|  |
| --- |
|  |
| 数据库应用多方安全计算的理论探究 |
| 作者： 齐风腾 侯柏韬  指导老师： 魏普文 |
|  |
|  |
|  |

|  |
| --- |
| Apr.20.2018 |

摘要

当前数据信息价值通过挖掘和分析而不断地被人们所重视，因此作为数据信息的载体，对数据库资源进行整合并对多数据库进行联合计算将会是未来信息发展的重点。本文介绍了数据扰乱、OT不经意传输、零知识证明等常用的数据隐私处理、交互证明和多方安全计算中一些经典算法，并通过这些技术构建了一个适用于数据库应用的多方安全计算算法，以期解决当数据库群体面临资源共享需求时进行共同计算并同时保护己方隐私的问题。同时我们对该算法进行了实践，详尽分析了数据安全性和计算效率。为该算法的

**关键词**：多方安全计算；不经意传输；零知识证明；数据库；联合计算；

**Abstract**

The value of current data information is continuously valued by people through mining and analysis. Therefore, as a carrier of data information, the integration of database resources and the joint calculation of multiple databases will be the focus of future information development. This paper introduces data privacy, OT careless transfer, zero-knowledge proofs and other commonly used data privacy processing, interaction verification and multi-party security calculations in some classic algorithms, and through these technologies to build a multi-party security calculation algorithm for database applications, in order to solve When the database community faces the need for resource sharing, it conducts common calculations while protecting the privacy of its own. At the same time, we implemented the algorithm and analyzed the data security and computational efficiency in detail.

**Keywords：**Multi-party security calculation; oblivious transfer; zero knowledge proof; database; Joint calculation

[摘要 1](#_Toc514082593)

[介绍 3](#_Toc514082594)

[背景 3](#_Toc514082595)

[OT不经意传输 3](#_Toc514082596)

[零知识证明 4](#_Toc514082597)

[双方求和安全计算 5](#_Toc514082598)

[多方安全求和计算算法 6](#_Toc514082599)

[核心构建 6](#_Toc514082600)

[安全度分析 7](#_Toc514082601)

[数据挖掘影响分析： 7](#_Toc514082602)

[数据挖掘的特征 7](#_Toc514082603)

[数据挖掘对隐私的要求 8](#_Toc514082604)

[数据挖掘的影响度 8](#_Toc514082605)

[现状分析 8](#_Toc514082606)

[数据扰乱技术 8](#_Toc514082607)

[数据加密技术 8](#_Toc514082608)

[多方安全计算技术 9](#_Toc514082609)

[应用 9](#_Toc514082610)

[修正 10](#_Toc514082611)

[调度节点S 10](#_Toc514082612)

[计数器counter 10](#_Toc514082613)

[简化矩阵组参数p 11](#_Toc514082614)

[数据检验 11](#_Toc514082615)

[结语 12](#_Toc514082616)

介绍

多方安全计算概念的提出源于1982年YAO A.C在23届IEEE上做出的关于《Protocols for secure computations》的报告[1]。随后在1987年由O.Goldreich、S.M icali和A.Wigderson三位学者提出了密码学安全的可以计算任意函数的多方安全计算协议[2]。自此之后，多方安全计算协议在这个思想引导下又得到了诸多有益的新进展，在现实生活中的应用也愈加成熟。在实现基本加法、乘法运算的基础上，在几何计算、隐私保护、统计分析方面都有着广泛的应用，并且形成了电子投票、群包含验证等应用协议。

下文我们将从多方安全计算的理论基础和基本算法入手，介绍经典的双方求和安全计算协议和双方比较安全计算协议，并推广为多方安全计算协议，其中将详细阐明所涉及的一些重要方法，并且我们将多方求和协议应用在数据库信息查询中，实现多数据库之间的共享计算。

