大数据访问控制研究

李 昊 张 敏 冯登国 惠 榛
(中国科学院软件研究所可信计算与信息保障实验室 北京 100190)

摘 要 大数据时代的到来,使得数据成为了重要的经济资产.为了更好地利用它们,有偿或无偿的共享数据将是一种趋势.作为确保大数据安全分享重要技术之一的访问控制技术也将在大数据时代发挥重要作用.该文首先对大数据及大数据应用的新特点进行分析,并提炼出这些新特点为访问控制领域带来的五个迫切需要解决的新问题:授权管理问题、细粒度访问控制问题、访问控制策略描述问题、个人隐私保护问题,以及访问控制在分布式架构中的实施问题.接着对相关访问控制关键技术的研究现状进行了梳理,包括角色访问控制、风险访问控制、半/非结构化数据的访问控制、针对隐私保护的访问控制、世系数据相关的访问控制、基于密码学的访问控制等.虽然这些现有技术不一定能直接应用于大数据场景,但是它们都可以被大数据访问控制的研究所借鉴,以解决大数据带来的上述访问控制的新问题.在此基础上,总结并提炼了若干大数据访问控制所呈现的新特点:判定依据多元化、判定结果模糊化、多种访问控制技术融合化.最后,对未来大数据访问控制的研究进行了展望,给出了一些有待研究的问题.该文认为大数据应用的发展将为访问控制技术的研究提出许多新的挑战,同时也将带来巨大的机遇,这必将引起访问控制技术的一次重大变革.

关键词 大数据;数据分享;大数据安全;访问控制;隐私保护
中图法分类号 TP309 **DOI 号** 10.11897/SP.J.1016.2017.00072

Research on Access Control of Big Data

LI Hao ZHANG Min FENG Deng-Guo HUI Zhen
(*Department of TCA, Institute of Software, Chinese Academy of Sciences, Beijing* 100190)

Abstract With the arrival of the era of big data, data has become a kind of important assets. In order to get a better utilization of big data, paid or unpaid data sharing will be a trend. And as one of key techniques to maintain security of data sharing, access control will play an important role in the era of big data. In this paper, five new problems of access control which are brought by big data are summarized by analyzing characteristics of big data and its applications. They are respectively described as authorization problem, fine-grained access control problem, policy description problem, privacy protection problem, and the problem of implementing access control over distributed architecture. Then some key technologies which can be exploited to solve these problems are described, including role-based access control, risk-based access control, access control of semi-structured data and unstructured data, privacy-aware access control, provenance-aware access control, cryptography based access control, etc. Although some of these technologies cannot be directly used in big data applications, they can afford us lessons for solving the access control problems mentioned above. Then three expected characteristics of access control for big data are also discussed. The first one is the multivariate basis for access control deciding. The second one is that the access control decision comes fuzzier than before. The last one is the combination

收稿日期:2016-06-30;在线出版日期:2016-09-20.本课题得到国家自然科学基金(61402456,61303248)资助.**李 昊**,男,1983 年生,博士,副研究员,主要研究方向为可信计算、数据安全、访问控制.E-mail:lihao@tca.iscas.ac.cn.**张 敏**,女,1975 年生,博士,副研究员,主要研究方向为数据隐私保护、可信计算.**冯登国**,男,1965 年生,博士,研究员,主要研究领域为信息安全与密码学、可信计算与信息保障.**惠 榛**,男,1987 年生,博士研究生,主要研究方向为数据安全、访问控制.

of different access control models. Finally, several challenging research problems for access control of big data are given. And in our view, the development of big data will bring both challenges and opportunities for access control, and result in a revolution in this area.

Keywords　big data; data sharing; big data security; access control; privacy protection

1　引　言

随着社会信息化和网络化程度的不断提高，以及云计算、物联网等新兴技术的发展和应用，一个大规模生产、分享和应用数据的时代已经悄然而至. 目前人们已经逐步认识到大数据的价值，并开始在公共卫生、商业、科学研究等领域应用大数据分析技术[1-3]. 例如，淘宝、天猫等电商为了应对“双 11 购物狂欢节”暴增的交易量，采用大数据分析技术对往年用户的消费情况以及购物节前用户浏览、关注、收藏的商品等数据进行分析，从而更有效地进行备货①. 再比如，IBM 与同仁医院合作，建立了对临床、运营、科研、考核等数据的大数据分析系统，以更好地服务患者②. 我们可以预测，随着大数据分析技术的应用推广，在当今时代数据会变得越来越重要，它将成为一种与矿产和石油同样的巨大经济资产[1].

然而人们在利用大数据创造价值的同时，也被大数据带来的安全问题所困扰. 相对于普通数据，大数据在收集、存储、共享和利用等环节都面临着更为严重的安全威胁. 这是由于大数据环境中数据和计算能力通常都是充足的，所以分散在大数据应用中的各种敏感信息可以被更轻松的获得，并进行关联分析[4]. 例如“棱镜门”③、“查开房记录”④等事件就充分说明了大数据时代数据安全问题可能对国家、组织、个人造成重大影响. 因此，人们对大数据安全和隐私保护的需求是非常迫切的.

访问控制不是解决数据安全问题的“银弹”，但它却是公认的确保数据安全共享的重要手段之一. 大数据时代，为了更好地从数据中挖掘出价值，数据被有偿或无偿的共享是一种必然趋势. 事实上，许多国家的政府部门已经认识到这点，正逐步公开其掌握的数据供学术界和产业界分析，以造福社会⑤. 因此，在大数据背景下，作为确保数据共享安全重要手段的访问控制技术仍然将发挥其作用[5]. 然而由于大数据所具有的一些新特征，传统的访问控制从模型到实施都需要进行一定的改进和创新以满足新应用场景下的新安全需求.

本文将重点关注大数据为访问控制带来的新问题和挑战，并对现有关键技术进行梳理，在此基础上归纳总结出大数据访问控制所呈现的一些新特点. 本文第 2 节介绍大数据访问控制面临的新问题；第 3 节选取部分重点相关技术进行了梳理，并归纳总结大数据访问控制所呈现的一些新特点；第 4 节对未来大数据访问控制的研究进行展望；第 5 节总结全文.

2　大数据访问控制面临的新问题

大数据应用通常是一个包含了数据收集、存储、共享和利用等多个环节、多种技术的庞大且复杂的系统. 并且由于大数据应用场景的不同，这些环节和技术也将具有较大的差异. 因此，大数据访问控制所要面对的安全问题也必然是各不相同的. 然而，如同云计算是大数据的基础平台和支撑技术那样[6]，大数据的访问控制研究也不应该抛开现有访问控制技术来重新研究. 目前一些诸如物联网的访问控制、云计算的访问控制等热点研究工作都应该作为大数据访问控制的关键支撑技术，但是这些问题和相关研究的归纳和总结不在本文范围内. 本文主要从大数据价值的体现——安全地分享和使用数据的角度，来对大数据访问控制所面临的问题和相关技术进行探讨.

在上述限定范围内，本节对大数据的概念和大数据应用的共性特征进行概述，并给出这些新特征为访问控制带来的一些新问题.

2.1　相关基础

(1) 大数据的概念

根据维基百科上的定义，大数据是指规模大且复杂以至于很难在合理时间内用现有的数据库管理工具或传统的数据处理应用来处理的数据集⑥. 这

① http://www.yicai.com/news/2013/11/3091893.html
② http://www.tuicool.com/articles/vEF7bu3
③ http://baike.so.com/doc/5685827.html
④ http://baike.baidu.com/view/10968867.htm
⑤ http://tech.it168.com/a2013/1009/1542/000001542502.shtml
⑥ http://en.wikipedia.org/wiki/Big_data

个定义是较为抽象的.目前,人们更倾向于通过对大数据特点的阐述和归纳来理解大数据的概念[6].其最为常见的特点是:大规模(Volume)、高速性(Velocity)和多样性(Variety),也就是人们普遍认可的 3V 特点.此外,还有一些组织机构或企业提出大数据的第 4 个特点:价值性(Value)或真实性(Veracity).而本文认为,与其拘泥于这些大数据概念或特点的描述是否准确完善,不如直接关注这些特点对所研究领域带来的问题和挑战.

(2)大数据的应用场景分析

本文从访问控制角度对大数据的应用场景进行分析,并提取与访问控制相关的共同特征,以进一步展开探讨.因为访问控制解决的是数据在用户之间安全分享的问题,所以本文关注的大数据场景特征可以用如下的问题描述:"什么样"的数据将被"什么样"的用户"以何种方式"分享.具体地,表 1 从数据来源、数据类型、数据规模、数据用户类型和数据分享的安全目标 5 个方面对若干经典大数据应用场景进行了分析.虽然这些大数据应用场景差异较大,但是从访问控制角度可以总结出如下共同特征:

(a)数据从多个渠道大量汇聚,往往达到 TB、PB 级;

(b)数据类型组成通常包括了结构化、半结构化、非结构化数据,且在社交网络大数据、工业大数据等场景下数据往往呈现出顺序、大量、快速、连续的流特点;

(c)在诸如社交网络、医疗、交通等大数据场景下,个人用户成为了数据的重要来源;

(d)大数据应用往往具有更加复杂的数据用户类型,并且在这些数据用户之间分享数据时,出现了诸如朋友关系、病人与医生相关性、业务流程等安全约束需求;

(e)为了充分利用大数据创造价值,大数据应用所收集到的数据除了会在企业或组织内部的不同部门之间分享,还可能提供给第 3 方分析者部分数据访问权限,使得访问控制的范围进一步扩大.但是数据的匿名化发布这种分享方式不在本文探讨范围内.

最后,正如大数据的定义那样,Volume、Velocity 和 Variety 中的任意一个单独的"V"都不是大数据集所独有的特点,本文所归纳的上述 5 个大数据应用的特征也不一定是大数据场景下新出现的,甚至也不是每个大数据应用全部都具备的.然而,这些特征给访问控制所带来的问题,却是大数据访问控制所必须解决的.

2.2 主要问题

本节我们将对大数据及其应用的特征为访问控制领域带来的若干问题进行阐述.

问题 1. 特征(a)增加了访问控制策略制定以及授权管理的难度,过度授权和授权不足现象越来越严重.

首先,在访问控制系统中,哪些资源能够被哪些用户访问通常是由安全管理员定义的[7].这种访问控制授权行为本身就是一种劳动密集型工作[8].而在面对规模巨大且来源复杂的大数据时,管理员更加难以实施这种劳动密集型的授权管理操作,因此往往授予用户过多甚至全部的访问权限以确保用户能够顺利使用系统.例如,在医疗大数据系统中,医学研究者是能够访问病人的部分临床数据的.由于"病人的哪部分临床数据能够被医学研究者查看"需要非常专业的医学知识才能把握,所以为了不影响研究工作,这些系统往往采用了过度授权.据 Ponemon 在 2010 年的统计,随着数据量的增加,过度授权的问题比 2008 年上涨了 9%.其次,由于大数据及其应用系统的复杂性,往往会造成一些访问需求没有被管理员预先考虑到,或者出现一些新的访问需求.因此,大数据应用中用户的授权不足现象也将越来越频繁.最后,该问题的解决对大数据应用场景具有重要意义.一方面能够将管理员从繁重的授权管理工作中解救出来;另一方面也能够减小授权工作对管理员个人专业知识的依赖,提高授权的准确性.

问题 2. 特征(b)增加了访问控制客体描述的困难,为细粒度访问控制的实施提出了挑战.

