Practical Techniques for Searches on Encrypted Data 阅读笔记

李忆诺

2024 年 3 月 4 日

1 基本信息

1.1 论文来源

Song D X, Wagner D, Perrig A. Practical techniques for searches on encrypted data[C]//Proceeding 2000 IEEE symposium on security and privacy. S&P 2000. IEEE, 2000: 44-55.

1.2 概述

本文提出了一种可搜索加密系统，实现在不丢失数据机密性的情况下，支持远程搜索功能。除此之外，该加密系统还支持受控搜索与隐藏搜索，以及查询隔离功能。

2 论文要点

2.1 背景

对于各类数据服务器而言，完全可信的服务器数量少之又少。而在不受信任的服务器上存储数据时，数据应当以加密方式存储，以降低安全风险和隐私风险，这意味着通常需要为了数据安全而牺牲其他功能，如数据检索等等。并且，在加密数据上进行计算非常困难。

2.2 价值

该论文解决了客户机如何在不受信任的服务器上进行信息检索问题，实现在不丢失数据机密性的情况下，让数据存储服务器执行搜索并回答查询。

其主要贡献有：

- 提出了一种可搜索加密方案 SWP。该方案为加密提供了可证明的保密性，不受信任的服务器不能从密文中获取关于明文的任何信息；
- 该方案实现了受控搜索，不受信任的服务器在没有用户授权时无法搜索单词；

- 该方案实现了隐藏查询，用户可以请求服务器查询一个秘密单词，而不向服务器透露该单词；
- 该方案实现了查询隔离，不受信任的服务器只能了解到关于明文的搜索结果，而不能获取关于明文的其他任何信息；

2.3 问题陈述

2.3.1 安全定义

如果一个攻击算法用资源 R 成功破坏一个密码原语，则我们称为一个攻击 R-**breaks** 这个密码原语；如果没有算法可以 R-**breaks** 一个密码原语，则我们说该密码原语是 R-**secure** 的。

定义 $A: \{0,1\}^n \to \{0,1\}$，$X, Y \in \{0,1\}^n$ 则算法 A 的区别概率为：

$$Adv\ A = |Pr[A(X) = 1] - Pr[A(Y) = 1]|$$

简而言之，本文采用了可证明安全性文献中安全的标准定义，并根据破解它们所需的资源来衡量加密原语的强度。

2.3.2 定义原语

根据上述定义，本文所需原语如下：

- 伪随机发生器 G：如果任意一个运行时间不超过 t 的算法 A，满足 $Adv\ A < e$，则 $G: \mathcal{K}_G \to s$ 是一个 (t, e)-secure 伪随机发生器；
- 伪随机函数 F：如果每个最多进行 q 次 oracle 查询，且最长运行时间不超过 t 的 oracle 算法 A 满足 $Adv\ A < e$，则 $F: \mathcal{K}_F \times \mathcal{X} \to \mathcal{Y}$ 是一个 (t, q, e)-secure 伪随机函数；
- 伪随机排列 E：如果每个最多进行 q 次 oracle 查询，且最长运行时间不超过 t 的 oracle 算法 A 满足 $Adv\ A < e$，则 $E: \mathcal{K}_E \times \mathcal{Z} \to \mathcal{Z}$ 是一个 (t, q, e)-secure 伪随机排列。

2.4 方法

可搜索加密方案 SWP 通过使用具有特殊结构的伪随机位序列来达到搜索加密数据的效果。其核心想法如下：

1. 加密算法
    - 选取一种确定性加密算法 E，使用密钥 k'' 对文本中的每个单词进行加密，要求该加密算法只能基于单词本身，而不能受到该单词在文本中的位置等因素影响；
    - 对所得单词密文进行分段。设位于位置 i 处的某单词长度为 n，则分为长度为 $n-m$ 的 L_i 和长度为 m 的 R_i；
    - 客户随机且均匀地选取一个主密钥 k'，额外定义一个伪随机函数 $f: \mathcal{K}_f \times \{0,1\}* \to \mathcal{K}_f$，用于选择每个位置的密钥，即 $k_i := f_{k'}(L_i)$；

- 使用伪随机生成器 G 生成序列 S_i，使用伪随机函数 F 生成 $F_{k_i}(S_i)$，然后将 S_i 和 $F_{k_i}(S_i)$ 拼接起来得到 $\langle S_i, F_{k_i}(S_i)\rangle$；
- 将单词加密结果与 $\langle S_i, F_{k_i}(S_i)\rangle$ 进行异或计算，得到最终密文，上传给服务器。

