基于时间约束的隐私保护数据查询方法研究

邓海生[1],刘 啸[2],李军怀[2],王珊歌[2]
(1. 西京学院 教务处,陕西 西安 710123;
2. 西安理工大学 计算机科学与工程学院,陕西 西安 710048)

摘 要:很多学者和机构在研究隐私保护的分布数据挖掘方法时,通过挖掘全局数据以保护各站点数据的隐私和安全。但是这些方法假设数据集成已经完成,隐私保护数据挖掘处理的是集成问题解决后的知识获取问题。因此,在隐私保护数据处理之前的数据集成中,如何保护来源数据的隐私信息,是一个必须解决的问题。文章在考虑数据的时效性因素下,提出了一种采用 Shamir's 秘密共享方法的时间约束隐私保护数据查询方法,重点介绍了时间约束下隐私保护数据集成与共享中的聚集操作方法。实验结果表明文中方法可以有效提高隐私保护数据查询的效率,大大降低隐私保护数据查询的响应时间。

关键词:隐私保护;时间约束;数据共享;数据查询
中图分类号:P309.2 **文献标识码**:A **文章编号**:1673-629X(2013)10-0119-04
doi:10.3969/j.issn.1673-629X.2013.10.030

Research on Privacy Preserving Data Query Method Based on Time-constrained

DENG Hai-sheng[1], LIU Xiao[2], LI Jun-huai[2], WANG Shan-ge[2]
(1. Dean's Office, Xijing University, Xi'an 710123, China;
2. College of Computer Science & Engineering, Xi'an University of Technology, Xi'an 710048, China)

Abstract: Many scholars and institutions protect the site data privacy and security through the global data mining in the research of distributed data privacy protection. But these methods assume that the data integration has been completed, processing of privacy preserving data mining is the problem of obtaining knowledge after the integrated problem solving. Therefore, before the privacy preserving data processing in data integration, how to protect the privacy information of data sources is a problem that must be solved. Propose a privacy-preserving data sharing method applied the Shamir's secret sharing method with time-constrained. Some basic operations utilized this method for the privacy preserving and data integration and sharing, such as intersection, join and aggregation, are emphatically illustrated. The experimental results demonstrate that this method is effective and greatly improves the efficiency and significantly reduce response time of the privacy-preserving data query.

Key words: privacy preserving; time-constrained; data sharing; data query

0 引 言

隐私保护是数据挖掘和数据集成面对的一个重要问题,不但在金融、医疗和卫生保健等传统领域[1-2],而且随着互联网的广泛应用,电子商务和社会化网络的隐私数据保护[3-5]也越来越为人们所关注。很多学者和机构在研究隐私保护的分布数据挖掘方法,通过挖掘全局数据以保护各站点数据的隐私和安全[6-7]。但是这些方法都是假设数据集成[8]已经完成,隐私保护数据挖掘处理的是集成问题解决后的知识获取问题。因此,在隐私保护数据处理之前的数据集成中,如何保护来源数据的隐私信息,也是一个必须解决的问题。

目前在隐私保护数据集成与共享研究方面,主要有基于信任第三方的方法、安全多方计算、随机混乱等方法[9-11]。信任第三方的方法存在第三方信任危机。在基于安全多方计算方法中,由于计算复杂性和数据库中大量元素参与计算使其代价较高。Naor 等人提出基于密码学技术的茫然传输和多项式进行交查询处

收稿日期:2012-11-28 **修回日期**:2013-03-28 **网络出版时间**:2013-07-24
基金项目:陕西省科技攻关项目(2009K08-24,2011NXC01-12);陕西省教育科技项目(09JK659);西安市科技项目(CXY09020)
作者简介:邓海生(1980-),男,硕士,讲师,研究方向为移动计算技术、数据挖掘。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130724.0953.015.html

理[12],这个方法仅仅能保证数据源得到交查询结果。Agrawal 等人提出的用最小信息共享模式进行交和并操作[13],也存在两个缺点:

1)在计算过程中,加密和解密具有非常高的操作成本,同时查询响应时间也非常高。

2)它不支持聚集的查询。

在上述隐私保护集成与共享方法中,几乎都基于这样一个假定,即所保护的对象数据库是永远有效的。在这种情况下,没有任何特性表明数据何时变得有效,何时又被认为无效。同样,目前无效的数据也没有说明它在过去或将来是否有效。而事实上,很多情况下,数据库的有效性是和时间紧密相关的,这样带有时间约束的隐私保护就显得非常有价值。基于这样的考虑,文中探索了一种采用 Shamir's 秘密共享方法[14]的时间约束隐私保护数据查询方法,重点介绍了隐私保护数据集成与共享中的聚集操作方法。