大数据时代,非结构化数据正以惊人的速度快速增长,它们是大数据集的重要组成部分.据 IDC(International Data Corporation)的一项调查报告中指出:企业中 80%的数据都是非结构化数据,这些数据每年都按指数增长 60%①.传统访问控制方案中往往采用数据类型来描述访问控制策略中的客体.例如,医疗系统中针对"病人 ID"这种数据类型的策略——"医学研究人员不能访问病人 ID".然而,相比于结构化数据,非结构化数据自身能够体现更丰富的信息.譬如,在医疗系统中,同样是"病例"这种类型的非结构化数据,患有不同病症的病人的病例在内容和表达方式上往往都是截然不同的.人们也期望这些信息能够体现在访问控制中,以实现更细粒度的访问控制.例如,"医学研究人员能够访问

① http://www.csdn.net/article/2011-07-15/301704

与其研究方向相关的病例”.在传统的访问控制方案中,虽然可以利用基于属性的访问控制(Attribute-Based Access Control,ABAC)来表达该策略中“与研究方向相关”的概念,但是在大数据应用中,由于数据规模大且增长速度快的缘故,管理员较难实施这种劳动密集型的标记操作,因此较难针对非结构化数据进行细粒度的访问控制策略描述.为了在大数据应用中实施细粒度的访问控制,就需要对访问控制策略中的非结构化和半结构化数据客体的描述方式展开进一步研究.

问题 3.　特征(d)使得访问数据的主体集合构成复杂化,同时也增加了分享数据时安全需求的描述难度.

以典型的大数据应用——区域医疗及基层医疗信息系统①为例,它所涉及的公共卫生、医疗服务、医疗保险、药品管理、运营管理等多方面的应用都将使用医疗大数据系统收集和存储的大数据集.在这种情况下,“医疗缴费通知单”这个客体,将可能被收费员工、药房护士、社保员工等多种用户访问,以支持各类不同业务.而这种粗粒度的授权是不满足最小权限原则[9]的,例如“医疗缴费通知单”在报销时还能被收费员工、药房护士访问是非常不合理的,会造成不必要的患者隐私泄漏.并且如果某客体在业务过程中会被大量用户在不同阶段访问,那么这种授权方式的结果很可能就是该客体能同时被系统中大多数主体访问,使得访问控制近乎“形同虚设”.例如,上述“医疗缴费通知单”就能够被医疗大数据系统中大部分主体访问.再譬如,社交网络应用中,数据访问者之间存在着复杂的朋友关系,而基于朋友关系来约束数据的访问也是该类应用的基本需求之一[10].因此,如何更好地表达大数据场景下复杂主体的多样化访问需求是迫切需要解决的问题.

问题 4.　特征(c)增加了访问控制对数据客体中个人隐私的保护难度.

在大数据时代,个人数据正以惊人的速度增长.据 IDC 统计,2011 年全球被创建和复制的数据总量为 1.8ZB,其中 75%来自个人[11].这些数据不仅包括用户主动发布的各种文字、图片信息,也包括各类传感器收集的用户数据.而随着大数据分析技术的发展,这些以往被忽视的个人数据的价值将越来越被重视.与此同时,在这些数据被收集和使用过程中所引发的隐私泄漏问题也会越来越严重.例如,医疗大数据系统中,好奇的医生可能会对治疗过程无关的病人数据进行访问.而传统的访问控制技术重点关注的是数据的安全管理——“医生能够访问自己病人的数据”,却对大数据应用中日益突出的个人隐私泄漏问题——“好奇医生对治疗无关的病人隐私的窥探”缺乏有效应对手段.在这种情况下,如何利用访问控制确保数据分享中的个人隐私安全将是大数据访问控制研究的重要问题之一.

问题 5.　特征(e)为访问控制增加了数据分析这一数据访问场景,而大数据分析架构本身缺乏对安全性的考虑.

与普通数据集相比,大数据集中的单个数据价值通常不大,但是聚集起来价值较大.或者是价值高的少量数据隐藏于价值较低的海量数据中.因此需

表 1　常见的大数据应用场景

	数据主要来源	主要数据类型	数据规模	数据用户类型	数据分享的安全目标
社交网络大数据	个人用户的网上社交行为	个人信息等结构化数据;社交关系等半结构化数据;文本、图片等非结构化数据	PB	社交网络个人用户;数据分析者	向社交网络中特定关系的朋友进行数据分享,阻止与数据属主之间关系不符合要求的访问者;向授权的数据分析者分享数据,阻止未授权的数据访问
医疗大数据	政府部门提供的基础数据(居民、医疗资源等);病人、医生等个人用户的就医过程	病人、医生等人员信息的结构化数据;电子病例等半结构化数据;PACS影像等非结构化数据	TB	医疗过程的病人、医生、护士、医院工作人员、社保服务机构人员等相关用户;政府卫生部门、医药研究机构等数据分析者	向医疗过程中与病人治病过程相关的参与者分享治病所需的必要数据,阻止不相关访问者对病人数据的访问;向授权的数据分析者分享数据,阻止未授权的数据访问
交通大数据	政府部门提供的基础数据(人、车、路、气象等);从车辆、监控等采集的动态交通行为数据	人员、车辆、道路信息等结构化数据;路况影像等非结构化数据	PB	参与交通的车辆、行人、管理人员等用户;交通相关部门、导航服务机构等数据分析者	在特定环境、特定目的、特定流程中向交通过程参与者分享其交通活动需要的数据,阻止对不相关数据的访问;向授权的数据分析者分享数据,阻止未授权的数据访问
工业大数据	生产过程所产生的数据	人员、设备信息等结构化数据;传感器采集到的半结构化、非结构化的生产过程数据	TB	参与工业生产过程的工人、管理人员、控制器等用户;数据分析者	在特定环境、特定目的、特定流程中向工业生产活动参与者分享生产活动需要的数据,阻止对不相关数据的访问;向授权的数据分析者分享数据,阻止未授权的数据访问

① http://baike.baidu.com/view/5418873.htm

要采用分布式的大数据处理架构进行数据分析以挖掘数据中潜在的价值.而这些处理架构通常是针对大数据的 3V 特点进行设计,较少考虑分析过程中的安全性.一些具有访问控制的分析架构,其访问控制实现往往也是临时或补丁式的,无法满足安全需求[12].与此同时,传统的针对每个数据客体进行授权,并对每个数据访问请求进行判定的访问控制方式也较难直接在大数据分析架构上实施.在这种情况下,结合大数据分析架构展开更加高效的访问控制和授权管理研究是非常必要的.

3 关键技术分析

为了确保大数据场景下数据的安全分享,亟需针对上述问题展开大数据访问控制关键技术的研究.虽然许多现有访问控制技术不是针对大数据场景提出的,但是却能够被用于应对大数据场景下访问控制存在的问题.本节选取了部分重点相关技术进行了梳理,并进一步从大数据访问控制需求角度对这些技术的特点进行了归纳总结.需要注意的是,现有这些关键技术所针对的都是上述大数据访问控制问题中的一个或若干个,而不是全部.实际系统中的大数据访问控制必然是多种访问控制技术的综合应用.

3.1 角色挖掘

角色是基于角色访问控制模型(Role-Based Access Control,RBAC)的核心概念[13].当系统较为复杂时,角色的设计也将变成一项困难任务.为了应对该问题,Kunhlman 等人[14]提出了角色挖掘的概念.角色挖掘主要用于解决如何产生角色,并建立用户-角色、角色-权限的映射问题.相比于"自上而下"进行人为的角色设计,角色挖掘是"自下而上"地从系统已有的用户-权限分配关系中来自动化地实现角色定义和管理工作,以减小对管理员的依赖.因此,它能够有效缓解采用 RBAC 的大数据应用中存在的"问题 1"——过度授权和授权不足越来越严重的现象.

在角色挖掘技术的研究初期,主要是针对已有的用户-权限分配关系数据进行角色挖掘.也就是,假定系统在初始情况下已经有了简单的访问控制——"哪些用户能够访问哪些资源".而角色挖掘将从这些已有的用户-权限的分配关系中寻找出潜在的"角色",并将角色与用户、角色与权限分别关联.文献[15]提出了利用聚类进行角色挖掘的方法,即在权限上进行分层聚类,最终获得一个树形结构的角色层次.角色层次的形成使得角色挖掘的效果更加贴近实际,也更方便后期的角色管理.而文献[16]提出了另一种子集枚举的角色挖掘算法,该算法能够标识出系统中所有可能的角色,其挖掘结果相比之前的工作更加完备.文献[17]则提出了一种角色挖掘的通用方法,将角色挖掘问题转化为了约束满足问题,从而支持自定义最优度量,并求解该度量下最优的角色挖掘结果.该方法的优势在于能够设定不同需求,并获取满足该需求的最优挖掘结果.可以看出,上述研究工作实现了从已有的用户-权限分配关系中自动获取角色,并进一步分配用户-角色、角色-权限的授权关系,能够有效降低对管理员的依赖,同时缓解过度授权或授权不足问题.

然而,上述研究工作存在一个缺陷,即这些研究方案都假设了系统已有的用户-权限分配关系是准确无误的.不幸的是该假设在复杂系统中往往是不成立的.针对已有的用户-权限分配可能存在错误的问题,文献[18]进一步在角色挖掘中引入了概率模型,可以缓解错误的用户-权限分配对角色挖掘结果的影响.但该研究仍然是基于封闭的、精确描述的用户-权限分配关系的数据集,所以仅对部分错误授权情景有效.近期,一些研究者开始基于更丰富的数据集进行角色挖掘,以期获得更好的挖掘效果.例如,文献[19]采用基于机器学习的算法对访问日志进行角色挖掘.该方案从访问日志中抽取了用户对权限的使用次数,期望基于权限的实际使用情况来生成角色.接着,将"用户"看作是"文档","权限"看作是"单词"、"权限使用次数"看作是"单词出现次数"、"角色"看作是"主题",从而将角色挖掘问题转变为一个文本分类问题来处理.其实验结果也表明这种考虑了权限实际使用情况的角色挖掘方法更加符合系统的实际安全需求.

总之,在采用了 RBAC 的大数据应用中,由于其系统和数据的规模与复杂性都远远超出了管理员的能力,所以为了减小过度授权和授权不足现象,自动化地对角色进行挖掘,并完成授权是亟待进一步研究的;与此同时,大数据应用往往具有更强大的数据聚集能力和计算能力,因此也为角色挖掘提供了更好的分析基础.那么,在大数据环境下如何利用好这些丰富的数据集和强大的计算能力来提升角色挖掘的效果,也是非常值得进一步深入研究和探讨的.

3.2 风险访问控制

根据前述问题分析,预先定义的严格不变的访

问控制策略是较难满足大数据应用场景的.由于大数据应用系统的复杂性,通常会存在一些特定的访问需求在设计策略时没有考虑,或者访问需求的变化引起访问控制策略不再适合等.如果严格按照预先定义的策略执行访问控制,将产生授权不足无法完成业务的情况,这对于诸如医疗、交通等要求高可用性的大数据场景是无法接受的.而基于风险的访问控制技术则提供了另一种解决“问题 1”中授权不足问题的途径.

文献[20]是较早将风险引入访问控制领域的工作,它定义了风险量化和访问配额等概念,并给出了基于风险的信息系统应该满足的一些指导性原则和建议.文献[21]提出了一种平衡信息共享风险和收益的访问控制模型(Benefit and Risk Access Control, BARAC).该模型将风险和收益与每个读写事务都进行了关联,并利用事务图来描述允许的事务,最后要求整个系统的收益要大于风险.这些研究工作不再严格地按照预先分配的权限进行访问控制,而是开始衡量访问行为所带来的风险是否为系统可接受的.因此,当发生一些未预料到的访问行为时,若其风险是可接受的,则仍然可以允许该访问.这对于大数据应用来说是非常必要的,极大地提高了其可用性.