背景

信息技术的快速发展使得信息数据库作为21世纪新的重要资源而得到了人们的重视，每个数据库中含有的信息就是数据库的核心所在，也是数据拥有者的资本。但单个数据库的存储量终究受限于存储条件、成本、信息源等因素。因此信息资源共享成为新技术的必要手段。但是各个数据库的拥有者都要求本身私有信息的不可泄露或知识版权等问题，在资源共享过程中完成公共查询要求并不泄露自身的保密数据，这正是安全多方计算绝佳的应用场景。

正是数据库资源共享的需求，使得实现数据库间的安全多方计算的要求应运而生。安全多方计算是对多个计算参与方共同为了获得目标结果进行求和、乘积、比较等操作的总称，安全性的定义是指在计算过程中，各参与方除己方具体数据和最终结果知道以外，无法获得额外的知识。我们称该过程为数据联合计算的隐私保护过程。

## OT不经意传输

OT不经意传输(oblivious transfer)提供了一种进行隐私信息交互的方式[3]。在这种方式下，存在含有n个输入值的发送方，而接收方希望接收到其中的某一个值；并且接收方收到信息值时不能拥有其他信息值，发送方也不知道接收方选择的信息值。在实际应用中多采用其发展出来的传输协议：

步骤1：发送方含有输出值，等待接收方的请求；

步骤2：接受方希望接受的输入值编号为，计算单向函数列发送给发送方，其中函数满足RAND函数性质[4]；

步骤3：发送者接受函数列并返回目标值；

步骤4：接受方计算，得到原数值。

## 零知识证明

零知识证明是指知识的掌握者通过交互或非交互的方式使验证者相信自己拥有知识，并且不能使验证者获取自己的知识信息的过程[5]。零知识证明在我们的生活中应用十分广泛，如我们常用的电子卡身份认证协议即采用Schnorr身份识别协议[6]。一般来讲，零知识证明的构造过程大多是以知识拥有者向验证者证明知识m的表现形式的方式，其表现形式函数可以是m的数学特征，也可以是m的一些物理或其他群体性特性。

以下为QR零知识证明系统的运作过程[7]：

步骤1：Alice随机选择整数r，并将发送给Bob；

步骤2：Bob在{0、1}中随机选择数字i，发送给Alice；

步骤3：Alice计算,发送给Bob；

步骤4：Bob验证成立；

步骤5：以上步骤重复m次，验证通过即Bob相信Alice中x为n的QR。

一份零知识证明协议具有完全性、合理性和零知识性，完全性是指如果证明结果命题为真，则验证方概率为1的接受证明成立；合理性是指如果证明结果命题为假，则验证方承认证明结果为真的概率最多为；零知识性是指，证明过程没有泄露包括知识在内的额外知识。零知识性的证明过程即：

通过构造模拟机M，其不知道是否x为n的QR，但是通过随机产生的方式参与到证明体系中来，模拟Alice完成验证过程，在m次步骤中产生的数据和真实协议中产生的数据统计不可区分，则可以认为零知识性成立。

## 双方求和安全计算

2001年Du W L和Atallah M J[8]在（Computer Security Foundations Symposium）报告中提出了以下双方安全线性方程组计算协议，满足Alice和Bob在不暴露自身隐私数据的情况下[9]，计算线性方程组（M1+M2）x=V1+V2。

步骤1：Bob生成可逆矩阵P、Q；

步骤2：成矩阵子列满足；向量子列, 满足；

步骤3：Bob生成矩阵子列满足；向量子列, 满足；

步骤4：生成随机数并组成矩阵序列和向量序列,并发送给Bob；

步骤5：Bob对矩阵序列遍历计算，并对向量序列计算并发送和；

步骤6：Alice得到与;

