大数据时代数据隐私安全研究

肖 洁,袁 嵩,谭 天
(武汉科技大学 计算机科学与技术学院,湖北 武汉 430065)

摘 要:近年来,利用数据分析的方法从大数据中挖掘出有价值信息的大数据应用发展极为迅速,为人们的日常生活带来了极大的便利。然而,随着隐私泄露事件的屡屡发生,隐私安全问题引起了社会的广泛关注。文中对当下已然产生的各种数据隐私问题进行分析,将一系列的数据隐私保护方式进行综合,从数据存储、数据处理以及数据共享的角度出发,结合现有的数据处理技术给出在数据加密、数据防护、匿名保护技术等不同方面的保护措施以应对数据在处理、传输、共享中存在的安全隐患。由于使用云计算平台来存储和分析大数据的方式被广泛应用,而云平台的流动性、跨界的融合性以及动态的变化特性增加了隐私泄露的风险,文中还研究了基于云存储及云处理的大数据保护,以期解决隐私泄露问题。

关键词:大数据;隐私保护;数据处理;数据存储;数据加密
中图分类号:TP309 **文献标识码**:A **文章编号**:1673-629X(2016)05-0091-04
doi:10.3969/j.issn.1673-629X.2016.05.019

Research on Data Privacy in Big Data Age

XIAO Jie, YUAN Song, TAN Tian
(College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China)

Abstract: In recent years, big data analysis and application used for digging out valuable information has developed rapidly and brought great convenience for people. However, the privacy leak incidents occur frequently in big data age, the privacy security has caused the extensive concern. In this paper, through analyzing the current data privacy problems, based on a series of privacy preservation ways, from the perspective of data storage, data processing and data sharing, combined with the current technology of data processing, the corresponding protection measures in data encryption, data protection and anonymous protection technology are proposed to deal with the safe hidden trouble in the data processing, transmission and sharing. The cloud computing platform has been widely used, and its liquidity, cross-border fusion and dynamic changes increase the risk of privacy. Therefore, the big data security based on cloud storage and cloud processing is studied in this paper, in order to solve the problem of privacy leak.

Key words: big data; privacy preservation; data processing; data storage; data encryption

0 引 言

随着互联网技术的不断发展,全球数据量呈现爆炸式增长。数据挖掘技术将这些之前无法聚合的数据聚集起来,从海量的、不完全的、有噪声的、模糊的、随机的大型数据库中更迅速并且精确地发现有价值的信息。通过分析这些信息然后做出归纳性的推理,从中挖掘出潜在的模式,帮助人们做出正确决策。然而,科学技术是把双刃剑,在为人类生活带来巨大便利的同时,大数据背后所隐藏的安全隐患也是不容小觑的。随着虚拟化、云计算等新技术的广泛应用,互联网隐私泄密事件屡见不鲜。如何能够在享受到大数据时代下便利生活的同时有效避免其所带来的威胁,也成了目前研究的热点。

1 数据隐私

无论是阅读网站还是购物网站,都存在根据对用户浏览页面停留时间、浏览内容等数据的分析后产生的用户可能感兴趣内容的推荐,这在很大程度上方便了用户在网上进行目标性极强的浏览与选择。可是在获得方便快捷的个性化服务的背后,却在某种程度上暴露了自己的隐私。使用互联网时,信息在不知不觉中就被记录下来;手机通话时,通话对象与通话时间,甚至通话地点均在运营商的掌控之中;发表言论或者分享照片时,互联网运营商便可获得用户喜好……随

收稿日期:2015-05-06 **修回日期**:2015-08-12 **网络出版时间**:2016-05-05
基金项目:湖北省高等学校 2014 年省级大学生创新创业训练计划项目(201410488037)
作者简介:肖 洁(1994-),女,研究方向为软件工程;袁 嵩,博士,副教授,研究方向为智能计算。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20160505.0814.020.html

着数据采集技术的发展,个人的兴趣习惯、身体特征等隐私信息可以在用户毫无察觉的情况下被更容易地获取。大数据时代产生的众多精细化的数据,可以用来描述各种物体、社会和整个环境的行为。通过分析处理这些数据,可以大大减少社会的复杂度,提高人们认识世界、改造世界的能力,辅助人们做出重要决策。这些信息若被有效地利用确实会给人类生活带来诸多便利,但是若对其无限制甚至恶意利用,所造成的后果将是无法估量的。

