

基于编辑距离的远程数据库安全搜索协议

仲红, 张守奇, 张瑞, 方兴, 李江华

(安徽大学 计算机科学与技术学院, 安徽 合肥 230039)

摘要: 远程数据库安全搜索作为安全多方计算的一项应用变得越来越重要, 它被广泛地应用到电子商务领域中。给出了基于编辑距离的远程数据库安全搜索协议, 回顾了编辑距离的定义及计算方法, 利用安全多方计算的相关知识构建了一系列基础安全协议以实现双方安全计算编辑距离, 并对这些协议的计算复杂度进行了分析。最后给出了基于编辑距离的远程数据库安全搜索协议和协议的代价, 该协议可以用在基于 DNA 序列匹配的远程数据库安全搜索中。

关键词: 编辑距离; 动态规划法; 同态公钥密码体制; 茫然第三方; 百万富翁协议; 相似模式匹配

中图分类号: TP393.08

文献标识码: A

文章编号: 1673-629X(2008)09-0134-04

A Protocol for Secure Remote Database Searching Based on Edit Distance Kind

ZHONG Hong, ZHANG Shou-qi, ZHANG Rui, FANG Xing, LI Jiang-hua

(School of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: Because of privacy protection, secure remote database searching, which is used as an application of secure multi-party computation, becomes more and more important and it is used in the e-commerce domain broadly. Gave a protocol for secure remote database searching based on edit distance kind, reviewed the definition of edit distance and how to compute that, then based on the knowledge of secure multi-party computation, constructed a set of basal secure protocols to make the edit distance computed securely by the two parties, and analysed the cost of them. Gave a protocol for secure remote database searching based on edit distance kind and its cost. This protocol can be used in secure remote database searching based on DNA sequence matching.

Key words: edit distance; dynamic programming; homomorphic encryption; oblivious third party; Yao's millionaire problem; approximate pattern matching

0 引言

自从1980年华裔计算机科学家、图灵奖获得者姚启智教授提出百万富翁协议^[1]以来,越来越多的人开始研究安全多方计算领域某些特定问题的高效的解决方案^[2,3],远程数据库安全搜索作为安全多方计算的一项应用变得越来越重要。一般意义上的远程数据库安全搜索^[4]是这样定义的: Alice 拥有一个字符串 q , Bob 拥有一个字符串数据库 $T = \{t_1, \dots, t_N\}$, Alice 想要知道 Bob 的数据库 T 中是否存在一个字符串跟 q 匹配, Alice 的查询 q 、返回结果不能泄漏给 Bob。这里的匹配包括精确模式匹配和相似模式匹配两种。精确模

式匹配的结果只有两种,要么存在,要么不存在,许多文章已经对精确模式匹配进行了广泛深入的研究^[5,6],它可以通过普通的安全多方计算协议解决。相似模式匹配则是给出一个比较结果,这个比较结果代表了两个序列的相似度,相似模式匹配在有些文章里也进行了初步的研究^[4]。不同的应用领域中两个序列相似度的表示形式也是不相同的,例如,在图像序列匹配中, $\sum_{i=1}^n (a_i - b_i)^2$ 和 $\sum_{i=1}^n |a_i - b_i|$ 经常用来表示两个序列 a 和 b 的相似度,而在 DNA 序列匹配中,编辑距离通常作为两个 DNA 序列的相似度的表示形式。文中给出了一个基于编辑距离的远程数据库安全搜索协议,它属于相似模式匹配。

收稿日期: 2007-12-25

基金项目: 国家自然科学基金资助项目(60773114); 安徽省自然科学基金资助项目(070412051); 安徽高校省级重点自然科学基金研究项目(KJ2007A43)

作者简介: 仲红(1965-),女,副教授,硕士生导师,研究方向为网络与信息安全、分布式计算。

1 编辑距离问题

假设 A 为一个长度为 n 的字符串 $A_1A_2 \cdots A_n$, B 为一个长度为 m 的字符串 $B_1B_2 \cdots B_m$, A 和 B 的元素都来自字母表 G 。所谓的编辑距离^[7]是指: 让字符串 A

