一种云计算环境下的加密模糊检索方案

梁 宇 路 劲 刘笠熙 张 驰
(云南大学软件学院 昆明 650091)

摘 要 随着移动终端的发展,云计算也越来越普及,很多敏感数据被集中存储到云存储器上。为了保证数据的私密性,这些数据在上传时应该经过加密,这就使得传统数据检索方案的可行性降低。一些在加密环境下的数据检索方案,只能处理准确的关键字检索,对于拼写错误、格式错误的情况则无法进行正常查询,因此不适用于云计算。提出了一种基于k-gram索引的模糊关键字在加密环境下的检索方案。这种方案可以提供一种安全、高效的数据存取服务。

关键词 云计算,加密,模糊检索
中图法分类号 TP393 **文献标识码** A

Vague-keyword Search under Encrypted Environment in Cloud Computing

LIANG Yu LU Jin LIU Li-xi ZHAGN Chi
(School of Software, Yunnan University, Kunming 650091, China)

Abstract As there is a significant development in potable devices market recently, cloud computing is more commonly implemented. Confidential datas are accumulated and stored in cloud servers. To ensure the datas are secured, they should be encrypted when being uploaded, which makes efficient searching among them a much complicated task. Traditional searching programmes in encrypted environments are able to find the data only if the keyword is exactly correct, but will fail if spelling errors or format errors occur. These programmes are hence not suitable for cloud computing. We introduced a high efficiency programme which can deal with even vague keyword searching in the encrypted environment.

Keywords Cloud computing, Encryption, Fuzzy search

1 引言

随着云计算的普及,越来越多的数据(私人邮件、个人文件、个人信息等)被集中存储到云存储器上。移动终端的发展也极大地展现了人们对云存储服务的需求。随时随地地享受到高存储量高效率的数据存取服务是用户十分期望的。

在得益于云服务便利的同时,考虑得最多的就是数据在云服务器上的安全问题。由于云存储器中会存储一些用户的敏感信息,或者重要的文件,数据的私密性就必须得到保证。在实际情况中,云服务器是不能被完全信任的。因此用户的数据在上传到云服务器之前是应该加密的。在这种加密的环境下,传统的数据检索方案已经不能满足其需求。

在本文中,我们在保证数据私密性的情况下,提出了一种基于k-gram索引的模糊关键字在加密环境下的检索方案。这种方案可以提供一种安全、高效的数据存取服务。本文结合实际,对提出的方案进行了详细的描述和深入分析。

2 概述

一种面向大众的服务需要保证安全性和便利性。云存储服务中,为了保证用户数据的安全,在用户上传数据到云服务器之前,数据是经过加密的。在加密环境下进行高效的模糊查找是一件比较困难的事情。一些传统的数据检索方案,比如基于关键字的模糊检索,在明文环境下,它就有着很好的表现。但是在加密的环境下,它就并不那么适用。在云计算的环境下,服务器上存储了大量的数据,同时数据的拥有者还可能将他的数据共享给大量的其他用户。这些用户中的一部分,可能只想从共享数据中取出一些他比较感兴趣的。现在对大量数据进行检索用的比较多的方案是使用基于明文关键字查询的文件检索,但是这种方式并不适用于云计算的环境。在云计算环境中,数据的私密性包括了关键字的私密性,关键字在很多时候包含了很多重要的信息。如果对关键字进行加密,那么在传统方式的下,关键字加密使得数据检索变得很麻烦,而且检索效果也不好。

近年来提出的一些在加密环境下的数据检索方案中,往往为每一个关键字建立了一个索引,然后将这个索引与含有其关键字的文件关联起来。这种索引包含关键字陷阱门的方式保证了检索的安全性。但是这种方案只支持特定关键字的检索,也就是说如果用户输的关键字中出现了哪怕是极小的拼写错误或格式上的不一致,使得关键字与系统预先设定的关键字不匹配(例如jake,jaky,或者tree,trea),就无法查询到相关文档。这种方式是无法支持加密环境下模糊关键字检索的。