2 安全威胁

2014 年 2 月,全球最大的比特币交易平台 Mt. Gox 由于交易系统出现漏洞,75 万个比特币以及 Mt. Gox 自身账号中约 10 万个比特币被窃,损失估计达到 4.67 亿美元,被迫宣布破产。2014 年 3 月,有安全研究人员在第三方漏洞收集平台上曝出携程安全支付日志可遍历下载导致大量用户银行卡信息泄露。2014 年 4 月 Heartbleed 漏洞被曝用于窃取服务器敏感信息,黑客利用 OpenSSL 漏洞发动攻击,非法获取了有些网站的用户信息。2014 年 9 月,大约有 500 万谷歌的账户和密码的数据库被泄露给一家俄罗斯互联网网络安全论坛。2014 年 12 月,索尼影业公司被黑客攻击,摄制计划、明星隐私、未发表的剧本等敏感数据都被黑客窃取并公布在网络上,甚至包括到索尼影业员工的个人信息。2014 年 12 月 25 日,大量 12306 用户数据在互联网疯传,内容包括用户账号、明文密码、身份证号码、手机号码和电子邮箱等[1]。

上述事件凸显了互联网金融在网络安全威胁面前的脆弱性,同时反映出信息若是遭遇入侵,不论是社会、企业还是个人都将遭受巨大的损失。基于云计算的网络化社会为大数据提供了一个开放的环境。正是由于平台暴露的原因,拥有巨大潜藏价值的大数据更容易遭到黑客的攻击。大数据一旦遭受攻击,失窃的数据量无疑将会是巨大的。以前,这些对人们的生活并不会造成很大的影响。因为面对海量冗杂的数据,即使刻意寻找也会消耗大量的时间和精力才能获得某些有价值的信息。如今,大数据的分析能力导致看似简单的信息也可能被挖掘出其中的隐私。这些隐私一旦遭到恶意使用,将会严重影响人们的正常生活。

3 安全防护

数据作为企业和公共组织越来越重要的资产,其安全防护也随之越发重要。近年来频发的安全问题让越来越多的人关注安全防护,隐私泄露问题已经令人无法忽视。生活在智能化的时代,避免数据的传输与分享从而切断隐私泄露根源显然是不可能事件,如何加强对数据的保护与加密成为了隐私保护的新命题。

3.1 数据存储防护

想要解决大数据的存储安全问题,数据加密必不可缺[2]。大数据安全服务设计根据安全存储的要求将大数据存储在数据集的任何存储空间,通过安全套接层(SSL)协议加密[3]的方式实现在数据集的节点和应用程序之间移动保护大数据。与应用层协议独立无关是 SSL 协议的最大优势,同时,高层的应用层协议能透明地建立于 SSL 协议之上,SSL 协议在应用层协议通信之前就已经将加密算法、通信密钥的协商以及服务器认证工作完成。为保证通信的私密性,在此之后应用层协议所传送的数据均得到了加密。如此便可在一定程度上减少数据被窃取与篡改的风险,使得数据安全得到保护。同时用软件或硬件设备对向网络上传或从网络下载的数据流进行有选择的控制。设置好规则指定哪些类型的数据包被允许通过,哪些类型的数据包将会被阻止,使得数据包在从英特网向内部网络传输数据以及从内部网络向英特网传输数据的过程中能被控制是否通过。一旦发现非常态数据,可以自动阻止并切断数据的传输,进一步提高了安全性。

目前,普遍采用虚拟化海量存储技术[4]来存储数据,大数据多被存储于云端。由于数据在云端集中,其巨大的流动性、跨界的融合性以及动态的变化等特点使得数据在传输时的保密性受到极大威胁。作为第三方的云平台在服务器故障的情况下,自身有可能将数据泄露;一旦被非法接入,数据将面临被窃取、篡改、伪造等的风险。因此,数据拥有者通过拆分、加密后才将数据上传存放在云端,用户下载后经解密方可使用。这样一来,即使数据在传输或存放的过程中意外丢失,也会因为实现加密避免发生机密信息泄露的情况。孙辛未等在文献[5]中提出,在上传数据前,将数据按照比特位进行拆分后重新组装形成多个数据文件之后再分别上传到云存储服务器。下载时,先将所有数据文件下载,通过位合并再恢复成原始文件。利用移位和扩散的基本思想设计出的位拆分技术对数据隐私具有一定的保护作用,同时该方法不依赖于密钥,通过汇编语言编写核心代码以及调整代码顺序的方式对 BSBC 隐私保护技术的代码进行了优化,加快了数据拆分和合并的速度,进一步提高了隐私保护技术的性能,对于存储在云端的数据有着很大的应用意义。

