基于同态加密算法的密文数据处理

许嘉晨[1]，王子旭[1]，夏萌[1]，王宏任[1]

（1．中南大学信息院计算机科学与技术系 湖南长沙 410000）

王建新教授，段桂华副教授

中文摘要：随着大数据时代的到来，数据挖掘和数据分析现在已经成为我们生活中必不可少的一个技术，它给我们带来了很大的益处的同时也给我们的隐私数据安全情况带来了很多的挑战。针对在数据分析中可能会造成的数据安全问题，我们结合了 Paillier 加密算法和 RSA 加密算法设计了一套全同态加密方案，能够直接完成对密文的加法和乘法运算操作，在保证安全性的情况下进行数据分析。随后我们在医疗数据分析方向进行了应用证明了我们提出的全同态加密方案的有效性和可靠性。

英文摘要：With the advent of the Big Data era, data mining and data analytics are now an indispensable technology in our lives, bringing great benefits to us while also bringing a lot of problems to our privacy data security situation. Facing the data security issues that may be caused, we design a set of fully homomorphic encryption scheme combining the Paillier encryption algorithm and RSA encryption algorithm, which can directly make addition and multiplication operations on the ciphertext under the premise of data security. Then we applied the direction of medical data analysis to prove the validity and reliability of our proposed homomorphic encryption scheme.

关键词：信息安全；同态运算；数据分析

一、引言

生活在当今的时代中，每天人们都会产生出大量的数据，而更是有不少的隐私数据，它们往往对每个人的意义重大。然而，总是会存在一些不法分子像通过某些非法的渠道，例如一些黑客技术，来窃取他人的隐私信息，从而获取利益，而他们的这种行为则极大地影响了人们的生活。尽管如此，这样的安全事件却屡见不鲜，至今仍然没有得到有效的解决。

随着大数据时代的到来，越来越多的决策需要通过对每个人的信息数据进行挖掘和统计来制定，而这些信息中更是包含了很多的隐私数据，但是正如我们所知，隐私数据在传输的过程中十分容易被某些不法分子所截获，造成很多不必要的麻烦和损失。因此，在对隐私数据进行数据挖掘和数据统计等操作时仍然面临着诸多不小的挑战。也正是因为频频发生的数据安全问题，导致了越来越多的人对数据安全表示十分担忧，同时拒绝提供自己的真实数据。而这种现象会造成很多有效的信息的浪费，严重的减慢了大数据时代中社会和科技的发展和进步。因此，究竟如何能够在不暴露个人的私密信息的条件下，对隐私数据进行适当的挖掘和统计，成为了时下大家所关心的也急需解决的问题。

在 1978 年，有一种被称为同态加密的概念曾被提出，它是一种能够直接对密文进行操作的加密技术，这项技术

的关键点就是在于“双盲”，即可以检测加密漏洞并可以对其修复，但是却不会造成任何的信息泄漏问题，而这也为解决上述对隐私数据挖掘的安全问题提供了一种思路。回顾以往的加密算法，我们发现它的一个很大缺点就是它仅仅是作为一个箱子将所有的明文信息直接存放在箱子内，而人们无法对箱子内的数据进行分析处理，除非拥有打开箱子的钥匙，再将明文信息从箱子中取出来，才可以对其进行计算统计。但是全同态加密算法则可以让人们在箱子仍未被打开的情况下对箱子中的数据进行操作分析，而不用打开箱子，而这样的操作方式就在很大程度上保证了数据的安全性，同时能够对数据进行分析和计算。

全同态加密算法可以被简单的定义为一种对于加法和乘法都能够找到其相对应的操作的加密算法。但是由于其效率尚未达到工业界应用的水准，所以知道目前仍然没有真正可以使用的全同态加密算法的在日常生活中被广泛的使用。

二、设计方法

为了实现上述的可以进行加密情况下的加法运算和乘法运算的全同态加密算法，我们结合了Paillier算法[1]和RSA算法[2]构成一个全同态加密算法。下面我们首先对 Paillier 算法以及 RSA 算法进行分别的介绍。

