计算机科学
COMPUTER SCIENCE

保护隐私的汉明距离与编辑距离计算及应用

窦家维

引用本文

窦家维. 保护隐私的汉明距离与编辑距离计算及应用[J]. 计算机科学, 2022, 49(9): 355-360.

DOU Jia-wei. Privacy-preserving Hamming and Edit Distance Computation and Applications[J]. Computer Science, 2022, 49(9): 355-360.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220100241

保护隐私的汉明距离与编辑距离计算及应用

窦家维
陕西师范大学数学与统计学院　西安 710119

摘　要　随着信息技术的快速发展,在保护数据隐私的条件下进行多方合作计算越来越普及,安全多方计算已成为解决这类问题的核心技术。在科学研究及实际应用中,人们常根据两个字符串之间的汉明/编辑距离度量其相似程度,研究汉明/编辑距离的保密计算具有重要意义。文中主要针对汉明距离与编辑距离的两方保密计算问题进行研究。首先将汉明距离的计算问题转化为向量内积计算问题,应用加密选择技巧以及 Okamoto-Uchiyama(OU)密码系统设计保密计算协议。然后通过对参与者字符串中各字符进行统一编号的方法,将编辑距离的计算问题转化为判定隐私数据的差是否为 0 的问题,应用 OU 密码系统设计编辑距离保密计算协议。应用模拟范例严格证明了协议的安全性,分析了协议的计算复杂性,测试了协议的实际执行效率,并与目前已有相关结果进行了分析比较。理论分析和实验结果都表明了协议的高效性。

关键词:安全多方计算;汉明距离;编辑距离;半诚实模型;模拟范例
中图法分类号　TP309.7

Privacy-preserving Hamming and Edit Distance Computation and Applications

DOU Jia-wei
School of Mathematics and Statistics,Shaanxi Normal University,Xi'an 710119,China

Abstract　With the rapid development of information technology,privacy-preserving multiparty cooperative computation is becoming more and more popular. Secure multiparty computation is a key technology to address such problems. In scientific research and practical applications,people measure the similarity of two strings with Hamming(edit) distance. It is of great significance to study privacy-preserving Hamming(edit) distance computation. This paper studies privacy-preserving Hamming(edit) distance computation. First,we transform Hamming distance computation to inner product computation of vectors,and then use Okamoto-Uchiyama(OU) cryptosystem and encryption-and-choose technique to design protocol for Hamming distance. Second,we give each alphabet of a string a number,transform edit distance to determine whether the difference of the number of two alphabets is 0,and use OU cryptosystem to design a privacy-preserving edit distance computation protocol. The security of the protocol is strictly proved,the computational complexity of the protocol is analyzed,the actual implementation efficiency of the protocol is tested and compared with the existing results. Theoretical analysis and experimental results show that our protocols are efficient.

Keywords　SMC,Hamming distance,Edit distance,Semi-honest model,Simulation paradigm

1　引言

科学技术的发展将我们带进了信息社会,信息社会数据成了最重要的财富。由于各方面的限制,任何机构或者个人都不可能获得自己所需要的全部数据,因此迫切需要共享其他机构或个人的数据。在共享数据上进行数据挖掘、机器学习等操作,挖掘数据的价值,发挥数据的作用,辅助决策,从而获取经济效益与社会效益。

例如,医药研发机构和医院共享药物疗效数据以及病人医疗数据可以加快医药的研发,降低研发成本,获得巨大的经济与社会效益。但其中涉及较多隐私数据(如病人的医疗数据),在数据共享过程中极易导致隐私泄露。如何在信息共享的同时避免隐私信息泄露,实现保护隐私的信息共享是信息安全研究人员需要研究的重要课题。

Yao[1]首次提出并研究了一个两方安全计算问题,由此开启了安全多方计算(Secure Multiparty Computation,SMC)的研究。随后 Goldreich 等对 SMC 问题进行了系统深入的研究[2-3],提出了 SMC 问题的通用解决方案,目前 SMC 的理论体系已基本完善。SMC 指多方参与者在保证自己数据隐私性的条件下进行的合作计算,SMC 是数据共享中隐私保护的关键技术。由于应用通用解决方案解决具体问题时计算效率一般都较低,因此对于具体问题需要寻求高效的解决方案。

到稿日期:2022-01-26　返修日期:2022-06-08
基金项目:国家自然科学基金(61272435);陕西师范大学教学改革研究项目(22JG37)
This work was supported by the National Natural Science Foundation of China(61272435) and Teaching Reform Research Project of Shaanxi Normal University(22JG37).
通信作者:窦家维(jiawei@snnu.edu.cn)

目前众多学者提出并研究了各种具体的SMC问题，包括保密的科学计算[4-5]、保密的计算几何[6]、保密的统计分析[7]、保密的数据挖掘[8]、保密的机器学习[9]、保密的集合计算[10]、保密的基于位置服务[11]、保密查询[12]等。随着人们越来越重视隐私保护，SMC也成为国际密码学界研究的热点。在近几年的信息安全顶级会议（如美密会、欧密会和CCS会议）上，SMC专题的论文都是最多或次多的。