大数据在存储阶段面临隐私泄露风险的主要原因是大数据的完整性验证协议采用了第三方审计机构。因此,大数据存储方面的主要隐私保护问题是如何设计一种安全高效的、能够阻止数据拥有者的数据泄露给第三方审计机构的大数据完整性验证协议[2]。曹夕等综合考虑云存储网络环境的特性以及安全需求,设

计了一种云存储数据完整性验证(CS—DIV)协议[6]。该协议通过随机抽查客户端上传到云端的数据文件及其校验标签的方式,让服务器生成指定数据块的验证证据并返回,之后再对数据文件的完整性进行判断。该协议对于不同类型的文件均具有良好的适应性。通过检查较小文件所有的数据块的方式来保证结果的有效性。而对于较大的文件,则通过检查其中的部分数据块以概率来保证数据的完整性,如此便可减小对系统资源以及网络带宽的消耗。同时,该协议的有效执行只需要系统少量的存储和通信开销,并且随着文件的增大,验证所花的时间也仍然可以保持在一个低值水平,这满足了云存储中海量数据对处理效率的要求。该协议能够以较低的存储、通信以及时间开销有效地验证云存储数据的完整性,同时又能抵抗恶意服务器欺骗和恶意客户端攻击,实现了对数据完整性的保护,提高了整个云存储系统的可靠性和稳定性。

3.2　数据处理防护

大数据是庞大而又复杂数据集的汇集,只有经过分析挖掘后才能产生有用信息,体现出其价值。由于大数据具有数据多样性、数据处理速度快、数据价值密度低等重要的特性,使得传统的数据分析与处理方式不再完全适用,因此目前对大数据的处理方式大多是在 Hadoop 的框架上采用 Mapreduce 的模式对海量数据进行分布式的处理[7]。这种数据处理方式在某种程度上讲能够适应大数据的特性,并且具有低成本、高可扩展性、可容错性的优势,也能最大限度地利用机器资源。但是 Mapreduce 的数据处理模式过于复杂灵活,有着很强的依赖性,并且运行效率较低,而 Hadoop 对数据的聚合也增加了数据泄露的风险。

Zhang K H 等提出的面向大数据的隐私感知混合云计算模式 Sedic[8]在开源的 Hadoop 的模式上增加了隐私模块。在用户指定敏感数据之后将计算任务分割,把隐私数据留在私有云中处理,其他数据交由公有云计算。这样一来既可以保证有效利用低计算成本的公有云,同时也可以保障敏感数据在私有云中的隐私性。陈志伟等提出了一种基于 RSA 和 Paillier 的同态云计算方案[9],该方案可实现公有云服务器的密文数据处理,无需解密密文可对其执行操作便能实现对明文数据的各种计算。对于某些用户不愿意公开的密文数据,云端只需完成相关计算便可将所需数据的密文值返回。云端服务器在此操作过程中不接触明文,在某种程度上保护了用户隐私。由于通信链路和公有云服务器数据都是以 RSA 或 Paillier 加密的密文形式存在的,而未采用填充方案的 RSA 和 Paillier 是抗选择明文攻击(Chosen Plaintext Attack,CPA)的,所以该方案的密文数据符合 CPA 安全。此外出现在通信链路中的数据仅仅只是整个密文数据以及用户操作的一部分,即使这部分数据被窃取,窃听者也无法根据某次窃取的数据将明文或用户的操作请求恢复,保证了用户数据和请求的安全。在计算过程中该方案采用的是同态加密,密文规模是可以调控的,具有很好的同态操作深度。同时与基于格的全同态方案相比,基于整数域上的更容易实现和理解。虽然该方案在耗时方面有所增加,但是却拥有更好的可行性和安全性,能够很好地保护用户的隐私安全。徐计等提出的基于粒计算的大数据处理方法[10]有助于提高数据处理的速度和效率,并且对隐私保护也有一定的作用。目前,粒计算已经成为发展迅速的一种信息处理方式,被很多学者列为处理大数据的首要方法。信息粒化的概念是建立基于外部世界的、有效的、以用户为中心,同时简化对物理世界和虚拟世界的认识,对于现今在大数据处理中面临的挑战有着十分重要的意义。粒化不仅可以实现对原始数据量的压缩,而且能够在一定程度上排除噪声和不精确数据的影响。更重要的是,信息粒结构可以隐藏细节信息。隐私信息一般是以最细粒度原始数据的形式存在,采用粒计算处理将数据粒化之后,在传输和处理的过程中,规避了隐私泄露的风险。

