(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110537183 B
(45)授权公告日 2023.07.07

(21)申请号 201880023325.8
(22)申请日 2018.04.10
(65)同一申请的已公布的文献号
申请公布号 CN 110537183 A
(43)申请公布日 2019.12.03
(30)优先权数据
15/488,304 2017.04.14 US
15/858,994 2017.12.29 US
(85)PCT国际申请进入国家阶段日
2019.09.30
(86)PCT国际申请的申请数据
PCT/IB2018/052511 2018.04.10
(87)PCT国际申请的公布数据
WO2018/189681 EN 2018.10.18
(73)专利权人 国际商业机器公司
地址 美国纽约阿芒克
(72)发明人 A·勒曼 M·C·奥斯伯内 M·厄斯特赖歇尔
(74)专利代理机构 北京市金杜律师事务所 11256
专利代理师 酆迅 李峥宇
(51)Int.Cl.
G06F 21/62(2013.01)
H04L 9/40(2022.01)
H04W 12/02(2009.01)
(56)对比文件
CN 102055760 A,2011.05.11
CN 103780386 A,2014.05.07
US 2009175442 A1,2009.07.09
US 2011289565 A1,2011.11.24
US 2015074407 A1,2015.03.12
US 2016006566 A1,2016.01.07
US 2016316365 A1,2016.10.27
审查员 史雪飞

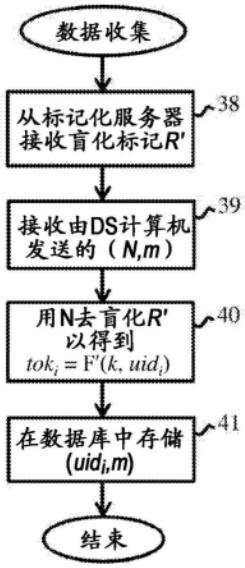
权利要求书4页 说明书9页 附图6页

(54)发明名称

数据标记化方法和系统

(57)摘要

数据源计算机提供具有相关联的id数据的、要发送到数据收集计算机的消息数据;通过用一个随机数来盲化所述id数据,生成一个盲化id;将所述盲化id发送到标记化计算机;通过网络发送所述随机数和所述消息数据,以便由数据收集计算机接收。作为响应,标记化计算机生成包括用所述随机数盲化的、所述id数据和标记化计算机的一个密钥的函数的一个盲化标记,将所述盲化标记发送到数据收集计算机。作为响应,数据收集计算机使用所述随机数来揭盲所述盲化标记以获得包括所述id数据和所述密钥的一个确定性函数的id标记。然后,数据收集计算机将所述id标记和所述消息数据存储在可操作地耦合到数据收集计算机的存储器中。



CN 110537183 B

1.一种数据标记化系统,包括适于通过网络进行通信的数据源计算机、标记化计算机和数据收集计算机,其中:

所述数据源计算机提供具有相关联的id数据的、要被发送到所述数据收集计算机的消息数据,并且适于通过用随机数盲化所述id数据来生成一个盲化id,将所述盲化id发送到所述标记化计算机,以及通过所述网络发送所述随机数和所述消息数据,以供所述数据收集计算机接收;

所述标记化计算机适于响应于接收所述盲化id而从其生成一个盲化标记,所述盲化标记包括用所述随机数盲化的、所述id数据和所述标记化计算机的密钥的函数,以及将所述盲化标记发送给所述数据收集计算机;和

所述数据收集计算机适于响应于接收到来自所述标记化计算机的所述盲化标记以及由数据源计算机发送的所述随机数和所述消息数据,用所述随机数来去盲化所述盲化标记,以获得包括所述id数据和所述密钥的一个确定性函数的id标记,并将所述id标记和所述消息数据存储在可操作地耦合到所述数据收集计算机的存储器中。

2.如权利要求1所述的系统,其中:

所述数据源计算机适于使得所述盲化id包括值R=F(N,h),其中F是一个预定函数,N是所述随机数,h是所述id数据的一个函数;

所述标记化计算机适于使得所述盲化标记包括值R′=F(k,R),其中k是所述密钥;

所述数据收集计算机适于使得所述id标记包括值F(n,R'),其中n是所述随机数N的一个函数;和

所述预定函数F使得F(n,R')=F'(k,h),其中F'是所述确定性函数。

3.如权利要求2所述的系统,其中所述预定函数F使得$F(x,y)=y^{x}$并且所述函数n包括值N^{-1}。

4.如权利要求3所述的系统,其中:

所述标记化计算机还适于周期性地生成一个新密钥k',以生成一个包括值k'/k的标记更新数据Δ,并将所述标记更新数据Δ发送到所述数据收集计算机;和

所述数据收集计算机还适于响应于接收所述标记更新数据Δ,将所述存储器中的所述id标记tok替换为包含一个值tokΔ的更新标记。

5.如权利要求2所述的系统,其中所述函数h包括所述id数据的散列。

6.如权利要求1所述的系统,其中:

所述数据源计算机还适于选择要发送到所述数据收集计算机的所述消息数据的一个会话标识符,将所述会话标识符与所述盲化id一起发送到标记化计算机,并将会话标识符、所述消息数据和所述随机数发送到所述数据收集计算机;和

所述标记化计算机还适于将所述会话标识符与所述盲化标记一起发送到所述数据收集计算机。

7.如权利要求1所述的系统,其中:

所述数据源计算机还适于加密所述消息数据和所述随机数以生成加密数据并将所述加密数据与所述盲化id一起发送到所述标记化计算机;

所述标记化计算机还适于将所述加密数据与所述盲化标记一起发送到所述数据收集计算机;和

所述数据收集计算机还适于解密所述加密数据以恢复所述消息数据和所述随机数。

8.如权利要求1所述的系统,包括一个以上的所述数据收集计算机。

9.一种用于通过网络向数据收集计算机提供数据的计算机实现的方法,该方法包括在一个适于与所述网络中的一个标记化计算机通信的数据源计算机:

提供具有相关联的id数据的、要发送到所述数据收集计算机的消息数据;

通过用一个随机数盲化所述id数据来生成一个盲化id;将所述盲化id发送到所述标记化计算机以从其生成一个盲化标记,该盲化标记包括一个用所述随机数盲化的、所述id数据和所述标记化计算机的一个密钥的函数,要由所述标记化计算机向所述数据收集计算机发送;和

通过所述网络发送所述随机数和所述消息数据,以便由所述数据收集计算机接收;

由此,所述数据收集计算机可以用所述随机数来去盲化所述盲化标记,以获得用于所述消息数据的、包括所述id数据和所述密钥的一个确定性函数的id标记。

10.如权利要求9所述的方法,包括:在所述数据源计算机处,生成所述盲化id,使得所述盲化id包括一个值R=F(N,h),其中N是所述随机数,h是所述id数据的一个函数,且F是一个预定函数,使得F(n,R')=F'(k,h),其中:

R'=F(k,R)是包含在所述盲化标记中的一个值,其中k是所述密钥;

F(n,R')是包含在所述id标记中的一个值,其中n是所述随机数N的一个函数;和

F'是所述确定性函数。

11.如权利要求10所述的方法,其中所述预定函数F使得$F(x,y)=y^{x}$并且所述函数n包括值N^{-1}。

12.如权利要求9所述的方法,包括在所述数据源计算机处:

选择要发送给所述数据收集计算机的所述消息数据的会话标识符;

将所述会话标识符与所述盲化id一起发送到所述标记化计算机,用于与所述盲化标记一起转发到所述数据收集计算机;和

将所述会话标识符、所述消息数据和所述随机数发送到所述数据收集计算机。

13.如权利要求9所述的方法,包括在所述数据源计算机处:

加密所述消息数据和所述随机数以生成加密数据;和

将所述加密数据与所述盲化id一起发送到所述标记化计算机,用于与所述盲化标记一起转发到所述数据收集计算机;

由此,所述数据收集计算机可以对所述加密数据进行解密,以恢复所述消息数据和所述随机数。

14.一种计算机实现的方法,用于标记与一个数据源计算机要经由一个网络向一个数据收集计算机提供的消息数据相关联的id数据,该方法包括在所述网络中的标记化计算机处:

从所述数据源计算机接收通过用一个随机数盲化所述id数据而生成的盲化id;

从所述盲化id生成一个盲化标记,该盲化标记包括一个用所述随机数盲化的、所述id数据和所述标记化计算机的一个密钥的函数;和

将所述盲化标记发送到所述数据收集计算机;

由此,所述数据收集计算机在接收到所述随机数和所述消息数据时,可以用所述随机

数来去盲化所述盲化标记,以获得用于所述消息数据的、包括所述id数据和所述密钥的一个确定性函数的id标记。

15.如权利要求14所述的方法,其中:

所述盲化id包括一个值R=F(N,h),其中F是一个预定函数,N是所述随机数,h是所述id数据的一个函数;

所述方法包括在所述标记化计算机处生成所述盲化标记,使得盲化标记包括一个值R'=F(k,R),其中k是所述密钥;和

所述预定函数F使得F(n,R')=F'(k,h),其中F'是所述确定性函数,n是所述随机数N的一个函数。

16.如权利要求15所述的方法,其中所述预定函数F使得$F(x,y)=y^x$并且所述函数n包括一个值N^{-1}。

17.如权利要求16所述的方法,包括:在所述标记化计算机处:

定期生成一个新密钥k';

生成标记更新数据Δ,其包括一个值k′/k;和

将所述标记更新数据Δ发送到所述数据收集计算机,用于将所述id标记更新为一个包括一个值tokΔ的更新标记。

18.如权利要求14所述的方法,包括:在所述标记化计算机处:

与所述盲化id一起从所述数据源计算机接收所述消息数据的一个会话标识符;和

将所述会话标识符与所述盲化标记一起发送到所述数据收集计算机。

19.如权利要求14所述的方法,包括:在标记化计算机处:

从所述数据源计算机与所述盲化id一起接收加密所述消息数据和所述随机数的加密数据;和

将所述加密数据与所述盲化标记一起发送到所述数据收集计算机。

20.一种用于通过一个网络从一个数据源计算机获得数据的计算机实现的方法,该方法包括在一个适于与所述网络中的一个标记化计算机通信的数据收集计算机处:

经由所述网络接收由所述数据源计算机发送的与所述数据源计算机处的id数据相关联的消息数据和一个随机数;和

从所述标记化计算机接收在数据源计算机处通过用所述随机数盲化所述id数据而生成的一个盲化标记,所述盲化标记包括用所述随机数盲化的、所述id数据和所述标记化计算机的一个密钥的函数;和

用所述随机数来去盲化所述盲化标记,以获得一个包括所述id数据和所述密钥的一个确定性函数的id标记;和

将所述id标记和所述消息数据存储在可操作地耦合到所述数据收集计算机的存储器中。

21.如权利要求20所述的方法,其中:盲化id包括值R=F(N,h),其中F是一个预定函数,N是所述随机数,h是所述id数据的一个函数;

所述盲化标记包括一个值R′=F(k,R),其中k是所述密钥;

该方法包括在所述数据收集计算机处生成所述id标记,使得所述id标记包括一个值F(n,R'),其中n是所述随机数N的一个函数;和

所述预定函数F使得F(n,R')＝F′(k,h)，其中F'是所述确定性函数。

22.如权利要求21所述的方法，其中所述预定函数F使得$F(x,y)=y^{x}$并且所述函数n包括一个值N^{-1}。

23.如权利要求22所述的方法，包括：在数据收集计算机处：

周期性地从所述标记化计算机接收包括一个值k'/k的标记更新数据Δ，其中k'是所述标记化计算机的一个新密钥；和

用一个包括一个值tokΔ的更新标记替换所述存储器中的所述id标记tok。

24.如权利要求20所述的方法，包括：在数据收集计算机处：

从所述数据源计算机与所述消息数据的一个会话标识符一起接收所述消息数据和所述随机数；和

从所述标记化计算机与所述盲化标记一起接收所述会话标识符。

25.如权利要求20所述的方法，包括：在数据收集计算机处：

从所述标记化计算机与所述盲化标记一起接收由所述数据源计算机生成的加密所述消息数据和所述随机数的加密数据；和

解密所述加密数据以恢复所述数据源计算机发送的所述消息数据和所述随机数。

26.一种计算机可读存储介质，包括随其体现的程序代码，所述程序代码适于在计算机上运行时执行权利要求9至25中任一项的方法。

数据标记化方法和系统

技术领域

[0001] 本发明一般涉及数据标记化。

背景技术

[0002] 数据标记化(tokenization)是一种用于在将数据移动到不太可信的环境时对数据进行脱敏的技术。例如,当数据集被外包时,或者出于某种目的收集或汇总数据(诸如交易数据)时,法律约束或安全顾虑通常要求在跨越边界转移数据或将数据转移到不受信任的环境中之前使用标记化技术。特别地,要通过网络传输的数据可能包括识别信息,例如社会安全号码、银行帐号、车辆识别号码或不应由数据提供者透露的其他唯一标识符。因此,这种id数据被其他—通常是随机查找的—数据(标记)(token)替换。为了保持数据的整体效用,必须通过标记化过程维护参照完整性(referential integrity)。也就是说,标记化操作必须是一个确定性过程,以便所有出现的相同id数据始终被相同的标记替换。

[0003] 已经提出了多种标记化技术,并且这些技术目前处于商业运营中。典型的方法要么依赖于非加密方法,诸如替换、扰动或转换表;要么使用加密机制,诸如加密钥的哈希函数或确定性加密。所有方法的共同之处在于它们要求标记化操作在可信环境中执行,即,由可信数据源本身或由数据源的信任域内的专用实体执行。这对标记化系统的实现施加了限制。此外,当从不同的、可能广泛分布的数据源收集数据时,难以以安全且有效的方式实现该假设。参照完整性要求标记化操作在所有数据源中保持一致,因此所有源必须共享相同的秘密标记化密钥,或者更糟糕的是,必须保持转换表的共享和一致版本。更实际的方法是将标记化任务集中在中央可信实体或TTP(可信第三方),后者处理所有标记化请求。然后,TTP提供将敏感id数据转换为安全标记的服务。当前的解决方案需要将id数据公开给TTP,这使得TTP成为安全和隐私瓶颈。例如,当响应于多个请求和/或针对多个源以动态方式执行标记化时,有一个能够识别和跟踪对应于id数据的用户或其他实体的活动的实体,显然是不期望的。

发明内容

[0004] 根据本发明的至少一个实施例,提供了一种数据标记化系统,包括适于经由网络进行通信的数据源计算机、标记化计算机和数据收集计算机。数据源计算机提供具有相关联的id数据的、要被发送到数据收集计算机的消息数据,并且适于通过用随机数盲化id数据来生成盲化id。数据源计算机还适于将所述盲化id发送到标记化计算机,并通过网络发送所述随机数和所述消息数据以供数据收集计算机接收。标记化计算机适于响应于对所述盲信息的接收而从其生成盲化标记(blinded token),所述盲化标记包括用所述随机数盲化的、所述id数据和所述标记化计算机的密钥的函数,并且将盲化标记发送给数据收集计算机。数据收集计算机适于响应于从标记化计算机接收到所述盲化标记以及接收到由数据源计算机发送的所述随机数和所述消息数据,用所述随机数来去盲化所述盲化标记,以获得包括所述id数据和所述密钥的确定性函数的id标记。然后,数据收集计算机将所述id标

