云平台下基于隐私保护的桶划分方案

张 浩[1] 黄 涛[1] 刘三女牙[1] 王丽娜[2]

[1](国家数字化学习工程技术研究中心(华中师范大学) 武汉 430072)

[2](武汉大学计算机学院 武汉 430072)

摘 要 DAAS(Database as a Service)模式以其管理便捷的特性受到大量组织机构的青睐,同时托管数据的安全也成了迫切需要解决的难题.数据加密对于外包数据的安全起着重要作用,这会降低数据查询效率,因此高效安全的密文查询成为解决数据机密性的突破口,然而,云计算环境下国内外针对 DAAS 模式密文查询的研究缺少攻击模式下对隐私的深度分析.针对该问题,该文提出了一种 DAAS 模式下基于隐私保护的桶划分算法.首先根据查询效率指标提出了一种基于遗传算法的桶划分方案;在此基础之上,针对查询的过程中隐私泄漏情况提出了信息泄露的隐私指标体系,并将该指标体系与查询效率进行结合,最后基于遗传算法的桶划分算法对隐私与效率的模型进行最优化,从而获得最优的桶划分方案来确保查询过程中的隐私与查询效率最优的平衡.该算法可以在提高范围查询精确度和系统效率的基础上,降低密文查询中隐私泄露的信息,从而提高云平台中隐私数据的可用性和隐私性.最后,为了验证文中所提方案的可行性,将文中的算法与目前采用的几种桶划分方案进行对比,发现文中的方案在查询精准度上以及在隐私的保护上均优于其他方案.

关键词 密文查询;桶划分;遗传算法;隐私指标;云计算

中图法分类号 TP309 **DOI 号** 10.11897/SP.J.1016.2016.00429

A Privacy-Preserving Bucket Partition Mechanism in Cloud

ZHANG Hao[1] HUANG Tao[1] LIU Sannv-Ya[1] Wang Li-Na[2]

[1](*National Engineering Research Center for E-Learning (Central China Normal University), Wuhan* 430072)

[2](*School of Computer, Wuhan University, Wuhan* 430072)

Abstract A large number of organizations and institutions have been attracted to the cloud platform for its features, such as convenient management. Thus, the security of the outsourced data become more and more important. Encryption is a useful approach to protecting the data, while the features of the encrypted data are vanished. To manage the data effectively, the efficient and secure ciphertext query approach is urgent. However, the existing ciphertext query technology fails to provide a deep analysis in privacy leakage under attack. To solve this problem, we propose a privacy-preserving bucket partition mechanism in Database as a Service (DAAS) model in cloud. First, this paper proposed a generation algorithm (GA) based bucked partition mechanism according to the query efficiency. Then this paper built a privacy index system for information disclosure during the query and combined the privacy index system with the query efficiency. Finally, this paper optimized the proposed model based on the GA to balance the privacy and the

收稿日期:2014-11-17;在线出版日期:2015-12-18.本课题得到国家自然科学基金(61373169,61272453,61103219)、"十二五"国家科技支撑计划(2015BAK07B03)、华中师范大学中央高校基本科研业务费资助项目(CCNU15GF001,CCNU15A05010)资助.**张 浩**,男,1986 年生,博士,中国计算机学会(CCF)会员,主要研究方向为云计算安全、虚拟化安全、大数据安全.E-mail:haozhang@whu.edu.cn.**黄 涛**(通信作者),男,1979 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为大数据分析.E-mail:tmht@mail.ccnu.edu.cn.**刘三女牙**,男,1973 年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为大数据安全、人工智能.**王丽娜**,女,1964 年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为云计算安全、虚拟化安全、大数据安全.

accuracy during the query. The algorithm maximized the query accuracy and efficiency, reduced the information leakage during the query, and consequently enhances the availability and privacy of sensitive data in cloud. To verify the effectiveness of the proposed mechanism, the comparison experiments of our proposed mechanism and other bucket partition mechanisms were done. The result shows that the proposed mechanism is superior to others.

Keywords ciphertext query; bucket partition; genetic algorithms; privacy indicators; cloud computing

1 引 言

在云计算环境下,虽然 DAAS(Database as a Service)数据库外包模式为企业带来了开销的节省等许多优势,但也对机密数据的安全与用户的隐私带来了挑战[1],例如组织机构将用户的个人医疗数据等隐私数据等进行外包,一旦这些隐私数据被非法窃取将会对用户或企业造成巨大的损失.加密机制为 DSP(Database Service Provider)端数据安全提供了重要思路,却也使得数据丧失了许多特性,如有序性、相似性等,这给 DAAS 系统的数据查询带来很大的困难和挑战.因此,针对数值型数据,如何构建一种既能提高密文的查询效率且能确保隐私的密文索引显得至关重要.

国内外针对数值型数据的密文查询进行了大量的研究.目前的研究主要分为两种:一种是直接对密文进行构建可搜索的加密算法,查询结果直接为需要的密文;另外一种是利用 hash 桶对数据进行分段,返回结果中不仅包含正确的结果也包含其他混淆结果.Qi 和 Atallah[2]研究了基于同态加密算法的密文 kNN 查询方法,其将时间开销从二次方变成了线性,然而在提高效率的同时泄露了一些信息(比如盲值的大小等).Agrawal 等人[3]提出了一种保持有序的加密方法,但加密数据保持有序性,容易遭到选择密文攻击.Ozsoyoglu 等人[4]提出了另一种保序加密方法,极大地减少了加/解密开销,提高了加密数据操作性能,但该方法容易遭受服务器端的统计攻击.IBM 研究员 Gentry[5]最近找到了一种全同态加密算法,在理论上取得了一定突破,但还存在很大的性能缺陷,实用尚有差距.

Hacıgümüş 等人[6]在 DAAS 模型基础上,构建密文数据分桶机制,从而缩小用户端解密范围,因返回客户端的记录集包含一些不满足查询条件的记录,需由客户端进行解密过滤处理;Wang 等人[7]提出了一种桶划分的自适应调整方法 STBucket,以减小查询的误检率.宋伟等人[8]通过分析不同查询分布对客户端查询精度的影响,提出了一种面向服务的加密数据高效查询方法 AEI,可以很好地适应数据的频繁更新操作,然而对于查询的隐私并没有过多考虑.Hore 等人[9]提出了启发式桶划分算法,通过优化算法获取最优的桶数目,并且提出信息熵等指标对算法进行隐私加强,提高对敏感属性的保护效率,然而该算法中并没有给出一个有效的均衡隐私与查询效率的方案.Ciriani 等人在文献[10]中针对关系型数据,采用数据分段、加密技术,提出一种系统查询消耗计算模型,并提出一种启发式搜索算法找出系统查询消耗最小的数据分段,但是该方法的时间复杂度较高,并没有很好的平衡安全性和效率的关系.

从上述的分析情况来看,国内外针对 DAAS 模型下密文范围查询的技术虽然已经做了大量的研究,然而它们缺少攻击模式下对隐私的深度分析.关于隐私问题目前还没有一个统一的定义,这是一个挑战性的问题.因此,如何设计一种平衡用户隐私和 DAAS 查询精度的密文查询算法对于数据的安全具有重要的意义.

