*****	4*****	维沃	900633**	腾讯大王卡	********
*	*	男	34	3**	3*****	3*****	小米	900633**	腾讯大王卡	********
*	*	男	51	5**	4*****	4*****	维沃	900633**	腾讯大王卡	********
*	*	女	51	5**	4*****	4*****	苹果	900633**	腾讯大王卡	********
*	*	男	84	8**	6*****	6*****	苹果	903505**	地王卡	********

(c) 第5轮算法迭代脱敏后

图4　用户静态属性

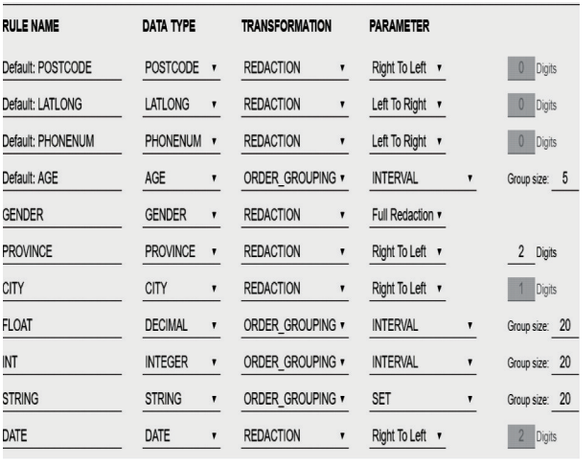
分字段被强制脱敏(province,product_name),有些本该被脱敏的字段没有脱敏(工作和和居住地),经过多轮迭代后,如图4(c)所示,精细到某一列某一个值被单独脱敏,如第3行的product_name,因为使用该product_name的用户较少。图5、图6分别是算法针对每个字段智能选择的脱敏规则以及脱敏规则库。

最后得到智能脱敏算法的脱敏结果,如表2所示,在保证information loss不降低太大的情况下保证了piracy risk保持较低水平。由于用户属性表和用户偏好表比另外2个表含有的敏感字段多,所以为了保证piracy risk下不得不降低information loss,表现出来就比套餐开通数据表和用户使用行为表的information loss高。



#	Feature	Type	Attribute	Rule
1	ref_id	INTEGER	QUASI_IDENTIFYING_ATTRIBUTE	No Rule
2	age	AGE	QUASI_IDENTIFYING_ATTRIBUTE	Default: AGE
3	gender	STRING	INSENSITIVE_ATTRIBUTE	No Rule
4	province_code	PROVINCE	INSENSITIVE_ATTRIBUTE	No Rule
5	city_code	CITY	QUASI_IDENTIFYING_ATTRIBUTE	CITY
6	work_area	INTEGER	QUASI_IDENTIFYING_ATTRIBUTE	DATE
7	house_area	INTEGER	QUASI_IDENTIFYING_ATTRIBUTE	DATE
8	terminal_brand	STRING	INSENSITIVE_ATTRIBUTE	No Rule
9	product_id	INTEGER	QUASI_IDENTIFYING_ATTRIBUTE	PROVINCE
10	product_name	NAME	QUASI_IDENTIFYING_ATTRIBUTE	No Rule
11	join_date	DATE	QUASI_IDENTIFYING_ATTRIBUTE	DATE

图5 静态属性表的智能选取的脱敏规则



RULE NAME	DATA TYPE	TRANSFORMATION	PARAMETER	
Default: POSTCODE	POSTCODE	REDACTION	Right To Left	0 Digits
Default: LATLONG	LATLONG	REDACTION	Left To Right	0 Digits
Default: PHONENUM	PHONENUM	REDACTION	Left To Right	0 Digits
Default: AGE	AGE	ORDER_GROUPING	INTERVAL	Group size: 5
GENDER	GENDER	REDACTION	Full Redaction	
PROVINCE	PROVINCE	REDACTION	Right To Left	2 Digits
CITY	CITY	REDACTION	Right To Left	1 Digits
FLOAT	DECIMAL	ORDER_GROUPING	INTERVAL	Group size: 20
INT	INTEGER	ORDER_GROUPING	INTERVAL	Group size: 20
STRING	STRING	ORDER_GROUPING	SET	Group size: 20
DATE	DATE	REDACTION	Right To Left	2 Digits