变成字符串 B 的操作序列的最小代价。这里有三种允许对字符串 A 进行的操作:

- 1) 把某个字符 a 变成 b , 其代价记为 $S(a, b)$;
- 2) 删除某个字符 a , 其代价为 $D(a)$;
- 3) 插入某个字符 a , 其代价为 $I(a)$ 。

每一种把字符串 A 变成字符串 B 的操作序列都对应一个代价, 编辑距离指的是操作序列中的最小代价, 我们的协议允许任意的 $I(a)$, $D(a)$ 和 $S(a, b)$ 。

动态规划求解编辑距离:

令矩阵元素 $M(i, j)$ ($0 \leq i \leq n, 0 \leq j \leq m$) 表示 A 的子串 $A_1 A_2 \cdots A_i$ 和 B 的子串 $B_1 B_2 \cdots B_j$ 的编辑距离, 则 $M(n, m)$ 代表字符串 A 跟字符串 B 的编辑距离。

根据编辑距离的定义, 存在下述等式:

$$M(0, 0) = 0$$

$$M(0, j) = \sum_{k=1}^j I(B_k) \quad (1 \leq j \leq m)$$

$$M(i, 0) = \sum_{k=1}^i D(A_k) \quad (1 \leq i \leq n)$$

当 $1 \leq i \leq n$ 并且 $1 \leq j \leq m$ 时满足下列等式:

$$M(i, j) = \min(M(i-1, j-1) + S(A_i, B_j), M(i-1, j) + D(A_i), M(i, j-1) + I(B_j))$$

编辑距离的计算具有重叠子问题和最优子结构性, 符合动态规划法的基本要素, 可以使用动态规划法把复杂度降低至多项式级别 $O(nm)$ 。

2 相关的基础安全协议

2.1 同态公钥密码体制

对于公钥加密算法 $E(\cdot)$, 如果给定 $E(x)$ 和 $E(y)$, 在没有私钥的情况下能够计算出 $E(x \cdot y)$, 则称该公钥加密算法具有同态性质^[8]。

2.2 包含茫然第三方的对双方共同拥有的向量求最小元素的协议

假设 Alice 拥有一个向量 $a = (a_1, a_2, \dots, a_k)$, Bob 拥有一个向量 $b = (b_1, b_2, \dots, b_k)$, 向量 $c = a + b = (c_1, c_2, \dots, c_k)$ 。现在 Alice 和 Bob 想要计算向量 c 的最小元素, 且该最小元素由双方共同拥有(任何一方都不能单独知道该最小元素)。在对文献[9]中的对双方共同拥有的向量求最小元素的协议进行修正后, 提出了包含茫然第三方^[10,11]的对双方共同拥有的向量求最小元素的协议。协议描述如下:

对于任意的 $1 \leq i, j \leq k$, 我们知道 $c_i = a_i + b_i$, $c_j = a_j + b_j$, 要比较 c_i 和 c_j 两个元素的大小关系可以通过比较 $(a_i - a_j)$ 跟 $-(b_i - b_j)$ 的大小关系来确定。因此, 可以通过执行百万富翁协议来比较 $(a_i - a_j)$ 跟 $-(b_i - b_j)$ 的大小关系得出向量 c 中所有元素之间的大小关系。但是, 协议只是要求得到向量 c 的最小元素, 并且该最小元素由双方共同拥有, 向量 c 中元素之间

的大小关系对 Alice 和 Bob 双方来说应该是保密的。要解决这个问题, 比较的双方必须在执行百万富翁协议之前, 对各自拥有的向量作出必要的改变, 使得双方不能通过观察百万富翁协议的比较结果推断出向量 c 中元素之间的大小关系。协议如下所述:

通信双方为 Alice 和 Bob, Alice 拥有一个向量 $a = (a_1, a_2, \dots, a_k)$, Bob 拥有一个向量 $b = (b_1, b_2, \dots, b_k)$ 。引入一个茫然第三方 C 。