2. 搜索过程

- 客户想要查询某个单词 W，则先对该单词进行加密，得到 $E_{k''}(W)$，进而计算得到 k_i，将 $E_{k''}(W)$ 和 k_i 发送给服务器；
- 服务器通过检查 $C_i \bigoplus E_{k''}(W)$ 是否为某种 $\langle S_i, F_{k_i}(S_i)\rangle$ 形式，来判断当前单词是否为搜索目标；
- 服务器返回搜索结果对应的密文，用户通过序列 S_i 与密文的前 $n-m$ 位异或得到 L_i，从而进行后续计算，成功解密出明文。

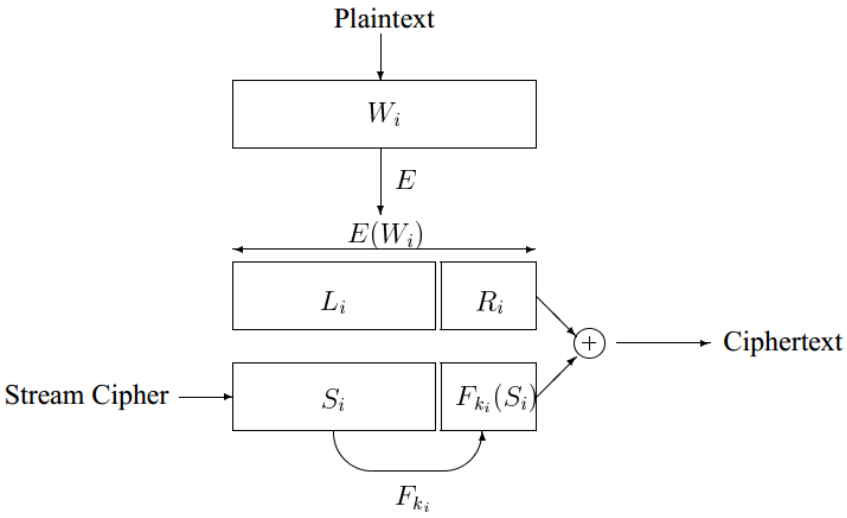

图 1: Encryption Scheme

2.5 结果

本文设计了一种使用不受信任的服务器对加密数据进行远程搜索的技术，并为生成的加密系统提供了安全性证明。

该技术简单快速。对于长度为 n 的文档，加密和搜索算法只需要 $O(n)$ 次流密码和分组密码操作，也几乎没有空间和通信开销。

该方案十分灵活，可以很容易地扩展到支持更高级的搜索查询，这为在不受信任的基础设施中构建安全服务提供了一个强大的新构建块。

3 评论

3.1 局限性

在该方案中，为了允许服务器搜索用户指定单词，用户潜在地泄露了指定单词可能出现的位置列表，如果该过程重复多次，可能会引起统计学攻击隐患。

另外，该方案基于服务器会诚实地返回所有查询结果，如果服务器只返回了部分结果，用户将无法检测到这一点。

3.2 扩展阅读

Curtmola R , Garay J , Kamara S ,et al.Searchable symmetric encryption: Improved definitions and efficient constructions[J].Journal of computer security, 2011, 19(5):p.895-934.

Sharma M K , Karpagam S .Privacy-preserving multi-keyword ranked search over encrypted cloud data[C]//IEEE Infocom.IEEE, 2011.DOI:10.1109/INFCOM.2011.5935306.

3.3 启示

通过阅读该论文，我对可搜索加密的基本方案有了基本的认识。

该论文从**如何进行远程搜索**着手，选择通过异或生成密文的方式实现，然后考虑**如何保证数据安全性**，引入了加密，最后思考**如何保证服务器返回的加密数据能成功解密**，以一种巧妙的方式实现：将加密数据分成两部分。

单独看这三个处理方式，前两个是经典加密思路，后一个处理在各类加密算法中也十分常见。在密码学课程学习过程中，我对其的理解或许仅仅停留在某个特定的加密方案中。而这些处理方法是根据什么思路组装，最终形成一种加密方案，在该论文中有鲜明的体现，也让我对加密中数据的各种处理方法有了新的理解。