本文提出了一种 DAAS 模式下基于隐私保护的桶划分算法.首先根据查询的过程提出了一套密文查询中信息泄露的隐私指标体系,并将该指标体系与查询效率进行结合,最后基于遗传算法的桶划分算法对隐私与效率的模型进行最优化,从而获得最后的桶划分方案来确保查询过程中的隐私与查询效率最后的平衡.该算法可以在最大化范围查询的精确度和提高系统效率基础上,降低密文查询中隐私泄露的信息,从而提高云平台中隐私数据的可用性和隐私性.

2 密文查询方案的基础模型

2.1 数据外包模型

图 1 中描述的是在云计算架构下数据外包程序

的简单架构. 本文假设客户端环境是可信的，服务端的环境是不可信的，服务端的数据是以密文的形式存放. 为了查询数据，客户端可以将整个数据表下载下来，然后解密筛选出正确的结果，服务端的功能缩减成了一个安全存储容器，然而这样违背了数据库外包的初衷. 为了提高查询的效率，需要服务端能够对加密数据进行查询处理.

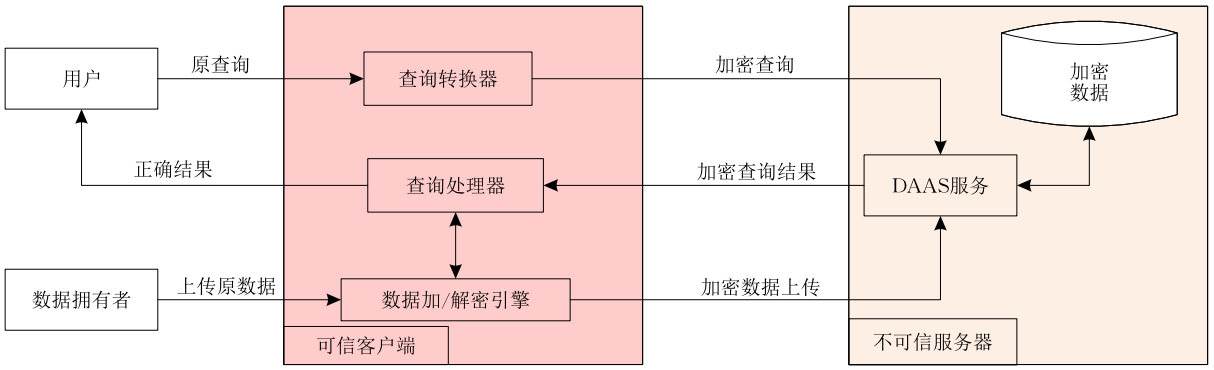


图 1　DAAS 密文查询架构

在图 1 的 DAAS 架构中，服务端通过 webservice 的方式将数据的管理功能（创建数据库、表等）、数据查询功能等发布出来，客户端通过 webservice 调用相应的数据库管理功能，客户端包含 3 个主要组件：数据加/解密引擎、查询转换器和查询处理器. 为了让服务端处理查询，每个加密记录属性值增加 1 个索引属性（桶 id）. 用户的明文需要桶查询转换器转换成针对桶 id 的查询. 这个转换是查询转化器通过“映射函数”来完成的. 服务端执行转换后的查询，返回 1 个密文结果集. 结果集中包含有所有的正确结果，同时也包含“假阳性”的错误结果. 随后系统通过查询处理器解密返回的结果集同时剔除“假阳性”结果. 另外，客户端需要负责密钥的管理和数据的加解密，由于对列或整个表加密将会导致数据的客户端的负担非常重，同时对每个值进行加密也会使得客户端的加密开销很大，因此，本文采用行加密方式对数据进行加密. 密钥的管理问题不属于本文讨论范围.

2.2　安全假设

在本文中，我们主要关注的攻击者具有比较强的攻击能力，其能够攻入数据库服务器，并能够有权访问 DSP 端的密文数据库. 本文的攻击与信任模型如下：

（1）假设服务器是半可信的，即其能够执行用户的请求及安全策略，但是会对用户的数据及用户的隐私比较好奇，但却不会破坏用户数据的完整性.

（2）我们假设用户与 DSP 之间的通信信道是安全的，比如现有的安全通信机制（SSL 或 IPsec），同时我们信任用户的前端，即信任前端是安全的.

（3）同时为了防止攻击者通过日志等其他方式间接获取用户的相关信息，我们对所有的日志信息、用户数据、数据库模式文件等相关文件进行加密存放.

（4）假设采用加密算法是不可能被强行暴力破解的.

2.3　区间查询的理论分析

接下来，本文着重研究范围查询的加密索引的建立. 在文献[6]中介绍了通过对敏感数据域进行划分，并为每一个条目分配一个桶 id，这就是该条目对应的加密索引值. 如原始表记为 R，服务器端的表将只包含数据库加密的元组（etuple）和相应的桶 id. 下面给出例子说明：

表 1 和表 2 分别展示了加密前的明文格式和加密后的密文存储格式. 表 2 中服务端只能看到数据段的加密数据和加密索引（如 A^S，B^S）. 针对 age 属性将[25，30]之间划分为 a_1 号桶，而将（30，60]划分为 a_2 号桶；针对 salary 属性将[4500，4900]划分 b_1 号桶，将[5000，6500]划分为 b_2 号桶. 因此，查询时一般的查询将被转换成针对加密索引的查询. 如下面的 SQL 查询语句

表 1　客户端明文表（R）

eid	name	age(A)	salary(B)
654	Jack	36	5000
700	Mary	25	4500
750	Tom	30	4700
775	John	56	6500

表 2　服务端数据表（R^S）

etuple	age(A^S)	salary(B^S)
01101010	a_2	b_2
01101110	a_1	b_1
01110001	a_1	b_1
11110010	a_2	b_2

Select name, age from R where $R.B>=5000$

经过查询转换转化成下面的查询语句

Select etuple from R^S where $R^S.B^S=b_2$

其中 b_1,b_2 是敏感数据属性 salary 的桶编号.

这种方法所返回的结果一般是正确结果的超集,意味着返回的结果中含有不满足查询条件的记录.为了过滤出这些记录,在客户端需要将返回的结果解密成明文,然后根据原始的查询标准对明文再次查询得到正确结果.

这种方式也会出现查询效率与隐私保护的平衡问题.例如当桶的总数量很少而每个桶内数值非常大的时候,只会泄露很少的信息,然而数据需要返回许多的假阳性的结果(极端情况下相当于将整个数据库返回给用户),客户端的通信开销和计算开销都较大.当桶的数量较多每个桶的数据量较少,虽然提高了查询精度,然而却增加了隐私泄露的风险,极端情况下相当于每个值分为一个桶.

因此,客户端的桶划分算法的目标就在于如何平衡隐私泄露风险与数据的查询精准性,从而提高查询效率和降低隐私泄露风险,增强对推理攻击的抵抗能力.