距离度量在科学研究以及实际应用中具有重要作用。对于不同问题往往需要选用不同的距离进行度量，通常使用欧氏距离、曼哈顿距离、测地距离等来度量物理对象之间的距离，而汉明距离、编辑距离等则常被用于度量数字对象之间的距离。由于两个物理对象之间的距离在基于位置的服务中具有重要应用，目前关于欧氏距离、曼哈顿距离、图上距离保密计算的研究成果较多。在实际应用中，人们也常根据数字对象之间的距离来度量两个对象的相似程度，因此对汉明距离、编辑距离保密计算的研究也具有重要的实际意义。如在医学研究中常根据两个DNA序列的汉明/编辑距离来度量两个DNA的相似程度，在通信编码中常用两个二进制字符串的汉明距离进行检错、纠错。由于DNA序列等信息大多属于个人或医疗机构的隐私信息，因此需要在保护隐私的条件下计算相似度。

本文主要研究汉明距离和编辑距离的安全两方计算问题，虽然这些问题能够用SMC通用解决方案解决，但通用解决方案的效率较低，因此人们不断寻求这些具体问题的高效解决方案。文献[13]设计了一个保密的近似汉明距离计算协议。文献[14]利用不经意传输设计了一个保密的汉明距离计算协议，该协议的计算复杂性与通信复杂性都较高。文献[15]将汉明距离的保密计算引入生物认证领域。文献[16]基于理想格和learning with error-LWE设计了安全汉明距离计算方案，其安全性基于LWE问题，但这类问题需要很长的密钥，导致计算和存储困难，且安全性也有待检验。文献[17]所计算的两个DNA序列的汉明距离实际上是编辑距离。该文所应用的方法是将长度为n的DNA序列编码为长度为$4n$的0-1序列，并证明转化后的两个0-1序列汉明距离的一半等于原DNA的汉明距离。文献[17]以GM异或同态密码系统为基础设计协议，在协议中两方参与者共需要进行$8n$次加密运算和$4n$次解密运算，由于其编码方法仅适用于DNA序列，因此协议的应用范围受到限制。长度为n的字符串x,y之间的汉明距离被定义为$h(x,y)=\sum(x_i\oplus y_i)$，文献[18]根据$x_i\oplus y_i=x_i+y_i-2x_iy_i$，将异或运算转化为加法和乘法运算，使其能够利用加法同态密码系统进行计算，并设计了汉明距离和编辑距离的计算协议。文献[18]的协议是目前已有的关于汉明距离和编辑距离计算协议中效率较高的。

在汉明距离和编辑距离保密计算研究中还存在以下问题：一方面，已有的汉明距离保密计算协议计算效率都较低，而编辑距离主要是应用汉明距离的计算方法来解决，因此，汉明距离和编辑距离计算方案的效率都有待提高；另一方面，目前已有的计算协议中两方参与者的计算量基本相同，参与者的在线计算量都较大，这样的方案无法适用于群智感知、众包等应用场景，因为在这些应用中用户的计算能力有限，无法进行大量的计算。

本文旨在设计效率更高、实用性更强的汉明距离与编辑距离的保密计算协议。本文设计的汉明距离保密计算协议的计算效率比文献[18]提高了13倍以上，设计的编辑距离保密计算协议效率比文献[18]的协议效率提高了70倍以上。在本文协议中，如果有一方参与者计算能力较弱，该参与方只需进行少量计算即可。本文协议能很好地适用于群智感知、众包等应用场景。本文的贡献如下：

(1)将汉明距离的计算问题转化为向量内积的计算，运用加密选择技巧实现向量内积的计算，以此为基础设计了汉明距离保密计算协议，应用模拟范例严格证明了协议的安全性。

(2)通过对参与者字符串中各字符进行统一编号，将编辑距离的计算转化为判定隐私数据的差是否为0的问题，设计了编辑距离的保密计算协议，应用模拟范例严格证明了协议的安全性。

(3)对本文设计的协议的计算效率进行了详细分析，与目前已有的效率较高的解决方案进行了分析比较，并应用Python编程语言对协议的实际执行进行了模拟测试，理论分析和实验结果都表明了本文协议的高效性。

(4)应用本文协议还易于解决两个数据比等问题以及文本的相似性问题。以本文协议作为基本模块，还可以解决其他相关隐私保护问题，例如，以汉明距离协议中Bob最后计算的密文作为中间结果，可以保密判定所计算的汉明距离是否大于某个保密的阈值、是否等于某个值等。

本文第2节介绍了构造安全协议需要的一些基本知识以及协议的安全性定义；第3节构造了半诚实模型下安全的汉明距离保密计算协议，并证明了其安全性；第4节构造了半诚实模型下安全的编辑距离保密计算协议并证明了其安全性；第5节给出了本文协议的几个推广应用；第6节分析了协议的效率并进行了实际测试，并与目前已有协议进行了比较；最后总结全文。

2 预备知识

本文设计的协议在半诚实模型下是安全的，在协议设计中需要用到OU密码系统。

2.1 OU公钥密码系统

一个密码系统由密钥生成算法、加密算法以及解密算法组成，如果密码系统具有某种同态性质，则该系统还有一个同态运算算法。OU密码系统具有加法同态性，对其4个算法的简述如下[19]。