（一）Paillier 算法

密钥生成：假设有两个大素数 p、q，$\mathrm{n}=\mathrm{pq}$，$\mathrm{g}\in Z_{N^2}^{*}$，再记$\mathrm{L}(\mathrm{x})=\frac{x-1}{n}$，公钥则是$\mathrm{pk}=(\mathrm{n},\mathrm{g})$，私钥$\mathrm{sk}=\lambda(\mathrm{n})=\mathrm{lcm}(\mathrm{p}-1,\mathrm{q}-1)$。

加密阶段：对于任意的明文$\mathrm{m}\in Z_n$，随机选择$\mathrm{r}\in Z_n^{*}$，那么我们就可以的到密文$\mathrm{c}=E_{pk}(m)=g^m r^n \ mod\ n^2$。

解密阶段：对于得到的密文 c，我们可以解密得到明文$\mathrm{m}=D_{sk}(\mathrm{c})=\frac{\mathrm{L}(c^{\lambda(n)}\ mod\ n^2)}{\mathrm{L}(g^{\lambda(n)}\ mod\ n^2)}\ \mathrm{mod}\,\mathrm{n}$。

而 Paillier 算法是一种安全性基于 DCR 假设的一种概率公钥加密算法，其具有加法同态性。

（二）RSA 算法

密钥生成：假设有两个大素数 p、q，$\mathrm{n}=\mathrm{pq}$，又根据欧拉定理易知$\varphi(\mathrm{n})=(\mathrm{p}-1)(\mathrm{q}-1)$，然后随机选择整数 d、e，使得$\gcd\big(\mathrm{e},\varphi(\mathrm{n})\big)=1$，$\mathrm{ed}\equiv 1\big(\mathrm{mod}\ \varphi(\mathrm{n})\big)$，则可得公钥$\mathrm{pk}=(\mathrm{n},\mathrm{e})$，私钥$\mathrm{sk}=\mathrm{d}$。

加密阶段：对于任意的明文$\mathrm{m}\in Z_n$，则对应的密文可得$\mathrm{c}=\mathrm{E}_{\mathrm{pk}}(\mathrm{m})=\mathrm{m}^{\mathrm{e}}(\mathrm{mod}\ \mathrm{n})$。

解密阶段：对于得到的密文 c，我们可以解密得到明文$\mathrm{m}=\mathrm{D}_{\mathrm{sk}}(\mathrm{c})=\mathrm{m}^{\mathrm{d}}(\mathrm{mod}\ \mathrm{n})$。

RSA 算法是目前应用较为广泛的同态加密算法，其满足乘法同态的特性。

（三）新的全同态加密算法

通过上述分别对两种加密算法的介绍，我们可以了到 Paillier 算法具有的是加法的同态性，即对于任意由 Paillier 算法加密得到的密文进行加法运算操作之后再进行解密得到的结果将会等于对原始明文直接进行加法运算操作之后得到的结果；而 RSA 则恰好具有乘法同态性，即对任意的经过 RSA 算法进行加密运算之后的密文进行乘法运算操作之后再解密得到的结果恰好就是直接对明文进行乘法运算得到的结果。因此，我们可以通过在需要进行对数据的乘法操作时采用 RSA 算法，而在需要进行加法运算操作的时候就采用 Paillier 算法。其具体的算法流程如图 1 所示。