在本文中我们只介绍单一属性上的桶划分算法,多属性的桶划分算法,只是分别对不同的属性分别进行桶划分算法.接下来,我们首先提出桶划分方案查询效率的模型,并对此模型提出查询优化算法,具体介绍单一属性上的桶划分算法,而后对查询过程中的可能出现隐私泄露分析,并提出相应的隐私指标.最后在此隐私指标的基础之上对查询精准度算法进行改进,从而达到在密文情况下数据查询的精准性与隐私安全的平衡.

3 面向精准度的智能化桶划分算法

3.1 基础模型构建

为了简化分析,本文的桶划分算法的研究对象主要是数值型的离散域,如整数 Z.

假设关系为 $R=(V,F)$,其中 V 是不同的数值,在敏感属性域中每个数值至少出现了一次.F 是每个数值对应的频数.P 是范围查询 Q 的分布.使用最优桶划分算法将 R 划分到 M 个桶中,确保所有的范围查询的"假阳性"错误总数最小.具体的概念详见表 3.

表 3 基于遗传算法的桶划分概念表

符号	符号含义
$V_{\min}$	给定属性的最小取值
$V_{\max}$	给定属性的最大取值
N	不同值的个数
R	明文关系
$\|R\|$	关系 R 的元组个数
V	$[V_{\min},V_{\max}]$区间内所有的出现的数值的集合
F	$[V_{\min},V_{\max}]$区间内所有出现的数值对应的频数
n	$n=\|V\|=\|F\|$
R^S	加密的关系表
M	桶的数目
Q	R 上的所有范围查询的集合
q	q 是 Q 中随机的一个查询$q=[low,high]$,其中 $low,high\in[V_{\min},V_{\max}]$
Q'	桶级别的查询的集合
q'	q'是桶级别查询集合 Q'中的随机查询
$T(q)$	$T(q)$是转换函数,它将原始查询 $q\in Q$ 转换成桶查询 $q'\in Q'$

为了简化模型,本文主要考虑查询分布为均匀分布的情况.因此,密文数据划分为 M 桶的情况下,所有的可能的查询概率相同且为$\frac{2}{N(N-1)}$,即所有的 $C_n^2=\frac{N(N-1)}{2}$种可能的查询均为同等概率.

假设每个桶的密文元组规模为 $N_1,N_2,\cdots,N_M$,针对任意 K 个查询,每次查询返回的正确结果为 $C_1,C_2,\cdots,C_K$,K 次查询返回的元组总数为 $T_1,T_2,\cdots,T_K$,因此,查询的平均查询准确度(Query Accuracy)如式(1)所示

$$QA=\frac{\sum_{i=1}^{K}C_i}{\sum_{i=1}^{K}T_i}=\frac{\sum_{i=1}^{K}C_i}{\sum_{i=1}^{K}C_i+TFP}=1-\frac{\sum_{i=1}^{K}T_i-\sum_{i=1}^{K}C_i}{\sum_{i=1}^{K}T_i}=1-\frac{TFP}{\sum_{i=1}^{K}T_i}\tag{1}$$

其中假阳性值(Total number of False Positive)错误总数如式(2)所示

$$TFP=\sum_{i=1}^{K}T_i-\sum_{i=1}^{K}C_i\tag{2}$$

因此,根据式(1)和(2)可知,为了提高查询精度,提高密文的查询效率需要最小化"假阳性"错误总数.

首先考虑关系的单个属性的情况,这个属性的值域是有序的离散域,在本文中,属性值域假设为非负整数域.针对桶 i,假设其中有 N_i个不同的值,V_i代表分布在桶 i 的数据集合,$F_i=\{f_1^i,f_2^i,\cdots,f_{N_i}^i\}$

为 V_i 中对应的 N_i 值的频数. 在查询的一端落在桶 i 的情况下，接下来建立桶 i 的“假阳性”错误总数（TFP）模型.

定义 1. 假设查询的最小端位于桶 i 的 q_i 的集合记 Q_i，其中 $q_i=[l,h]$，$l\in V_i$，即 $v^i_{\min}\le l\le v^i_{\max}$，$Q_i$ 中查询的右端 h 位于第 i 个到第 M 个桶之间的任何一个桶中. 在该情况下，任意两个桶之间没有范围交集，即任意桶中不存在值 v^i_k 同时满足：(1) $v^i_{\min}\le v^i_k\le v^i_{\max}$；(2) $v^j_{\min}\le v^i_k\le v^j_{\max}$ 两个条件（例如：数字 2 同时满足桶[1,3]，[2,4]的范围），整个划分的 TFP 是相对最小的.

针对上述命题，用反证法证明，假设在桶 i,j 之间存在空间交集的情况下，即存在 $l\le v^i_k\le h$ 同时满足上述条件 a 和 b，且 TFP 的值最小. 针对 l 落在桶 i 的任意查询 q，其 h 端位于桶 x 中时，因此需要对桶 i 和 x 之间的桶进行查询，同时还要取出 v^i_k 所在桶 j（桶 j 不位于桶 i 和 x 之间），所以针对查询的整个划分方案的 TFP 为 $U_i+U'_x+TFP(j)$，其中 U_i 为桶 i 中 l 离桶 i 中最小值之间的值的频数之和，U'_M 为 h 离桶 M 最大值之间的值的个数（频数之和），$TFP(j)$ 为桶 j 的 TFP. 然而当将 v^i_k 挪动到桶 i 中时，TFP 仅为 $U_i+U'_x$，因此，假设不成立，定义 1 成立.

当任意两个桶之间不存在范围交集的情况下，且桶按照由小到大的顺序排列. 为了计算整个方案的 TFP，我们以查询为着手点，分析在这些查询的情况下每个桶内部的 TFP 情况. 假设 Q_i 为长度为 k 的查询集，其中 $q_k=[left,right]$ 为查询区间为 k 的任意随机查询，因此，所有长度为 k 且与桶 i 有交集的个数为 N_i+k-1. 然而针对其中任意元素 e 而言，仅有 k 个查询与元素 e 有交集，对于另外的 N_i-1 个查询而言，其作为假阳性值返回. 然而，对于同时跨越两个不同的桶的 q_k，其对假阳性值的贡献也分为两部分. 因此，当遍历所有的查询请求 q_k 时，其对应假阳性值如表 4 所示.

从表 4 中可以看出每个查询的条件概率均为 $\frac{1}{N_i+k-1}$，这是由于所有长度为 k 且与桶 i 有交集的个数为 N_i+k-1，且这些查询又是等概率. 因此，在查询长度固定的情况下，对于所有交集的 q_k 查询的假阳性值之和为 $\frac{(N_i-1)F_i}{N_i+k-1}$. 对于所有查询长度情况下桶 i 的假阳性值如式(3)所示.