记和所述消息数据存储在可操作地耦合到数据收集计算机的存储器中。

[0005] 在本发明的实施例中,可以在不可信的域中安全地生成与要发送到数据收集计算机的其他数据(在此一般称为“消息数据”)相关联的敏感id数据的id标记。标记化计算机提供一个中央标记化点,但以不经意的方式执行其标记化操作。它不会习得有关被标记化的id数据的任何信息,也不会了解有关盲目计算的id标记的任何信息。此外,标记化计算机甚至不能确定两个标记请求是否针对相同的id数据。这防止了对通过联系给定id的请求的活动的跟踪,从而防止旨在利用事件的可联系性的干扰攻击。类似地,数据收集计算机不了解有关被标记化的id数据的任何信息,并且数据源不了解有关生成的标记的任何信息。因此,本发明的实施例提供了安全且非常实用的数据标记化系统。

[0006] 体现本发明的系统可以容易地容纳一个以上的数据源计算机,每个数据源计算机适于与如上所述的标记化计算机和数据收集计算机通信。不需要跨多个数据源同步标记化密钥或其他安全标记化信息,因为安全标记化操作是以不经意的方式集中执行的,以便为所有源提供确定性标记。

[0007] 在优选实施例的有效实现中,数据源计算机适于使得盲化id包括值R=F(N,h),其中F是一个预定函数,N是随机数,h是id数据的一个函数。标记化计算机适于使得盲化标记包括值R’=F(k,R),其中k是密钥。数据收集计算机适于使得id标记包括值F(n,R’),其中n是随机数N的函数。这里,所述预定函数F使得F(n,R’)=F’(k,h),其中F’是上述确定性函数。在一个特别有效的实现中,所述预定函数F使得$F(x,y)=y^{x}$并且函数n包括值N^{-1}。这还允许使用简单有效的密钥更新过程。标记化计算机还可以适于周期性地生成一个新密钥k’,以生成包括值k’/k的标记更新数据Δ,并将所述标记更新数据Δ发送到数据收集计算机。响应于接收到标记更新数据Δ,数据收集计算机可以简单地用包括值tokΔ的更新标记替换其相关联的存储器中的(由tok表示的)id标记。以这种方式,可以根据需要刷新安全标记化密钥,同时保持在新密钥k’下生成的新标记与先前存储的在旧密钥k下生成的标记之间的参照完整性。

[0008] 本发明的各个进一步的实施例提供由如上所述的数据源标记化系统的数据源计算机、标记化计算机和数据收集计算机执行的方法。

[0009] 根据一个方面,提供了一种用于经由网络向数据收集计算机提供数据的计算机实现的方法,该方法包括在适于与所述网络中的一个标记化计算机通信的数据源计算机处:提供具有相关联的id数据的消息数据,以发送到所述数据收集计算机;通过用随机数盲化所述id数据来生成盲化id;将所述盲化id发送到所述标记化计算机以从其生成一个盲化标记,所述盲化标记包括用所述随机数盲化的、所述id数据和所述标记化计算机的密钥的函数,所述盲化标记要由标记化计算机发送到数据收集计算机;通过网络发送随机数和所述消息数据,以供数据收集计算机接收;由此,数据收集计算机可以用所述随机数来去盲化所述盲化标记,以获得包括所述id数据和所述密钥的一个确定性函数的id标记,用于所述消息数据。

[0010] 根据另一方面,提供了一种计算机实现的方法,用于标记化与要由一个数据源计算机经由一个网络向一个数据收集计算机提供的消息数据相关联的id数据,该方法包括:在所述网络中的一个标记化计算机:从数据源计算机接收一个通过用随机数盲化所述id数据而生成的盲化id;从所述盲化id生成一个盲化标记,该盲化标记包括一个用所述随机数

盲化的、所述id数据和所述标记化计算机的一个密钥的函数;并将所述盲化标记发送到数据收集计算机;由此,数据收集计算机在接收到所述随机数和所述消息数据时,可以用所述随机数来去盲化所述盲化标记,以获得一个包括所述id数据和所述密钥的一个确定性函数的id标记,用于所述消息数据。

[0011] 根据另一方面,提供了一种用于经由一个网络从一个数据源计算机获得数据的计算机实现的方法,该方法包括在适于与所述网络中的一个标记化计算机通信的数据收集计算机处:经由所述网络接收与数据源计算机处的id数据相关联的消息数据,以及数据源计算机发送的随机数;并且从标记化计算机接收通过用所述随机数盲化所述id数据从数据源计算机处生成的一个盲化id生成的一个盲化标记,所述盲化标记包括用所述随机数盲化的、所述id数据和标记化计算机的一个密钥的函数;用所述随机数来去盲化所述盲化标记,以获得包括所述id数据和所述密钥的一个确定性函数的一个id标记;将所述id标记和所述消息数据存储在可操作地耦合到数据收集计算机的存储器中。

[0012] 本发明还可以实现为计算机程序。

附图说明

[0013] 现在将仅通过示例的方式参考附图描述本发明的实施例,附图中:

[0014] 图1是根据本发明实施例的数据标记化系统的示意图;

[0015] 图2是根据本发明实施例的图1系统中的计算机的概括示意图;

[0016] 图3a至3c表示根据本发明实施例的在数据标记化操作中在系统1中执行的步骤;

[0017] 图4a到4c表示在标记化系统的一个实施例中执行的更详细的步骤;

[0018] 图5表示在一个实施例中执行的密钥更新过程的步骤;和

[0019] 图6a到6c表示在标记化系统的另一实施例中执行的步骤。

具体实施方式

[0020] 图1的框图示出了体现本发明的示例性标记化系统。系统1包括多个数据源计算机2、在此由服务器3实现的标记化计算机、以及在此由服务器4实现的数据收集计算机。数据源计算机2、标记化服务器3和数据收集服务器4适于经由网络5进行通信,网络5一般可以包括一个或多个组件网络和/或包括因特网的互联网络。

[0021] 在系统1的操作中,数据收集(DC)服务器4收集由数据源(DS)计算机2通过网络5提供的数据。标记化服务器3在该过程中向DS计算机2提供标记化服务。特别地,DS计算机2在任何给定时间发送的数据可以表示为元组(uid_i,m),其中uid_i表示不应向DC服务器显示的安全敏感id数据,m表示另一个数据,即与id数据uid_i相关联的伴随数据(“消息数据”)。在DS计算机2处提供的数据元组(uid_i,m),可以包括在DS计算机处例如在交易操作中动态生成的数据,和/或存储在可操作地耦合到DS计算机的存储器中的数据,例如DS计算机2中的本地存储器或DS计算机可访问的存储器中的数据。由DS计算机2提供并由DC服务器4收集的数据,可以与多个标识符uid_i相关联,i=1,2,3.....,每个标识符必须通过标记化过程一致地标记化。由标记化服务器3执行的标记化操作使用对标记化服务器来说是秘密的加密密钥k。对于数据元组(uid_i,m)中的uid_i,由tok_i表示的最终的id标记是根据下问详述的过程在DC服务器4处导出的。DC服务器将得到的标记化数据(tok_i,m)存储在可操作地耦合到

DC服务器4的、此处由数据库6代表的存储器中。数据存储器6通常可以包括任何方便的、包括一个或多个数据存储介质的数据存储装置。典型的实现包括盘存储装置，其包括一个或多个盘，诸如磁盘或光盘，它们可以在计算机内部，例如在硬盘驱动器中，或由外部可访问的磁盘装置提供，例如在诸如RAID(独立磁盘冗余阵列)阵列的磁盘驱动器阵列中中提供。