密钥生成：给定安全参数k，密钥生成算法选择两个k比特的大素数p和q，令$N=p^2q$。随机选择$g\in Z_N^*$使得$g^{p-1}\bmod N$的阶为p，计算$h=g^N \bmod N$。公钥为(g,h,N)，私钥为p,q。加密算法和解密算法分别记为$E(\cdot)$和$D(\cdot)$。

加密算法。要加密消息$m(0\leqslant m\leqslant 2^{k-1})$，选择随机数$r\in Z_N^*$，计算密文：

$$C=E(m)=g^m h^r \bmod N$$

解密算法。令$L(x)=(x-1)/p$。对于密文C，计算下式可解密得到明文：

$$m=D(C)=\frac{L(C^{p-1} \bmod p^2)}{L(g^{p-1} \bmod p^2)} \bmod p$$

加法同态性。对于密文 $C_1=E(m_1)$,$C_2=E(m_2)$,容易证明下面性质成立:

$$C_1C_2=E(m_1+m_2)$$

因此 OU 密码系统具有加法同态性。

2.2 密文重随机化

OU 密码系统是概率密码系统,它是语义安全的,即同一个明文可以加密成多个不同的密文形式,所有密文都是计算不可区分的。此外,根据密码系统的加法同态性,在不需要解密的情况下,可以将密文 $C=E(m)$ 转换为 m 的一个新的密文 C',此时只要计算 $C'=C\cdot E(0)$ 即可,这个计算过程被称为密文的重随机化。

2.3 协议的安全性

理想协议。SMC 协议的安全性需要借助一个理想协议进行描述。在理想协议中有一个完全可信的第三者(Trusted third party,TTP),所有参与者都将自己的隐私数据交给 TTP,TTP 计算协议规定的函数,并告知每个参与者规定得到的输出结果。虽然理想协议简单安全且高效,但在实际中要找到所有参与者都信任的 TTP 也很困难。因此人们要设计没有 TTP 参与的 SMC 协议来解决具体的计算问题,以实现理想协议的功能。

半诚实模型下协议的安全性。半诚实模型又被称为诚实但好奇的模型。在半诚实模型中所有参与者都是半诚实的,即他们严格按照协议要求执行协议,但他们可能保留协议执行过程中收到的所有信息,试图在协议执行后利用保留的信息推算出其他参与者的隐私信息。显然,如果一个 SMC 实际协议 π 没有比理想协议泄露更多的信息,则 π 是安全的。在半诚实模型下要证明一个协议是安全的,通用的证明方法是模拟范例方法,简述如下:

设 $f(x,y)=(f_1(x,y),f_2(x,y))$ 是一个多项式可计算函数。Alice 拥有私密数据 x,Bob 拥有私密数据 y,他们希望计算 $f(x,y)$ 而不愿意泄露 x,y。计算结束后 Alice 和 Bob 分别得到 $f_1(x,y)$ 和 $f_2(x,y)$。设 π 是一个计算 $f(x,y)$ 的两方协议,在执行协议过程中将 Alice 得到的消息序列记为 $view_1^{\pi}(x,y)=(x,r,m_1^1,\cdots,m_t^1,f_1(x,y))$,其中 r 表示 Alice 在执行协议时选择的随机数,m_i^1 表示 Alice 收到的第 i 条消息。Bob 得到的消息序列 $view_2^{\pi}(x,y)$ 可以进行类似定义。

定义 1 对于计算多项式函数 $f(x,y)$ 的两方协议 π,如果存在概率多项式时间算法 S_1 和 S_2(被称为模拟器),使下面两式成立:

$$\{S_1(x,f_1(x,y))\}_{x,y}\overset{c}{\equiv}\{view_1^{\pi}(x,y)\}_{x,y} \quad (1)$$

$$\{S_2(y,f_2(x,y))\}_{x,y}\overset{c}{\equiv}\{view_2^{\pi}(x,y)\}_{x,y} \quad (2)$$

则称 π 保密计算了 f,其中 $\overset{c}{\equiv}$ 表示计算不可区分。

要证明一个 SMC 协议在半诚实模型下是安全的,就需要构造出使得式(1)和式(2)成立的模拟器 S_1 和 S_2。

3 汉明距离保密计算

汉明距离:对于两个长度相同的二进制字符串 x,y,将它们之间的汉明距离 $h(x,y)$ 定义为两个字符串中对应位置字符不同的位置数。对两个字符串进行异或运算,并统计结果为 1 的个数,这个值就是汉明距离。例如 10101 和 00110 分别有第 1 位、第 4 位、第 5 位不同,它们的汉明距离为 3。

问题描述:假设 Alice 和 Bob 分别有隐私的二进制字符串 $x=x_1\cdots x_n$ 和 $y=y_1\cdots y_n$,他们希望计算 x,y 之间的汉明距离 $h(x,y)$ 而不泄露自己隐私数据的任何其他信息。

3.1 计算原理

首先把汉明距离的计算转化为向量内积的计算,如此即可利用具有加法同态性的 OU 密码系统进行保密计算。

由于 $x=x_1\cdots x_n$ 与 $y=y_1\cdots y_n$ 对应位置的字符 x_i 和 y_i 要么相同要么不同,从字符串的总位数 n 中减去对应位置具有相同字符的位数就是汉明距离。因此汉明距离可以转化为两个字符串中对应位置具有相同字符的位数的计算。