表 4　区间查询的假阳性值

High 位置	查询的条件概率值	带来的假阳性值
为 i 桶的第 1 个值	$\frac{1}{N_i+k-1}$	$f^i_2+f^i_3+\cdots+f^i_{N_i}$
为 i 桶的第 2 个值	$\frac{1}{N_i+k-1}$	$f^i_3+\cdots+f^i_{N_i}$
为 i 桶的第 3 个值	$\frac{1}{N_i+k-1}$	$f^i_4+\cdots+f^i_{N_i}$
…	…	…
为 i 桶的第 $k+1$ 个值	$\frac{1}{N_i+k-1}$	$f^i_1+f^i_{k+2}+\cdots+f^i_{N_i}$
…	…	…
为 i 桶的第 N_i 个值	$\frac{1}{N_i+k-1}$	$f^i_1+\cdots+f^i_{N_i-k}$

$$TFP(i)=p_Q\cdot p_i\cdot\frac{(N_i-1)F_i}{N_i+k-1}=\sum_{k=1}^{N}\frac{N-k+1}{N^2}\cdot\frac{N_i-k+1}{N-k+1}\cdot\frac{(N_i-1)F_i}{N_i+k-1}=\frac{(N_i-1)F_i}{N}\tag{3}$$

其中 p_Q 为长度为 k 的查询占所有的查询的比例，p_i 为与桶 i 有交集的且查询区间为 k 的查询占所有长度为 k 的查询的比例. 式(3)中的假阳性主要包含两部分，其中一部分是整个查询请求位于桶 i 内部；另外一部分是查询请求跨越桶 i 和其他的桶 j，在式(3)中仅计算了对桶 i 的影响，另外一部分对桶 j 的影响将在桶 j 的 TFP 中体现出来. 因此，整个方案的假阳性值等价为所有长度的查询请求带来的假阳性值，其为所有桶的假阳性值相加，具体如式(4)所示

$$TFP=\sum_{i=1}^{M}TFP(i)=\sum_{i=1}^{M}\frac{(N_i-1)F_i}{N}=\frac{\sum_{i=1}^{M}(N_i-1)F_i}{N}\approx\frac{\sum_{i=1}^{M}N_iF_i}{N}\tag{4}$$

通过式(4)可知，由于 N 为总的值的数目，针对具体的数据集而言是一个固定值，因此为了降低查询的“假阳性”错误以提高密文查询的准确率，需要对所有桶的 N_iF_i 之和进行最小最优化.

3.2　面向精准度的智能桶划分算法

为了降低查询中的“假阳性”错误，需要对 $\sum_{i=1}^{M}N_iF_i$ 进行最小化处理. 该问题可以归结为是集合划分的问题，将元素划分在不同的数据集合中确保“假阳性”目标函数最小. 集合划分问题在数学上是十分困难的问题，也是经典的 NP-hard 问题[11-14]，

因此不可能在多项式的时间内计算出最优值,因此本文采用遗传算法启发式思想设计了面向精准度的智能桶划分算法(Precision-Oriented Intelligent Partitioning Algorithm,POIPA)来优化该假阳性指标,进而优化查询精准度,给出查询方案.

(1)适应度函数

由于遗传算法适合求目标函数的最大值,同时本文研究的范围为正整数范围,$N_i \cdot F_i$一定为正数因此,本文的假阳性模型进行取倒,适应值的范围在[0,1]之间,具体形式如式(5)所示

$$Fit(X) = 1 \Big/ \sum_{i=1}^{M} N_i \cdot F_i \tag{5}$$

其中 $X=\{x_1, x_2, \cdots, x_N\}$ 表示桶的划分方案,x_i 表示第 i 个 V_i 值所对应的桶的标号,且集合 V 中数值均是按照由小到大的顺序进行排列.

(2)基因表示

POIPA 算法的最优化问题 $MaxFit(X)$,其变量可以采用二进制编码来表示染色体.数字串的位数取决于变量的取值范围以及期望达到的精度.

针对整数区域的最大值求解问题,除了二进制编码之外,还有一种十进制编码,这两种编码方式各有优缺点.然而,在本模型中,在桶的数目不为 2^n 的情况下,迭代的过程中会消耗大量的计算资源和空间进行判断交叉变异后种群的合法性,因此,本文采用十进制编码.

在基因中采用一个十进制位表示一个变量,从而确保每个 V_i 只会出现在一个桶中,具体的表现方式如图 2 所示.

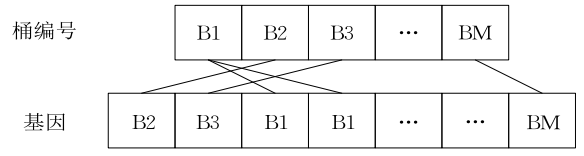


图 2 基因编码方式

(3)算法过程

遗传算法的初始种群是均匀分布,若单纯利用 POIPA 方法调整桶的划分需要较长时间,不利于服务质量的稳定.因此,为加快密文分段的划分,种群初始化时将 equi-width 方法划分得到的划分机制作为初始种群的一个染色体,再利用 POIPA 方法进行动态调整和优化,在优化的过程中利用精英保留策略,将父代最优的值传递子代中直到目标函数达到最优.具体的算法如算法 1 所示.

算法 1. 面向精准度的智能桶划分算法.

函数原型 *Best BucketPartition* (*int bucketNum*, *int*[] *value*, *int*[] *freq*)

输入:划分后桶的数目 *bucketNum*、敏感属性的不同取值 *value*、不同值对应频数 *freq*

输出:最佳分桶方案

```
1.  initialPopulation(PopSize);
                          //将等宽方案作为初始种群
2.  for(i=0; i<MaxPopulation; i++){
                          //直到最后的值稳定
3.    select(ipop,lastBest,curBest);
                          //选择并保留子代最好值
4.    Crossover(ipop,crossPara);
                          //对选择出来的群体交叉
5.    Mutation(ipop,MutationPara);
                          //对选择出来的群体变异
6.    Elite();//利用父代的最佳值替换子代中最差值
7.    Return Best;}
8.  void select (String[] ipop,Best lastBest,
                                    Best curBest){
9.    for(int i=0; i<popNum; i++)
10.     evals[i]=calculatefitnessvalue(ipop[i]);
                          //计算适应值
11.   generateNextPop(evals,nextpop);//轮盘选择法
12.   ipop=nextpop;
13.   if(generation>=1) { //运行的代数大于 1
14.     Worst=FindWorst(ipop);//找出本代最差值
15.     Worst=Best; }
                          //利用全局最优替换本代中最差
16. void Elite (){//
17.   curBest=findcurBest(ipop);
                          //查找后子代中最好方案
18.   if(Best.fitness<curBest.fitness)
19.   Best=CurBest; }
```

注. Best 为全局最优划分,curBest 为本代最优划分.

POIPA 算法中的种群初始化为了加快收敛速度,构建 *PopSize* 个个体.初始化函数是利用随机数生成器,为每个个体生成 *popNum* 个[0, *bucketNum*)之间的正整数,形成染色体方案.在初始化的过程中,将 equi-width 方案作为种群的一个染色体,即将 *value* 集合的值按照顺序进行均匀划分到 *bucketNum* 个桶中.

初始化种群之后,计算本代中所有染色的适应值(参照式(5)),然后根据计算出每个个体的适应值与整个种群的适应值的比值,并随机产生(0,1)之间的浮点数确定选用哪个染色体.同时为了确保收敛

速度，采用了精英保留策略，通过将选择出来的新种群中的最差的染色体方案替换为全局最优的染色体方案，这样将全局最优的染色体方案能够进行杂交和变异，进而在此个体之上产生更加优秀的子代个体.