[0022] 用于标记化操作的典型应用场景包括从银行、商店等收集/聚合交易数据，从车辆池收集车辆使用/性能数据，从用户设备收集网络浏览数据等。根据应用场景，DS计算机2可以由诸如台式计算机、膝上型计算机、平板电脑、笔记本电脑、掌上电脑、移动电话、PDA(个人数字助理)、个人音乐播放器等用户计算机实现，或由车载计算机或银行、商店或其数据要被收集的其它实体处的计算机实现。

[0023] 一般来说，系统1的计算机2、3、4可以由通用或专用计算机实现，该计算机可以包括一个或多个提供用于实现本文所述操作的功能的(真实或虚拟)机器。该功能可以由以硬件或软件或其组合实现的逻辑提供。可以在由计算装置执行的计算机系统可执行指令(例如程序模块)的一般上下文中描述这种逻辑。通常，程序模块可以包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、逻辑、数据结构等。计算设备可以在分布式云计算环境中实践，其中任务由通过通信网络链接的远程处理设备执行。在分布式云计算环境中，数据和程序模块可以位于包括内存存储设备在内的本地和远程计算机系统存储介质中。特别地，标记化服务器3的功能可以方便地实现为云计算环境中的服务。DC服务器4可以完全由从多个源收集数据的专有实体操作，或者可以全部或部分地实现为云计算环境中的服务。例如，DC服务器4可以将基于云的存储用于数据库6。

[0024] 图2的框图示出了用于实现系统1的计算机2,3的示例性计算装置。这里以通用计算设备10的形式示出了该装置。计算机10的组件可以包括诸如由处理单元11、系统存储器12和将包括系统存储器12的各种系统组件耦合到处理单元11的总线13代表的一个或多个处理器的处理装置。

[0025] 总线13表示几类总线结构中的一种或多种，包括存储器总线或者存储器控制器，外围总线，图形加速端口，处理器或者使用多种总线结构中的任意总线结构的局域总线。举例来说，这些体系结构包括但不限于工业标准体系结构(ISA)总线，微通道体系结构(MAC)总线，增强型ISA总线、视频电子标准协会(VESA)局域总线以及外围组件互连(PCI)总线。

[0026] 计算机10通常包括多种计算机可读介质。这些介质可以是任何能够被计算机10访问的可用介质，包括易失性和非易失性介质，和可移动的和不可移动的介质。例如，系统存储器12可以包括易失性存储器形式的计算机可读介质，例如随机存取存储器(RAM)14和/或高速缓存存储器15。计算机系统/服务器10可以进一步包括其它可移动/不可移动的、易失性/非易失性计算机系统存储介质。仅作为举例，存储系统16可以用于读写不可移动的、非易失性磁介质(通常称为“硬盘驱动器”)。尽管图中未示出，可以提供用于对可移动非易失性磁盘(例如“软盘”)读写的磁盘驱动器，以及对可移动非易失性光盘(例如CD-ROM,DVD-ROM或者其它光介质)读写的光盘驱动器。在这些情况下，每个驱动器可以通过一个或者多个数据介质接口与总线13相连。

[0027] 存储器12可以包括至少一个程序产品，该程序产品具有被配置为执行本发明的实施例的功能的一个或多个程序模块。举例来说，具有一组(至少一个)程序模块18的程序/实用工具17，可以存储在存储器12中，操作系统、一个或者多个应用程序、其它程序模块以及

程序数据也如此。这些示例中的每一个或某种组合中可能包括网络环境的实现。程序模块18通常执行本发明所描述的实施例中的功能和/或方法。

[0028] 计算机10也可以与一个或多个外部设备19(例如键盘、指向设备、显示器20等)通信,还可与一个或者多个使得用户能与该计算机系10交互的设备通信,和/或与使得该计算机系统/服务器10能与一个或多个其它计算设备进行通信的任何设备(例如网卡,调制解调器等等)通信。这种通信可以通过输入/输出(I/O)接口21进行。并且,计算机10还可以通过网络适配器22与一个或者多个网络(例如局域网(LAN),广域网(WAN)和/或公共网络,例如因特网)通信。如图所示,网络适配器22通过总线13与计算机系统/服务器10的其它模块通信。应当明白,尽管图中未示出,可以结合计算机10使用其它硬件和/或软件模块,包括但不限于:微代码、设备驱动器、冗余处理单元、外部磁盘驱动阵列、RAID系统、磁带驱动器以及数据备份存储系统等。

[0029] 图3a至3c指示在标记化操作中在系统1中执行的基本步骤。(在将要描述的实施例中,计算机2、3、4之间的所有通信可以使用标准PKI(公钥基础设施)的加密/签名方案以通常的方式加密和/或签名。这些过程在本领域是众所周知的。为简单起见,下面省略了技术和细节。)标记化操作由具有要发送到DC服务器4的数据的DS计算机2启动。图3a表示DS计算机2为其数据元组(uid_i,m)执行的步骤。在步骤30中,DS计算机选择一个随机数N。在步骤31中,DS计算机通过使用随机数N对id数据uid_i进行盲化(blinding)来产生盲化id,其在此用R表示。可以通过对uid_i本身或对它的函数应用一个盲化函数,而生成盲化id R,如下面的实施例所示。在步骤32中,DS计算机经由网络5向标记化服务器3发送一个包括该盲化id R的标记请求。在步骤33中,DS计算机通过网络5发送随机数N和消息数据m以供DC服务器4接收。如下面的例子所示,该通信可以通过不同的信道进行。DS计算机2的过程然后完成。

[0030] 图3b示出了在标记化服务器3处的标记化操作。标记化服务器在步骤35中从DS计算机2接收盲化id R。在步骤36中,标记化服务器3从该盲化id R为id数据uid_i产生一个盲化标记,其在此用R'表示。盲化标记R'包括id数据uid_i和标记化服务器的密钥k的用随机数N盲化的函数。然后,标记化服务器在步骤37中将盲化标记R'发送到DC服务器4,并且该过程完成。

[0031] 图3c示出了在DC服务器4处获得数据的过程。DC服务器在步骤38中从标记化服务器3接收盲化标记R'。步骤39表示由DC服务器4接收由DS计算机2在图3a的步骤33中发送的随机数N和消息数据m。在步骤40中,DC服务器用随机数N将盲化标记R'去盲化(unblind)以获得id数据uid_i的id标记tok_i。该去盲化(unblinding)操作用随机数N反转该盲化。所得到的标记tok_i包括id数据uid_i和标记化服务器3的密钥k的一个确定性函数(由F'表示)。在步骤41中,DC服务器在数据库6中存储标记化的数据,其包括id标记tok_i和消息数据m的元组,然后标记化操作完成。对来自DS计算机2的所有数据传输执行标记化操作,由此DC服务器可以从数据库6中的所有源收集标记化数据池。

[0032] 利用上述系统,标记化服务器3可以在DS计算机2的信任域之外,同时向DS计算机提供安全的标记化服务。标记化服务器仅被信任以正确执行其标记化操作,并不学习id数据uid_i或最终的id标记tok_i。它仅接收用计算机2为每次数据传输选择的随机数N生成的盲化id R。因此,标记化服务器甚至不能确定两个标记请求是否针对相同的uid_i,因此不能将与任何给定身份(identity)相关联的活动联系起来。DC服务器不学习标记化的uid_i,并且