随后,一些研究者开始在传统访问控制模型中扩展风险概念.文献[22-23]将风险扩展到 RBAC 模型中,在进行访问控制决策时利用了风险分析方法.文献[24-25]则将风险引入多级安全模型中以增加其访问控制的灵活性.在对传统访问控制模型进行扩展时,一些研究者注意到只有基于风险的访问控制判定是不够的.也就是,某访问行为的风险较大,但是收益可能也非常可观.如果直接拒绝该访问,则就完全无法获得收益.尤其是在医疗系统中,当出现紧急情况时,一些风险较大的访问请求也是不能简单拒绝的.因此,文献[24]进一步提出了一种风险自适应访问控制实施方法.它针对多级安全模型中访问控制判定过于严格的问题(只有允许/拒绝两种结果),定义了风险带(Risk Band)的概念,并给出了一种根据访问行为的风险值在风险带中的位置来实施访问控制的方法.也就是说,该研究不但通过在访问控制条件中引入“风险”提高了访问控制判定的灵活性,而且通过为访问行为划分“风险带”的方式提高了访问控制实施的灵活性.如图 1 所示,在严格拒绝和弹性拒绝之间有着一个细分的风险容忍区域,可以根据访问行为的风险在其中的位置来调整允许的权限,这对于大数据场景下的访问控制是非常有利的.

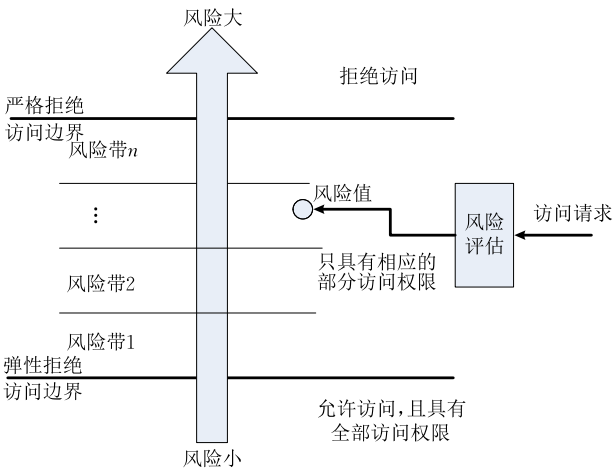


图 1　风险带示意图

近期,随着大数据技术的兴起,一些数据分析技术也被应用于风险访问控制中.例如,文献[7]针对医疗信息系统提出了一种基于风险的访问控制方案来保护系统中存储的病人隐私信息.该方案明确给出了医疗信息系统中诚实医生与好奇医生的区别:诚实医生只访问正常治疗过程所必需的病人数据;好奇医生除了访问必需的病人数据外,还会由于好奇访问一些额外的病人隐私数据.利用信息熵来描述医生访问行为时,会发现好奇医生由于访问了更多病人数据,因而具有更高的熵值.在此基础上,系统将所有医生访问行为的熵作为可容忍的风险配额分配给每个医生.在治疗过程中,每个医生的访问行为都会被评估风险值,并在其风险配额中进行扣减.当一个医生的风险配额被扣为零时,则不能再进行数据访问.而系统每隔一段时间会重新为医生分配风险配额.在这种情况下,好奇医生会由于经常窥探病人隐私而很快消耗完风险配额,进而被管理员注意到.该方案一方面能够有效抑制好奇医生的非正常访问行为,另一方面也不会在医生具有风险配额时限制其一些高风险的访问行为,因此是一种较为宽松的访问控制方式.而文献[26]则进一步采用了 EM 算法对所有医生的历史访问行为进行分析,区分了诚实医生和好奇医生访问行为的概率分布,并以诚实医生访问行为的熵作为系统可承受风险的基准值.实验表明,该方案能够进一步提高风险评估和实施的准确性.

从上述研究工作可以看出,大数据在为访问控制带来问题的同时,一些数据分析技术也可以被用来评估不同访问行为的风险,从而更好地实现风险访问控制.因此,相关研究将是未来大数据应用场景下访问控制的研究方向之一.但是由于大数据应用系统的规模和复杂性,其风险的定义和量化以及系

统能容忍的风险阈值的设定将比一般信息系统更加困难,这也是后续研究所需要进一步关注的问题.

3.3 半/非结构化数据的访问控制

半结构化数据是指那些既不是完全无结构化的,也不是传统数据库系统中那样有严格结构的数据[27].其特点为:(1)没有明确的数据模式;(2)结构不规则;(3)没有严格的类型约束[28].而非结构化数据,则是指那些完全无结构的数据,譬如文本数据.由于这些数据在结构上不够严格,所以存在"问题2"所述的细粒度访问控制难以实施的问题.针对于此,一些研究者展开了大量相关研究工作,具体如下.

(1)XML 数据的访问控制

在金融、医疗等领域应用最为广泛的一种半结构化数据就是可扩展标记语言(Extensible Markup Language,XML)数据,它采用一系列简单的标记来描述数据,使得数据更容易被理解和交换.并且这些应用领域也对 XML 数据的细粒度访问控制有着迫切需求.例如,对于一个 XML 格式的医疗文档,保险公司能够查看其中的账单条目,但是却不应该能查看治疗细节.目前,已经有了许多针对此类需求的访问控制研究.

文献[29]提出了基于 DTD(Document Type Definition)的针对 XML 文档内容的细粒度访问控制方法和系统.为了高效地进行授权管理,授权的对象就必须是 XML 文档的格式,而非每个 XML 文档的具体内容.由于 DTD 定义了合法的 XML 文档构建模块,所以可以看作是合法的 XML 文档所必须遵循的格式的描述.因此,采用 DTD 或其他能够描述 XML 文档格式的机制进行授权和访问控制是 XML 数据访问控制的一种基本方式.

随后,针对 XML 细粒度访问控制的效率和授权管理复杂性,研究者又进一步展开了许多相关研究工作.文献[30]提出了针对 XML 文档的强制访问控制方法,同时也给出了强制访问控制标记的授予和继承规则.这种标记继承规则能够极大减少标记管理的工作量,所以对大数据这种大规模、复杂的应用场景是非常必要的.文献[31]则提出了另一种基于授权图的 XML 文档访问控制方法.该方案通过图匹配的方式能够分析出查询请求是全部允许、全部拒绝和部分允许中的哪一种,从而避免了进一步处理前两种情况的计算代价.更进一步,由于实际系统中的 XML 文档类型数量和用户角色数量都较多,所以访问控制规则的数量可能是百万级的[32].这对访问控制的扩展性和性能都提出了较高要求.文献[32]采用了与具体文档无关的规则功能集来提高对 XML 文档访问控制的效率.也就是,每个规则功能都是一个被同一种文档类型的所有文档共享的,封装了所需访问规则的代码段.在运行时,访问请求所对应的规则功能(代码段)将直接决定 XML 文档片段的可访问性,因此可以在一定程度上提高访问控制的效率.文献[33-34]提出了 XML 文档的简化授权管理方法,并解决了祖先节点与子孙节点之间的授权冲突问题,可以有效降低授权管理工作量.最近,文献[35]还针对远程医疗应用环境,提出了一种 XML 文档的访问控制系统,并将其在实际环境中进行了有效性和效率的验证.

可以看到,在大数据兴起之前已经出现了许多 XML 数据的细粒度访问控制研究,并且在实际系统中也已经有了较多应用.然而在大数据场景下,XML 数据的结构复杂性以及规模将为细粒度的授权管理和访问控制效率提出挑战,是未来有待研究的实际问题.

(2)图数据的访问控制

除了 XML 这类树型半结构化数据外,大数据集中还存在图型的半结构化数据,例如社交网络中的朋友关系.针对这类半结构化数据,文献[10]提出了基于关系的访问控制模型(Relationship-Based Access Control,ReBAC),也就是将访问请求者与资源所有者之间的关系作为了访问控制判定的一个依据.而商业信息系统中的大多数访问控制机制主要依据访问请求者的身份或角色来实施访问控制,这与现实生活中人们分享隐私数据的模式是非常不同的.例如,一个人可能只希望将某个照片与自己最亲密的朋友分享,而不让朋友的朋友看到.如果用 RBAC 来实现这种需求,就要为最亲密的朋友和朋友的朋友分别建立角色.当这种基于关系的访问控制需求更加复杂时,RBAC 就会由于角色过于繁多而无法适用.因此,基于关系来描述这种访问控制需求就成为了最直接的选择.

随后,文献[36]规范化地描述了 ReBAC,并提出了一种基于模态逻辑的策略语言来描述此类基于关系的访问控制需求.文献[37]则进一步针对策略语言是否完备的问题展开了研究.它指出文献[36]提出的策略语言不足以表达所有基于关系的访问控制需求,并给出了相应扩展来确保策略语言的完备性.更进一步,文献[38]认为模态逻辑在表达特定关

系时存在不足，并采用混合逻辑对上述策略语言进行了改进.经过改进，该策略语言在表达诸如“至少3个朋友”这种复杂关系时是非常高效的.

此外，由于图结构非常擅长描述各种关系，也包括了访问控制中的授权关系，因此文献[39]扩展了ReBAC模型，用于解决数据客体存在多个属主时的访问控制问题.最近，文献[40]又提出了使用ReBAC管理ReBAC的方法，并结合医疗系统OpenMRS实现了这种管理模型.文献[41]则进一步解决了ReBAC管理模型中的级联删除等问题，完善了该模型的实施细节.可以看出，这种半结构化数据的访问控制研究对于其他访问控制的授权管理有着重要意义，因此将是大数据访问控制的重要研究方向.

(3)文本数据的访问控制

文本数据是一种典型的非结构化数据，它在结构上是近乎无规则的.对于此类数据实施细粒度访问控制的最大难点就在于缺少恰当的方式来描述客体.如前所述，采用ABAC是解决客体描述问题的一种方式.然而在大数据场景下，ABAC中的属性标记工作将由于繁多的客体数量、属性种类等缘故变得较为困难.

针对该问题，文献[8]提出了基于内容的访问控制模型(Content-Based Access Control，CBAC)，可以将用户已拥有文本的内容与新请求文本的内容进行相似度分析，并依据分析结果进行访问控制判定，从而达到“用户只能够访问与其相关的文本内容”的安全目标.更进一步，文献[42]给出了CBAC的形式化定义，将CBAC分为两个阶段：初始授权和基于内容的授权.在初始授权阶段，用户将被预先分配一个base set，描述了用户初始能够访问的客体；而基于内容的授权阶段则会根据base set与请求的客体之间的相似度进行访问控制判定，并根据判定结果更新base set.该研究实现了对文本数据客体的细粒度描述，同时也避免了通常细粒度访问控制中管理员的繁重授权管理工作.

综上所述，本小节展示的这些研究工作虽然能够用于解决半/非结构化数据的基本访问控制问题，但是在大数据环境下，如何对大量且不断增加的半/非结构化数据进行授权管理仍将是一项非常具有挑战性的工作.而诸如文献[8,42]中提到的自动化授权方式将是解决该问题的有效途径之一.

3.4　针对隐私保护的访问控制

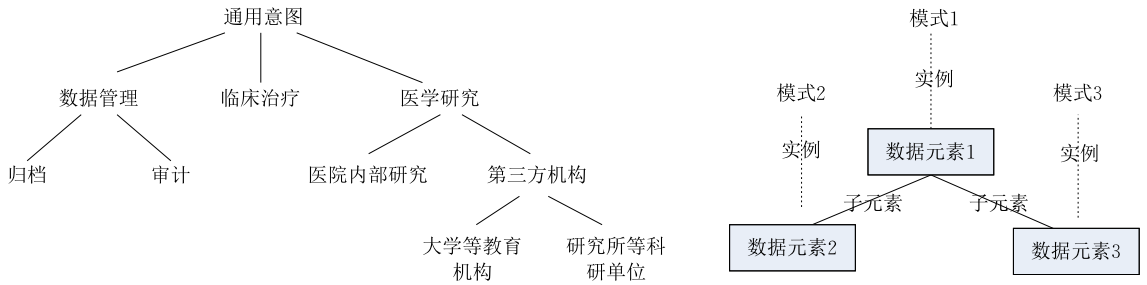
如“问题4”所述，随着大数据应用对个人数据的大量采集、存储和分析，用户的个人隐私正面临着前所未有的严重威胁.在早期数据库的访问控制研究中，文献[43]就提出了十大隐私原则，首次将隐私保护结合到关系型数据库系统中.在该系统架构中，构建了隐私策略和隐私授权两张系统表：隐私策略中定义了每个表的属性的使用意图(usage purpose)、外部接受者、保留时间；隐私授权则定义了每个用户可以利用的意图(purpose).