两个字符串中对应位置字符 x_i 和 y_i 相同时又分为两种情况:$x_i=y_i=1$ 以及 $x_i=y_i=0$。由于当 $x_i=y_i=1$ 时,$x_iy_i=1$,因此 $x_1y_1+\cdots+x_ny_n$ 就表示字符串中对应位置同为 1 的总位数。为了计算两个字符串中对应位置同为 0 的总位数,可将 $x_i=y_i=0$ 转变为 $\bar{x}_i=\bar{y}_i=1$ 进行计算,此时 $\bar{x}_1\bar{y}_1+\cdots+\bar{x}_n\bar{y}_n$ 就表示字符串中对应位置同为 0 的总位数,因此 $(x_1y_1+\cdots+x_ny_n)+(\bar{x}_1\bar{y}_1+\cdots+\bar{x}_n\bar{y}_n)$ 就表示字符串 x,y 中对应位置具有相同字符的位数,即 $h(x,y)=n-(\sum x_iy_i+\sum\bar{x}_i\bar{y}_i)=n-(\sum_{y_i=1}x_i+\sum_{y_i=0}\bar{x}_i)$

例如,设 $x=10110011011$,$y=11011100001$。$h(x,y)$ 的计算方式为:首先将所有 $y_i=1$ 所对应的 x_i 相加,再将所有 $y_i=0$ 对应的 $\bar{x}_i$ 相加。

$x=10110011011$

$y=11011100001$

$\bar{x}=01001100100$

由于 $y_1,y_2,y_4,y_5,y_6,y_{11}$ 的值为 1,在 x 中选择 $x_1,x_2,x_4,x_5,x_6,x_{11}$ 相加得到 3;y_3,y_7,y_8,y_9,y_{10} 的值为 0,在 $\bar{x}$ 中选择 $\bar{x}_3,\bar{x}_7,\bar{x}_8,\bar{x}_9,\bar{x}_{10}$ 相加得到 1。由此可知,$h(x,y)=11-3-1=7$。

3.2 汉明距离保密计算协议

协议 1 汉明距离保密计算协议

输入:Alice 输入 $x=x_1\cdots x_n$,Bob 输入 $y=y_1\cdots y_n$

输出:$h(x,y)$

准备:Alice 生成 OU 密码系统的公钥(g,h,N)及相应的私钥 p,q

1. Alice 应用公钥加密 $x,\bar{x}$(逐位加密),得到两个密文向量:

   $\mathbf{E}(\mathbf{x})=(E(x_1),\cdots,E(x_n))$,$\mathbf{E}(\bar{\mathbf{x}})=(E(\bar{x}_1),\cdots,E(\bar{x}_n))$

   Alice 将公钥以及 $\mathbf{E}(\mathbf{x})$,$\mathbf{E}(\bar{\mathbf{x}})$ 发送给 Bob。

2. Bob 进行如下操作:对于每一个 $i=1,\cdots,n$,若 $y_i=1$ 则选择 $\mathbf{E}(\mathbf{x})$ 的第 i 个密文 $E(x_i)$;若 $y_i=0$ 则选择 $\mathbf{E}(\bar{\mathbf{x}})$ 的第 i 个密文 $E(\bar{x}_i)$。并将选择的 n 个密文相乘得到一个乘积密文 V,Bob 将 V 重随机化后发送给 Alice。

3. Alice 应用私钥解密 V 得到 $v=D(V)$,进一步计算 $z=n-v$。

4. Alice 将 z 告诉 Bob,并输出 z。

根据协议的计算原理以及 OU 密码系统的加法同态性容易证明协议的输出结果 $z=h(x,y)$。

在协议中 Bob 只得到 $x,\bar{x}$ 的密文,由于 Bob 没有私钥解密,根据 OU 密码系统的语义安全性,Bob 从协议执行中

得不到 x 的任何信息，因此 Alice 的隐私数据 x 是安全的；在协议中 Alice 得到密文 V，解密后仅能计算出 $h(x,y)$，这是协议规定的输出结果，因此 Alice 从协议中也得不到 Bob 隐私数据 y 的任何额外信息。关于协议 1 的安全性，我们有以下定理。

定理 1 保密计算汉明距离的协议 1 在半诚实模型下是安全的。

证明：通过构造满足式(1)和式(2)的模拟器证明定理 1。在协议中 Alice 的 $view_1^{\pi}(x,y)=\{x,r_1,V,h(x,y)\}$，Bob 的 $view_2^{\pi}(x,y)=\{y,r_2,\boldsymbol{E}(\boldsymbol{x}),\boldsymbol{E}(\bar{x}),h(x,y)\}$，其中 r_1 和 r_2 分别是 Alice 和 Bob 在协议中选择的随机数。