在选择出适应值好的父代之后，根据交叉因子确定杂交对数，然后随机选择一对个体，进行单点交叉，形成中间种群，而后在中间种群中，随机抽取(变异因子 $*PopSize$)个个体，然后在选择的个体中随机选择一个基因位在$[0, bucketNum)$之间进行随机变异. 由于初始化的过程将 *equi-width* 方案加入了种群，因此需要加大变异因子，以免陷入局部的最优方案中. 在完成交叉变异之后，查找出子代中最好的个体并与全局最优方案比较，选出两者中最好的个体作为新全局最优方案.

POIPA 算法的时间复杂度是 $O(n*popNum)$，外层循环是遗传算法的经历的代数，内层循环次数是总群的个数(总群个数是常值). 每一代中的选择、交叉、变异等操作都会花费 $O(popNum)$时间. 算法的空间复杂度是 $O(n)$，主要是来自于存放染色体的空间.

4　面向隐私保护的智能化桶划分算法

4.1　攻击模型

上一节中提出的介绍的提高精准性的查询的优化算法，虽然能够有效地提高查询的精准度，然而，在一定程度上威胁到了数据的隐私安全. 其构建的索引是建立明文与密文之间的桥梁，这种对应于映射关系很可能为推理攻击以及数据分析等攻击技术提供了数据库内容重构及破解索引含义的可能性. 因此，评估和量化公开的信息或索引的信息泄露的系数是非常重要的，这将为桶划分的方案的查询效率与隐私安全的均衡提供很重要的依据.

针对 DAAS 的推理攻击主要分为两大类[15]，其主要是根据攻击者的先验知识程度不同的进行分类，然而在这两种的情况下攻击者对密文数据均具有绝对的控制权.

第 1 种推理攻击在攻击者仅知道数据库中的数据频数以及数据库的密文来进行攻击，这种称为 Freq-attack. 这种场景下，攻击者对数据库中明文数据的分布规律非常了解，这种分布可以是具体的，比如存放用户薪资的数据库的薪资频数，当然也可以是近似的. 在本文中，我们考虑攻击者知道数据库的具体频数这种比较强的攻击方式. 在这种情况下，攻击者攻击的对象主要是两种：第 1 种是对明文数据的攻击，即通过掌握的数据库的分布，去猜测具体某个密文的值；第 2 种是对索引的攻击，即根据映射关系，推导出明文与索引之间的对应关系，从而分析出数据库的密文以及用户的隐私.

第 2 种推理攻击在攻击者知道数据库的明文与密文来进行攻击，即攻击者知道原始数据库的明文以及密文数据库，却不知道明文与密文之间的对应关系(索引). 因此，这种情况下，攻击者需要攻击的是对索引与明文的对应关系进行攻击，从而来明确掌握用户新插入数据的含义及意义. 然而这种攻击方式的条件太强，不属于本文的考虑范围.

Freq-attack 推理攻击也分为简单的推理攻击和利用统计的推理攻击. 简单的推理攻击的依据主要是通过索引的相同个数以及密文的相同个数来确定. 例如：针对那些直接对密文进行一一对应建立索引的情况，如果某敏感属性的索引值完全相同，推断出这两个数字的明文相同，再结合数据的分布情况就可以推断出其中出现次数较多的数据.

密文数据库的简单推理攻击等价于根据索引出现次数将其归为同一个等价类中. 下面以 2.3 节中的案例进行推理分析.

$age.2=\{a_1, a_2\}=\{36, 25, 30, 56\}$,

$salary.2=\{b_1, b_2\}=\{5000, 4500, 4700, 6500\}$.

在这种情况下，由于明文与索引之间是多对一的关系(等价于将 4 个数据划分到 2 个集合中)，同样索引与密文之间也是多对一，因此直接根据明文出现的次数进行分析对这种桶划分机制的攻击较弱. 每个值的信息泄露概率就是属性值的基数的倒数，如表 5 所示.

表 5　攻击概率分布图

属性 *age*	属性 *salary*
1/4	1/4
1/4	1/4
1/4	1/4
1/4	1/4

整个表格的索引推理攻击信息泄露为每个值信息泄露概率的乘积，具体如式(6)所示

$$infoexpose=\prod_{t=1}^{n}\prod_{c=1}^{k}p_{t,c} \tag{6}$$

其中 t 遍历所有的元组即行数，c 遍历所有敏感属性即列数，其中 $p_{t,c}$ 表示每个数值的泄露概率.

由式(6)可知，上述案例中的整个表格的索引推理攻击信息泄露就是 $\left(\frac{1}{4}\right)^8$，尤其对于规模庞大的数据库而言，期望于利用简单推理攻破这种索引是不切实际的.

因此，针对本文的桶划分方案，本文主要考虑的攻击场景为 Freq-attack 的统计攻击对数据库内容进行攻击. 客户端使用非确定性加密算法加密用户的敏感属性的值，假设该加密算法是不会被破解的，因此攻击者不能通过解密数据来获取敏感属性的明文. 对于给定的元组，攻击者能做的就是用比较高的概率去估计这个密文的真实明文值. 在这种情况下，攻击者需要获得整个桶的划分机制(这个可以通过大量的统计分析获得，简单推理无法获得)，并根据这个划分模式，加上利用统计估计技术和大量统计实验，不断的缩小数据的范围，最终以比较高的概率猜测出该值的最小范围，这种方式的代价非常大.

为了简化分析，本文假设处在最坏的情况下即：攻击者已获得整个桶划分机制，同时通过统计的方式知道每个值的频数. 但是攻击者不能将每个加密元组与值进行对应，因为同一个桶中的加密属性的值是一样的. 为了对统计攻击下的信息泄露进行度量，本文提出两个隐私指标：方差以及方差之方差.

4.2 隐私指标

由于攻击者采用统计攻击的方式，因此，桶划分方案中数据的分布规律指标如方差、方差之方差需要进行增强以抵抗攻击者的统计攻击. 本文定义估计误差方差 SEE(Squared Error of Estimation)以及方差之方差.

定义 2. 估计误差方差：假设桶 K 内部元素的分布律为随机分布 X_K，P_K 是对应值的概率分布值，那么其对应的离散值概率为 p_k. 攻击者为了猜测加密元组的真实值，通过大量的统计拟合出桶 K 的元素的评估分布律 Y_K，对应的概率函数为 p'_k.