1 时间约束下的隐私保护数据查询方法

隐私保护数据查询:$D_1, D_2, \cdots, D_n$ 分别是存储在数据源 $P=\{P_1, P_2, \cdots, P_n\}$ 中的数据表,q 是数据表 D_1 到 D_n 上的一个查询。查询的目的是在没有泄露额外信息给其他数据源的同时,计算出 q 的查询结果。

从隐私保护查询处理角度来看,交、并、聚集查询是特别具有挑战性的。隐私保护查询处理方法一般使用单向加密散列函数,如 SHA 或者 MD。在这种方法中,数据源对各自列使用散列函数加密并且发送散列值给第三方。第三方比较这些散列值并得到查询结果,发送给各个数据源。然而,在实际中,单向散列函数的数量是有限的,可信第三方可能会提取散列值中的原始数据。如果数据库区域小的话,第三方可以遍历得到各个数据源的原始数据。同时,这种方法难以解决多数据源的隐私保护聚集查询操作,而 Shamir's 秘密共享方法可以较好地解决上述问题。为了提高查询效率,文中引入了时间约束机制,如果数据库中的每个元组均有其有效时间,那么在数据库中所发现的知识也必然有相应的时间约束,以表明所发现的知识何时是有效的。因此,对于隐私保护数据集成与共享中的数据查询操作,引入时间约束因素,将更加有利于系统的安全以及隐私的保护,也更符合实际的情况。

文中在进行隐私保护数据查询操作中,采用如图 1 所示的信任第三方模型。在此模型中,隐私保护查询处理主要通过信任第三方有效地完成,如果第三方不和任意数据源串通泄露隐私,那么它就是可信的。用户提出查询,数据源首先对数据集做预处理,即对数据进行时间约束处理,然后使用加密函数对私密信息进行加密,发送数据给第三方。随后,第三方根据查询要求,对各个数据源发过来的数据进行计算,将其最终结果发送给用户。每个数据源不会知道其他数据源的额外信息,第三方也不知道每个数据源的其他信息,用户也仅仅看到最后的查询结果。

2 基于 Shamir's 秘密共享算法的隐私保护查询操作

2.1 Shamir's 秘密共享算法

秘密共享是指在一个由若干个($n>0$)不同实体所构成的集合中,允许一个被称为密钥分发者(Dealer)的人专门处理密钥,而原始密钥也称为主密钥,被这若干个实体也称作秘密分享者或者玩家(shareholder/Player)分享,结果只有特定的若干密钥分享者构成的子集才能恢复出原始密钥。特别地,若原始密钥必须由所有秘密分享者才能恢复出,这时提出的方案就是一种最简单形式的一般秘密共享方案;若给出一个门限值 t,$t<n$,只需 t 个分享者就能恢复出原始密钥,而任何一个含有元素个数比 t 少的子集都不能得出有关密钥的任何信息,此时得出的方案就是(t,n)门限方案。

下面重点介绍采用 Shamir's 秘密共享算法,加入时间属性的基础上进行隐私保护聚集查询的方法。

图 1 为隐私保护数据模型。

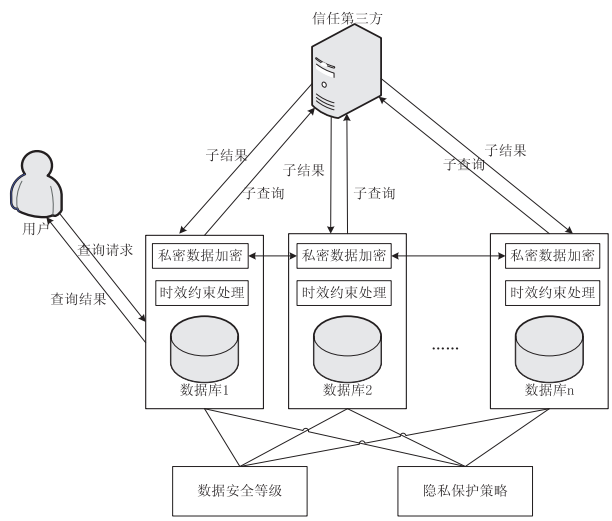

图 1 隐私保护数据模型

2.2 时间约束下隐私保护聚集查询操作

传统的聚集操作通常是多个表中列的聚集,比如:求和、平均值或者最小/最大值。然而,隐私保护聚集考虑的数据源隐私性,即不同数据源的聚集,除了知道最终查询结果外,并不知道其他额外信息。