首先构造 S_1。S_1 接收到输入 $(x,h(x,y))$ 后，按下面的方式进行模拟：1) S_1 随机选择一个 $y'=y'_1\cdots y'_n$ 使得 $h(x,y')=h(x,y)$；2) S_1 加密 $x,\bar{x}$ 得到 $\boldsymbol{E}(\boldsymbol{x}),\boldsymbol{E}(\bar{x})$，并按照协议第 2 步的操作方式对每一个 $i=1,\cdots,n$ 根据 y_i' 的值在 $\boldsymbol{E}(\boldsymbol{x})$ 或 $\boldsymbol{E}(\bar{x})$ 中选择适当的分量，并将得到的 n 个密文相乘，将得到的乘积密文记为 V'；3) S_1 解密 V' 后计算得到 $h(x,y')$。令：

$$S_1(x,h(x,y))=\{x,r_1',V',h(x,y')\}$$

因为随机数都是计算不可区分的，故有 $r_1\overset{c}{\equiv}r_1'$。由于 OU 密码系统是语义安全的，Bob 对密文 V 进行了重随机化，Alice 在解密前无法区分 V 与 V'，在解密后也仅能得到协议规定的输出结果，因此 $V\overset{c}{\equiv}V'$。又由于 $h(x,y)=h(x,y')$，故有：

$$\{S_1(x,h(x,y))\}_{x,y\in\{0,1\}^n}\overset{c}{\equiv}\{view_1^{\pi}(x,y)\}_{x,y\in\{0,1\}^n}$$

下面构造 S_2。S_2 接收到输入 $(y,h(x,y))$ 后，按下面的方式模拟：1) S_2 随机选择 $x'=x_1'\cdots x_n'$，使得 $h(x',y)=h(x,y)$；2) S_2 生成 OU 密码系统的公钥及私钥，应用公钥加密 x' 和 $\bar{x}'$ 得到 $\boldsymbol{E}(\boldsymbol{x}')$ 和 $\boldsymbol{E}(\bar{x}')$，并按照协议第 2 步的操作方式对每一个 $i=1,\cdots,n$ 根据 y_i 的值在 $\boldsymbol{E}(\boldsymbol{x}')$ 和 $\boldsymbol{E}(\bar{x}')$ 中选择适当的分量，并将得到的 n 个密文相乘，将得到的乘积密文记为 V'；3) S_2 解密 V' 得到 $h(x',y)$。令：

$$S_2(y,h(x,y))=\{y,r_2',\boldsymbol{E}(\boldsymbol{x}'),\boldsymbol{E}(\bar{x}'),h(x',y)\}$$

由于随机数都是计算不可区分的，因此 $r_2\overset{c}{\equiv}r_2'$。又由于 Bob 没有私钥解密，根据 OU 密码系统的语义安全性，对于 Bob 来说 $\boldsymbol{E}(\boldsymbol{x})\overset{c}{\equiv}\boldsymbol{E}(\boldsymbol{x}')$，$\boldsymbol{E}(\bar{x})\overset{c}{\equiv}\boldsymbol{E}(\bar{x}')$。又由于 $h(x',y)=h(x,y)$，故有：

$$\{S_2(x,h(x,y))\}_{x,y\in\{0,1\}^n}\overset{c}{\equiv}\{view_2^{\pi}(x,y)\}_{x,y\in\{0,1\}^n}$$

4 编辑距离计算协议

问题描述：假设 Alice 和 Bob 分别有私密字符串 $x=x_1\cdots x_n$ 和 $y=y_1\cdots y_n$，$x_i,y_i\in\Gamma$，其中 Γ 表示某个字母表。他们希望计算 x,y 之间的编辑距离 $b(x,y)$ 而不泄露任何额外信息。这里的编辑距离被定义为两个(相同长度)字符串中对应位置具有不同字符的位数，如 $x=abcd$ 和 $y=acbd$ 的编辑距离为 2。

4.1 计算原理

下面通过对字母表中每个字符进行编号的方法，将每个字符对应为一个数字，此方法可使每个参与者的字符串对应于一个向量，再通过计算两个向量中有多少个对应分量不相等得到 $b(x,y)$。

由于在 $x=x_1\cdots x_n$ 和 $y=y_1\cdots y_n$ 中，x 的每一位 x_i 和 y 的对应位 y_i 要么相同要么不同，如果给每个字符分配一个编号，将对字符串进行加密转化为对字符对应编号的加密，进一步利用 OU 密码系统的加法同态性，在密文上对两个向量对应的编号分量进行减法运算，如果相减得到的差值为 0，则对应字符相同，如果差值不为 0，则对应字符不同。如果差值不为 0 的字符个数为 l，l 即为所求的编辑距离。在这一过程中，要解决的关键问题是当对应字符的编号不同时不能泄露其差值，也不能泄露具体有哪些位的差值为零，因为泄露差值后参与者就可根据自己的字符编号以及差值推算出对方在相应位上是什么字符，泄露哪些分量差值为零也会推断出相应位置字符相同，从而造成信息泄露。这些问题可以采用在计算中添加随机因子以及应用随机置换的方法来解决。

4.2 编辑距离协议

协议 2 编辑距离保密计算协议

输入：Alice 和 Bob 分别输入各自的私密字符串 $x=x_1\cdots x_n$ 和 $y=y_1\cdots y_n$ $(x_i,y_i\in\Gamma)$

输出：b(x,y)

准备：Alice 生成 OU 密码系统的公钥(g,h,N)和私钥 p,q。并设 $w=g^{-1}$ 为 g 在 Z_N^* 中的逆元。双方商定好对 Γ 中每个字符的编号。为简单起见，仍以 x_i(或 y_i)表示其编号值。