随后，出现了大量基于意图的访问控制(Purpose-Based Access Control，PBAC)研究，进一步把意图概念作为了访问控制中确保隐私安全的重要基础.例如，文献[44]提出了关系型数据库中的基于意图的访问控制方法，并给出了4种意图标记模式来支持不同的意图标记粒度.文献[45]则进一步提出了包含角色的基于意图的访问控制模型(Role-involved Purpose-based Access Control，RPAC).该模型在基于角色访问控制模型的基础上增加了对访问意图的支持，实现了对数据隐私的细粒度保护.上述研究主要针对的是结构规则的数据客体，因此意图标记可以被扩展到数据类型上.然而对于大数据中的复杂数据客体，例如XML数据，细粒度的意图标记扩展就会存在问题.

针对于此，文献[46]提出了面向XML数据或对象数据等有复杂数据结构的基于意图访问控制模型.该方案将意图分为预期意图和访问意图两种.预期意图被用于隐私需求的规范表达，即“数据客体能够在何种意图下被访问”，所以是关联在数据客体上的；访问意图则表达了某次访问请求的目的性，所以应该被关联在发起请求的主体上.然而为了提高意图的管理效率，该方案又提出了角色属性和系统属性两个概念，并将角色相关的访问意图绑定到角色属性上，角色无关的访问意图绑定到系统属性上.因此，主体在发起访问请求时，会自动获得这两部分访问意图，并用于和被访问数据的预期意图进行比较来实施访问控制.进一步，为了提高意图的授权效率，这种意图之间的比较是可以在树结构上进行推导的.图2(a)展示了不同意图之间的关系，其推导方法为：意图被表示为一棵树，越靠近树根的意图就越泛化，即上层意图被允许/拒绝，则下层意图也将被允许/拒绝.而图2(b)则展示了复杂数据客体的层次结构(类似于树)，其意图的推导方法为：祖先节点允许的意图也应该被子孙节点所允许；祖先节点禁止的意图也应该被子孙节点所禁止.通常情况下，子孙节点是祖先节点的子元素，或者子孙节点是祖先节点的实例.类似地，角色层次上的意图也可以进

行推导.而这种推导方法就使得意图的标记不必详细到每个意图、客体、角色上,极大地简化了管理员的隐私权限管理工作.更进一步,文献[47]将基于意图的访问控制扩展到了流数据的隐私保护中.该方案提出了一个遵循有限采集原则(limited collection)以及有限公开原则(limited disclosure)的数据流管理框架.在该框架下隐私策略的实施被分为两步:(1)查询请求被提交到系统时,就要进行数据流级别和属性级别的访问控制.例如,限定数据流中的某些属性只能在特定的访问意图下才能被访问;(2)元组级别的访问控制则是通过将查询请求进行改写来实现的.例如,通过查询改写对数据流中具体元组的取值范围进行限制.

综上所述,个人隐私问题已经得到了普遍重视,并出现了许多相关研究.这些研究工作引入了意图的概念来保护个人隐私,并解决了许多具体实施中的问题.然而在大数据应用中,个人用户(数据所有者)和数据的规模都将非常庞大,因此隐私策略的制定和管理将是非常繁重的工作.虽然文献[46]给出了根据数据客体层次关系的意图推导方法,减少了意图标记的工作量,但是只适用于特定的具有层次结构的数据集,不具有通用性.在这方面,文献[48]给出了一种解决思路,通过建立描述数据之间语义关系的数据模型,实现隐私策略在数据客体上的推理,进而提高授权的效率.该方案不再仅限于具有层次关系的数据集,所以其应用范围更加广泛.与此同时,大数据环境下,数据隐私的利益相关者可能更加复杂.例如,社交网络中用户发布了一张照片,并且圈出了多个人.在这种情况下,这张照片就涉及到多个人的隐私.如何处理这种复杂的隐私利益关系,制定恰当的访问控制策略仍是一个具有挑战性的问题.这些都将是未来可供进一步深入探讨的研究方向.



(a) 意图的树形结构示例 (b) 数据客体层次结构示例

图 2 意图及复杂数据客体的结构示意图

3.5 世系数据相关的访问控制

世系数据描述了数据产生、并随时间推移而演变的整个过程[49],因此在大数据分析中能够发挥重要作用,也是大数据集的重要组成部分.近年来,世系数据的安全性也已经开始得到重视.相比于传统数据,世系数据不但要保护数据项本身,还要保护数据项之间的演变关系[50-51].虽然描述世系数据的模型存在多种,但是它们都表达了从源数据到目的数据的演变关系.这种数据演变关系以及参与其中的实体之间的关系,本质上仍是一个有向无环图(Directed Acyclic Graph,DAG).因此,对于 DAG 这种图的访问控制是较难通过独立的边或节点的访问控制来实现的.这也是传统访问控制无法直接用于世系数据保护的主要原因.针对与此,文献[52]提出采用 Dependency Path Patterns 作为实施 DAG 数据访问控制的基础单位.由于 Dependency Path Patterns 可以高效地描述用户、进程、数据对象之间的因果依赖关系,所以 DAG 数据的访问控制就可以转化为 Dependency Path Patterns 的模式匹配问题,从而实现了对 DAG 这种数据客体的精确描述.在此基础上,一些传统的访问控制技术就可以基于 Dependency Path Patterns 来实施了.具体地,世系数据访问控制的重要研究包括:首先,文献[53]探讨了 SOA(Service Oriented Architecture)架构的世系系统的安全问题,提出了世系数据的访问控制需求.随后,文献[54]针对来自不同数据源、不同层次的多种世系数据访问控制策略难以描述和实施判定的问题,提出了一个较为通用的支持细粒度访问控制的策略描述语言以及一个灵活的访问控制决策机制.文献[52]提出了 Dependency Path Patterns,解决了世系数据访问控制中客体描述的问题.而文献[55]则进一步提出了针对世系数据对象特点的访问控制策略语言.最近,文献[56]又提出了在软件全生命周期中对世系数据相关的访问控制策略的识别、描述和精化的方法.

与此同时,由于世系数据能够描述业务流程中数据客体的状态变化,因此能够利用它对数据客体进行更加精确的访问控制,也是解决“问题 3”——

系统具有复杂主体构成时造成的安全需求描述困难的一种有效途径.这方面的研究工作主要包括：

文献[57]提出了基于世系数据的访问控制(Provenance-Based Access Control，PBAC)的概念和基础模型，将数据客体的状态作为了访问控制判定的依据，能够进一步表达数据客体处于不同状态时的安全需求.例如，医疗大数据系统中，“医疗缴费通知单”这个客体处于报销后的状态时，是不能再被医院的收费人员或药房护士访问的.图 3 给出了 PBAC 和其他访问控制模型在安全需求描述粒度上的差别，即大多数访问控制模型对客体的访问控制粒度为“主体能够访问客体集合中的哪些客体”，而 PBAC 对客体的访问控制粒度为“主体能够访问处于什么状态下的某客体”.文献[58]提出了对 PBAC 模型的扩展，采用了事务的上下文信息来实施多种动态职责分离(Dynamic Separation of Duties，DSOD)，并基于 XACML(Extensible Access Control Markup Language)架构实现了一个原型来对扩展后的模型进行验证.此外，由于在大数据应用中存在大量的世系实体类型和复杂的世系依赖关系，所以手工编写基于世系数据的安全规则是非常困难的.针对于此，文献[59]给出了一个采用决策树从世系数据中获取访问控制规则的方法.

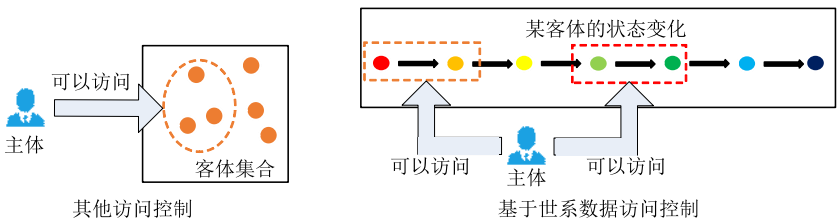


图 3　基于世系数据访问控制与其他访问控制在安全需求描述粒度上的对比

可以看出，一方面，世系数据作为大数据集的重要组成部分，针对其内部关联性保护的访问控制研究是非常必要的；另一方面，世系数据所带来的丰富信息也为更好地实施大数据访问控制提供了良好基础.因此，相关技术的研究也将是未来大数据访问控制的发展方向之一.

3.6　基于密码学的访问控制

传统访问控制方案的安全性依赖于可信的引用监控器，而基于密码学的访问控制方案的安全性则依赖于密钥的安全性，因此能够有效解决前述问题 5——大数据分析架构自身缺乏安全性考虑的问题.一方面，由于大数据分析架构的复杂性，很难建立可信的引用监控器；另一方面，部分大数据场景下，数据是处于所有者控制范围外的.因此，不依赖于可信引用监控器的基于密码学的访问控制研究对于大数据的一些特定场景具有重要意义.

密文访问控制的概念最早由文献[60]提出，首次将访问控制的安全性建立在密钥安全的基础上.该方案采用了双层加密机制，每个文件都会采用一个对称密钥加密，在共享时这些文件会被组织为“组”，并产生一个组密钥负责对每个文件的加密密钥进行加密.文件密文和对应的加密密钥的密文被存储在服务器上，而组密钥则被单独分发给需要共享的用户.该方案的许多基本概念被许多后续研究广泛采用，但是随着“组”的增长，其密钥数量也将线性增长.针对该问题，文献[61]提出了基于代理重加密技术的访问控制方案，即将每个文件用对称密钥加密，再将该加密密钥用文件属主的主密钥加密.文件属主在进行文件分享时，需要用自己的主密钥与目的用户的公钥一起产生一个代理重加密密钥，而服务器将利用该代理重加密密钥对密文进行转换，使得密文只有目的用户才能解密访问.这些研究工作虽然一定程度上满足了数据在缺少可信引用监控器环境下的访问控制需求，但是在实现细粒度和灵活的访问控制时，密钥管理的代价仍然较大，因此影响了其进一步应用推广.

为了解决该问题，一些研究者提出了基于属性加密算法[62-63](Attribute Based Encryption，ABE)的访问控制.如图 4 所示，属性加密是一种一对多的加密模式，能够对加密后的数据进行细粒度的访问控制.只有当用户拥有的属性超过数据所有者加密时设置的属性门槛时，才能解密出数据.文献[64]和文献[65]分别基于 KP-ABE(Key Policy Attribute Based Encryption)和 CP-ABE(Ciphertext Policy Attribute Based Encryption)给出了一套完整的密文访问控制方法.它们假定存储服务器是“诚实而好奇”的，即服务器会忠实执行用户发起的操作，但是却可能泄露数据的内容.基于此前提，数据所有者在持有核心机密的前提下，可以将一部分机密程度较低的权限管理工作(如权限撤销等)委托给服务器执

行.随后,文献[66]又针对文献[64]中使用了固定的ABE随机参数这一安全隐患提出了新的修正方案,解决了原有方案中权限撤销不完全的问题.上述这些方案利用了ABE算法的细粒度特性,实现了灵活的访问控制.