步骤7：Alice计算线性方程组的解，并发送给Bob；

步骤8：Bob反解可逆矩阵计算结果并公布。

# 多方安全求和计算算法

参考以上知识，我们将双方安全求和技术扩展为多方安全求和技术，并且考虑到数据库数据信息大、对通讯和计算等要求较高的特点，进行了改进，其核心构建和安全性分析如下。

核心构建**：**

在隐私数据拥有者彼此之间可分布式交换流量的（i>2）环境下，实现数据交换并计算所拥有隐私数据矩阵之和的多方安全求和计算协议：

步骤1：各节点拥有隐私数据矩阵，矩阵规模，协商参数，使得穷搜个矩阵之和是不可能的；

步骤1.1：各节点生成矩阵子列满足；

步骤1.2：各节点生成随机矩阵列，满足 和可逆矩阵；

步骤2：首节点生成扩展矩阵组，其中；而为同规格随机矩阵，标记矩阵记录扩展矩阵分组中所在位置；

步骤2.1：首节点将扩展矩阵发送给下一节点，节点对扩展矩阵、矩阵子列和随机矩阵列进行矩阵运算；

步骤2.2：节点将从节点收到的扩展矩阵和本节点矩阵子列、可逆矩阵、随机矩阵按步骤2.1进行矩阵运算；

步骤3.1：首节点收到遍历参与节点后的扩展矩阵列，逆序依次接收节点的随机矩阵，并计算，计算完成后发送；

步骤3.2：节点反解可逆矩阵；并向节点返回扩展矩阵；

步骤3.3：循环直至首节点接收到遍历参与节点后的扩展矩阵列，通过自身矩阵还原原矩阵；

步骤4：公布求和结果。

安全度分析：

对于数据库节点来说，其安全性依靠茫然传输的保护。节点的信息经过分解矩阵处理后，真实的数据矩阵隐藏在扩展矩阵组中，通过茫然传输发送给下一节点，使得在给出的矩阵组中，只有一半的数据是真实的数据矩阵。根据协商参数我们构建的矩阵组是的矩阵组，在不知道标记矩阵的情况下，其数据被还原的概率为，所以安全性依靠于参数，我们定义首节点安全度。

对于所有参与节点来说，安全性依赖于矩阵分解的保护。每个节点的真实数据矩阵被分解为矩阵子列，我们的子矩阵是随机生成的，保证了子矩阵不易被轻易还原为父矩阵，

共谋节点下，节点信息安全性依赖于随机矩阵R的数据扰乱：首先我们假设共谋节点为间隔为一的相邻节点和，则对扩展矩阵组进行操作得到，因为对子矩阵进行了随机子矩阵R的扰乱，隐藏了子矩阵的真实数据，但是由于随机矩阵具有良好的随机性，存在着子矩阵的数据分布可以被显示出来的问题，因此我们引入了可逆矩阵，修正单纯随机矩阵加和所带来的分布可显的问题。其次我们假设共谋节点为调度节点和非调度节点，这部分内容会在章节:应用>修正>4中说明。

数据挖掘影响分析：

数据挖掘的特征**：**数据挖掘就是对数据库中的信息进行查询、查找或挖掘并获取有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程的一种行为[10]，这也是数据库中信息发挥价值的主要方式。通过对数据的统计特征、物理特征、社会特征等进行提取，依靠分类、聚类、关联规则等逻辑方式进行整理。

数据挖掘对隐私的要求：简单来说数据挖掘就是对数据的查询和处理，一般来说，大数据挖掘关心的是群体数据所体现出来的总体特征（我们称之为共同隐私），但并不会涉及个人隐私，在目的上数据挖掘和个人隐私之间并不是冲突对立的关系[11]；但是在对数据特征的查询和分析方法上，必然会涉及用户的具体数据，而我们隐私保护工作要做的就是，在数据的查询过程中使查询过程不直接接触用户的具体数据或者对用户数据进行掩盖，完成数据探查和分析工作。

数据挖掘的影响度：可以看到，主流数据挖掘下的挖掘方式均是基于数据实体所含有的数学统计特征、关联特征进行挖掘分析。至于数据的统计特征和关联特征是底层技术，数据挖掘的分析工作只要求查询的结果，而对于特征的获取途径并没有要求。因此多方安全求和计算算法对于数据挖掘工作并没有影响。