1. Alice 应用公钥加密 x 得到 $\mathbf{E}(x)=(E(x_1),\cdots,E(x_n))$，Alice 将公钥和 $\mathbf{E}(x)$ 发送给 Bob。

2. 对于每个 $i=1,\cdots,n$，Bob 计算如下：

$$c_i=(E(x_i)w^{y_i})^{r_i}\bmod N$$

其中 r_i 是随机数。对 $(c_1,\cdots,c_n)$ 中各分量进行随机置换(随机置换为 τ)，得到 $(e_1,\cdots,e_n)=\tau(c_1,\cdots,c_n)$，并将 $(e_1,\cdots,e_n)$ 发送给 Alice。

3. Alice 解密每个 $e_1,\cdots,e_n$，统计解密结果不等于 0 的密文个数，记为 l。

4. Alice 将 l 告诉 Bob，输出 l。

根据协议的计算原理以及 OU 密码系统的加法同态性，易知协议的输出结果 l 即为 x,y 之间的编辑距离 $b(x,y)$。

在协议 2 中 Bob 仅得到密文数据 $\boldsymbol{E}(\boldsymbol{x})$。由于 Bob 没有私钥解密，根据 OU 密码系统的语义安全性，Bob 无法得到 Alice 的字符串 x 的任何信息。在协议 2 中 Alice 仅得到随机置换后的密文向量 $(e_1,\cdots,e_n)=\tau(c_1,\cdots,c_n)$。由于 OU 密码系统具有语义安全性，Alice 在解密前从这些密文中得不到 y 的任何信息；解密后如果 e_i 的解密结果不为零，由于 Bob 在计算 c_i 时加入了随机数 r_i，因此解密结果中含有随机因子 r_i，对于 Alice 来说解密结果即为随机数。虽然 Alice 可获知有哪些 e_i 解密结果为 0，但由于在协议中 Bob 应用了随机置换 τ，Alice 无法获知解密结果为零的 e_i 具体对应的 c_i，因此解密后 Alice 仅能获得有 l 个解密结果为零，这是协议的输出结果。Alice 从协议 2 中无法得到 Bob 字符串 y 的任何额外信息。

定理 2 保密计算编辑距离的协议 2 在半诚实模型下是安全的。

证明：通过构造满足式(1)和式(2)的模拟器证明定理 2。在协议 2 中 Alice 和 Bob 得到的信息序列分别为：

$view_1^{\pi}(x,y)=\{x,t_1,(e_1,\cdots,e_n),b(x,y)\}$

$view_2^{\pi}(x,y)=\{y,t_2,(E(x_1),\cdots,E(x_n)),b(x,y)\}$

其中,t_1,t_2 分别为 Alice 和 Bob 选择的随机数。

首先构造 S_1。S_1 接收到输入 $(x,b(x,y))$ 后,按下面的方式进行模拟:1)S_1 随机选择 $y'=y_1'\cdots y_n'$ 使得 $b(x,y')=b(x,y)$;2)对于每个 $i=1,\cdots,n$,S_1 选择随机数 s_i 并计算 $c_i'=(E(x_i)w^{y_i'})^{s_i} \bmod N$,将 $(c_1',\cdots,c_n')$ 随机置换后得到 $(e_1',\cdots,e_n')$;3)S_1 解密所有 $e_1',\cdots,e_n'$,并得知解密结果不为零的密文个数为 $b(x,y')$。令:

$S_1(x,b(x,y))=\{x,t_1',(e_1',\cdots,e_n'),b(x,y')\}$

由于 OU 密码系统是语义安全的,在解密前 Alice 无法区分 $(e_1,\cdots,e_n)$ 与 $(e_1',\cdots,e_n')$。尽管 Alice 可以解密,但由于协议执行中 Bob 添加了随机数并进行了随机置换,Alice 解密后也只能得知解密结果不为零的密文个数,这是协议的输出结果,因此对 Alice 来说 $(e_1,\cdots,e_n)\overset{c}{\equiv}(e_1',\cdots,e_n')$。又由于 $t_1\overset{c}{\equiv}t_1'$ 以及 $b(x,y)=b(x,y')$,因此有:

$\{S_1(x,b(x,y))\}_{x,y\in\Gamma}\overset{c}{\equiv}\{view_1^{\pi}(x,y)\}_{x,y\in\Gamma}$

下面构造 S_2。S_2 接收到输入 $(y,b(x,y))$ 后,按下面的方式进行模拟:1) S_2 随机选择 $x'=x_1'\cdots x_n'$ 使得 $b(x',y)=b(x,y)$;2) S_2 生成 OU 密码系统的公钥及私钥(记公钥为 (g',h',N'),并记 u 为 g' 在 $Z_{N'}^*$ 中的逆元),应用公钥加密 x' 得到 $\boldsymbol{E}(\boldsymbol{x}')=(E(x_1'),\cdots,E(x_n'))$;3)对于每个 $i=1,\cdots,n$,S_2 选择随机数 s_i' 并计算 $c_i'=(E(x_i)u^{y_i'})^{s_i'} \bmod N'$,将 $(c_1',\cdots,c_n')$ 随机置换后得到 $(e_1',\cdots,e_n')$;4) S_2 解密所有 $e_1',\cdots,e_n'$,并得知解密结果不为零的密文个数为 $b(x',y)$。令:

$S_2(y,b(x,y))=\{y,t_2',(E(x_1'),\cdots,E(x_n')),b(x',y)\}$

由于随机数都是计算不可区分的,因此 $t_2\overset{c}{\equiv}t'_2$。由于 Bob 没有私钥,根据 OU 密码系统的语义安全性,对于 Bob 来说 $(E(x_1),\cdots,E(x_n))\overset{c}{\equiv}(E(x_1'),\cdots,E(x_n'))$。又由于 $b(x',y)=b(x,y)$,故有:

$\{S_2(x,b(x,y))\}_{x,y\in\Gamma}\overset{c}{\equiv}\{view_2^{\pi}(x,y)\}_{x,y\in\Gamma}$

5 协议的推广应用

5.1 社会主义百万富翁问题

社会主义百万富翁问题[20]指两个百万富翁 Alice 和 Bob 想知道他们的财富是否相等,若不等,则不泄露任何额外信息。

此问题本质上就是比较两个隐私数据是否相等。数据比等是一种基本运算,保密的数据比等协议作为基本模块可用于解决很多其他 SMC 问题。如果将 Alice 和 Bob 的数据用二进制字符串表示,当字符串较长时,一种简单的方法就是选择一个单向散列函数,通过比较字符串的单向散列函数值是否相同即可知字符串是否相等。当字符串较短时,应用单向散列函数方法进行比较容易泄露具体的财富数据,此时可调用汉明距离保密计算协议判断 Alice 和 Bob 的财富是否相等。当且仅当两个字符串的汉明距离为 0 时两方财富相等。

5.2 相似性检测

应用编辑距离保密计算协议可以保密计算两个 DNA 序列(或两个文档)的相似度。一般用 $d=1-b(x,y)/n$ 度量两个具有 n 个字符的 DNA 序列(或文档)x,y 的相似度。在判定两个文档的相似度时,通常用词频-逆文档频率提取文档中若干个重要的关键词来构成关键词向量,通过计算文档关键词向量(可将其对应成字符串)之间的编辑距离即可计算两个文档的相似度。

6 协议效率

6.1 复杂性分析

在以公钥密码系统为基础构造的保密计算协议中,通常以协议中需要进行的模指数运算次数衡量协议的计算复杂性,乘法运算和其他简单运算与模指数运算相比可以忽略不计。应用 OU 密码系统,生成公钥需要 1 次模指数运算,加密与解密各需要 1 次模指数运算(解密时的 $g^{p-1} \bmod p^2$ 可以一次计算,重复使用)。在协议 1 中 Alice 需要生成公钥,并进行 $2n$ 次加密和 1 次解密,Alice 共需 $2n+2$ 次模指数运算;在协议 1 中 Bob 进行 1 次密文重随机化,仅需 1 次模指数运算。所以协议 1 共需要 $2n+3$ 次模指数运算。

在协议 2 中 Alice 生成公钥需要 1 次模指数运算,并分别需要 n 次加密与解密,Alice 共需 $2n+1$ 次模指数运算;在协议 2 中 Bob 需要进行 n 次模指数运算;协议 2 共需 $3n+1$ 次模指数运算。

文献[18]中的汉明距离和编辑距离计算协议是目前已有相关协议中效率较高的,其需要的模指数运算次数与本文协议 1 需要的模指数运算次数相同,但由于采用了不同的密码系统,在同样的安全水平上(同样的安全参数 k),本文协议 1 的效率明显较高。本文协议 1 的重要优势在于 Bob 仅需要做很少的计算,而且可以得到汉明距离的密文,这使得协议 1 可以适用于群智感知、众包等应用场景并作为其他协议的子模块使用。

如果编辑距离计算中每个字符需要用 m 比特表示,文献[18]中的编辑距离协议需要 $m(2n+3)$次模指数运算,本文协议 2 需要 $3n+1$ 次模指数运算,而且由于采用的是不同的密码系统,因此本文协议 2 进行一次模指数运算的计算成本明显较低。

6.2 实验结果

应用实际数据测试了本文协议和文献[18]中协议的运行时间。实验环境如下:Intel(R) Core(TM) i5-9400CPU @ 2.90GHz,2.90GHz,内存 8GB,64 位 Windows10 操作系统,Python 3.8+Pycharm。在实际测试中选择密码系统的安全参数 $k=1024$,实验测得汉明/编辑距离协议运行时间随字符串长度的变化规律如表 1 和表 2 所列。

表 1 汉明距离计算协议 1 与文献[18]协议 1 效率比较

Table 1 Efficiency comparison between our protocol 1 and that of literature [18] for computing Hamming distance

(单位:s)

字符串长度	200	400	600	800	1000
本文协议 1	1.01	2.01	3.02	4.10	5.12
文献[18]协议	13.70	27.30	40.90	54.50	68.10

表 2 本文协议 2 与文献[18]协议 2 效率比较

Table 2 Efficiency comparison between our protocol 2 and that of literature [18] for privately computing edit distance

(单位:s)