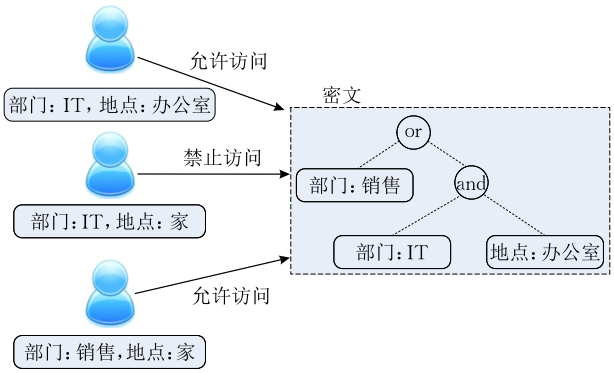


图4 基于属性加密的访问控制

随后,在实际应用中又进一步对基于ABE的访问控制提出了更多目标:多属性权威、加解密外包等.

首先,单属性权威是不满足分布式环境中灵活访问控制需求的,可能成为访问控制的瓶颈.针对该问题,文献[67]提出了支持多属性权威机构的ABE密文访问控制方案.该方案中无中央属性权威机构,能够支持任意线性秘密共享机制(Linear Secret Sharing Scheme,LSSS)访问结构的多属性权威CP-ABE算法,同时还保持了足够低的通信代价.进一步,针对多属性权威时可能出现的合谋攻击,文献[68]又提出了一种能够容忍至多$N-2$个属性权威机构合谋的访问控制方案,提高了多属性权威的基于ABE的访问控制系统的安全性.其系统模型如图5所示,整个属性集被划分为多个部分,分别由不同的属性权威维护.每个属性权威只知道自己维护的那部分属性,并且需要联合起来计算公钥.数据所有者可以从这些属性权威中的任意一个获得公钥,并利用该公钥进行数据加密,然后将数据密文存储在外包服务器上.数据访问者需要向属性权威请求私钥,然后才能解密数据.

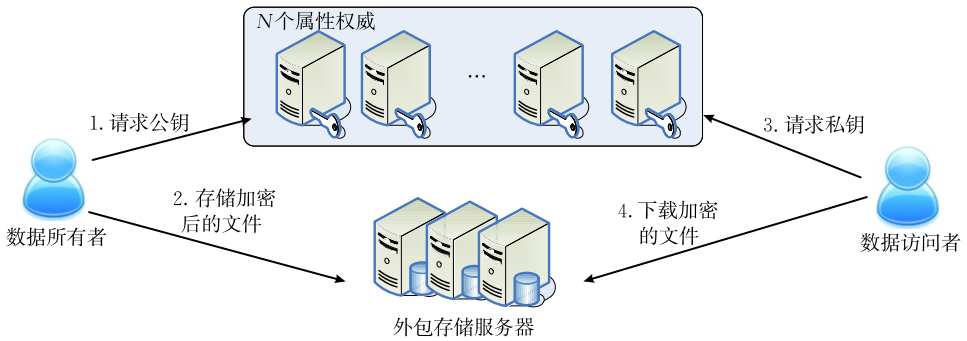


图5 多属性权威的基于属性加密的访问控制

其次,为了提高方案的效率,加解密操作也应该适当地被外包.文献[69]提出了一种可外包解密的密文访问控制方案.该方案中,用户在解密时只需要向外包服务器提供一个转换密钥,外包服务器就可以将任何一个满足该用户解密条件的密文转换为一个El Gamal风格的密文.这样在不降低安全性的前提下,将用户的大部分解密操作转移给了外包服务器.类似地,文献[70]提出了一种可外包加密的密文访问控制方案.该方案将密文加密过程分为两部分,一部分由数据属主执行,一部分由"诚实而好奇"的外包服务器执行.最近,文献[71]又提出了一种满足多属性权威、外包解密目标的ABE方案框架,能够利用已有的单属性权威ABE方案构建出支持外包解密的多属性权威ABE方案,极大地减少了重新设计ABE方案,并证明其安全性的工作量.

最后,虽然基于属性加密的访问控制能够在缺失可信引用监控器的场景下实现可信的访问控制,但是由于属性加密自身的效率较低,所以如何适应大数据应用场景的数据规模和数据增长速度,仍然是有待研究的实际问题.

3.7 访问控制的新特点

随着应用需求和人们认识的发展,访问控制技术也在不断演进变化着:从早期的访问控制矩阵[72-73]等自主访问控制模型(Discretionary Access Control,DAC)和BLP(Bell LaPadula)[74-75]等强制访问控制模型(Mandatory Access Control,MAC)发展到现在仍广泛使用的RBAC[76]模型,以及互联网、分布式系统发展壮大后出现的ABAC[77-80].每一种模型和技术都有其适用的特定应用场景.而随着大数据时代的到来,我们认为访问控制技术也会相应地呈现出新的特点.在本小节中,我们对上述大数据访问控制的关键技术进行了分析,提炼出了3个大数

据访问控制所具有的新特点：判定依据多元化、判定结果模糊(或不确定)化、多种访问控制技术融合化.

(1) 判定依据越来越多元化.

表 2 展示了部分现有访问控制技术的判定依据. 可以看出，早期 DAC 模型的访问控制判定依据是主体的身份，即访问控制的判定是基于“主体身份是否为客体的属主或具有访问客体的相应权限”. MAC 则是基于“主体所具有的安全标记与客体具有的安全标记之间的支配关系”进行判定. RBAC 是基于“主体当前激活的角色是否具有访问客体的相应权限”进行判定. 而从 ABAC 开始，访问控制的依据变得更加灵活，即可以基于主体属性、资源属性、环境属性等进行授权，并实施访问控制.

表 2　判定依据

访问控制	判定依据
自主访问控制	主体的身份是否为客体属主或具有客体相应权限
强制访问控制	主体的标记是否支配客体的标记
基于角色的访问控制	主体当前激活的角色是否具有访问客体的相应权限
基于属性的访问控制	主体属性、资源属性、环境属性等是否相符
基于意图的访问控制	访问请求的意图与客体标记的意图是否相符
基于风险的访问控制	客体访问所带来的风险是否可以被接受
基于世系数据的访问控制	数据客体的状态是否符合预置条件
基于关系的访问控制	访问主体与客体属主之间的关系是否符合预置条件
位置敏感的访问控制	数据存储的位置，以及数据与访问者的位置距离等是否符合预置条件
基于内容的访问控制	主体已拥有的文本内容与客体文本内容之间的相似度是否符合安全要求
基于行为的访问控制	主体历史行为分析结果是否符合安全要求

除了上述常见访问控制模型所采用的判定依据外，文献[43-44]等基于意图的访问控制引入了“意图(Purpose)”作为判定依据；文献[20]等基于风险的访问控制则引入了“风险”作为判定依据；文献[57]等基于世系数据的访问控制引入了“数据的世系”作为判定依据；而文献[36]等基于关系的访问控制引入了“关系”作为判定依据；文献[81]位置敏感的访问控制引入了“数据存储位置”作为判定依据，更进一步文献[82]甚至引入了数据与用户之间的距离作为判定依据. 例如，某重要数据客体在被访问时，安全管理员必须在客体附近等. 上述这些多样化的判定依据可以极大地丰富访问控制策略的描述能力，支持更复杂的访问控制需求. 例如，基于意图的访问控制可以更好地表达隐私保护策略；基于世系数据的访问控制则能够将访问控制粒度控制到客体的某个状态，更好地表达客体处于不同状态时的安全需求；基于关系的访问控制则能够允许用户按照目标用户与自己的关系种类和关系远近来分享数据.

近期，随着大数据分析技术的发展，一些研究人员开始将大数据分析的结果作为访问控制判定的依据融入访问控制模型中. 例如，文献[8,42]中提出的基于内容的访问控制模型 CBAC，将用户已经拥有的文档内容与请求文档的内容之间的相似性分析结果作为判定条件；文献[83]提出的基于行为的访问控制，则采用机器学习算法对企业参与者的多种日常行为进行分析，并将该分析结果作为访问控制的判定条件. 在引入这种动态的数据分析结果后，访问控制策略的判定条件已经变得越来越复杂化.

可以预见，由于大数据应用的复杂性和安全需求的多样性，未来会出现更多具有多元化访问控制判定依据的安全模型.

(2) 判定结果由简单二值决策向模糊、不确定性决策发展.

从判定结果来看，大数据兴起之前的访问控制模型对访问请求的判定结果多为二值“允许/拒绝”. 目前，这种简单的判定结果开始逐渐无法适应大数据场景.

首先，由于大数据应用的复杂性将导致授权不足的现象频繁发生，即用户的访问请求仅满足部分访问条件. 若简单地采用“允许/拒绝”作为判定结果，将极大地影响业务的正常进行. 针对这种问题，文献[24]给出了基于风险带(Risk Band)的解决方法，其访问控制判定的结果是“访问行为处于哪个风险带”，然后再根据风险带的不同，给予其不同的权限. 因此，该模型实现了“符合部分访问控制条件的请求，获得部分访问权限”的访问控制. 解决该问题的另一种思路，是文献[84]给出的一种属性值缺失情况下的基于属性访问控制的判定方法. 该方法假设判定条件内的各属性之间是无关的，然后通过各属性值出现的概率计算出允许、拒绝、不适用三种判定结果各自的概率区间. 这样最后的判定结果就变为不确定的决策值，而访问控制执行组件就可以进一步基于这些值实现更加灵活的访问控制，以提高系统的可用性.

其次，如前所述，大数据应用的访问控制判定条件中可能包含数据分析的结果. 而相比于其他访问控制判定条件，数据分析结果往往具有较高的误报率，因此基于这种分析结果来给出“允许/拒绝”的二

值判定,将影响整个系统的可用性.针对该问题,文献[8,42]在利用用户已有文档与请求文档的内容相似度分析结果作为判定条件时,采用了 top-K 的方式来模糊访问控制的判定结果.即该方案会将文档内容之间的相似度进行排序,并根据管理员预先设定的 K 值选出最相关的 K 个文档返回给用户,进而减小相似度分析结果误报率对访问控制结果的影响.

由此可见,未来大数据场景下的访问控制判定结果将不仅是允许/拒绝的二值形式,还会呈现模糊、不确定的特征.

(3)多种访问控制技术的融合.

如前所述,随着大数据分析技术的发展,出现了一批在访问控制中结合数据分析的方案,例如,文献[7,26]对医生访问行为的历史数据进行分析,并在此基础上评估访问风险;文献[8,42]对用户持有数据和访问请求数据之间的相似度进行分析,并在此基础上实施访问控制等.虽然它们具有传统的基于静态规则的访问控制所没有的一些优势,但是这些访问控制技术通常是不能独自应用于大数据系统的.

首先,这些访问控制技术中数据分析结果的准确性受被分析数据集的影响很大.而在系统初始情况下,用于支持访问控制中数据分析的数据集往往是不完整的.例如,文献[7,26]的方案在系统初始时,会由于缺少用于评估风险值的历史数据而造成风险评估不准确,进而影响访问控制的准确性.其次,数据分析是存在误报率的,这会使得访问控制存在一定错误.这种错误如果发生在一些需要严格遵守的访问控制规则上,就会对系统安全将造成严重影响.而传统的严格执行预先定义规则的访问控制系统则恰恰能够弥补这点.因此,将这两类技术进行结合是非常有意义的.