现状分析：

对数据库中信息的数据特征进行保护的技术有很多，就目前已经做的工作来说，针对不同方面大致从以下几个角度出发：

数据扰乱技术[12]:数据库中的数据多为真实有效的数据，可以直接反映出数据对象的全部特征，但在范围性查询过程中，往往只需要提取数据整体的某些数据特征，因此通过向数据添加噪声来增加扰乱，使得扰乱后的数据和原始数据的某些基本性质保持不变，而其他信息特征的失真使得攻击者无法重构原始数据。

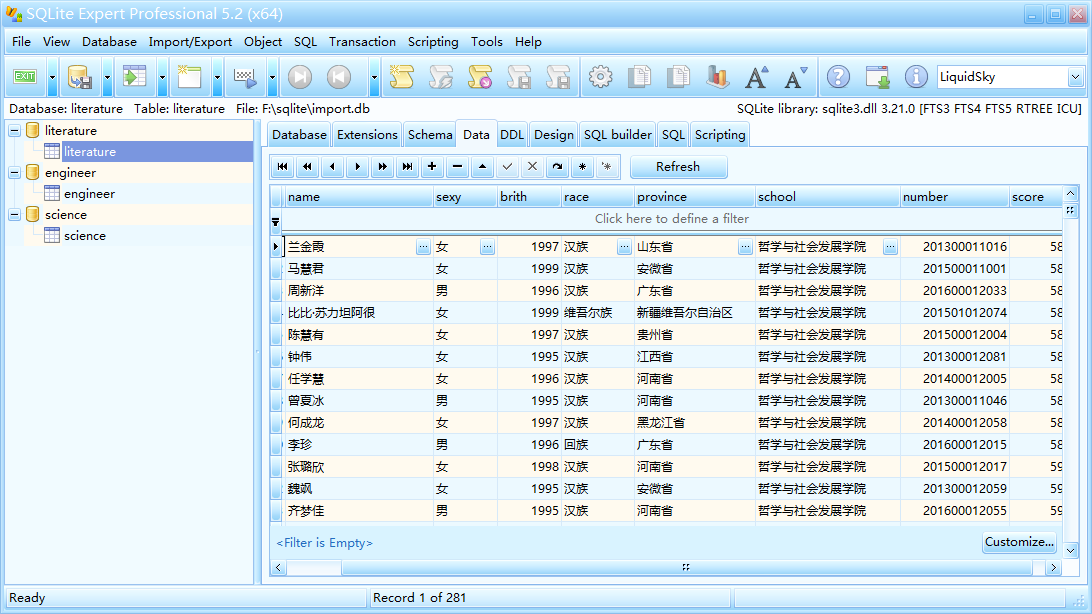
但数据扰动技术的缺点在于使用者必须提前知晓应用的全部特征结构，同时需要对某次查询的目标特征十分清楚，保证产生的噪音不会影响某些不易察觉的数字特征，并且针对每个应用的具体特征的差异，每次不同的应用都需要重建。显然不利于多样环境的推广。

数据加密技术[13]：一些加密算法的优良特性使得对加密过的数据的计算不会影响原始数据的计算结果，如RSA加密算法、ElGamal同态加密方案等算法的同态加密特性可以满足对数据的乘积运算同态。

但数据加密技术的缺点在于运算过程时间长、消耗大，对大数据库之间的交互计算来说，是不现实的。在通常的应用中，哪怕是简单的加密通信，也常使用对称加密算法进行通讯，而只使用非对称加密算法进行密钥协商。

多方安全计算技术：相较而言，通过多方安全计算算法所实现的数据库数据传输过程。对比数据扰乱技术，适用性强，可针对不同环境下的数据实体进行操作，计算的结果与对数据原型直接计算的结果一致，区别只在于计算过程不直接暴露数字实体，因此计算结果不会出现偏差，并且计算过程中不需要针对某一数据特征对数据进行修改，因而不会影响对结果的再分析；对比密码加密技术，不需要对选取加密算法的特征有特别的要求，加密过程不涉及或少涉及对数据的数学性操作，多为加减简单运算，考虑到数据库操作多为大数据，计算总量巨大，对提供多用户查询的数据库来说可以有效削减加密方式所带来的负载过重的问题。