然而猜测值与真实值之间存在误差，为了衡量该误差，提出估计误差方差 SEE，即真实值 x_i 与猜测值 y_i 之间的欧式距离的平方. 具体表现形式如式(7)所示

$$\begin{aligned}SEE&=E(\|X_K-Y_K\|^2)\\&=\sum_{i=1}^{N_K}\sum_{j=1}^{N_K}p(x=x_i)p(y=y_j\mid x=x_i)(x_i-y_j)^2\\&=\sum_{i=1}^{N_K}\sum_{j=1}^{N_K}p(x=x_i)p(y=y_j)(x_i-y_j)^2\\&=\sum_{i=1}^{N_K}\sum_{j=1}^{N_K}p(x=x_i)p(y=y_j)(x_i^2+y_j^2-2x_iy_j)\\&=\sum_{i=1}^{N_K}\sum_{j=1}^{N_K}p(x=x_i)p(y=y_j)x_i^2+\\&\quad\sum_{i=1}^{N_K}\sum_{j=1}^{N_K}p(x=x_i)p(y=y_j)y_j^2-\\&\quad 2\sum_{i=1}^{N_K}\sum_{j=1}^{N_K}p(x=x_i)p(y=y_j)x_iy_j\\&=E(X^2)+E(Y^2)-2E(X)E(Y)\\&=Var(X_K)+Var(Y_K)+(E(X)-E(Y))^2\end{aligned} \tag{7}$$

从式(7)可以得知，攻击者 A 为了尽量减小估计均差误差 SEE，其可以通过以下两种方式：(1)减少 A 估计的数据分布的方差值 $Var(Y_K)$；(2)减小桶内数据的真实分布均值与猜测分布均值之间的差值($E(X_K)-E(Y_K)$)，其中桶 K 内数据的分布均值为 $E(X_K)=\sum_{i=1}^{M}p(x=x_i)x_i, x_i\in K$. 因此，可以让 A 在统计上获得与真实值最接近的猜测值的统计分布形式为两个分布的均值要相等，$E(X_K)=E(Y_K)$，且 $Var(Y_K)=0$.

因此，客户端可以通过提高 $Var(X_K)$ 增加攻击者的攻击难度，即尽量保持桶内数据分布保持无序状态，从概率角度提高密文数据的安全性. 因此为了确保整个方案的泄露信息量尽量少，需要确保所有桶的分布方差的期望尽量大，$\sum_{K=1}^{M}P_K\times Var(X_K)$.

定义 3. 桶的估计方差：数据划分为 M 个桶，整个划分的机制的随机分布为 X，桶 i 对应的分布为 X_i，其对应的概率为 p_i. 通过式(7)可知为了保证每个桶内部密文数据的安全需要提高 $Var(X_i)$，然而当一个桶的 $Var(X_i)$ 增加很有可能导致其他桶的 Var 值降低，这样容易导致一部分桶的 Var 值很高，而另一部分桶的 Var 值很低时，那么 Var 值低的桶内部的密文值就很容易被攻击. 因此为了确保所有的桶的 Var 都相差不大，本文提出了方差之方

差 $VarofVar$ 的概念，具体形式如式(8)所示

$$VarofVar=\sum_{i=1}^{M}p(x=x_i)(var(x_i)-E(Var(X)))^2$$
$$=Var(Var) \tag{8}$$

为了衡量方案的信息保护能力系数，我们将 Var 指标与 $VarofVar$ 进行归一化处理

$$procoe=\frac{\beta}{\pi}\left(\arctan(Var)+\frac{\pi}{2}\right)+$$
$$\frac{\gamma}{\pi}\left(\arctan(1/VarofVar)+\frac{\pi}{2}\right),$$

其中 β,γ 是由用户定义的系数.

4.3　隐私与性能的权衡

假设原始的数据集合为 $\{x_1,x_2,\cdots,x_N\}$，并将数据集合划分为 M 个桶中并保证数据只出现在唯一的一个桶，$x_i\in B_k,i\in[1,N],k\in[1,M],B_i\cap B_j=\varnothing,i,j\in[1,M]$. 为了保证划分后密文数据的查询的准确率，同时保证桶的划分机制泄露的隐私最少，需要对这两组相互矛盾的指标进行最优化处理，即对第 3 节中 TFP 指标进行最小化，对桶内数据分布方差指标 $\sum_{K=1}^{M}Var(X_K)$ 最大化处理，同时需要对 $VarofVar$ 指标进行最小化处理，这是一个多目标最优化的问题. 为了简化将多目标最优化问题进行统一指标转换成用于寻求隐私和准确率的单目标 $priv\_eff$ 的最优化问题，即将对应的指标进行反正切转换($\arctan(X)+\pi/2$)将值域映射到$[0,\pi]$中，并根据用户的定义权重来确定两组指标的重要性，具体形式见式(9).

$$priv\_eff=\frac{\alpha}{\pi}\left(\arctan(1/TFP)+\frac{\pi}{2}\right)+procoe$$
$$=\frac{\alpha}{\pi}\left(\arctan(1/TFP)+\frac{\pi}{2}\right)+$$
$$\frac{\beta}{\pi}\left(\arctan(Var)+\frac{\pi}{2}\right)+$$
$$\frac{\tau}{\pi}\left(\arctan(1/VarofVar)+\frac{\pi}{2}\right)$$
$$=\frac{\alpha}{\pi}(\pi-\arctan(TFP))+$$
$$\frac{\beta}{\pi}\left(\arctan(Var)+\frac{\pi}{2}\right)+$$
$$\frac{\tau}{\pi}(\pi-\arctan(VarofVar))$$
$$\alpha,\beta,\tau\geq 0,\ \alpha+\beta+\tau=1 \tag{9}$$

由式(9)可知，为了确保准确率和划分机制可能泄露的隐私信息最小，需要对隐私和准确率均衡指标 $priv\_eff$ 进行最大化优化. 该问题依然归结为是集合划分的问题，经典的 NP-hard 问题，因此针对该问题本文采用精英保留遗传算法的思想来优化该指标，并利用该思想设计面向隐私保护的智能桶划分算法(PRIPA)，进而确保查询精准度和可能泄露的隐私均达到最优.

PRIPA 算法的核心就是将指标 $priv\_eff$ 进行最大化最优处理. PRIPA 算法为加快密文分段的划分，与 PROPA 算法类似将初始种群，将 equi-width 方法划分得到的划分机制作为初始种群的一个染色体，然后在初始种群进行迭代，在迭代中不断寻求本代中最优值进而找出全局最优解，同时为了加快收敛，算法利用精英保留策略对，将父代最优的值传递子代直至全局最优解收敛到固定值. 具体的算法如算法 2 所示.

算法 2.　面向隐私保护的智能桶划分算法.

函数原型 *Best BucketPartition* (*int bucketNum*, *int*[] *value*, *int*[] *freq*, *float*[] *para*)

输入：划分后桶的数目 *bucketNum*、敏感属性的不同取值 *value*、不同值对应频数 *freq*，*para* 为隐私指标和准确率指标的权重系数数组

输出：最佳分桶方案

```
1.  initialPopulation(PopSize);
                          //将等宽方案作为初始种群
2.  for(i=0; i<MaxPopulation; i++){
                          //直到最后的值稳定
3.    select(ipop,lastBest,curBest);//保留子代最好值
4.    Crossover(ipop,crossPara);
                          //对选择出来的群体交叉
5.    Mutation(ipop,MutationPara);//变异操作
6.    Elite();//利用父代的最佳值替换子代中最差值
7.    Return Best;}
8.    Void caculate(){
9.      TFP=caculateTFP();
10.     VarSum=caculateVar();
11.     VarofVar=caculateVarofVar();
12.     pri-per=caculatePriv_Eff(TFP,VarSum,
                    VarofVar,para)} //计算 priv_eff
```

PRIPA 算法主要的核心算法与 PROPA 主要思想基本一致，均是利用迭代遗传算法对初始均匀的种群进行迭代，在指标 $priv\_eff$ 最优的约束下使得种群不断朝着 $priv\_eff$ 最优的方向进化. 为了加快迭代收敛速度，采用了精英保留策略，通过将选择出来的新种群中的 $priv\_eff$ 最差的染色体方案替

换为全局 $priv\_eff$ 最优的染色体方案，这样将最优的染色体方案能够进行杂交和变异将最优的基因遗传到后代中. 对选择出来的种群进行单点交叉和单点变异，形成新种群. 在完成交叉变异之后，比较子代中最优方案与全局最优方案，选出两者中最好的个体作为新全局最优方案，直到局部最优方案与全局最优方案的 $priv\_eff$ 基本一致，即两者的之差的绝对值小于 10^{-n}.