本文受云南省自然基金项目(2008CD0084)资助。

梁 宇(1964—),男,硕士,高级工程师,主要研究方向为计算机网络、移动计算,E-mail:yuliang@ynu.edu.cn;路 劲(1986—),男,硕士生,主要研究方向为分布式计算;刘笠熙(1989—),男,本科生,主要研究方向为计算机安全等;张 驰(1990—),男,本科生,主要研究方向为信息安全、密码技术等。

在部分方案中可能会采用一些对关键字进行拼写纠错就的方法，但是还是不能完全解决这个问题，比如用户将 hat 输入成 cat，hat 与 cat 同样是正确的单词，拼写纠错就无法对这类关键字问题进行纠错。

在本文中，把问题集中于解决在加密环境下进行高效的模糊关键字查询。在我们的方案中，模糊关键字检索能够在用户输入关键字后，查找到与之符合或最为接近的结果。使用编辑距离来衡量关键字的相似度，并且生成了一个模糊关键字集，其中使用到 k-gram 技术。然后返回用户检索到的全部结果。对这种方案进行了模拟实现，实验表明，在这种方案下，模糊检索的效果很好，效率也很高。

3 系统构想以及设计目标

3.1 系统模型

提出一种系统模型，由数据拥有者、数据使用者以及服务器组成。

给出一个加密后的文件集合 $C=\{F_1,F_2,F_3,\cdots,F_n\}$，这些文件存储在云服务器上，再给出一组预先定义好的关键字集合 $K=\{K_1,K_2,K_3,\cdots,K_m\}$，云服务器为授权用户提供在加密文件集合 C 上的数据检索服务。

系统运行时，被数据拥有者授权的数据使用者，向云服务器提交检索请求，并向其发送检索陷阱门，云服务器利用索引（陷阱门和文件 ID 的集合加密生成）将用户的请求与文件相匹配。系统执行时有以下两个原则：1）如果找到完全匹配的关键字集，则返回与该关键字集相匹配的文件 ID；2）如果没有找到完全匹配的关键字集，则寻找与请求关键字集最接近的关键字集，同样返回对应的文件 ID。此种情况下包括用户输入的关键字有拼写错误或格式不一致。

3.2 系统威胁

尽管文件是经过加密的，但是因为服务器是不完全受信任的，它仍可能试图从用户的检索请求中获取信息，比如使用统计攻击的方式。所以我们的系统必须在让授权用户取得文件的同时，让服务器得到尽可能少的信息。

3.3 设计目标

给出一种在加密环境下进行模糊关键字检索的方案，该方案满足 1）对模糊关键字集的存储量是可接受的；2）查询效率较高；3）没有对系统的安全性造成破坏。

4 系统概念定义

4.1 编辑距离

给出两个字符串 S_1 和 S_2，它们之间的编辑距离指的就是从 S_1 转变到 S_2 所需要的最少的编辑操作。这些操作通常是指：1）插入一个字母；2）删除一个字母；3）修改一个字母。

4.2 模糊关键字检索

给出一个加密文件集合 $C=\{F_1,F_2,F_3,\cdots,F_n\}$，该文件集合存储于云服务器上，同样给出预定的关键字集合 $K=\{K_1,K_2,K_3,\cdots,K_m\}$，以及编辑距离 d，用户输入关键字 (K_s,s)，s 为用户输入的编辑距离，用来衡量用户期待的关键字模糊程度。在系统执行模糊关键字查询时，如果用户输入的关键字 $K_s=K_i\in K$，则返回包含 K_i 的文件 id，Fid；如果 K_s 不属于关键字集合 K，则计算 K 与 K_i 的编辑距离 e，找到 $e<\min(s,d)$，然后返回包含 K_i 的文件 id。