图6 智能脱敏算法规则库

表2 4个表的智能脱敏结果

指标	privacy_risk /%	information loss/%	k-anoymize/ k
用户静态属性表	2	21	5
套餐开通数据表	1	0	4
用户使用行为列	1	10	2
用户偏好列	1	30	10

5 讨论

针对大数据智能化背景下的数据脱敏内涵、工作流程、系统机制阐述和相关技术研究分析,数据系统的脱敏是一个极其复杂、系统性的工作。特别是随着目前大数据与人工智能的飞速发展,数据脱敏将面临诸多挑战:(1) 如何将数据安全脱敏技术与人工智能的自主学习和强大的数据分析能力有机结合;(2) 如何利用大数据智能分析及人工智能建模算法从传统的静态脱敏方式到自适应的动态脱敏模式转变,有效满足多模态数据交互流量的不断增长和复杂多变的安全处理业务场景需求,例如在电力、运营商这种关乎国计民生的行业的应用[7,13,21];(3) 如何应对大数据智能化场景下用户信息的透明导致的数据所有权及使用权的伦理问题,仅仅依赖智能化的技术是否能使数据脱敏评价机制的可靠性、敏感数据准确识别方法多样性及数据治理体系的全面性达到预期目标。

在大数据与智能技术时代,如何应对新的技术、新的方法对数据脱敏带来的挑战,持续可靠地完成数据脱敏,都值得深入思考和研究。从技术、需求和法治、管理机制方面提出数据脱敏的3点思考和建议。

1) 辩证地认识数据脱敏安全问题,随着大数据多源融合与人工智能技术的快速发展,已有的数据脱敏方法或手段可能会失效,因此需以大数据为基础、人工智能等先进前沿技术为支撑,持续开展数据脱敏技术的研究。

2) 多媒体数据的隐私保护与脱敏会成为未来数据脱敏领域的重点方向。目前的数据脱敏多针对文本、数值等业务数据,随着未来大数据平台的日益完善,多媒体数据的采集和存储日益成熟,对诸如音频、视频等多媒体数据如何完成脱敏,进而应用于实际问题将成为未来数据脱敏领域的重要方向之一。

3) 数据脱敏问题还需要与法律法规、管理机制等层面进行联动,促进这个领域从技术到应用的合规发展,界定数据的隐私点、各种所有权及使用权,从而让政府、企业、社会和个人透明化地利用数据、知晓数据,减少各种不必要的数据纠纷、违法行为和非法事件等。

6 结论

从数据脱敏技术的方法现状、工作流程、典型脱敏案例、大数据与人工智能对现有脱敏方法的启示和挑战等方面,探讨了大数据与智能时代下的数

据脱敏中的若干问题,技术、需求和法治、管理机制方面提出数据脱敏技术发展建议。目前的脱敏方法虽然一定程度上可以保护数据隐私防止信息泄露,但是也面临着目前多源数据感知、融合推断、智能算法与分析等新技术、新方法的融合问题[24-26]。同时,要做到完整意义上的数据脱敏,除了技术发展外,还需要与法律法规、管理机制联动[27]。

参考文献(References)

[1] Bakken D E, Rarameswaran R, Blough D M, et al. Data obfuscation: Anonymity and desensitization of usable data sets[J]. IEEE Security and Privacy Magazine, 2004, 2(6): 34-41.

[2] Lake B M, Salakhutdinov R, Tenenbaum J B. Human-level concept learning through probabilistic program induction [J]. Science, 2015, 350(6266): 1332.

[3] Li G L, Wu J Li J H, et al. Service popularity-based smart resources partitioning for fog computing-enabled industrial internet of things[J]. IEEE Transactions on Industrial Informatics, 2018, 14(10): 4702-4711.

[4] Lecun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436.

[5] 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(1): 33-38.

[6] 卞超轶, 朱少敏, 周涛. 一种基于Spark的大数据匿名化系统实现[J]. 电信科学, 2018, 34(4): 156-161.

[7] 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(13): 17-20.

[8] 包英明. 大数据平台数据安全防护技术[J]. 信息安全研究, 2019, 5(3): 60-65.

[9] 徐焕, 查志勇, 罗弦, 等. 大数据环境下配电系统测试数据脱敏技术研究[J]. 机械设计与制造工程, 2019, 48(3): 54-58.