3.3　数据共享防护

对于大数据中的结构化数据(或称关系数据)而言,数据发布匿名保护是实现其隐私保护的核心关键技术与基本手段,目前仍处于不断发展与完善阶段[11]。而对于云共享中的数据来说,采用数据加密技术与数据水印技术相结合的方式,不仅能够监控数据防止其被篡改或伪造,而且能够保护隐私不被窥探与窃取。早期 k 匿名保护技术[12]使用最为普遍,不过其容易产生对某个属性匿名处理不足的现象而被攻击者利用。针对这种情况,l-diversity 模型匿名保护技术[13]被提出。基于聚类的 l-diversity 匿名保护方法在满足 l-diversity 模型的约束条件下,采用基于距离的层次化聚类算法划分元组,对不同类型的准标识符使用不同的概化策略,并依据数据概化前后属性值不确定性程度的变化描述数据概化带来的信息损失。同现有的 l-diversity 模型相比,该方法不仅能够较好地保护用户的敏感信息,而且在一定程度上降低了概化处理带来的信息损失。针对在数据共享中所需的敏感属性的保护,王智慧等在文献[14]中提出的 L-Clustering 不仅满足结果数据集符合 l-diversity 模型,而且消除了传统数据概化处理时的概念层次结构限制。在数据共享中对数据进行匿名保护,防止与个体相关的敏感属性值泄漏。同时采取更为灵活的数据概化策略,利用基于聚类的思想来寻找合适的概化方案,从而有效地减少在实现匿名保护时概化处理所带来的信息损

失。通过数据匿名化实现隐私保护,为数据在传输过程中的隐私问题提供一定的保障。

对于云共享而言,访问权限控制与数据加密是安全防护的关键。访问权限控制确保合法用户才能访问云存储数据,数据加密限制拥有解密密钥的用户才能对存储在云端的数据进行下载并解密。刘孟占等提出的基于密文规则的属性基加密技术的云存储数据共享机制[15]通过制定合适的访问结构来实现细粒度访问权限控制。只需修改访问结构的撤销操作机制解决了公钥基础设施(Public Key Infrastructure,PKI)机制中用户撤销操作需要重复执行大量非对称加密操作带来的系统扩展性问题。数据使用公钥加密技术加密,解密密钥使用属性加密(Ciphertext - Policy ABE,CP_ABE)技术加密,数据拥有者在共享数据时对共享用户发放 CP_ABE 私钥。当共享用户的私钥满足密文的访问结构时便可获得解密密钥,而后方能解密加密数据。CP_ABE 加密技术具有灵活的访问权限控制、简单的用户撤销操作以及无需获取用户的公钥证书等优势,在一定程度上避免了 PKI 机制存在的系统扩展性问题。在用户进行数据共享时,访问权限控制和用户撤销操作不会向云存储服务提供商泄露任何机密数据,确保了数据在不可信域中的机密性,达到了保护用户数据隐私安全的目的。云平台作为第三方,存在遭受外部攻击以及系统故障等安全风险,除此之外对于参与计算的动态数据,云服务提供商可能窥探用户在使用服务过程中产生的数据流和隐私信息。面对云服务下数据的机密性、隐私性、可靠性等方面可能存在安全风险的情况,数字水印技术被用于监控数据,保障数据安全。数字水印技术是将一些标识信息直接嵌入数字载体当中,但不影响原载体的使用价值,也不容易被人的知觉系统(如视觉或听觉系统)觉察或注意到,是信息隐藏技术的一个重要研究方向。通过这些隐藏在载体中的信息,可以判断并确认信息是否被篡改。作为标识信息的数字水印应在保证不会被篡改或伪造的同时保证极低的误检率,从而使得在被保护内容发生变化时做出相应的变化,以便检测出被保护内容的变更。鲁棒数字水印[16]目前广泛用于在数字作品中标识著作权信息,利用这种水印技术在多媒体内容的数据中嵌入创建者、所有者的标识信息。在发生版权纠纷时,可用于确认数据的版权所有者,并能通过序列号追踪违反协议的用户。将这种水印技术移植到数据保护中可以监控数据,防止数据被恶意篡改,同时鲁棒数字水印还能够抵抗一些恶意攻击。