(1) Alice 生成同态公开加密密钥 E_A , 保密密钥 D_A , Alice 把 E_A 发送给 C , Bob 生成同态公开加密密钥 E_B , 保密密钥 D_B , Bob 把 E_B 发送给 C 。

(2) Alice 用自己的公钥 E_A 对她所拥有的向量 a 加密, 得到 a' , $a' = E_A(a)$, 然后 Alice 把 a' 发送给 C 。

(3) Bob 用自己的公钥 E_B 对她所拥有的向量 b 加密, 得到 b' , $b' = E_B(b)$, 然后 Bob 把 b' 发送给 C 。

(4) C 生成一个随机向量 r 和随机置换 π , 根据同态加密的性质, C 得出向量 $\theta_a = E_A(r) \cdot a' = E_A(r + a)$ 和 $\theta_b = E_B(-r) \cdot b' = E_B(b - r)$ 。然后令 $a'' = \pi(\theta_a)$, $b'' = \pi(\theta_b)$, 并且把 a'' 发送给 Alice, 把 b'' 发送给 Bob。

(5) Alice 用自己的私钥 D_A 解密 a'' , Bob 用自己的私钥 D_B 解密 b'' 。

上面的协议完成了对 Alice 和 Bob 各自拥有的向量的改变, 他们现在拥有向量的元素下标顺序相对于原来拥有向量的元素下标顺序已无任何参照关系, Alice 和 Bob 都不知道顺序是如何被打乱的, 他们也就不能通过观察百万富翁协议的比较结果来推断出向量 c 中元素之间的大小关系。

在改变双方拥有的向量的协议中, Alice 的主要代价包括 $E_A(a)$ 、 $D_A(a'')$, Bob 的主要代价包括 $E_B(b)$ 、 $D_B(b'')$, 茫然第三方 C 的主要代价包括 $E_A(r) \cdot a'$ 、 $E_B(-r) \cdot b'$, 因此时间复杂度为 $O(k)$ 。

双方在茫然第三方 C 的协助下执行上述协议后, 就可以通过执行百万富翁协议来比较向量 c 中元素之间的大小关系进而得出向量 c 的最小元素了。在这里, 百万富翁协议的执行数量级满足 $O(k^2)$ 。

2.3 $S(a, b)$ 的安全计算协议

假设字母表为 $G = \{1, 2, \dots, \partial\}$, $1 \cdots h \cdots \partial$ 代表 ∂ 个不同的字符, a, b 都来自字母表 G 。假设 Alice 拥有一个字符 a , Bob 拥有一个字符 b , 协议的目标是 Alice 和 Bob 各得到一个分量 F_A, F_B , $F_A + F_B = S(a, b)$, a, b 的信息不能泄漏给对方, 任何一方都不知道所要计算的 $S(a, b)$ 。协议执行如下:

(1) Alice 生成同态可交换加密密钥 E_A , 保密密钥

D_A , Bob 生成同态可交换加密密钥 E_B , 保密密钥 D_B 。

(2) Alice 选择一个随机的数值 k , 计算得到 $l_h = E_A(S(a, h) + k) (1 \leq h \leq \partial)$, 并把 l_h 发送给 Bob (按顺序发送), Alice 令 $F_A = -k$ 。

(3) Bob 取第 b 个元素 $l_b = E_A(S(a, b) + k)$, 加密 l_b 得到 $l'_b = E_B(E_A(S(a, b) + k))$, Bob 发送 l'_b 给 Alice。

(4) Alice 计算 $x = D_A(l'_b) = D_A(E_B(E_A(S(a, b) + k))) = D_A(E_A(E_B(S(a, b) + k))) = E_B(S(a, b) + k)$, Alice 发送 x 给 Bob。

(5) Bob 计算 $F_B = D_B(x) = D_B(E_B(S(a, b) + k)) = S(a, b) + k$, 满足 $F_A + F_B = S(a, b) + k - k = S(a, b)$ 。