5 加密环境下模糊关键字检索方案

在加密环境下模糊关键字检索的核心包含了两个方面：1）建立一个模糊关键字的索引，用于对用户输入的检索关键字进行拼写纠错和模糊检索；2）在模糊关键字集的基础上，设计一套高效的安全搜索方案。

5.1 基于 k-gram 索引的模糊关键字集的建立

使用 k-gram 索引来生成关键字的模糊集，k-gram 是一串 k 个字符，比如在单词 ballet 中，bal，all，let 都属于 3-gram。为了使任意词中的字母都可以分到一个 gram 里面，我们在一个词的前面和后面添加上 $，以 ballet 为例，它的所有 3-gram 的集合为{$ba，bal，all，lle，let，et$}。

接下来，要为数据拥有者上传的文件的关键字集生成基于 k-gram 的索引。这里以 3-gram 为例，使用链式结构来建立这个索引，以 3-gram：etr 为例，建立如下索引：

这样每一个包含 ine 串的单词都被关联到了一起，便于进行检索。

下面以 ad * ce 为例来说明如何利用 3-gram 进行检索。我们的目标是检索含有 ad，ce 的单词，先分别对 ad 和 ce 进行检索，使用 3-gram 生成 $re 与 ve$，然后在系统中预先设定好的索引中对 $re 与 ve$ 分别进行检索，将返回的结果进行‘与’运算，得到结果，例如 advice，advance 等。然后根据每一个查询到的关键字，找到与之相关联的文件，返回文件 ID。

对于可能存在的一些特殊情况，例如，对 red * 进行检索，我们使用 3-gram 方法，生成 $re，red。我们先对 $red 进行检索，得到例如 retired，然后对 red 进行检索，将得到的结果进行‘与’运算，最后的结果中 retired 仍旧保持了下来，但是这个结果并不符合我们要检索的 red *（前 3 个字符不完全匹配），对于这种情况，我们在检索完成后，将结果再次进行一下字符串匹配的操作，将 red * 与 retired 前面字母部分进行对比，可将这个结果去除。

这种方式同样支持多关键字的索引，例如用户输入 ad * e，af * d，先分别对这两个单词进行 k-gram 索引的检索，然后将结果进行布尔与运算，得到用户期望的结果。

5.2 安全高效的关键字检索方案

如图 1 所示，系统运行时：

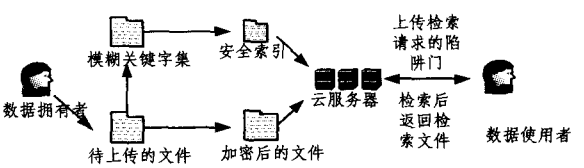

图 1 模糊关键字检索框架

1. 用户上传文件到云服务器前，使用 k-gram 索引方法对文件的关键字生成模糊关键字索引集合和编辑距离 d，$S_{(kd)}$。然后使用用户密钥 k_s 为每个文件的关键字集生成一个陷阱门集合 $T_{(kd)}$，k_s 是数据拥有者和授权的数据访问者所共有的。数据拥有者将文件 id 以及与文件相关联的模糊关键字集合使用 k_s 加密，将先前生成的陷阱门集合同加密后的文件 id 集合一起上传到服务器，如图 2 所示。

（下转第 105 页）

$$\Pr[Sucess]\geqslant\frac{1}{2^l}\cdot\Pr[\wedge_{i=1}^{q_E}(H(ID_i)\neq 0(\bmod\lambda_1))]\cdot\Pr[\wedge_{j=1}^{q_V}(H(ID_j)\neq 0(\bmod\lambda_1)\vee F(M_j)\neq 0(\bmod\lambda_2))]\cdot Pr[(F(M^*)\equiv 0(\bmod p))]\cdot\varepsilon'$$