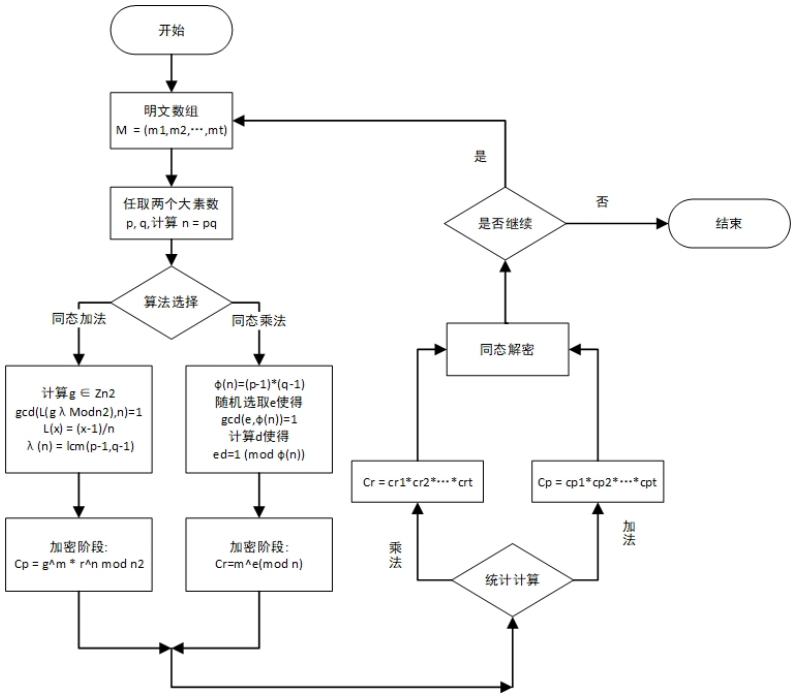

图1 算法流程图

1. 加密算法

首先将明文比特分组（分组长度根据安全需求来确定），然后对每个明文分组长度做加密运算，最后将加密得到的分组密文依次合并，得到加密密文。具体的过程分为如下几步 ：

（1）选取随机产生的两个安全大素数 P 和 Q（目前两个数的长度都接近 512bit 是安全的）；

（2）计算乘积 $N=P\times Q$；并且生成一个随机数 R_i ；

（3）把消息 M 分组为若干长度为 L（L 的长度应该小于 P）的消息分组 $M=m_1m_2...m_t$；

（4）使用加密算法 $c_i=(m_i+P\times R_i)\ mod\ N$，同时计算出密文 $C=c_1c_2...c_t$。

2. 解密算法

将接收到的密文分组，依次采用解密算法对分组密文进行解密，得到分组明文，然后将分组明文合并得到解密后的明文。具体步骤如下：

1）接收方收到密文 C 并把密文 C 分组得到 $C=c_1c_2...c_t$；

2）使用密钥 P 和解密算法 $m_i=c_i\ mod\ P$ 计算 m_t ；

3）得到明文消息 $M=m_1m_2...m_t$。

3. 对该算法进行同态性分析

（1）同态加法特性验证

假设有两组明文 M_1 和 M_2，分别对他们用上述的加密算法进行加密得到 C_1 和 C_2。

$$C_1 = (P_1 + P\times R_1) \bmod N$$

$$C_2 = (P_2 + P\times R_2) \bmod N$$

则对于明文$M_3 = M_1 + M_2$，有：

$$C_3 = C_1 + C_2 = \left(M_1 + M_2 + P\times(R_1 + R_2)\right) \bmod N = (M_3 + P\times R_3) \bmod N$$

对C_3进行解密：

$$M_3 = C_3 \bmod P = (C_1 + C_2) \bmod P = \left(M_1 + M_2 + P\times(R_1 + R_2)\right) \bmod P = M_1 + M_2$$

2）同态乘法特性验证

对于明文$M_4 = M_1 + M_2$，有：

$$C_4 = C_1\times C_2 = ((M_1 + P\times R_1)\times(M_2 + P\times R_2)) \bmod N = (M_1 + P\times R_3) \bmod N$$

对C_4进行解密：

$$C_4 = C_4 \bmod P = (C_1 \times C_2) \bmod P = ((M_1 + P \times R_1) \times (M_2 + P \times R_2)) \bmod P = M_1 \times M_2$$