该协议的主要代价是 Alice 执行 $O(\partial)$ 数量级同态加密的代价, Bob 执行 $O(1)$ 数量级同态加密的代价, 因此时间复杂度为 $O(\partial)$ 。

2.4 安全计算编辑距离的协议

假设 Alice 拥有一个字符串 $A = A_1A_2 \cdots A_n$, Bob 拥有一个字符串 $B = B_1B_2 \cdots B_m$, A 和 B 的元素都来自字母表 $G = \{1, 2, \cdots, \partial\}$, $1 \cdots h \cdots \partial$ 代表 ∂ 个不同的字符, 编辑距离矩阵为 M 。现在双方要安全计算编辑距离^[9], 要求双方不知道任何一个 $M(i, j)$, 因此编辑距离矩阵 M 应该由 Alice 和 Bob 双方共有。令 Alice 拥有一个矩阵 M_A , Bob 拥有一个矩阵 M_B , $M = M_A + M_B$, 协议的执行结果是 Alice 得到一个分量 $M_A(n, m)$, Bob 得到一个分量 $M_B(n, m)$, $M(n, m) = M_A(n, m) + M_B(n, m)$, 初始化:

Alice 令 $M_A(0, j) = 0 (0 \leq j \leq m)$, $M_A(i, 0) = \sum_{k=1}^i D(A_k) (1 \leq i \leq n)$

Bob 令 $M_B(i, 0) = 0 (0 \leq i \leq n)$, $M_B(0, j) = \sum_{k=1}^j I(B_k) (1 \leq j \leq m)$

这样就满足:

$$M_A(0, 0) + M_B(0, 0) = 0 = M(0, 0)$$

$$M_A(0, j) + M_B(0, j) = \sum_{k=1}^j I(B_k) = M(0, j) (1 \leq j \leq m)$$

$$M_A(i, 0) + M_B(i, 0) = \sum_{k=1}^i D(A_k) = M(i, 0) (1 \leq i \leq n)$$

对于任意的 $1 \leq i \leq n$ 并且 $1 \leq j \leq m$, $M(i, j)$ 按下面的协议计算:

(1) Alice 和 Bob 执行 $S(a, b)$ 的安全计算协议, Alice 得到一个分量 ∂_a , Bob 得到一个分量 ∂_b , 满足 $\partial_a + \partial_b = S(A_i, B_j)$ 。

(2) Alice 计算得到 $V_A = M_A(i-1, j-1) + \partial_a$, Bob 计算得到 $V_B = M_B(i-1, j-1) + \partial_b$, 满足 $V_A +$

$$V_B = M(i-1, j-1) + S(A_i, B_j)。$$

(3) Alice 计算得到 $U_A = M_A(i-1, j) + D(A_i)$, Bob 计算得到 $U_B = M_B(i-1, j)$, 满足 $U_A + U_B = M(i-1, j) + D(A_i)$ 。

(4) Alice 计算得到 $W_A = M_A(i, j-1)$, Bob 计算得到 $W_B = M_B(i, j-1) + I(B_j)$, 满足 $W_A + W_B = M(i, j-1) + I(B_j)$ 。

(5) Alice 得到一个向量 $X_A(V_A, U_A, W_A)$, Bob 得到一个向量 $X_B(V_B, U_B, W_B)$, 满足 $X_A + X_B = (M(i-1, j-1) + S(A_i, B_j), M(i-1, j) + D(A_i), M(i, j-1) + I(B_j))$ 。Alice 和 Bob 通过执行对双方共同拥有的向量求最小元素的协议各自得到一个分量 Z_A, Z_B , 满足 $Z_A + Z_B = \min(M(i-1, j-1) + S(A_i, B_j), M(i-1, j) + D(A_i), M(i, j-1) + I(B_j))$ 。