4 结束语

大数据使人类生活变得方便而又高效,但是频发的隐私泄露问题却给在享受便利生活的人们敲响了警钟,隐私安全问题得到了社会的普遍关注。文中从安全威胁事件切入,分别给出在数据存储、数据处理与数据共享方面的相应保护措施,采用多种加密方式对数据进行层层加密来保护数据,利用水印技术监控数据是否被篡改。运用多种方法保护使数据在存储和传输过程中不被窃取,希望能对隐私防护有所帮助。随着隐私防护技术的不断发展,相信人们能更安心地享受大数据时代的智能化生活。

参考文献:

[1] CCTIME. 2014 年全球 14 大网络安全事件 5 个在中国[EB/OL]. 2015 - 01 - 20. http://www. cctime. com/html/2015-1-20/2015120161127366. htm.

[2] 黄刘生,田苗苗,黄 河. 大数据隐私保护密码技术研究综述[J]. 软件学报,2015,26(4):945-959.

[3] 钟 军,吴雪阳,江一民,等. 一种安全协议的安全性分析及攻击研究[J]. 计算机工程与科学,2014,36(6):1077-1082.

[4] 刘正伟,文中领,张海涛. 云计算和云数据管理技术[J]. 计算机研究与发展,2012,49(S):26-31.

[5] 孙辛未,张 伟,徐 涛. 面向云存储的高性能数据隐私保护方法[J]. 计算机科学,2014,41(5):137-142.

[6] 曹 夕,许 力,陈兰香. 云存储系统中数据完整性验证协议[J]. 计算机应用,2012,32(1):8-12.

[7] 孙彦超,王兴芬. 基于 Hadoop 框架的 MapReduce 计算模式的优化设计[J]. 计算机科学,2014,41(11A):333-336.

[8] Zhang K H,Zhou X Y,Chen Y Y,et al. Sedic:privacy-aware data intensive computing on hybrid clouds[C]//Proceedings of the 18th ACM conference on computer and communications security. Chicago,U. S. :ACM,2011:515-525.

[9] 陈志伟,杜 敏,杨亚涛,等. 基于 RSA 和 Paillier 的同态云计算方案[J]. 计算机工程,2013,39(7):35-39.

[10] 徐 计,王国胤,于 洪. 基于粒计算的大数据处理[J]. 计算机学报,2015,38(8):1497-1517.

[11] 冯登国,张 敏,李 昊. 大数据安全与隐私保护[J]. 计算机学报,2014,37(1):246-258.

[12] 刘 斐,樊 华,金松昌,等. 一种新型 k 匿名隐私保护算法[J]. 信息网络安全,2012(8):199-202.

[13] 刘雅辉,张铁赢,靳小龙,等. 大数据时代的个人隐私保护[J]. 计算机研究与发展,2015,52(1):229-247.

[14] 王智慧,许 俭,汪 卫,等. 一种基于聚类的数据匿名方法[J]. 软件学报,2010,21(4):680-693.

[15] 刘孟占,印凯泽. 基于密文规则的属性基加密技术的云存储数据共享机制[J]. 计算机应用,2013,33(S2):133-135.

[16] 史宝明,李恒杰,贺元香,等. 基于微遗传算法与 SVD 的鲁棒性数字水印技术研究[J]. 兰州文理学院学报:自然科学版,2014,28(6):45-49.