应用:

为了验证我们给出的多方安全计算算法的实际可行性和安全性，我们参照一些实际参数模拟生成了1500个用户数据，并将这些用户数据依照数据特征(school)的不同分成了三个独立的数据库，以模拟分布式数据库环境。图1为我们给出的数据库部分数据截图，可以看出该模拟数据立足于真实数据，对真实数据库中的数据具有代表性。

## 修正

考虑到实际情况，我们在应用中对我们给出的多方安全求和计算算法做了一些修正：

调度节点S。

1. 调度节点S的作用是协调不同服务器之间数据格式、输出数据矩阵大小不唯一、计算格式判断的问题。我们知道不同服务器之间信息存储的方式不同，因此在联合计算请求抵达的时候，需要调度节点D向各查询服务器节点发布数据及格式查询，并在确定参与计算节点和计算格式规范时向各参与节点发布格式要求。

2. 除此以外，为保证能抵抗共谋节点，在计算过程中的通讯顺序由调度节点随机决定并只告知参与节点的下一个节点地址，因此共谋节点并不能确定中间节点，也不能确定中间节点的数据信息。

3. 调度节点只负责对计算数据规范、通讯顺序进行确定，在计算过程中只处理最后结果的完整性对比（参见计数器章节内容）和结果公布，因此不涉及计算的具体过程。

4. 对于调度节点与首节点的共谋：只能获取到首节点自身的真实数据矩阵位置，而与非首节点的数据信息无关；对于调度节点与非首节点的共谋：并不能得到目标节点数据，因为调度节点可以获得目标节点的随机矩阵R，但是共谋节点无法获知之前的节点数据，因此无法确定目标节点的接收扩展矩阵数据，因此无法恢复出目标节点的具体数据值；调度节点与多个非首节点共谋，可以获得目标节点的中间状态并去除随机矩阵R的干扰，但是不能消除可逆矩阵的影响，仍然无法获得目标节点原始数据。

计数器counter。

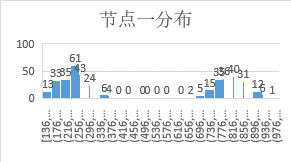
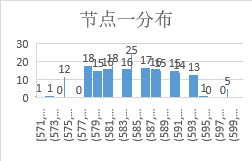
counter由调度节点S提供，构造方式为，其中是随机数，对应每个数据库。计数器的存在为完成一次计算轮的最后一个数据库向调度节点S证明一次计算轮完成，提供了零知识证明参数。

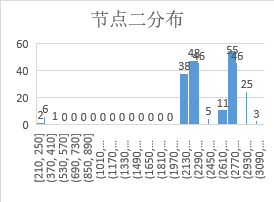
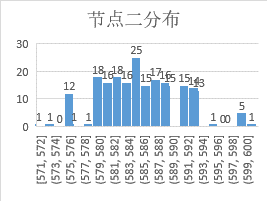
每次计算而随机产生的计数器，总数由调度节点S生成并由调度节点S构成子计数器，在通知各参与计算的数据库标准矩阵信息和相邻节点地址等信息时，计数器由S单独下发给每个参与计算的数据库节点。数据库节点在完成自己的计算操作后将计数器与前数据库节点计数器加和运算并向下传递。最后一个数据库节点计算完成后将计数器传递给调度节点S进行验证。同时对于数据库中的半诚实节点，依靠计数器counter，与总数C匹配则证明此次计算全节点参与，数据可信；否则认为存在节点参与偏差，偏差节点可通过S保存计数器进行还原。