字符串长度	200	400	600	800	1 000
本文协议 1	1.51	3.01	4.51	6.01	7.51
文献[18]协议	109.60	218.20	327.00	436.00	545.00

从表 1 和表 2 可以看出,本文协议的效率与文献[18]的协议相比提高了 10～70 倍,说明了本文协议非常高效,而且协议执行时间与字符串长度的关系基本是线性关系。这是因为协议 1 执行的模指数运算次数是 $2n+3$ 次,与字符串长度的关系是线性关系,执行一次模指数运算的时间是常数,所以协议的运行时间和字符串长度也是线性关系。协议 2 与之类似。

结束语 本文设计了一个二进制字符串汉明距离的保密计算协议和任意字母表上字符串的编辑距离的保密计算协议。编辑距离协议可用于保密计算数字对象的相似度。所设计的协议在半诚实模型下是安全的,协议的效率与现有协议相比提高了 10～70 倍。本文协议特别适用于群智感知和众包等应用场景。未来将进一步研究更多保护隐私的距离计算协议以及设计抗主动攻击的汉明距离安全计算协议。

参 考 文 献

[1] YAO A C. Protocols for secure computations [C]// The 23rd Annual Symposium on Foundations of Computer Science. New York: ACM Press, 1982: 160-164.

[2] GOLDREICH O, MICALI S, WIGDERSON A. How to play any mental game[C]// The 19th Annual ACM Symposium on Theory of computing. New York: ACM Press, 1987: 218-229.

[3] BEN-OR M, GOLDWASSER S, WIGDERSON A. Completeness theorems for non-cryptographic fault-tolerant distributed computation(extended abstract) [C]// Proceedings of the STOC. New York: ACM Press, 1988: 1-10.

[4] YANG X Y, LI S D, KANG J. Private substitution and its applications in private scientific computation [J]. Chinese Journal of Computers, 2018, 41(5): 1132-1142.

[5] FAGIN R, NAOR M, WINKLER P. Comparing information without leaking it [J]. Communications of the ACM, 1996, 39(5): 77-85.

[6] LIU C, ZHU L H, HE X J, et al. Enabling privacy-preserving shortest distance queries on encrypted graph data [J]. IEEE Transaction on Dependable Secure Computing, 2021, 18(1): 192-204.

[7] WU X T, WU T T, KHAN M, et al. Game theory based correlated privacy preserving analysis in big data [J]. IEEE Transactions on Big Data, 2021, 7(4): 643-656.

[8] LEE A S, JUN S P. Privacy-preserving data mining for open government data from heterogeneous sources [J]. Government Information Quarterly, 2021, 38(1): 101544.

[9] LI Y, ZHOU Y P, JOLFAEI A, et al. Privacy-preserving federated learning framework based on chained secure multiparty computing [J]. IEEE Internet Things Journal, 2021, 8(8): 6178-6186.

[10] FREEDMAN M J, HAZAY C, NISSIM K, et al. Efficient set intersection with simulation-based security [J]. Journal of Cryptology, 2016, 29(1): 115-155.

[11] ZHU X J, AYDAY E, VITENBERG R. A privacy-preserving framework for outsourcing location-based services to the cloud [J]. IEEE Transactions on Dependable Secure Computing, 2021, 18(1): 384-399.

[12] DING X F, WANG Z, ZHOU P, et al. Efficient and privacy-preserving multi-party skyline queries over encrypted data [J]. IEEE Transactions on Information Forensics Security, 2021, 16: 4589-4604.

[13] FEIGENBAUM J, ISHAI Y, MALKIN T, et al. Secure multiparty computation of approximations [J]. ACM Transaction on Algorithms, 2006, 2(3): 435-472.

[14] BRINGER J, CHABANNE H, PATEY A. Shade: Secure hamming distance computation from oblivious transfer[C]// Proceedings of the International Conference on Financial Cryptography and Data Security. Berlin: Springer, 2013: 164-176.

[15] KULKARNI R, NAMBOODIRI A. Secure hamming distance based biometric authentication[C]// Proceedings of The International Conference on Biometrics. Piscataway: IEEE Press, 2013: 1-6.

[16] YASUDA M. Secure Hamming distance computation for biometrics using ideal-lattice and ring-LWE homomorphic encryption[J]. Information Security Journal: A Global Perspective. 2017, 26(2): 85-103.

[17] MA M Y, XU Y, LIU Z. Privacy preserving Hamming distance computing problem of DNA sequences[J]. Journal of Computer Applications, 2019, 39(9): 2636-2640.

[18] JARROUS A, PINKAS B. Secure computation of functionalities based on hamming distance and its application to computing document similarity [J]. International Journal of Applied Cryptography, 2013, 3(1): 21-46.

[19] OKAMOTO T, UCHIYAMA S. A new public-key cryptosystem as secure as factoring[C]// Proceedings of the EUROCRYPT. Berlin: Springer, 1998: 308-318.

[20] BOUDOT F, SCHOENMAKERS B, TRAORE J. A fair and efficient solution to the socialist millionaires' problem [J]. Discrete Application Mathematics, 2001, 111(1/2): 23-36.

DOU Jia-wei, born in 1963, Ph.D, associate professor. Her main research interests include applied mathematics, cryptography and information security.

(责任编辑:何杨)