5 实验设计与分析

为了验证基于遗传算法的隐私增强算法，进行相关的性能测试和功能验证. 下面简单介绍实验环境.

5.1 数据集介绍

为了更加贴合实际，整个实验在开源云管理平台 opennebula 搭建的云平台中进行，实验环境中包含 6 台物理机和 12 虚拟机. 云平台中所有服务器为同等配置，均为 2×AMD 皓龙 2370×4 核，16GB 内存，千兆以太网网卡. 在云平台中，每个物理机的系统为 Linux，底层利用 Xen3.3 来实现对物理平台的虚拟化，平台中虚拟机的版本为 Windows 系统. 其中一台虚拟机作为 DSP 用来存放外包的数据. 其余 11 台虚拟机作为用户端.

5.2 数据集介绍

为了测试算法的可用性、效率、安全性性能以及相互之间的制约关系，本文分别采用随机函数模拟数据集均值为 5000 方差为 100 的正态分布数据集、均匀分布的合成数据集(数据区间为[1,10000]，数据规模为 40 万条)和真实的数据集 TPC-H 中的表格 partsupp 中的 ps_availqty 属性(区间为[1,10000]，数据规模为 80 万条)等数据集对算法进行测试，数据规模为 80 万条记录.

本文根据不同的数据集分别利用随机数生成器均匀产生不同的查询集，每个查询集规模为 10000. 每个查询根据对应的数据集合范围随机产生两个不同的数值组合成一个查询，这样确保了查询的随机性与公平性.

针对上述的数据集，本文分别针对 QoB 算法、CB 算法、equi-depth 算法、PRIPA 算法、PROPA 这 5 种算法，测试相应的准确率指标和隐私保护指标. 下面具体介绍相应的指标及相应的测试结果.

5.3 性能及结果分析

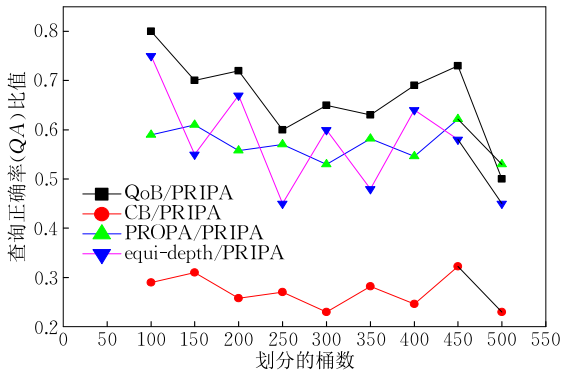
(1) 准确率 QA

为了评估本文算法对密文数据的查询准确率在实验中分别利用合成数据集和真实数据集测试了上述 5 种算法的 QA 值，其中 CB 算法的退化率为 2、PROPA 算法的用户隐私参数分别为 $\alpha=0.25$，$\beta=0.5$，$\gamma=0.25$，并以 PRIPA 的 QA 值为基准值，计算出其他 3 种算法的值与该值的比值关系，具体见图 3 所示.

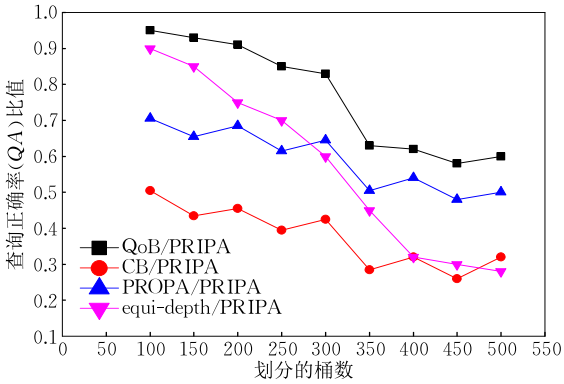
在图 3(a)中可以看出在数据服从均匀分布且用户查询均匀时，其他几种算法的查询精准率 QA 与本文算法 QA 的比值均明显小于 1，这说明 PRIPA 算法相对于其他几种算法而言其效果要优异许多，尤其随着划分桶的数目增加，其效果体现的更加明显. 从图 3(b)中可以看出，当数据服从正态分布时，所有的算法与 PRIPA 算法的查询精度比值随着桶的数量增加而降低，在桶为 450 的时候所有算法与 PRIPA 精准度比值达到最低. 在图 3(c)中，刚开始的效果比较稳定，当桶的数量为 350 的时候查询精度比值骤降，这可能与 ps_availqty 的数据分布有关，在桶为 500 时其他算法的查询精度仅为 PRIPA 算法的 30%左右. 从以上的 3 个子图中可以看出，PRIPA 算法的桶划分方案查询精度优于其他几种方案.

(2) 查询性能与隐私的均衡

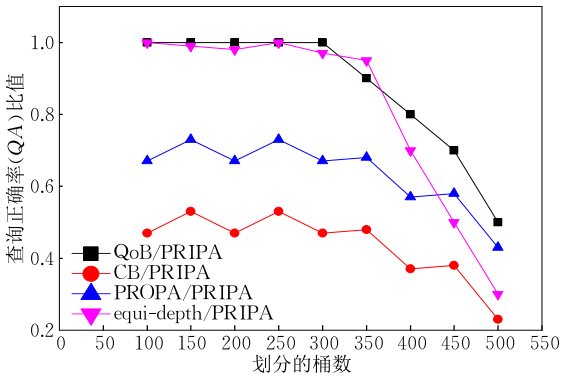
图 4 描述了查询精度与保护能力系数的均衡效果. 我们取每种算法(QoB 算法、CB 算法、equi-depth 算法、PRIPA 算法、PROPA 算法)的 10 组数据，这 10 组数据分别表示桶的划分数量在 100，150，200，…，550 的情况下，查询精度与 $procoe$(该值越大说明隐私保护能力越强)坐标值. 从图中可以看出，整体趋势上测试过程中的 5 种算法的查询精度与 $procoe$ 值是相互制约的，查询精度高则 $procoe$ 低，反之则 $procoe$ 高查询精度低. 在图 4 的右侧是 PRIPA、QoB、equi-depth 这 3 种算法的查询精度和 $procoe$ 值的坐标值，其查询精度较高，然而这 3 种算法的 $procoe$ 值较小，意味着他们对统计攻击能力较弱很容易泄露桶内数据的安全. CB 算法在 QoB 算法的基础上对数据进行随机化提高了数据安全性同时也兼顾了对查询精度但是没有达到最优的均衡效果，本文的 PROPA 算法的方案相对其他几种算法而言，其安全性与查询精度均是比较优异的，达到了对安全性与查询精度的有效均衡.