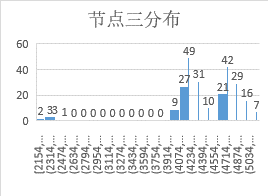
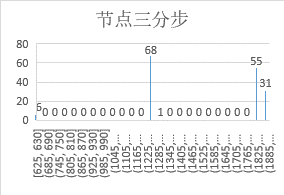
简化矩阵组参数p。根据章节多方安全求和计算算法>安全度分析>茫然传输中的分析，参数p与首节点的信息安全度息息相关，理论上p越大，安全度越高。但是过大的p会增加运算负担和通讯流量，而大幅度拉低算法的计算效率，因此从一般意义上讲，参照密码加密体系对安全性的要求，我们建议p的取值为256较为合适。

## 数据检验

为验证实验数据的准确性，我们对多组数据进行了检验。我们分别输入600组、1500组数据进入算法，在三个不同的数据库节点截取数据流量，并分析了每个节点流量数据与原始数据区别。在此我们定义了节点总数据掩盖值，公式1：







图表2：

图表为三个节点在不同组数据输入时的数据掩盖值，图表2为三个不同节点的原始数据与掩盖后的矩阵数据分布状况。可以看到我们的计算的结果数据包括计算过程中的数据都与原始数据有较大的数据掩盖率，大的数据掩盖率可有效的抵抗数据还原。而图三的数据分布图显示数据分布得到有效掩盖。

# 结语

综上，我们对多方安全计算算法及一些相关的常用隐私数据保护方式进行了探讨，对比并总结了多方安全计算的优劣之处，并以此为基础构建了一种较为合适的多方安全求和算法。我们对该算法在数据库大数据流量运算方面进行了适应性改进，并对其协同计算中的隐私保护效果、数据挖掘操作的影响等方面进行了分析，证明了其安全性和可用性。在对模拟数据库的交互运算中，该算法也体现出了足够的保护能力和可用性，为多方安全算法在数据库联合计算推广提供了可用的借鉴意义。

参考文献：

[1] YAO A C. Protocols for secure computations. In: Proceedings of 23th Annual IEEE Symposium on Foundations of Computer Science. Chicago, USA, 1982.160-164.

[2] Goldreich O, Micali S, Wigderson A. How to prove all NP-statements in zero-knowledge, and a methodology of cryptographic protocol design [J]. On Advances in Cryptology-crypto, 1987, 263(1):171-185.

[3] Rabin M. How To Exchange Secrets with Oblivious Transfer[J]. Iacr Cryptology Eprint Archive, 2005.

[4] 魏晓超 蒋瀚 赵川.一个高效可完全模拟的n取1茫然传输协议 [J]. 计算机研究与发展, 2016, 11(11):2475-2481

[5] Quisquater, Jean-Jacques; Guillou, Louis C.; Berson, Thomas A. (1990). "How to Explain Zero-Knowledge Protocols to Your Children". Advances in Cryptology - CRYPTO '89: Proceedings 435: 628–631.

[6]Tongqing W, Jun L, Jin H, etal. Double-factor System of Identity Authentication Based on MD5 Algorithm and Schnorr Protocol基于MD5算法和Schnorr协议的双因素身份认证系统 [J]. 计算机应用研究, 2004, 21(12):137-139.

[7] 冯国登. Cryptography: Theory and Practice, Third Edition][M]. 加拿大:Douglas R. Stinson, 2008. 157-158

[8] Du W L, Atallah M J., Privacy-preserving cooperative scientific computations,In CSFW 2001，Nova Scotia, Canada 2001,pp.273-282.

[9] 安全多方计算及其应用研究[D]. 北京邮电大学:孙茂华, 2013. 46-51.

[10] 姚钦锋. 基于数据挖掘的网络安全态势分析[D]. 上海交通大学:信息安全工程学院, 2012.

[11] 周水庚, 李丰, 陶宇飞, 肖小奎. 面向数据库应用的隐私保护研究综述[J]. 计算机学报, 2009, 32(5): 847-861

[12]李锋. 面向数据挖掘的隐私保护方法研究[D]. 上海交通大学, 2008.

[13]M.JM.Feingold and M.Leahy, "Data-mining reporting act of private" U.S. Senate Bill(proposed)2003.