DS计算机2不学习最终的id标记tok_i。标记是以确定的方式生成的:对同一uid_i的两个标记请求将产生相同的标记tok_i,但仅在DC服务器4处的去盲化操作之后。因此,尽管以完全盲目的方式执行标记化,但是所需的参照完整性得到保证。只需要标记化服务器来存储协议特定的密钥k,从而提供简单的密钥管理。由于DS计算机不需要为标记化协议存储标记化密钥或其他状态,因此不存在跨多个源分发安全状态的风险,并且系统可以容易地扩展到大型数据源组。

[0033] 图4a到4c指示标记化系统的实施例中的更详细步骤。图4a表示由DS计算机2为数据元组(uid_i,m)执行的步骤。在步骤45中,DS计算机如前所述地选择随机数N,并且还选择用于数据传输的会话标识符sid。在步骤46中,DS计算机将散列函数H应用于uid_i以获得散列值hi。在步骤47,DS计算机生成盲化id,其为R=F(N,h),其中F是用于标记化操作的预定函数。在步骤48,DS计算机向标记化服务器3发送标记请求。这里的标记请求包括如前所述的盲化id R,以及用于由标记化服务器转发到DC服务器4的会话ID sid。在步骤49,DS计算机将会话ID sid、随机数N和消息数据m发送到DC服务器4,该过程完成。

[0034] 图4b表示由标记化服务器3执行的步骤。标记化服务器在步骤50中从DS计算机2接收(sid,R)。在步骤51,标记化服务器3生成盲化标记,其为R'=F(k,R)。在步骤52中,标记化服务器将盲化标记R'与会话标识符sid一起发送到DC服务器4。然后该过程完成。

[0035] 图4c表示DC服务器4执行的步骤。DC服务器在步骤55中从DS计算机接收(sid,N,m),并在步骤56中从标记化服务器接收(sid,R')。收到这两个元组与匹配的会话标识符sid后,操作进行到步骤57。这里,DC服务器将id标记计算为tok_i=F(n,R'),其中n是在步骤55中接收的随机数N的函数。在该实施例中,选择图4a至4c中使用的预定函数F,使得F(n,R')=F'(k,h_i),其中F'是前述确定性函数。在步骤58中,DC服务器将(tok_i,m)添加到数据库6中的数据池,标记化操作完成。

[0036] 下面详细描述上述方案的基于离散对数问题的示例性实现。在该示例中,预定函数F使得F(x,y)=y^x并且在图4c的步骤57中使用的函数n是n=N-1。确定性函数F'=F。系统参数包括安全参数τ和阶数q的循环群$\mathbb{G} = \langle g \rangle$(其为τ位素数),以及映射到$\mathbb{G}$的散列函数H的描述。我们假设,例如通过SSL/TLS(安全套接字层/传输安全层)协议和服务器证书,所有各方之间的通信信道安全。遗忘标记化过程包括以下四个过程。

[0037] 设置。标记化服务器选择一个随机密钥$k \overset{\$}{\leftarrow} \mathbb{Z}_p$,其中$表示随机选择,$\mathbb{Z}_p$是整数模p的组。

[0038] DS计算机的标记请求。具有数据元组(uid_i,m)的数据源执行以下操作:

[0039] 1.选择一个随机会话标识符sid;

[0040] 2.通过选择一个随机数$N \overset{\$}{\leftarrow} \mathbb{Z}_q$并计算R←[H($uid_i$)]^{N}而计算盲化$(R,N) \overset{\$}{\leftarrow} \text{blind}(uid_i)$

[0041] 3.将标记请求(sid,R)发送到标记化服务器,并将(sid,N,m)发送到DC服务器;

[0042] 4.删除sid、N。

[0043] 标记化服务器处的响应。输入标记请求(sid,R)后,标记化服务器执行以下操作:

[0044] 1.计算$R' \leftarrow R^k$;

[0045] 2.发送响应(sid,R′)到DC服务器;

[0046] 3.删除sid、R.

[0047] DC服务器上的去盲化和装配。在从数据源接收(sid,N,m)并从(相同sid的)标记化服务器接收一个响应(sid,R′)后,DC服务器执行以下操作:

[0048] 1.通过计算$tok_i \leftarrow R'^{1/N}$来将所述响应去盲化为标记

[0049] $$tok_i \xleftarrow{\$} \text{unblind}(R', N);$$

[0050] 2.存储输出(tok_i,m);

[0051] 3.删除sid、N、r′.

[0052] 可以看出,最终的加密标记是确定性地从uid_i和密钥k导出的:

[0053] $$tok_i \leftarrow R'^{1/N} = (R^k)^{1/N} = (\mathrm{H}(uid_i)^N)^k)^{1/N} = \mathrm{H}(uid_i)^k = \text{token}(k, uid_i)。$$

总的来说,各方共同计算具有以下安全性和功能属性的标记。

[0054] 遗忘计算。标记化服务器作为不经意的第三方(OTP)运行。OTP既不学习传入的标识符uid_i也不学习盲目计算的标记tok_i,不能联系两个请求uid_i。

[0055] 伪随机标记。对于任何不知道密钥k的实体来说,盲目生成的标记tok_i与随机生成的无法区分。也就是说,获得标识符$\{uid_i\}$和标记$\{tok_j\}$的对手无法确定它们之间的关系。特别是,数据源不会了解有关所产生的标记的任何信息,数据收集器也不会了解有关标记背后的标识符的任何信息。

[0056] 确定性推导。标记是从唯一标识符确定性地导出的。因此,即使OTP的输入和输出是随机的,向数据池的最终输出也是一致的值。

[0057] 可以看出,以上提供了高效的动态数据标记化方案,由此可以“在运行中”对数据进行标记化,为在大型和多样化的环境中的多个分布式数据源提供安全且一致的标记化。不经意的标记化服务器盲目地计算加密强化的标记而不会成为隐私风险。此外,上述方案可以方便地适应许多安全关键的应用所需的周期性密钥更新过程。密钥更新过程如图5所示。

[0058] 图5的步骤60至63由标记化服务器3执行。在步骤60中,标记化服务器通过选择$k' \xleftarrow{\$} \mathbb{Z}_p$来生成一个新密钥k’。在步骤61中,标记化服务器按Δ=k′/k来生成标记更新数据Δ。标记化服务器在步骤62中将标记更新数据Δ发送到DC服务器4,然后在步骤63中删除旧密钥k和标记更新数据Δ。新密钥k’变为当前标记化密钥k并被用于来自DS计算机2的所有后续标记请求。

[0059] 图5的步骤64至67由DC服务器4执行。DC服务器在步骤64中接收标记更新数据Δ。在步骤65中,对于存储在数据库6中的每个标记tok_i,DC服务器按$tok'_i = tok_i^{\Delta}$来计算更新标记tok'_i。在步骤66中,DC服务器用相应的更新标记tok'_i替换每个旧标记tok_i。DC服务器在步骤67中删除所有旧标记tok_i和标记更新数据Δ,更新过程完成。

[0060] 密钥更新过程可以根据需要周期性地执行,定期或不定期地执行,和/或响应于检测到系统1中的恶意干预而执行。在新密钥k’下生成的新标记与在旧密钥k下生成的以前存储的标记之间保留参照完整性:

$tok'_i = tok_i^{\Delta} = tok_i^{k'/k} = (\mathrm{H}(uid_i)^k)^{k'/k} = \mathrm{H}(uid_i)^{k'} = \mathrm{token}(k', uid_i)$。随着每次密钥更新，任何先前丢失或受损的数据都变得与新密钥k'不兼容。因此，在任何给定时间，攻击者必须同时窃取数据并破坏用于标记该数据的当前密钥才破坏安全性。