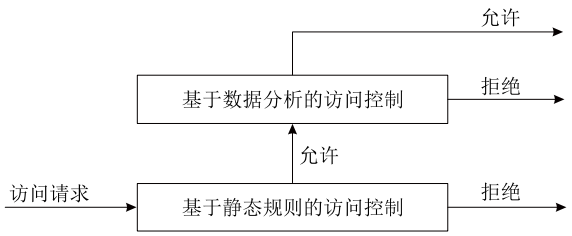


图 6 访问控制技术的结合方法

我们可以借鉴文献[83]提出的一种修正控制的架构来融合这两类访问控制技术,但是在修正方式上有所不同.文献[83]采取的方式是:将一些初始判定流程所禁止的访问请求,通过进一步判定可以修正为允许.而本文认为大数据访问控制更加需要的修正方式是:先进行粗粒度的初始访问控制判定流程,在该流程中阻止一些严格禁止的访问请求,随后通过进一步的细粒度访问控制流程将初始判定流程所允许的部分访问请求修正为禁止,从而实现访问控制的逐步精化.具体实施方法如下:

访问控制阶段 1(粗粒度、严格):传统的基于静态规则的访问控制适合用来描述和实施粗粒度的,或需要严格遵守的访问控制规则.一方面,这是由于前文所阐述的大数据场景下,精细地进行策略设计和授权将是非常繁重的任务;另一方面,基于静态规则的访问控制能够在重要访问控制规则上消除数据分析带来的错误率.

访问控制阶段 2(细粒度、宽松):采用了数据分析的访问控制适合用来描述和实施更细粒度的访问控制规则,以解决人工难以预先详细描述所有规则的问题.

这两个阶段的结合方式如图 6 所示,访问请求先由基于静态规则的访问控制模块进行判定,然后再由采用数据分析的访问控制模块在更细粒度上进一步判定.只有当两个模块判定结果都为"允许"时,才能让该访问请求通过.

4 研究展望

综上所述,大数据领域已经出现了一些有针对性和可借鉴的访问控制技术.然而这些技术都不能完美地解决大数据访问控制问题,还存在许多具有挑战性的工作.本节首先对一些有待研究的问题进行归纳总结,再抛砖引玉地提出一个未来可能的大数据访问控制框架供大家探讨.

4.1 大数据价值的评估

"哪些数据是有价值的"是访问控制甚至是信息安全首先要解决的基本问题.也就是说,它关系到确定"哪些数据要进行访问控制保护,用什么策略保护,保护到什么程度".目前人们已经充分认识到了数据的价值,并开始在公共卫生、商业、科学研究等领域应用大数据分析技术来解决问题.然而,区别于传统信息系统,大数据应用往往较难识别数据的价值.主要原因在于,一方面,大数据应用往往是先采集和存储大数据,然后才有针对性地寻找挖掘数据价值的方法.也就是说,相比于传统数据分析系统在开始阶段就明确了数据的价值,大数据应用在数据采集和存储时往往对数据价值的了解不足.另一方面,由于数据的规模性和价值稀疏的特征,有价值的

数据往往隐藏在海量低价值数据中，或者是低价值的数据聚集到一定规模才能体现出真正价值.在这种情况下，大数据应用中哪些数据需要被保护，采用什么样的策略来保护是非常难以准确把握的.因此，如何评估大数据价值，或者如何在缺少对大数据价值准确认识的情况下确保数据的安全分享和使用，将是大数据访问控制领域未来非常具有挑战性和意义的工作.

4.2　大数据相关者的利益平衡

相比于传统信息系统，大数据应用系统不仅仅是技术架构上更加复杂，其利益相关者也更加复杂.企业或组织采集和存储的数据，其所有权和使用权归谁，仍然是一个争论较多的问题.例如，在社交网络中，个人的属性信息以及所发布的文字、图片内容都会被社交网站所存储.社交网站是否可以任意使用这些数据，甚至授权第三方访问这些数据？社交网络中一个用户发布了他与朋友们的合影，并圈出了一些朋友，那么这张照片的所有权归谁，应该由谁决定这张照片能够被谁访问?这些问题涉及了道德、法律法规和技术等多个层面.在道德、法律法规未明确给出答案的时候，从技术角度我们是可以提前考虑一些模型、方案来支持数据相关者之间利益的平衡.例如，在众多数据利益相关者中，由谁来制定策略，策略如何实施，以及利益相关者相互之间的信任问题等，都是未来大数据访问控制应该关注的.此外，这种利益关系的平衡方案需要综合考虑多种因素，例如数据的访问意图、数据访问者之间的关系、数据世系等，因此将是非常具有挑战性的研究工作.

4.3　大数据架构的安全与效率

大数据从采集、存储，一直到处理和分享等环节，是对现有的物联网、云计算等技术的综合应用.而这些技术架构的主要目的在于解决传统技术方案所无法胜任的数据集规模太大、种类复杂、增长速度快等问题，进而高效地完成数据采集、存储、使用等任务.因此，这些架构注重的是效率，而对于安全问题的考虑往往是滞后的.例如，现有的关系数据库的安全策略通常是经过详细设计，甚至是形式化验证过的.相比之下，用于存储半结构化、非结构化数据的 NoSQL 数据库中，大多仅有基本的访问控制功能，其策略也是非常简陋的，甚至有些产品自身没有访问控制功能，而交由访问数据库的应用程序来实现.这就导致了这些大数据架构在功能和性能需求方面是满足大数据应用的，但是在安全需求方面却相差甚远.

本文第 3 节所列举的仅是一些可以被大数据应用场景借鉴的访问控制技术，但是其实施效率如何能够满足大数据架构的需求，仍然是非常值得研究的问题.在这方面，可以借鉴文献[85]提到的基于视图的访问控制，这种访问控制方式允许预先定义数据的安全视图，然后对安全视图的访问进行控制.文献[47]在实际的流数据访问控制中采用了这种方式，在数据流和属性层次采用安全视图进行数据过滤，并限制访问者只能通过安全视图对数据进行访问.这种实施方式满足了流数据处理系统对数据访问的效率要求.但是正如文献[85]所述那样，基于视图的访问控制的一大缺陷就是安全视图的定义较为困难.在仅有数据的时候，安全管理员很难预先设计恰当的安全视图，既能满足访问者的访问需求，又能保护数据的安全.因此，如何在大数据架构下获得安全与效率的平衡，是未来访问控制领域非常具有现实意义的研究工作.

4.4　大数据服务于访问控制

大数据在为访问控制带来问题的同时，也带来了发展机遇.如前所述，在庞大而复杂的大数据应用中“谁能访问什么”的设计，已经不是单个管理员能够胜任的.该工作不但是劳动密集型工作，而且需要大量专业知识.而由于大数据分析技术能够帮助人们洞察隐藏在数据之间的关联、规律等信息，所以非常适合在访问控制中用来挖掘访问控制规则，甚至是实时调整访问控制权限.这方面的研究可以分为两类：基于协同的访问控制和基于数据联系的访问控制.

基于协同的访问控制是指利用大量用户的访问历史来实施访问控制.虽然管理员缺少专业知识，无法进行细粒度的权限管理，但是从大量正常用户的访问历史数据中，可以采用数据分析技术获得“谁经常访问什么”.在假定这些历史数据中包含的非正常访问行为是较少量的情况下，“谁经常访问什么”就能够作为访问控制策略“谁能访问什么”.例如，一个医疗大数据系统中，管理员已经很难识别出“医生可以访问哪些病人资料”了，但是数据分析技术能够从正常治疗流程中大量医生对病人资料的访问历史，医生与病人之间的关系，发起访问请求的医生的历史行为等多种数据中找出答案.

与基于协同的访问控制思路不同，基于数据联系的访问控制是采用数据之间广泛存在的联系来简化访问控制中的授权管理工作.通常情况下，大数据集中包含了大量解释性的数据.这些解释性数据可

以供我们在数据之间建立各种关联,例如世系关联、语义关联、包含或引用等结构关联.在此基础上,可以定义访问权限在这些关联关系上的推导规则(授权规则).这样就可以减少管理员的权限管理操作,管理员只需要针对部分数据进行权限的管理.而未明确指定访问控制规则的数据或新增数据的访问控制规则将根据它们与其他有明确规则的数据之间的关系进行推导.

目前,无论是基于协同的访问控制,还是基于数据联系的访问控制,都已经有了一些研究工作,但是这些研究才刚刚开始,还未充分挖掘和利用大数据的价值.因此,哪些数据是对访问控制有用的,应该如何利用等问题仍然是有待研究的重要方向.

4.5 大数据的访问控制框架

针对前述大数据访问控制的一些问题和若干关键技术,本文提出一种未来可能的大数据访问控制框架.该框架目前仅限于概念层面,还有许多问题有待未来进一步研究.它主要包括4个核心概念:安全视图、访问控制分层、策略挖掘、访问控制反馈.在大数据的不同应用场景下,这些概念可以根据需要有选择的实施.

安全视图在本文中不局限于数据库视图,可以认为是一组限定如何从数据集中抽取部分数据的过滤规则.其好处在于:(1)可以灵活定义客体的保护粒度;(2)过滤的方式简化了访问控制判定过程,能够极大提高效率[47].

访问控制分层是指将基于安全视图的访问控制方式与其他访问控制方式结合起来分层实施.也就是说,在面对处理效率要求高,数据的结构复杂,且数据集规模大甚至没有边界(例如流数据)的情况,采用基于安全视图的方式;而在面对处理效率要求不高,数据的结构较为规范,且数据集规模可接受的情况,则采用直接针对数据客体的访问控制方式.由于安全视图本身也可以看作是一种数据客体,因此这两种访问控制方式是可以根据大数据的实际应用需求进行分层部署的.

策略挖掘是针对前述大数据场景下访问控制的过度授权和授权不足问题而提出的.它支持用数据挖掘或机器学习的算法从各类系统信息中获取访问控制所需要的策略.例如,可以基于世系数据[57]或访问日志进行策略挖掘[19]等.

访问控制反馈是指主体的访问请求在经过访问控制判定和实施后,会改变部分访问控制信息,进而动态地影响访问控制的决策结果,使整个访问控制过程呈现自适应的特点.例如,前文中描述的基于行为的访问控制[83]和风险自适应的访问控制[7,26],都将访问控制结果作为后续访问控制的输入返回给了访问控制系统.

具体地,框架的组成部分以及这些组成部分的相互关系如图7所示.

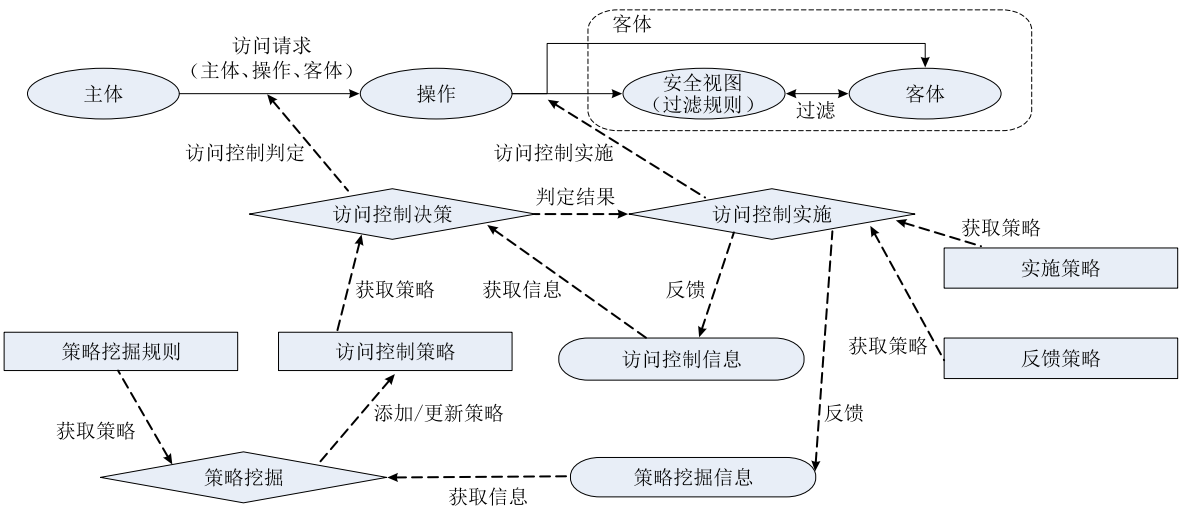


图7 一种大数据访问控制框架

主体负责发起对客体执行操作的访问请求.操作则是预先在系统中定义好的一组操作类型的实例.需要注意的是,基于分层实施的概念,访问请求中的客体既可以是能够被用户直接操作的普通客体,也可以是一些安全视图.一个安全视图就是一组过滤规则的集合.过滤规则描述了哪些客体能够通过安全视图的过滤而被用户访问到,所以主体对安全视图的访问请求也可以看作是主体对安全视图过滤后的普通客体的访问请求.