[10] 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.

[11] 胡荣磊, 何艳琼, 曾萍, 等. 一种大数据环境下医疗隐私保护方案设计与实现[J]. 信息网络安全, 2018(9): 54-60.

[12] 王冬, 李文, 徐高升, 等. 一种大数据环境下的数据隐私保护策略及其实践[J]. 微型电脑应用, 2013(6): 6-8.

[13] 叶水勇. 数据脱敏技术的探究与实现[J]. 电力信息与通信技术, 2019, 17(4): 23-27.

[14] Kawahara J, Saitoh T, Yoshinaka R. The time complexity of permutation routing via matching, token swapping and a variant[J]. Journal of Graph Algorithms and Applications, 2019, 23(1): 29-70.

[15] Mohseni M, Banani S A, Eckford A W, et al. Scheduling for VoLTE: Resource allocation optimization and low-complexity algorithms[J]. IEEE Transactions on Wireless Communications, 2019, 18(3): 1534-1547.

[16] Li J H. Cyber security meets artificial intelligence: A survey[J]. Frontiers of Information Technology & Electronic Engineering, 2018, doi: 10.1631/FITEE.1800573.

[17] 李盼, 赵文涛, 刘强, 等. 机器学习安全性问题及其防御技术研究综述[J]. 计算机科学与探索, 2018, 12(2), doi: 10.3778/j.issn.1673-9418.1708038.

[18] Guan Z, Bian L, Shang T, et al. When machine learning meets security issues: A survey[C]//2018 IEEE International Conference on Intelligence and Safety for Robotics (ISR). Piscataway N J: IEEE, 2018, doi: 10.1109/IISR.2018.8535799.

[19] 张雪芹, 张立, 顾春华. 社交网络中社会工程学威胁定量评估[J]. 浙江大学学报(工学版), 2019, 53(5): 24-29.

[20] 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.

[21] 彭小圣, 邓迪元, 程时杰, 等. 面向智能电网应用的电力大数据关键技术[J]. 中国电机工程学报, 2015, 35(3): 503-511.

[22] 吕军, 杨超, 王跃东, 等. 基于多业务场景的大数据脱敏技术研究及其在电力用户隐私信息保护中的应用[J]. 电力大数据, 2018, 21(7): 34-40.

[23] 冉冉, 李峰, 王欣柳, 等. 一种面向隐私保护的电力大数据脱敏方案及应用研究[J]. 信息安全与技术, 2018, 21(7): 105-113.

[24] 徐宗本. 用好大数据须有大智慧[N]. 人民日报, 2016-03-15(07).

[25] 程学旗, 靳小龙, 杨婧, 等. 大数据技术进展与发展趋势[J]. 科技导报, 2016, 34(14): 49-59.

[26] 马茜, 谷峪, 张天成, 等. 一种基于数据质量的异构多源多模态感知数据获取方法[J]. 计算机学报, 2013, 36(10): 2120-2131.

[27] 马新强, 刘勇, 范婧, 等. 大数据驱动下智慧城市建设的若干思考[J]. 科技导报, 2017, 35(21): 133-139.

Data desensitization meets big data and artificial intelligence: Some food for thought

WANG Hongkai[1], GONG Xiaogang[1], YE Wei[1], CHEN Chao[1], MA Xinqiang[2,3*], YAO Jinqiang[4], LIU Yong[2*]

1. State Grid Zhejiang Electric Power Company's Information Communication Company, Hangzhou 310027, China
2. Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China
3. School of Artificial Intelligence, Chongqing University of Arts and Sciences, Chongqing 402160, China
4. Zhejiang Communications Group Inspection Technology Co., Ltd., Hangzhou 310030, China

Abstract How to avoid the privacy and information leakage in the massive data is a widespread concern in the field of the big data and the information security. The data desensitization technology is one of the important means to solve this problem. In recent years, with the rapid development of the artificial intelligence technology driven by the big data and the high-performance computing, a large number of innovative methods were proposed, with many challenges to the existing data desensitization methods. This paper reviews the current situation of the big data and intelligent technology development, and the data desensitization technology, as well as the future development trends of the data desensitization technology.

Keywords big data; AI; data desensitization; information security

（责任编辑　刘志远）