(a) $U(1,10\,000)$数据集



(b) $N(5000,100)$数据集



(c) *ps_availqty*数据集

图 3　查询值正态分布正确率

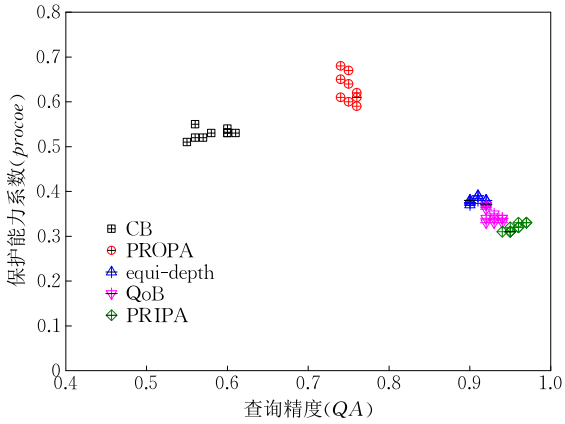


图 4　查询精度与保护能力系数的均衡

6　结束语

为了均衡云环境下外包数据的隐私性与查询效率，本文首先提出了桶划分模式下查询效率指标，并提出了一种基于遗传算法的 PRIPA 算法对该指标进行优化. 为了提升 PRIPA 算法的隐私保护能力，本文分析了查询过程中可能存在的信息泄露情况，提出了一套隐私保护指标，并将该指标与查询效率进行结合，提出了用于均衡查询效率与隐私性的 PROPA 算法. 为了验证本文方案的正确性，本问题与目前的几种常用的算法进行对比，最后结果发现，本文的方案相对于其他的几种方案而言，在查询效率与隐私均衡上均有较好的效果. 虽然本文的方案提出了一种均衡查询效率和隐私保护能力的桶划分方案，然而其只能适用于单一敏感属性，并没有涉及组合属性的情况. 同时，本文的研究过程中没有深入的研究查询的分布与数据分布的之间的关系. 这些问题是本文下一步研究工作的重点.

参　考　文　献

[1] Abadi D J. Data management in the cloud: Limitations and opportunities. Data Engineering Bulletin Issues, 2009, 32(1): 3-12

[2] Qi Yi-Nian, Atallah M J. Efficient privacy-preserving k-nearest neighbor search//Proceedings of the International Conference on Distributed Computing System(ICDCS). Quebec, Canada, 2009: 311-319

[3] Agrawal R, Kirenan J, Srikant R, Xu Yi-Rong. Order-preserving encryption for numeric data//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD). Pairs, France, 2004: 563-574

[4] Ozsoyoglu G, Singer D A, Chung S S. Anti-tamper databases: Querying encrypted databases//Proceedings of the 17th

Annual IFIP WG 11.3 Working Conference on Database and Application Security. Colorado, USA, 2003: 133-146

[5] Gentry C. Fully homomorphic encryption using ideal lattices//Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC). Maryland, USA, 2009: 169-178

[6] HacıgümüşH, Lyer Bala, Li Chen, et al. Executing SQL over encrypted data in the database-service-provider model//Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD). Wisconsin, USA, 2002: 216-227

[7] Wang Hao-Cong, Du Xiao-Yong, Wang Jie-Ping, Yang Ping-Ping. STBucket: A self-turning bucket index in DAS paradigm//Proceedings of the 4th ChinaGrid Annual Conference. Yantai, China, 2009: 102-109

[8] Song Wei, Peng Zhi-Yong, Cheng Fang-Quan, et al. A service-oriented adaptive search method over encrypted data in trusted database. Chinese Journal of Computers, 2010, 33(8): 1324-1338(in Chinese)
(宋伟, 彭智勇, 程芳权等. 可信数据库环境下面向服务的自适应密文数据查询方法. 计算机学报, 2010, 33(8): 1324-1338)

[9] Hore B, Mehrotra S, Tsudik G. A privacy-preserving index for range queries//Proceedings of the 30th International Conference on Very Large Databases (VLDB). Toronto, Canada, 2004: 720-731

[10] Ciriani V, De Capitani di Vimercati S, Foresti S, et al. Fragmentation design for efficient query execution over sensitive distributed databases//Proceedings of the ICDCS'09. Quebec, Canada, 2009: 32-29

[11] Pereira J, Averbakh I. The robust set covering problem with interval data. Annals of Operations Research, 2013, 207(1): 217-235

[12] Wang N F, Zhang X M, Yang Y W. A hybrid genetic algorithm for constrained multi-objective optimization under uncertainty and target matching problems. Applied Soft Computing, 2013, 13(8): 3636-3645

[13] Crawford B, Soto R, Monfroy E. Cultural algorithms for the set covering problem//Tan Ying, Shi Yuhui, Mo Hongwei eds. Advances in Swarm Intelligence. Berlin: Springer, 2013: 27-34

[14] Cacchiani V, Hemmelmayr V C, Tricoire F. A set-covering based heuristic algorithm for the periodic vehicle routing problem. Discrete Applied Mathematics, 2014, 163(1): 53-64

[15] Cai L, Chen H. On the Practicality of Motion Based Keystroke Inference Attack. Berlin: Springer, 2012

ZHANG Hao, born in 1986, Ph. D. His main research interests include security of cloud computing, security of virtualization and security of big data.

HUANG Tao, born in 1979, Ph. D., associate professor. His main research interest is analysis of big data.

LIU Sannv-Ya, born in 1973, Ph. D., professor. His main research interests include security of big data and artificial intelligence.

WANG Li-Na, born in 1964, Ph. D., professor. Her main research interests include security of cloud computing, security of virtualization and security of big data.

Background

This research dedicates to improving the privacy of ciphertext query of numerical data. The existing ciphertext query technology fails to provide a deep analysis in privacy leakage under attack. We propose a privacy-preserving bucket partition mechanism to maximize the query accuracy and efficiency and reduce the information leakage during the query. This research is supported by the National Natural Science Foundation of China under Grant Nos. 61373169, 61272453, 61103219 (Research of cloud edge security protective model and methods based on trusted virtual domain).

This program will be the first to propose a unified Cloud Edge Security Protective Model, enlightened by the concept of Trusted Virtual Domains. Grounded in this theory, this program extends its focus towards several aspects of current information security researches: Focusing on threats caused by the share of computing resources between multiple VMs, the project comes up with a Memory Coloring solution to better manage the resource allocation and improve VM isolation. To secure the weak and unstable Domain migration process, this project will research on the methods of Domain migration and policy synchronization, ensuring the confidentiality and integrity of VM migration procedures. To prevent unauthenticated data access between Domains of different security levels, this project will pursue proper access control and secure communication by new Attribute-Based Access Control and Group Key Agreement methods. The project also includes a system prototype of Secure Cloud Edge showing the feasibility of our models and methods. This research will seek breakthroughs in the Cloud Edge Security models and methods. It will surely add new bones into the studies of Cloud Security.