聚集查询定义:$T_1, T_2, \cdots, T_n$ 分别是存储在数据源 $P=\{P_1, P_2, \cdots, P_n\}$ 中的数据表,每个数据表包含一个 Key 和 Value 属性,$Q=\{Q_1, Q_2, \cdots, Q_n\}$ 是 n 个第三方。数据源 P_i 想要查询在所有数据库中具有相同属性 Key

的 Value 值的聚集,就是要在第三方 Q 的协助下,获得查询 q,并且不泄露任何额外的信息给第三方 Q,第三方仅仅提供查询结果给数据源 P_i。

如前面所述,时间特性是数据库的固有属性,因此在隐私保护数据查询操作中引入时间约束,更符合现实实际的情况。引入时间约束后,加入第三方 Q_0,表示接收的公开数据。隐私保护聚集操作分以下几种情况(以聚集和为例,N 为数据源)。

首先遍历表中的所有元组,比较当前时间与有效结束时间,在有效时间之前的元组,则认为是有效数据,即不能公开的数据,否则视为无效,是公开数据。

(1)N 个数据源中具有相同标识符 Key 的 Value 值全是公开数据:

Q_0收到每个数据源发送的无效数据,Q_0计算出 N 个表中具有 N 个相同标识符 Key 的 Value 值之和 Public Sum=$v_1+v_2+\cdots+v_n$;

(2)N 个数据源中具有相同标识符 Key 的 Value 值全是私密数据,即采用不加时间约束的处理方法。查询被提出之后,数据源使用 $k-1$ 个散列函数构造多项式,并且选择 n 个任意值 $X=\{x_1,x_2,\cdots,x_n\}$,X 的取值描述在上述交查询中。然后数据源构造两个多项式命名为 $q(x)$ 和 $r(x)$,分别对每个数据表中的 Key 和 Value 属性加密。

多项式 $q(x)$ 和 $r(x)$ 被构造如下:

$$q(x)=H_{k-1}(\text{Key})x^{k-1}+\cdots+H(\text{Key})x+\text{Key}$$
$$r(x)=H_{k-1}(\text{Key})x^{k-1}+\cdots+H(\text{Key})x+\text{Value} \tag{1}$$

对 Sum 和 Average 的查询,每个数据源选择不同的 $k-1$ 阶的多项式 $r(x)$,因为每个数据源有相同关键字 Key,因此不能使用相同的 $r(x)$。

例如:对于和查询计算,假设 n 个数据源共同具有 L 个具有相同 Key 的私密值 Value 值,v_1 到 v_l,如下所示是具有相同属性 Key 的 Value 私密值经加密后的总和:

$$\begin{aligned}&a_1x_i^{k-1}+b_1x_i^{k-2}+\cdots+v_1+\\&a_2x_i^{k-1}+b_2x_i^{k-2}+\cdots+v_2+\cdots\\&a_lx_i^{k-1}+b_lx_i^{k-2}a_1x_i^{k-1}+v_l\end{aligned} \tag{2}$$

因此 Q_i发送总和是 $\text{RES}_i=(a_1+a_2+\cdots+a_l)x_k^{k-1}+\cdots+\text{Sum}$,即具有相同属性值 Key 的 Value 私密值的总和。每个数据源收到 n 个第三方的结果:

$$\begin{aligned}&\text{RES}_1=(a_1+a_2+\cdots+a_l)x_1^{k-1}+\cdots+\text{Sum}\\&\text{RES}_2=(a_1+a_2+\cdots+a_l)x_2^{k-1}+\cdots+\text{Sum}\\&\cdots\\&\text{RES}_n=(a_1+a_2+\cdots+a_l)x_k^{k-1}+\cdots+\text{Sum}\end{aligned} \tag{3}$$

因为每个数据源都知道 $X=\{x_1,x_2,\cdots,x_n\}$,在这些方程组中有 k 个未知数,包括 Sum,并且 $n>k$。因此原始数据总和 Sum 可以从方程组中计算出来,但是数据源各方并不知道彼此的单个数据。

(3)N 个数据源中具有相同标识符 Key 的 Value 值,其中一些 Value 值可以公开,而其他 Value 值不可以公开:

假定 N 个数据源中具有相同标识符 Key 的 Value 值只有 P_1为私密项,则认为 P_1数据被泄漏,所以聚集和为其他数据源公开的数据之和加上此 Value 值;

假定 N 个数据源中具有相同标识符 Key 的 Value 值为私密数据的个数 S,并且 $S\geqslant i(i=2,3,\cdots,n-1)$,则 S 个数据源对其数据加密后,发送给 S 个第三方 Q_s,Q_s收到 s 个数据源经过加密后的数据,第三方 Q_s计算出 S 个表中具有相同标识符 Key 的私密值 Value 的总和 RES_s;

$$\begin{aligned}&a_ix_s^{k-1}+b_ix_s^{k-2}+\cdots+v_i+\cdots+\\&a_jx_s^{k-1}+b_jx_s^{k-2}+\cdots+v_j\end{aligned} \tag{4}$$

因此 Q_s发送给 S 个数据源结果是 $\text{RES}_i=(a_i+\cdots+a_j)x_i^{k-1}+\cdots+\text{Sum}$,即具有相同标识符 Key 的私密值之和。$S$ 个数据源收到 S 个第三方的结果如下所示:

$$\begin{aligned}&\text{RES}_i=(a_i+\cdots+a_j)x_i^{k-1}+\cdots+\text{Sum}\\&\cdots\\&\text{RES}_s=(a_i+\cdots+a_j)x_k^{k-1}+\cdots+\text{Sum}\end{aligned} \tag{5}$$

2.3　时间约束下并、交操作

交查询处理定义:$L_1,L_2,\cdots,L_n$分别是存储在数据源 $P=\{P_1,P_2,\cdots,P_n\}$ 中的私密项,$Q=\{Q_0,Q_1,Q_2,\cdots,Q_n\}$是 $n+1$ 个第三方,Q_0为各个数据源发送的可以公开的数据。P_1提出查询 $q=L_1\cap L_2\cap\cdots\cap L_n$,每个数据源使用 Shamir's 秘密共享对不能公开的数据进行加密,在第三方 $Q_i(i=0,1,2,\cdots,n)$ 的协助下,获得查询结果 q(包括 Q_0发送公开数据的聚集查询 q),并且不泄露任何额外的信息给第三方 $Q_i(i=1,2,\cdots,n)$,第三方 Q(除 Q_0外)仅仅提供查询结果给数据源 P。

并查询处理定义:假设 P_1存在数据表 T_1和 P_2存在数据表 T_2,每个表中存在特定属性 A,这个查询处理计算 $T_1\infty T_2$,即数据源 P_1仅仅知道元组 t,$t\in P_2$,$t.A\in T_1.A$。

并查询处理包含两个阶段:交阶段和并计算阶段。首先 P_1和 P_2根据上述交查询处理过程计算出交查询结果;然后在并计算阶段,P_2仅仅发送与 P_1相交行的所有数据信息,同样 P_1也仅仅发送与 P_2相交行的所有数据信息,在此过程中,并没有任何其他额外的信息泄漏。

3 实验分析

如图 2 所示,由于基于时间约束的隐私保护交、并查询中,有许多的数据处于保密时间之外,因此可以不需要隐私保护处理,可以直接共享,使得其查询处理时间明显低于不带时间约束的交、并查询操作。

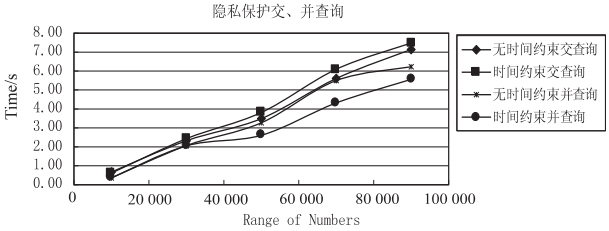

图 2 时间约束隐私保护数据交、并查询响应时间对比

如图 3 所示,基于时间约束的聚集查询时间比不带时间约束的查询响应时间稍小。

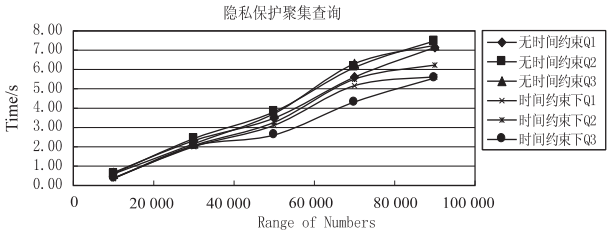

图 3 时间约束前后隐私保护数据聚集查询响应时间对比

不带时间约束的聚集查询,当数据量为 90 000 时,所有第三方发送给用户 P_1 具有相同 Key 的加密聚集和结果为 40 000 多条记录,需要解三元方程组求出原始数据聚集和,即用户解密过程,这将花费很长的时间。