(6) Alice 把 Z_A 赋值给矩阵元素 $M_A(i, j)$, Bob 把 Z_B 赋值给矩阵元素 $M_B(i, j)$ 。

根据动态规划算法, 上面的协议需要执行的数量级是 $O(mn)$, 最终 Alice 得到矩阵元素 $M_A(n, m)$, Bob 得到矩阵元素 $M_B(n, m)$, 满足 $M(n, m) = M_A(n, m) + M_B(n, m)$ 。在 $M(i, j)$ 的计算中, 主要代价包括 Alice 和 Bob 双方在 (1) 执行 $S(a, b)$ 的计算协议, 数量级是 $O(\partial)$, 在 (5) 中执行对双方共同拥有的向量求最小元素的协议, 数量级是 $O(1)$, 安全求解编辑距离的协议的主要代价就是 $O(mn)$ 数量级 $M(i, j)$ 的计算, 因此时间复杂度是 $O(mn\partial)$ 。

3 基于编辑距离远程数据库安全搜索协议

文中给出了一个基于编辑距离的远程数据库安全搜索协议, 它属于相似模式匹配。协议模型如图 1 所示。

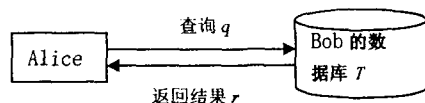


图 1 基于编辑距离的远程数据库安全搜索协议图

Alice 拥有一个字符串 q , Bob 拥有一个字符串数据库 $T = \{t_1, \cdots, t_N\}$, Alice 想要知道数据库中字符串与查询字符串 q 最小的编辑距离, 同时满足 Alice 的查询 q 和返回结果以及 Bob 的数据库信息都不能泄漏给对方。出于安全考虑, 协议不允许 Alice 知道查询字符串 q 跟 Bob 数据库中字符串 (返回结果除外) 的编辑距离, 同样 Bob 也不能知道查询字符串 q 跟自己数据库中字符串的编辑距离, 因此, 不能简单地把上面安全求解编辑距离的协议执行 N 次得出编辑距离然后比较大小。在协议中引入一个茫然第三方 C' , 假设查询字

字符串为 q , 长度为 n , Bob 的字符串数据库 $T = \{t_1, \dots, t_N\}$, t_i 长度为 k_i , 协议执行如下:

1) Bob 生成一个随机数 R 并把 R 发送给 Alice。

2) 对于 q 和任一个 t_i , Alice 和 Bob 执行安全求解编辑距离的协议, Alice 发送 $M_A(n, k_i) + R$ 给 C' , Bob 发送 $M_B(n, k_i) + R$ 给 C' 。

3) C' 计算 $r_i = M_A(n, k_i) + R + M_B(n, k_i) + R = M(n, k_i) + 2R$ 。

4) C' 计算 $r' = \min_{1 \leq i \leq N} M(n, k_i) + 2R$ 并把 r' 发送给 Alice。

5) Alice 计算 $r = r' - 2R$ 。

在 r 的计算中, 主要代价为 N 次执行安全计算编辑距离的协议, 时间复杂度是 $O(mn\partial N)$ 。

4 结束语

基于编辑距离的远程数据库安全搜索协议中, Alice 既得到了 Bob 数据库中字符串与查询字符串最小的编辑距离, Bob 又不能得到任何关于查询字符串和返回结果的信息。它可以应用到现实生活的以下场景中: 假设 Alice 身体出现了某种不适, 她想确认一下自己是否生病, 并且她知道 Bob 拥有一个 DNA 序列匹配特征数据库, 因此她发送自己的 DNA 序列信息给 Bob, Bob 把返回结果发送给 Alice, Alice 可以根据返回的结果判断自己是否生病。然而, 出于隐私保护的考虑, Alice 不想让 Bob 知道自己的 DNA 序列信息以及返回的结果, 同时, 除了返回结果外 Bob 也不想让 Alice 知道任何关于 DNA 序列匹配特征数据库的信息。然而, 该协议只是数据库安全搜索相关领域最基础的应用, 前提是基于安全信道 (Alice、Bob、茫然第三方之间的数据传递都是安全的), 并且 Alice、Bob、茫然第三方都是严格执行协议 (他们之间不存在作弊行为)。