访问控制决策组件负责对访问请求进行判定.

在进行访问控制判定时，需要获取访问控制策略和访问控制信息.访问控制策略描述了访问请求的主体、客体、系统环境满足何种条件时，应该具有何种访问权限.而访问控制信息则记录了主体、客体、系统环境的相关状态.因此，访问控制决策组件只需要根据访问控制策略和对应的访问控制信息就能够对访问请求做出判定.其判定结果可以是简单的“允许/拒绝”，也可以是 3.7 节中描述的更加复杂的判定结果，例如此次访问的风险值.

访问控制实施组件负责根据判定结果来对主体发出的访问请求进行控制.与传统访问控制不同的是，本框架的访问控制实施组件具有根据实施策略来允许主体访问部分客体的能力，而不仅仅是实施简单的允许/拒绝.实施策略应该与判定结果相配合，例如实施策略可以规定“访问风险位于某风险带时，仅能读取数据，而不能修改数据”.在这种情况下，主体对某数据客体的读写请求，将被访问控制实施组件根据该请求的风险在风险带中的位置来进一步灵活控制.在实施访问控制后，访问控制实施组件还需要获取反馈策略，根据该策略的定义去更新访问控制信息和策略挖掘信息，从而影响下一次的访问控制决策以及策略挖掘结果.其中，反馈策略描述了“某访问控制实施的结果应该如何更新访问控制信息和策略挖掘信息”.例如，用户的访问请求具有风险，且该风险小于用户剩余的访问配额时，则访问控制决策组件允许该请求，而访问控制实施组件在实施了决策结果后，必须从用户剩余的访问配额中减去该请求的风险评估值.那么，随着用户不断地消耗访问配额，他最后就无法再完成高风险的访问请求.这种反馈过程就使得访问控制系统具有了自适应的特点.

最后，策略挖掘组件负责根据策略挖掘规则从策略挖掘信息中分析挖掘出新的访问控制策略，并将其更新到系统的访问控制策略集合中.其中，策略挖掘信息是指策略分析的数据对象，包括系统的业务数据和访问控制的历史数据.而策略挖掘规则则描述了如何对这些数据对象进行分析.通常情况下，不同类型的访问控制策略的挖掘方法差异较大，因此这些策略挖掘规则往往是和具体的策略挖掘任务相关的，并在进行策略挖掘时由管理员定义和选择.

5　总　结

大数据时代的到来，为访问控制技术提出了新的问题和挑战.本文基于大数据及其应用的新特点，分析归纳出了 5 个大数据访问控制迫切需要解决的问题.接着从这些问题出发，梳理了当前大数据相关的一些访问控制关键技术.然后在此基础上提炼了大数据访问控制所呈现的一些新特点：判定依据多元化、判定结果模糊(或不确定)化、多种访问控制技术融合化.最后，对未来大数据访问控制的研究进行了展望，给出了若干未来值得进一步研究的问题，并提出了一个针对前述大数据访问控制问题和特点的概念性访问控制框架供大家探讨.总体而言，当前大数据访问控制相关的研究还处于一个初级阶段，许多问题的研究还不够充分.大数据对于访问控制领域即是一个挑战，也同样是一个机遇.这方面的研究还有很长的路要走，我们期望本文能够抛砖引玉，为同行的专家学者提供一定参考，共同推进大数据时代的访问控制技术发展.

参　考　文　献

[1] Mayer-Schonberger V, Cukier K. Big Data: A Revolution That Will Transform How We Live, Work and Think. Boston, USA: Houghton Mifflin Harcourt, 2013

[2] Li Guo-Jie, Cheng Xue-Qi. Research status and scientific thinking of big data. Bulletin of Chinese Academy of Sciences, 2012, 27(6): 647-657(in Chinese)
(李国杰，程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域. 中国科学院院刊，2012，27(6)：647-657)

[3] Gong Xue-Qing, Jin Che-Qing, Wang Xiao-Ling, et al. Data-intensive science and engineering: Requirements and challenges. Chinese Journal of Computers, 2012, 35(8): 1563-1578(in Chinese)
(宫学庆，金澈清，王晓玲等. 数据密集型科学与工程：需求和挑战. 计算机学报，2012，35(8)：1563-1578)

[4] Zhang X, Liu C, Nepal S, et al. Privacy-preserving layer over MapReduce on cloud//Proceedings of the Second IEEE International Conference on Cloud and Green Computing (CGC). Xiangtan, China, 2012: 304-310

[5] Feng Deng-Guo, Zhang Min, Li Hao. Big data security and privacy protection. Chinese Journal of Computers, 2014, 37(1): 246-258(in Chinese)
(冯登国，张敏，李昊. 大数据安全与隐私保护. 计算机学报，2014，37(1)：246-258)

[6] Meng Xiao-Feng, Ci Xiang. Big data management: Concepts, techniques and challenges. Journal of Computer Research and Development, 2013, 50(1): 146-169(in Chinese)
(孟小峰，慈祥. 大数据管理：概念、技术与挑战. 计算机研究与发展，2013，50(1)：146-169)

[7] Wang Q, Jin H. Quantified risk-adaptive access control for patient privacy protection in health information systems//

Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security (ASIACCS). Hong Kong, China, 2011: 406-410

[8] Zeng W, Yang Y, Luo B. Access control for big data using data content//Proceedings of the 2013 IEEE International Conference on Big Data. Santa Clara, USA, 2013: 45-47

[9] Bishop M. Computer Security: Art and Science. Boston, USA: Addison-Wesley Professional, 2002

[10] Gates C E. Access control requirements for Web 2.0 security and privacy//Proceedings of the IEEE Web 2.0 Privacy and Security Workshop (W2SP). Oakland, USA, 2007: 249-256

[11] Li Guo-Jie. Scientific value of research on big data. Communications of the CCF, 2012, 8(9): 8-15(in Chinese)
(李国杰. 大数据研究的科学价值. 中国计算机学会通讯, 2012, 8(9): 8-15)

[12] Hu V C, Grance T, et al. An access control scheme for big data processing//Proceedings of the 2014 International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom). Miami, USA, 2014: 1-7

[13] Zhai Zhi-Gang, Wang Jian-Dong, Cao Zi-Ning, Mao Yu-Guang. Hybrid role mining methods with minimal perturbation. Journal of Computer Research and Development, 2013, 50(5): 951-960(in Chinese)
(翟志刚, 王建东, 曹子宁, 毛宇光. 最小扰动混合角色挖掘方法研究. 计算机研究与发展, 2013, 50(5): 951-960)

[14] Kuhlmann M, Shohat D, Schimpf G. Role mining-revealing business roles for security administration using data mining technology//Proceedings of the 8th ACM Symposium on Access Control Models and Technologies (SACMAT). Villa Gallia, Italy, 2003: 179-186

[15] Schlegelmilch J, Steffens U. Role mining with ORCA//Proceedings of the 10th ACM Symposium on Access Control Models and Technologies (SACMAT). Stockholm, Sweden, 2005: 168-176

[16] Vaidya J, Atluri V, Warner J. RoleMiner: Mining roles using subset enumeration//Proceedings of the 13th ACM Conference on Computer and Communications Security (CCS). Alexandria, USA, 2006: 144-153

[17] Jafarian J H, Takabi H, Touati H, et al. Towards a general framework for optimal role mining: A constraint satisfaction approach//Proceedings of the 20th ACM Symposium on Access Control Models and Technologies (SACMAT). Vienna, Austria, 2015: 211-220

[18] Frank M, Basin D, Buhmann J M. A class of probabilistic models for role engineering//Proceedings of the 15th ACM Conference on Computer and Communications Security (CCS). Alexandria, USA, 2008: 299-310

[19] Molloy I, Park Y, Chari S. Generative models for access control policies: Applications to role mining over logs with attribution//Proceedings of the 17th ACM Symposium on Access Control Models and Technologies (SACMAT). Newark, USA, 2012: 45-56

[20] Office J P. Horizontal integration: Broader access models for realizing information dominance. The MITRE Corporation, Virginia: Technical Report JSR-04-132, 2004

[21] Zhang Lei, Brodsky A, Jajodia S. Toward information sharing: Benefit and risk access control (BARAC)//Proceedings of the 7th IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY). London, Canada, 2006: 45-53

[22] Dimmock N, Belokosztolszki A, Eyers D, et al. Using trust and risk in role-based access control policies//Proceedings of the 9th ACM Symposium on Access Control Models and Technologies (SACMAT). Yorktown Heights, USA, 2004: 156-162

[23] Nissanke N, Khayat E J. Risk based security analysis of permissions in RBAC//Proceedings of the 2nd International Workshop on Security in Information Systems (WOSIS). Porto, Portugal, 2004: 332-341

[24] Cheng P C, Rohatgi P, Keser C, Karger P A, et al. Fuzzy multi-level security: An experiment on quantified risk-adaptive access control//Proceedings of the IEEE Symposium on Security and Privacy (S&P). Oakland, USA, 2007: 222-230

[25] Ni Q, Bertino E, Lobo J. Risk-based access control systems built on fuzzy inferences//Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security (ASIACCS). Beijing, China, 2010: 250-260

[26] Hui Zhen, Li Hao, Zhang Min, Feng Deng-Guo. Risk-adaptive access control model for big data in healthcare. Journal on Communications, 2015, 36(12): 190-199(in Chinese)
(惠榛, 李昊, 张敏, 冯登国. 面向医疗大数据的风险自适应的访问控制模型. 通信学报, 2015, 36(12): 190-199)

[27] Abiteboul S. Querying semi-structured data//Proceedings of the 6th International Conference on Database Theory (ICDT). Delphi, Greece, 1997: 1-18

[28] Meng Xiao-Feng. An overview of web data management. Journal of Computer Research and Development, 2001, 38(4): 385-395(in Chinese)
(孟小峰. Web 数据管理研究综述. 计算机研究与发展, 2001, 38(4): 385-395)

[29] Damjani E, Vimercati S, Paraboschi S, et al. Design and implementation of access control processor for XML documents. Computer Network, 2000, 33(1-6): 59-75

[30] Zhu Hong, Lu Kevin, Jin Renchao. A practical mandatory access control model for XML databases. Information Sciences, 2009, 179(8): 1116-1133

[31] Chebotko A, Chang S, Lu S, Fotouhi F. Secure XML querying based on authorization graphs. Information Systems Frontiers, 2012, 14(3): 617-632

[32] Qi N, Kudo M, Myllymaki J, Pirahesh H. A function-based access control model for XML databases//Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM). Bremen, Germany, 2005: 115-122

[33] Finance B, Medjdoub S, Pucheral P. The case for access control on XML relationships//Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM). Bremen, Germany, 2005:107-114

[34] Bravo L, Segovia R. Simplified access control policies for XML databases//Proceedings of the 6th Alberto Mendelzon International Workshop on Foundations of Data Management (AMW). Ouro Preto, Brazil, 2012: 20-34

[35] Jo Sun-Moon, Chung Kyung-Yong. Design of access control system for telemedicine secure XML documents. Multimedia Tools and Applications, 2015, 74(7): 2257-2271