[0061] 图6a至6c表示另一个也基于离散对数问题的实施例的步骤。图6a表示由DS计算机2执行的步骤。在步骤70，DS计算机选择一个随机数N；在步骤71中，DS计算机如前所述地计算散列值$h_i=H(uid_i)$。在步骤72中，DS计算机按$R=h_i^N$生成盲化id。在这里的步骤73中，DS计算机通过一个加密方案Enc对消息数据m和随机数N进行加密，以生成加密数据Enc(N,m)。可以通过DC服务器4已知的一个密钥下的对称加密方案进行加密，或者通过DC服务器已知其密钥sk的公钥-私钥对(pk,sk)的一个公钥pk下的非对称加密方案来进行加密。在步骤74，DS计算机向标记化计算机3发送包含该加密数据Enc(N,m)和该盲化id R的标记请求，然后该过程完成。

[0062] 图6b表示由标记化服务器3执行的步骤。标记化服务器在步骤76接收标记请求(R,Enc(N,m))。在步骤77中，标记化服务器3按$R'=R^k$生成盲化标记。在步骤78中，标记化服务器将盲化标记R'与加密数据Enc(N,m)一起发送到DC服务器4，然后该过程完成。

[0063] 图6c表示DC服务器4执行的步骤。DC服务器在步骤80中从标记化服务器接收(R',Enc(N,m))。在步骤81中，DC服务器用上述密钥解密加密数据Enc(N,m)，以恢复消息数据m和随机数N。在步骤82，DC服务器按$tok_i=R'^{1/N}$计算id标记。在步骤83中，DC服务器存储(tok_i,m)，然后该标记化操作完成。

[0064] 虽然盲化id被计算为上面的值$R=F(N,h_i)$，但是盲化id另外还可以包括值$R=F(N,h_i)$的一个函数，例如，在其他实施例中，是$F(N,h_i)=h_i^N$的倍数或幂。类似地，盲化标记、id标记toki、函数n、标记更新数据Δ和更新的标记$tok'_i=tok_i^{\Delta}$可以包括一个函数，例如上述各值的倍数或幂。而且，系统也可以基于除$F(x,y)=y^x$之外的预定函数F。例如，可以容易地设想基于乘法(例如$F(x,y)=x\times y$)或除法(例如$F(x,y)=x/y$)的系统。

[0065] 可以使用各自不同的函数来计算盲化id、盲化标记和id标记。以下实施例提供了一个示例。除了秘密标记化密钥k之外，该实施例中的标记化服务器3保持一个公钥/私钥对(epk,esk)用于加性同态加密方案HEnc。这种加密方案具有同态特性，从而存在一个在加密公钥epk下加密消息m时对密文$C=HEnc_{epk}(m)$的有效操作⊙，使得：

[0066] 如果$C_1=HEnc_{epk}(m_1)$且$C_2=HEnc_{epk}(m_2)$

[0067] 则$C_1\odot C_2=HEnc_{epk}(m_1+m_2)$

[0068] 且$(HEnc_{epk}(m))^r=HEnc_{epk}(r\odot m)$.

[0069] 该实施例可以采用作为方案HEnc的示例的Paillier加密。标记化服务器3将公共密钥epk与加密秘密标记化密钥k的密文$Ck=HEnc_{epk}(k)$一起发布。因此，公共密钥epk和密文Ck可用于所有DS计算机2。如图3a至3c所示地进行系统操作，盲化id由DS计算机2在步骤31中计算为

[0070] $R=(HEnc_{epk}(h_i)Ck)^N=(HEnc_{epk}(h_i)HEnc_{epk}(k))^N=HEnc_{epk}(N(h_i+k))$

[0071] 其中⊙对应于乘法且$h_i=H(uid_i)$和散列函数在此映射到$\mathbb{Z}_q$。

[0072] 标记化服务器3通过经由加密方案的解密算法HDec用密钥esk解密R来在步骤36中

生成盲化标记以获得

[0073] $v=HDec_{esk}(R)=N(h_i+k)$然后计算$R'=g^{1/v}=g^{1/N(h_i+k)}$作为盲化标记，其中g是如前定义的循环群G的生成器。然后，DC服务器4在步骤40中去盲化该盲化标记，计算

[0074] $tok_i=R'^N=g^{1/(h_i+k)}$。

[0075] 盲目计算的(blindly-computed)确定性函数$F'(k,h_i)=g^{1/(h_i+k)}$也是一个伪随机函数。然而，该方案通过用于计算盲化id的密文Ck引入秘密标记化密钥k，效率低于先前的优选实施例。

[0076] 当然可以对所描述的示例性实施例进行许多其他改变和修改。例如，如果需要，可以使用加密散列函数H来计算值h_i，在这种情况下，散列密钥可以是特定于特定数据源组的。值h_i也可以被计算为id数据uid_i的其他确定性函数。此外，虽然上面使用散列函数H来将uid_i映射到所描述的详细协议的正确组，但是在其他实施例中，可以通过使uid_i本身而不是其函数盲化，例如通过在协议操作所需的组中选择uid_i，来计算盲化id。

[0077] 虽然已经描述了包括多个DS计算机2的系统1，但是在其他实施例中，标记化过程可以用于来自单个数据源的数据传输。

[0078] 一般来说，流程图的步骤可以以与所示顺序不同的顺序执行，并且一些步骤可以适当地同时执行。

[0079] 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

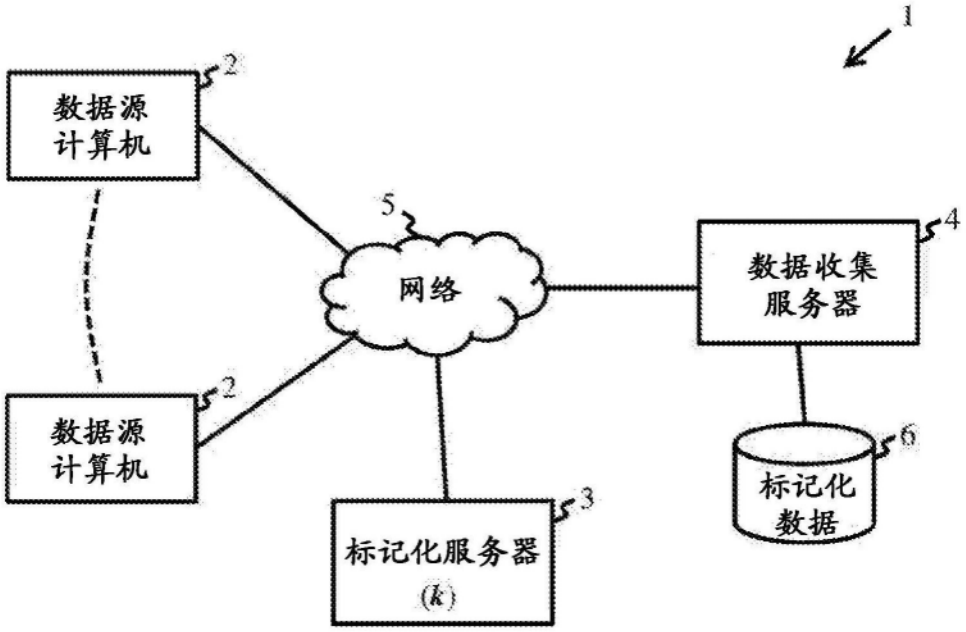

1
2
数据源
计算机
2
数据源
计算机
5
网络
4
数据收集
服务器
6
标记化
数据
3
标记化服务器
(k)