因此，该算法既具有同态加法特性也具有同态乘法特性。

通过以上的算法介绍与分析，我们知道了将 Paillier 算法和 RSA 算法结合起来，通过特定的运算方式，就可以使其既满足了加法同态性，又满足了乘法同态性。因此根据前面提到的定义我们就可以将该方案称作“全同态加密方案”。有了上述的性质，那么我们就可以将其应用在统计计算之中，因为大多数的统计计算基本上就是一些简单的四则运算的组合。而在现如今常见的统计计算的实现中，往往都是直接对明文进行操作，因此很多重要的隐私数据的安全性无法得到足够的保证，被泄漏的概率很高，所以为了解决这个问题，最好的方式就是使统计计算是在密文上进行的。因此我们可以利用上述的加密方案，分别进行 Paillier 加密和 RSA 加密，然后再直接在经过上述加密后的密文上进行相应的计算在进行解密，这个效果就相当于直接对明文进行操作一样，但是这样能够使数据的安全性得到极大的提高，有效地方式了隐私数据泄漏的问题。

三、算法实现及应用

有了上述的算法简介以及整个全同态加密方案的具体操作流程之后我们就通过编码实现了上述全同态加密方案，其中我们还根据实际的应用需求进行了进一步的改进。加入了一个比较普通的通用的对称加密算法 AES，因为 AES 是一种对称加密算法，其对应的运算速度较快，但是与之前提出的两种加密算法的不同之处在于其不具备加法和乘法的同态性，因此我们认为现实中所以需要加密的数据并不是全部能够被用来统计分析的，如果采用上述两种公钥加密算法，必然会导致加密的速度变慢，所以我们引入了 AES 加密算法，对那些普通的数据进行加密提高加密速度。

为了验证我们实现的上述全同态加密算法的有效性，我们将其应用在了医疗数据的分析中，首先我们获取了如图 2 所示的数据，然后对该数据利用全同态加密方案进行加密，得到了如图 3 的数据，从图中我们可以看到对于每一个身份证号，即对于每一个医院病人，都有两行数据，其中有一些数据列的内容是一样的，那么就代表其将不会被用作进行数据分析，仅仅只是通过简单的 AES 加密运算，保证安全性，但是也有一些数据列的内容是不一样，这是因为我们对需要进行加密情况下数据分析的数据分别进行了 Paillier 加密运算和 RSA 加密运算用以达到加法同态和乘法同态的特点。

身份证号	年龄	病种	主治医师	病情持续时间 / 天	入院总花费
532925198505196000	32	房颤	孙永水	57	13598
640323198710054000	30	上呼吸道感染	李行云	45	8876
410324199410295000	23	心肌梗塞	周糖	21	3346
350128197708307000	40	房颤	孙永水	12	4534
360982197610258000	41	上呼吸道感染	李行云	32	12355
152530197606104000	41	支气管炎	李炳熙	25	6544
451425198902046000	28	房颤	李亚明	23	756
450404198011182000	37	支气管炎	丁水	12	453
522631198506252000	32	过敏性鼻炎	丁水	24	563
510724198804273000	29	房颤	孙永水	23	454
441224199312303000	24	支气管炎	李行云	12	35
140824198712097000	30	上呼吸道感染	丁水	32	42
542429198505291000	32	房颤	李东东	35	3232
610625197606184000	41	过敏性鼻炎	李芳	56	1341
150785198808157000	29	脑震荡	方起航	87	1234
510101197810142000	39	支气管炎	李炳熙	98	4522
310118198706180000	30	支气管炎	周糖	77	342
331023198006062000	37	脑震荡	周明	34	3241
360827198403074000	33	上呼吸道感染	李芳	23	123
141081198409189000	33	脑震荡	蒋新	23	132
350424197708263000	40	房颤	李亚明	45	423
361129198505048000	32	脑震荡	方起航	6	675
330801199206268000	25	过敏性鼻炎	王丽	56	243
341004198411226000	33	支气管炎	康成	76	658
431122197602156000	41	心肌梗塞	李东东	78	234
440704198701149000	30	脑震荡	周明	89	8675
653024199311202000	24	脑震荡	周明	55	2345
650106197512042000	42	过敏性鼻炎	王丽	34	2341
440200197505055000	42	心肌梗塞	李亚明	3	1235
120000198605140000	31	脑震荡	方起航	22	4321