[36] Fong P W. Relationship-based access control: Protection model and policy language//Proceedings of the 1st ACM Conference on Data and Application Security and Privacy (CODASPY). San Antonio, USA, 2011: 191-202

[37] Fong P W, Siahaan I. Relationship-based access control policies and their policy languages//Proceedings of the 16th ACM Symposium on Access Control Models and Technologies (SACMAT). Innsbruck, Austria, 2011: 51-60

[38] Bruns G, Fong P W, Siahaan I, Huth M. Relationship-based access control: Its expression and enforcement through hybrid logic//Proceedings of the 2nd ACM Conference on Data and Application Security and Privacy (CODASPY). Antonio, USA, 2012: 117-124

[39] Mehregan P, Fong P W L. Policy negotiation for co-owned resources in relationship-based access control//Proceedings of the 21st ACM Symposium on Access Control Models and Technologies (SACMAT). Shanghai, China, 2016: 125-136

[40] Rizvi S Z, Fong P W, Crampton J, Sellwood J. Relationship-based access control for an open-source medical records system//Proceedings of the 20th ACM Symposium on Access Control Models and Technologies (SACMAT). Vienna, Austria, 2015: 113-124

[41] Cheng Y, Bijon K, Sandhu R. Extended ReBAC administrative models with cascading revocation and provenance support//Proceedings of the 21st ACM Symposium on Access Control Models and Technologies (SACMAT). Shanghai, China, 2016: 161-170

[42] Zeng W, Yang Y, Luo B. Content-based access control: Use data content to assist access control for large-scale content-centric databases//Proceedings of the 2014 IEEE International Conference on Big Data (Big Data). Washington DC, USA, 2014: 701-710

[43] Agrawal R, Kiernan J, Srikant R, Xu Y. Hippocratic databases//Proceedings of the 28th International Conference on Very Large Databases (VLDB). Hong Kong, China, 2002: 143-154

[44] Byun J, Li N. Purpose based access control for privacy protection in relational database systems. The VLDB Journal, 2008, 17(4): 603-619

[45] Kabir E M, Wang H, Bertino E. A role-involved purpose-based access control model. Information Systems Frontiers, 2012, 14(3): 809-822

[46] Byun J, Bertino E, Li N. Purpose based access control of complex data for privacy protection//Proceedings of the 10th ACM Symposium on Access Control Models and Technologies (SACMAT). Stockholm, Sweden, 2005: 102-110

[47] Ng W, Wu H, Wu W, et al. Privacy preservation in streaming data collection//Proceedings of the 18th IEEE International Conference on Parallel and Distributed Systems. Singapore, 2012: 810-815

[48] Paci F, Zannone N. Preventing information inference in access control//Proceedings of the 20th ACM Symposium on Access Control Models and Technologies (SACMAT). Vienna, Austria, 2015: 87-97

[49] Gao Ming, Jin Che-Qing, Wang Xiao-Ling, et al. A survey on management of data provenance. Chinese Journal of Computers, 2010, 33(3): 373-389(in Chinese)
(高明，金澈清，王晓玲等. 数据世系管理技术研究综述. 计算机学报，2010，33(3)：373-389)

[50] Braun U, Shinnar A. A security model for provenance. Harvard University Computer Science, Cambridge: Technical Report TR-04-06, 2006

[51] Braun U, Shinnar A, Seltzer M. Securing provenance//Proceedings of the 3rd USENIX Workshop on Hot Topics in Security (HotSec). San Jose, USA, 2008: 41-45

[52] Nguyen D, Park J, Sandhu R. Dependency path patterns as the foundation of access control in provenance-aware systems//Proceedings of the 4th USENIX Conference on the Theory and Practice of Provenance (TaPP). Boston, USA, 2012: 1-4

[53] Tan V, Groth P, Miles S, et al. Security issues in a SOA-based provenance system//Proceedings of the International Provenance and Annotation Workshop (IPAW). Chicago, USA, 2006: 203-211

[54] Ni Qun, Xu Shouhuai, Bertino E, et al. An access control language for a general provenance model//Proceedings of the 6th VLDB Workshop on Secure Data Management (SDM). Lyon, France, 2009: 68-88

[55] Cadenhead T, Khadilkar V, Kantarcioglu M, Thuraisingham B. A language for provenance access control//Proceedings of the 1st ACM Conference on Data and Application Security and Privacy (CODASPY). San Antonio, USA, 2011: 133-144

[56] Sun Lianshan, Park J, Sandhu R. Engineering access control policies for provenance-aware systems//Proceedings of the 3rd ACM Conference on Data and Application Security and Privacy (CODASPY). San Antonio, USA, 2013: 285-292

[57] Park J, Nguyen D, Sandhu R. A provenance-based access control model//Proceedings of the 10th Annual International Conference on Privacy, Security and Trust (PST). Paris, France, 2012: 137-144

[58] Nguyen D, Park J, Sandhu R. A provenance-based access control model for dynamic separation of duties//Proceedings of the 11th Annual International Conference on Privacy,

Security and Trust (PST). Tarragona, Spain, 2013: 247-256

[59] Pei J, Ye X. Towards policy retrieval for provenance based access control model//Proceedings of the 13th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Beijing, China, 2014: 769-776

[60] Kallahalla M, Riedel E, Swaminathan R, et al. Plutus: Scalable secure file sharing on untrusted storage//Proceedings of the Usenix Conference on File and Storage Technologies (FAST). San Francisco, USA, 2003: 29-42

[61] Ateniese G, Fu K, Green M, et al. Improved proxy re-encryption schemes with applications to secure distributed storage. ACM Transactions on Information and System Security, 2006, 9(1): 1-30

[62] Sahai A, Waters B. Fuzzy identity-based encryption// Proceedings of the 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques. Aarhus, Denmark, 2005: 457-473

[63] Goyal V, Pandey O, Sahai A, et al. Attribute-based encryption for fine-grained access control of encrypted data// Proceedings of the 13th ACM Conference on Computer and Communications Security (CCS). Alexandria, USA, 2006: 89-98

[64] Yu S, Wang C, Ren K, et al. Achieving secure, scalable, and fine-grained data access control in cloud computing// Proceedings of the 29th International Conference on Computer Communications(INFOCOM). San Diego, USA, 2010: 534-542

[65] Hur J, Noh D K. Attribute-based access control with efficient revocation in data outsourcing systems. IEEE Transactions on Parallel and Distributed Systems, 2011, 22(7): 1214-1221

[66] Hong Cheng, Zhang Min, Feng Deng-Guo. Achieving efficient dynamic cryptographic access control in cloud storage. Journal on Communications, 2011, 32(7): 125-132(in Chinese)
(洪澄, 张敏, 冯登国. 面向云存储的高效动态密文访问控制方法. 通信学报, 2011, 32(7): 125-132)

[67] Yang K, Jia X. Attributed-based access control for multi-authority systems in cloud storage//Proceedings of the IEEE 32nd International Conference on Distributed Computing Systems (ICDCS). Macau, China, 2012: 536-545

[68] Jung T, Li X Y, Wan Z, et al. Privacy preserving cloud data access with multi-authorities//Proceedings of the International Conference on Computer Communications (INFOCOM). Turin, Italy, 2013: 2625-2633

[69] Green M, Hohenberger S, Waters B. Outsourcing the decryption of ABE ciphertexts//Proceedings of the 20th USENIX Security Symposium. San Francisco, USA, 2011: 34-34

[70] Zhou Z, Huang D. Efficient and secure data storage operations for mobile cloud computing//Proceedings of the 8th International Conference on Network and Service Management (CNSM). Las Vegas, USA, 2012: 37-45

[71] Sherman S M C. A framework of multi-authority attribute-based encryption with outsourcing and revocation//Proceedings of the 21st ACM Symposium on Access Control Models and Technologies (SACMAT). Shanghai, China, 2016: 215-226

[72] Lampson B. Protection. Operating Systems Review, 1974, 8(1): 18-24

[73] Graham G, Denning P. Protection: Principles and practice// Proceedings of the 1972 Spring Joint Computer Conference on American Federation of Information Processing Societies (AFIPS). Atlantic City, USA, 1972: 417-429

[74] Bell D, LaPadula L. Secure computer systems: Mathematical foundations. MITRE Corporation, Virginia: Technical Report MTR-2547, 1973

[75] Bell D, LaPadula L. Secure computer systems: Unified exposition and multics interpretation. MITRE Corporation, Virginia: Technical Report MTR-2997, 1975

[76] Sandhu R S, Coyne E J, Feinstein H L, et al. Role-based access control models. IEEE Computer, 1996, 29(2): 38-48

[77] Wang Ling-Yu, Duminda W, Sushil J. A Logic-based framework for attribute based access control//Proceedings of the 2004 ACM Workshop on Formal Methods in Security Engineering (FMSE). Washington DC, USA, 2004: 45-55

[78] Zhang Xin-Wen, Li Ying-Jiu, Divya N. An attribute-based access matrix model//Proceedings of the 2005 ACM Symposium on Applied Computing (SAC). Santa Fe, USA, 2005: 359-363

[79] Hu V, Kuhn D, Ferraiolo D. Attribute-based access control. IEEE Computer, 2015, 48(2): 85-88

[80] Liu C, Hsien W, Yang C, Hwang M. A survey of attribute-based access control with user revocation in cloud data storage. Network Security, 2016, 18(5): 900-916

[81] Li J, Squicciarini A, Lin D, et al. SecLoc: Securing location-sensitive storage in the cloud//Proceedings of the 20th ACM Symposium on Access Control Models and Technologies (SACMAT). Vienna, Austria, 2015: 51-61

[82] Oluwatimi O, Midi D, Bertino E. A context-aware system to secure enterprise content//Proceedings of the 21st ACM Symposium on Access Control Models and Technologies (SACMAT). Shanghai, China, 2016: 63-72

[83] Cleveland J, Mayhew M, Adler A, Atighetchi M. Scalable machine learning framework for behavior-based access control// Proceedings of the 6th International Symposium on Resilient Control System (ISRCS). San Francisco, USA, 2013: 181-185

[84] Crampton J, Morisset C, Zannone N. On missing attributes in access control: Non-deterministic and probabilistic attribute retrieval//Proceedings of the 20th ACM Symposium on Access Control Models and Technologies (SACMAT). Vienna, Austria, 2015: 99-109

[85] Rosenthal A, Sciore E. Content-based and view-based access control//van Tilborg H C A, Jajodia S eds. Encyclopedia of Cryptography and Security. Boston, MA: Springer US, 2011: 249-253

LI Hao, born in 1983, Ph. D., associate professor. His main research interests include trusted computing, data security, access control.

ZHANG Min, born in 1975, Ph. D., associate professor. Her main research interests include data privacy protection, and trusted computing.

FENG Deng-Guo, born in 1965, Ph. D., professor. His main research interests include information security and cryptology, trusted computing and information assurance.

HUI Zhen, born in 1987, Ph. D. candidate. His main research interests include data security, access control.

Background

Access control deals with the problem how to share big data in security which is very important for big data. However, the traditional access control technologies are unsuited to secure the big data sharing. The problems about access control for big data are less discussed in recent research.

This paper discusses the security issues brought by big data, and summarizes five problems about access control for big data. Moreover, we describe several access control technologies which can be used in big data systems. Although these technologies are not presented for big data, they can solve the problems brought by big data mentioned above. Finally, we give three important characteristics of access control for big data. The first one is the multivariate basis for access control deciding. The second one is that the access control decision comes fuzzier than before. The last one is the combination of different access control models. And in our view, the development of big data will bring both challenges and opportunities for access control, and result in a revolution in this area.

This work is supported by the National Natural Science Foundation of China under Grant Nos. 61402456 and 61303248.