图1

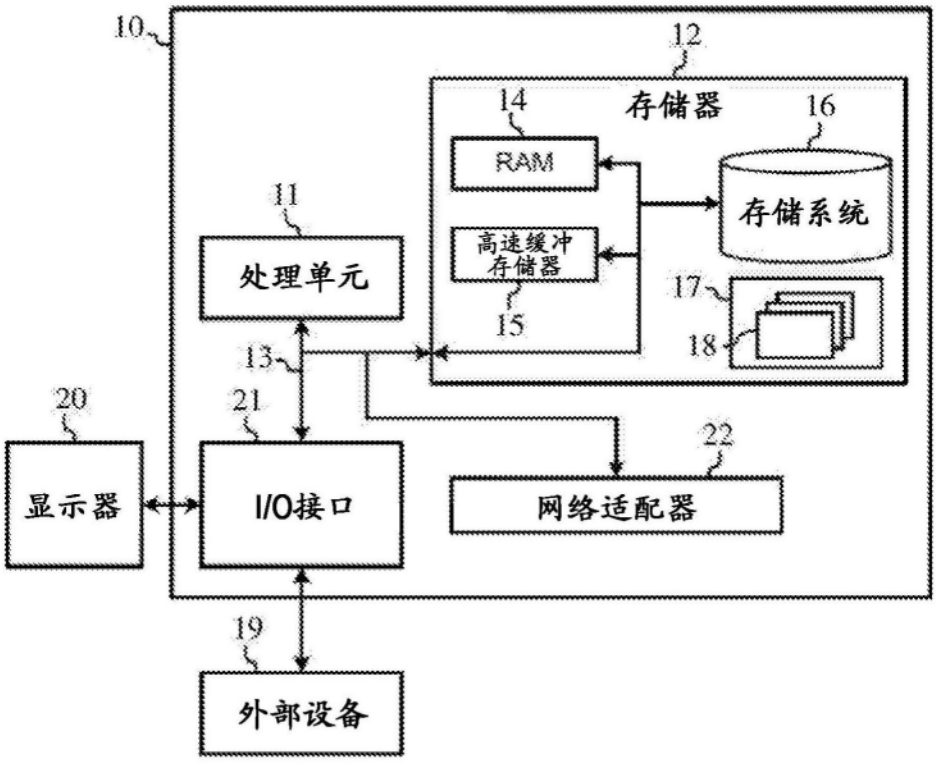

10
12
存储器
14
RAM
16
存储系统
11
处理单元
高速缓冲
存储器
15
17
18
13
20
21
22
显示器
I/O接口
网络适配器
19
外部设备

图2

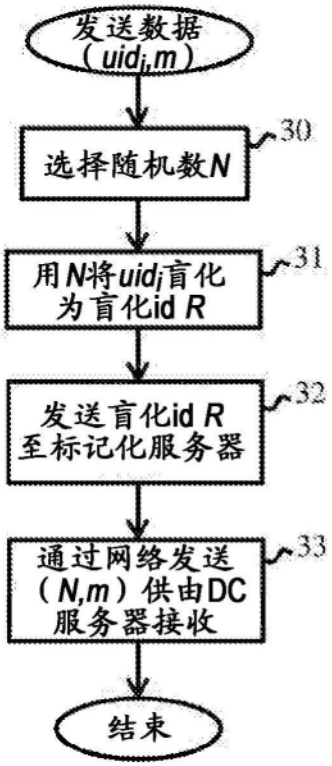

发送数据
(uid_i,m)
选择随机数N
30
用N将uid_i盲化
为盲化id R
31
发送盲化id R
至标记化服务器
32
通过网络发送
(N,m)供由DC
服务器接收
33
结束

图3a

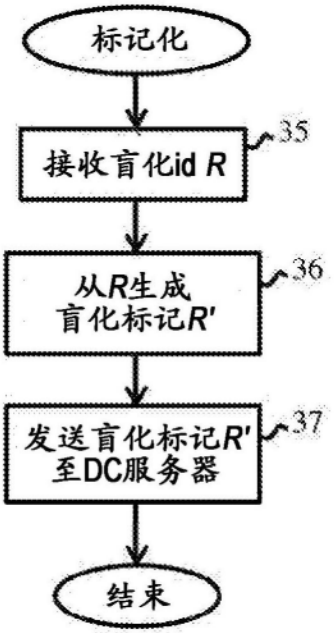

标记化
接收盲化id R
35
从R生成
盲化标记R'
36
发送盲化标记R'
至DC服务器
37
结束

图3b

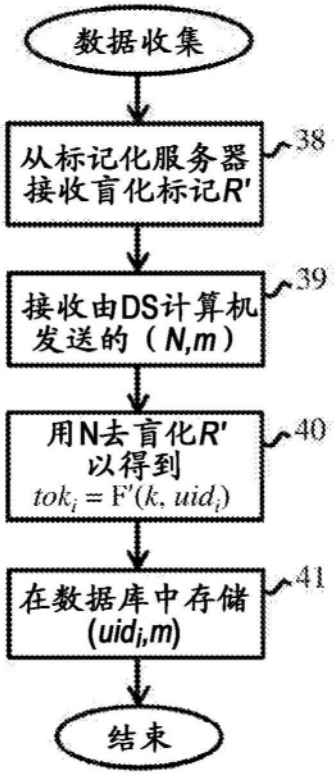

数据收集
从标记化服务器接收盲化标记R'
38
接收由DS计算机发送的（N,m）
39
用N去盲化R'以得到
$tok_i = F'(k, uid_i)$
40
在数据库中存储
(uid_i,m)
41
结束

图3c

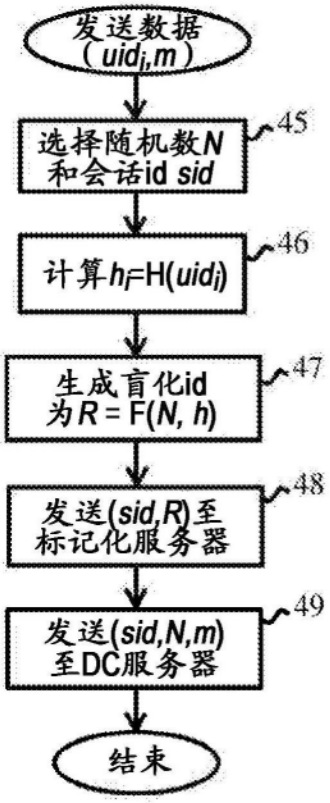

发送数据
（uid_i,m）
选择随机数N和会话id sid
45
计算h_F=H(uid_i)
46
生成盲化id为R = F(N, h)
47
发送(sid,R)至标记化服务器
48
发送(sid,N,m)至DC服务器
49
结束

图4a

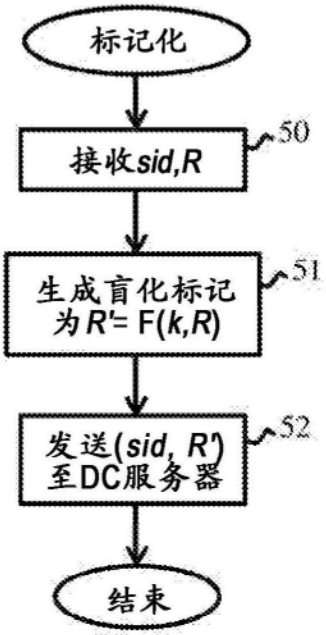

标记化
接收sid,R
50
生成盲化标记
为R'= F(k,R)
51
发送(sid, R')
至DC服务器
52
结束

图4b

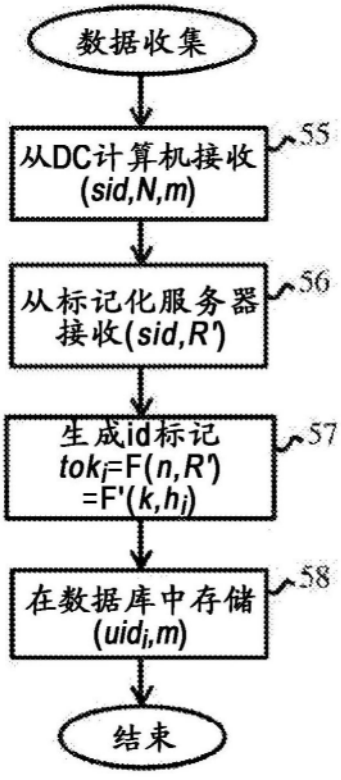

数据收集
从DC计算机接收
(sid,N,m)
55
从标记化服务器
接收(sid,R')
56
生成id标记
tok_j=F(n,R')
=F'(k,h_j)
57
在数据库中存储
(uid_i,m)
58
结束