信道不安全或者参与的多方存在作弊时如何完成搜索, 以及与之相关的许多应用, 仍然需要进一步研究。

(上接第 133 页)

参考文献:

- [1] 斯廷森. 密码学的原理与实践[M]. 冯登国译. 北京: 电子工业出版社, 2003.
- [2] Rivest R L, Shamir A, Adleman L. A method for obtaining digital signatures and public key cryptosystems[J]. Communications of the ACM, 1978, 21(2): 120-126.
- [3] 朱文余, 孙琦. 计算机密码应用基础研究[M]. 北京: 科学出版社, 2000.
- [4] 李荣森, 秦杰, 宴文华. RSA 系列算法在工程中的应用研究[J]. 计算机科学, 2007, 34(2): 86-90.

参考文献:

- [1] Yao A. Protocols for Secure Computations[C]//Proceedings of the Annual IEEE Symposium on Foundations of Computer Science. [s.l.]: [s.n.], 1982: 160-164.
- [2] 李顺东, 戴一奇, 游启友. 姚氏百万富翁问题的高效解决方案[J]. 电子学报, 2005(5): 3-4.
- [3] 仲红. 安全多方计算的关键技术分析[J]. 安徽农业大学学报, 2007, 34(2): 291-295.
- [4] Du W, Atallah M J. Protocols for Secure Remote Database Access with Approximate Matching[C]//In Proc. of the 7th ACM Conference on Computer and Communications Security, the First Workshop on Security and Privacy in E-Commerce. Greece: [s.n.], 2000: 2-22.
- [5] Kushilevitz E, Ostrovsky R. Replication is not needed: Single database, computationally private information retrieval[C]//In Proceedings of the 38th annual IEEE computer society conference on Foundation of Computer Science. Miami Beach, Florida, USA: [s.n.], 1997: 2-4.
- [6] Chor B, Gilboa N. Computationally private information retrieval (extended abstract)[C]//In Proceedings of the twenty-ninth annual ACM symposium on Theory of computing. El Paso, TX, USA: [s.n.], 1997: 2-3.
- [7] Wagner R A, Fischer M J. The String to String Correction Problem[J]. Journal of the ACM, 1974, 21(1): 168-173.
- [8] Goldwasser S, Micali S. Probabilistic Encryption[J]. Journal of Computer and System Sciences, 1984, 28(2): 270-299.
- [9] Atallah M J, Kerschbaum F, Du W. Secure and private sequence comparisons[C]//In proceedings of Workshop on Privacy in the Electronic Society. Washington, D C, USA: [s.n.], 2003: 3-4.
- [10] Du W, Atallah M J. Secuer multi-party computation problems and their applications: A review and open problems[C]//In New Security Paradings Workshop. Cloudecoft, New Mexico, USA: [s.n.], 2001: 11-12.
- [11] Cachin C. Efficient private bidding and auctions with an oblivious third party[C]//In Proceeding of the 6th ACM conference on Computer and communications security. Singapore: [s.n.], 1997: 120-127.

- [5] Rabin M. Probabilistic Algorithms for Testing primality[J]. Journal of Number Theroy, 1980, 12: 128-138.
- [6] Montgomery P L. Modular Multiplication Without Trial Division[J]. Mathematics of Computation, 1985, 44(170): 519-521.
- [7] 王冕, 周玉洁. 分割式 Montgomery 模乘运算的线性高基心动阵列新结构[J]. 计算机科学, 2006, 33(1): 184-187.
- [8] 李占才, 王许书, 涂序彦. RSA 快速硬件实现研究[J]. 计算机研究与发展, 2001, 38(11): 1360-1365.
- [9] 齐晓虹, 刘冬, 赵岳松. RSA 公开密钥密码体制的密钥生成研究[J]. 武汉理工大学学报, 2001, 23(6): 37-40.