图2 明文待加密医疗数据

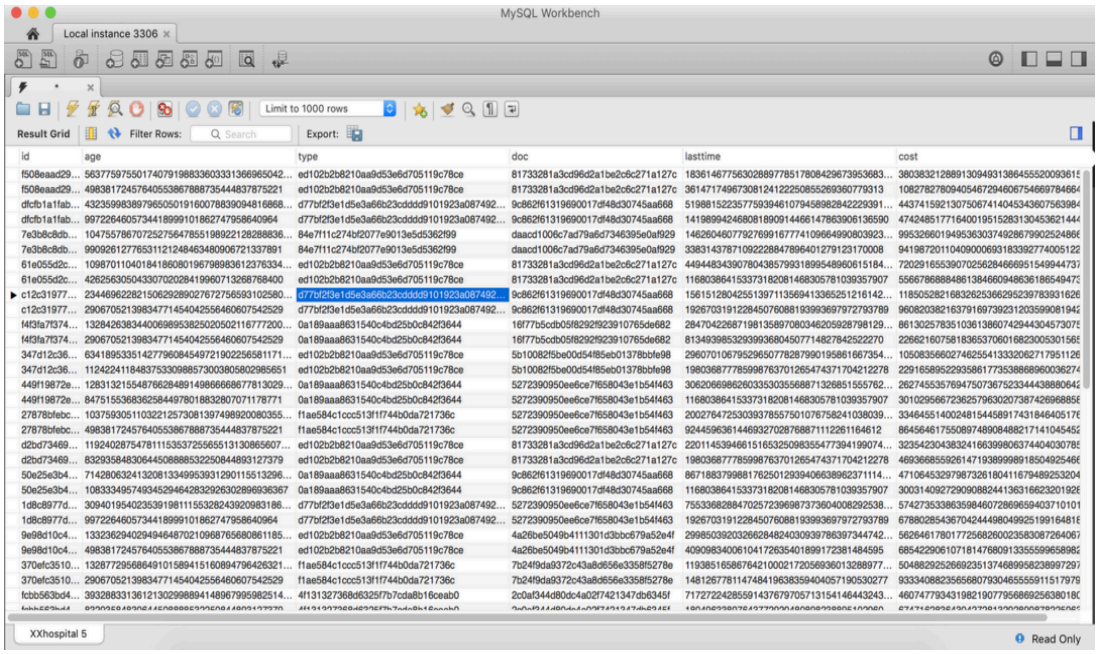

图3 加密之后的数据

有了如图所示的数据之后，我们可以通过选择要分析的数据以及要对其进行的分析类型来验证我们的全同态加密算法的正确性，其中在该应用中我们提供的可以进行分析的数据种类如图 4 所示，我们利用我们提出的全同态加密方案可以得到对应的一组数据之和，数据之积，数据平均值，如图 5 所示，用以分析，同时通过验证与原始数据运算所得值得到两个值相等，可知我们提出的全同态加密算法是正确可行的。

图4 选择数据分析的数据类别

图 5 对密文进行统计运算之后所得到的结果

四、算法性能分析

通过实验证明了我们提出的全同态加密方案的正确性和有效性之后，我们对其进行了性能的分析，具体的性能分析我们通过一下两个方面进行。

（1）安全性：保护伞云医疗平台采用 RSA 算法、AES 算法等常用的安全性很高的算法，因此，具有良好的保密性。

（2）高效性：在配置方面，用户只需几分钟时间即可轻松获取一个或若干个高性能计算实例；可按需灵活定制，一键升级到更高性能和容量的实例规格，实现快速、平滑扩容，满足业务快速发展需要。

在计算方面，透传 GPU 性能，极致发挥 GPU 性能；单机峰值计算能力突破 14T Flops 单精度浮点运算，0.4T Flops 双精度浮点运算。同时，不同用户间资源全面隔离，数据安全有保障；完善的安全组和网络 ACL 设置让您能控制进出实例和子网的网络入出站流量并进行。

参考文献：

[1]万泓伶. 基于口令的认证系统设计与实现[D]. 电子科技大学, 2013.

[2]林远辉. 基于口令的三方认证密钥交换协议研究[D]. 山东大学, 2014.

[3]李雪松. 基于同态加密的统计数据处理[D]. 贵州大学, 2015.