图4c

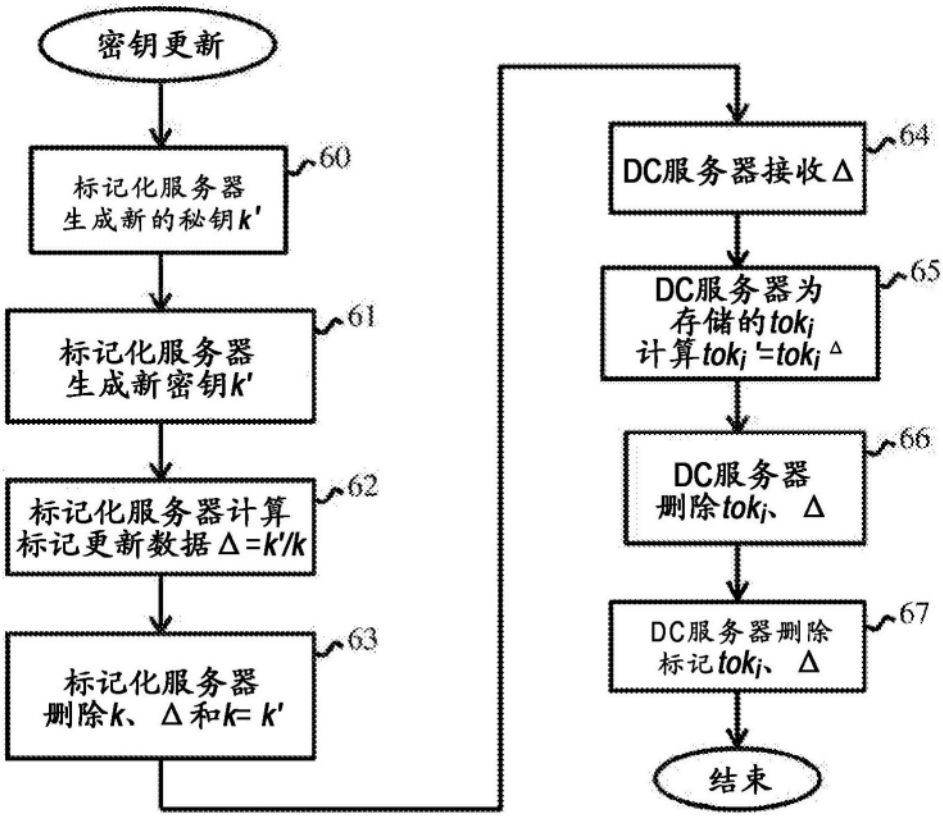

密钥更新
标记化服务器
生成新的秘钥k'
60
标记化服务器
生成新密钥k'
61
标记化服务器计算
标记更新数据Δ=k'/k
62
标记化服务器
删除k、Δ和k= k'
63
DC服务器接收Δ
64
DC服务器为
存储的tok_i
计算tok_i '=tok_i Δ
65
DC服务器
删除tok_i、Δ
66
DC服务器删除
标记tok_i、Δ
67
结束

图5

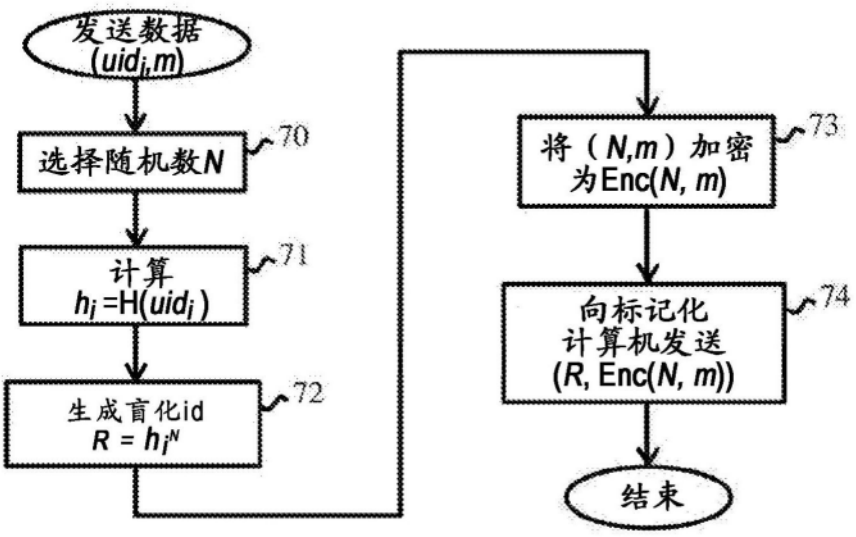

发送数据
(uid_i,m)
选择随机数N
70
计算
h_i =H(uid_i)
71
生成盲化id
R = h_i^N
72
将（N,m）加密
为Enc(N, m)
73
向标记化
计算机发送
(R, Enc(N, m))
74
结束

图6a

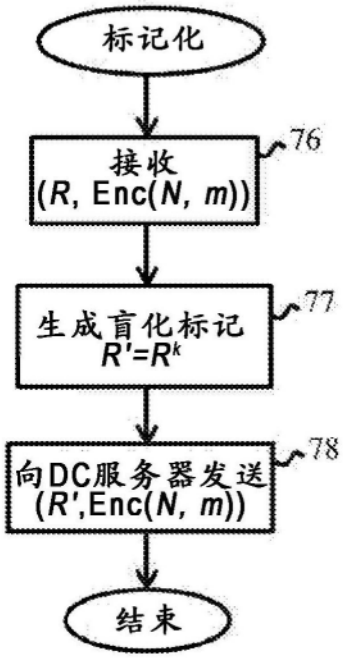

标记化
接收
(R, Enc(N, m))
76
生成盲化标记
R'=R^k
77
向DC服务器发送
(R',Enc(N, m))
78
结束

图6b

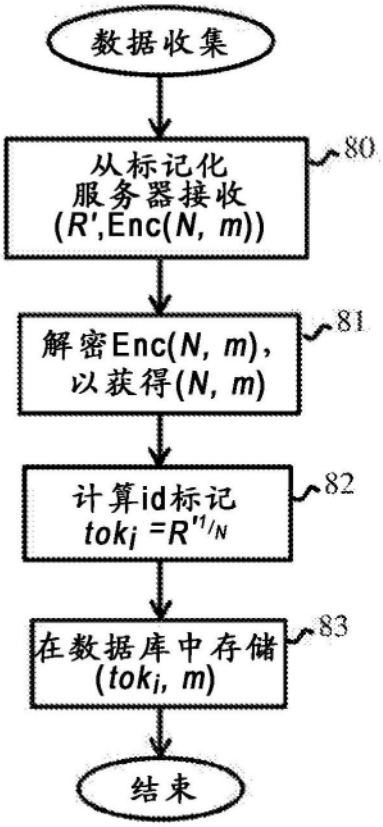

数据收集
从标记化
服务器接收
(R',Enc(N, m))
80
解密Enc(N, m),
以获得(N, m)
81
计算id标记
tok_i =R''^1/N
82
在数据库中存储
(tok_i, m)
83
结束

图6c