与不带时间约束的聚集查询相比,带时间约束的聚集查询解密时间相对短很多,当数据量为 90 000 时,所有第三方发给用户 P_1 具有相同 Key 的加密聚集和结果为 10 000 多条记录,即需要解三元三次方程组求出聚集和;收到 Q_2,Q_3 具有相同 Key 的加密聚集和为 800 多条记录,需解二元一次方程组求出原始数据聚集和;并且收到 Q_2 和 Q_3 具有相同 Key 的聚集和的个数为 10 000 条记录,这些数据不需要加密,只要加上 P_1 具有相同标识符的 Key 的 Value 值(可能包括私密项),即可求出这些数据聚集和。

因此,解密的过程明显也优于不带时间约束的聚集查询。所以带时间约束的聚集查询处理时间优于传统的查询方式。

4 结束语

隐私保护数据集成和共享中的数据时效性往往被人们所忽视,文中在利用 Shamir's 秘密共享方法进行隐私保护数据查询处理中,引入了数据的时效特性,探索了基于时间约束的隐私保护数据交、并和聚集操作方法,并进行了一系列的实验分析和比对。实验结果表明,文中方法可以有效提高隐私保护数据查询的效率,并大大降低隐私保护数据查询的响应时间。进一步的工作将集中在时间约束下的隐私保护数据泄漏度量和分析方面。

参考文献:

[1] Mohammed N. Privacy-preserving data mashup[C]//Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology. Saint Petersburg, Russia: [s. n.], 2009.

[2] Solanas A. Privacy protection with genetic algorithms[M]// Success in Evolutionary Computation. Berlin: Springer, 2008: 215-237.

[3] Fienberg S E. Privacy and confidentiality in an e-commerce world: Data mining, data warehousing, matching and disclosure limitation[J]. Statistical Science, 2006, 21(2): 143-154.

[4] Felt A, Evans D. Privacy protection for social networking platforms[C]//Proc of Web 2.0 Security & Privacy (W2SP). Oakland, CA: [s. n.], 2008.

[5] Zheleva E, Getoor L. Privacy in social networks: A survey [M]//Social Network Data Analytics. [s. l.]: [s. n.], 2011: 277-306.

[6] 李军怀,刘海玲,彭 军,等. 一种基于时态约束的关联规则隐私保护方法[J]. 计算机科学,2009,36(9):201-204.

[7] 王智慧. 信息共享中隐私保护若干问题研究[D]. 上海:复旦大学,2007.

[8] 李玉华,卢正鼎,孙小林,等. 基于隐私保护的语义数据集成[J]. 华中科技大学学报,2005,33(Sup):128-130.

[9] Emekci F. Privacy preserving query processing using third parties[C]//Proc of ICDE. Washington, DC: IEEE Computer Society, 2006: 27-27.

[10] 汤 琳,何 丰. 隐私保护的数据挖掘方法的研究[J]. 计算机技术与发展,2011,21(4):156-158.

[11] 罗永龙,徐致云,黄刘生. 安全多方的统计分析问题及其应用[J]. 计算机工程与应用,2005(24):141-143.

[12] Naor M, Pinkas B. Oblivious transfer and polynomial evaluation[C]//Proc of STOC. New York: ACM, 1999: 245-254.

[13] Agrawal R. Information sharing across private databases[C]// Proc of SIGMOD. New York: ACM, 2003: 86-97.

[14] Shamir A. How to share a secret[J]. Communications of the ACM, 1979, 22(11): 612-613.

基于时间约束的隐私保护数据查询方法研究

作者： 邓海生， 刘啸， 李军怀， 王珊歌， DENG Hai-sheng， LIU Xiao， LI Jun-huai， WANG Shan-ge

作者单位： 邓海生,DENG Hai-sheng(西京学院 教务处,陕西 西安,710123)， 刘啸,李军怀,王珊歌,LIU Xiao,LI Jun-huai,WANG Shan-ge(西安理工大学 计算机科学与工程学院,陕西 西安,710048)

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2013(10)