$$=\frac{1}{2^l}\cdot(1-\Pr[\vee_{i=1}^{q_E}(H(ID_i)\equiv 0(\bmod\lambda_1))])\cdot(1-\Pr[\vee_{j=1}^{q_V}(H(ID_j)\equiv 0(\bmod\lambda_1)\wedge F(M_j)\equiv 0(\bmod\lambda_2))])\cdot\Pr[(F(M^*)\equiv 0(\bmod p))]\cdot\varepsilon'$$

$$\geqslant\frac{1}{2^l}\cdot(1-\frac{q_E}{\lambda_1})\cdot(1-\frac{q_V}{\lambda_2})\cdot\frac{1}{n\lambda_2}\cdot\varepsilon'$$

$$=\frac{\varepsilon'}{2^{l+3}nq_V}$$

(3) 抵抗仲裁者攻击

一个不诚实的仲裁者 T,其目的是借助预言机 O_{VESig},O_{Ext} 的帮助,在没有相应的可验证加密签名的情况下,直接产生有效的普通签名(M,σ)。由于 T 拥有仲裁私钥 tsk,它可以由任意可验证加密签名得到普通签名。对于 T 来说,预言机 O_{VESig} 可以看作普通签名预言机。因此,T 对方案 Σ 的伪造签名就是对 BW 方案的有效伪造签名。

综上所述,在(t,ε)-CDH 假定下,方案 Σ 是安全的。

结束语 本文应用 Waters 方法,提出了一个基于身份的可验证加密签名方案,并在标准模型下,证明了方案的安全性可以归约为双线性群中 CDH 问题的难解性。本文方案满足可验证性和可恢复性,这是其区别于普通签名方案的主要特性,在电子商务中具有重要的应用价值。

参 考 文 献

[1] Shamir A. Identity-based cryptosystems and signature schemes [C]// Proceedings of the Advances in Cryptology-Crypto'84. LNCS 196. Berlin, Heidelberg: Springer-Verlag, 1984: 47-53

[2] Boneh D, Franklin M. Identity-based encryption from the Weil pairing[C]// Proceedings of the Advances in Cryptology-Crypto' 2001. LNCS 2139. Berlin, Heidelberg: Springer-Verlag, 2001: 213-229

[3] Hess F. Efficient identity based signature schemes based on pairings[C]// Proceedings of the SAC 2002. LNCS 2595, Berlin, Heidelberg: Springer-Verlag, 2003: 310-324

[4] Boneh D, Boyen X. Secure identity based encryption without random oracles[C]// Proceedings of the Advances in Cryptology—Crypto' 2004. LNCS 3152. Berlin, Heidelberg: Springer-Verlag, 2004: 443-459

[5] Paterson K G, Schuldt J C N. Efficient identity-based signatures secure in the standard model[C]// Proceedings of the ACISP 2006. LNCS 4058, Berlin. Heidelberg: Springer-Verlag, 2006: 207-222

[6] 李进,张方国,王燕鸣. 两个高效的基于分级身份的签名方案[J]. 电子学报,2007,35(1):150-152

[7] Zhang Z F, Feng D G, Xu J, et al. Efficient ID-based optimistic fair exchange with provable security[C]// Proceedings of the 7th International Conference on Information and Communication Security. LNCS 3783. Berlin, Heidelberg: Springer-Verlag, 2005: 14-26

[8] Xu J, Zhang Z F, Feng D G. Constructing optimistic ID-based fair exchange protocols via proxy signature[J]. Journal of Software, 2007, 18(3): 746-754

[9] Waters B. Efficient identity-based encryption without random oracles[C]// Cramer R, ed. EUROCRYPT' 2005. LNCS 3494. Berlin, Heidelberg: Springer-Verlag, 2005: 114-127

[10] Boyen X, Waters B. Compact group signatures without random oracles[C]// Berlin, LNCS 4004. Heidelberg: Springer-Verlag, 2006: 427-444

[11] 李继国,姜进平. 标准模型下可证明安全的基于身份的高效签名方案[J]. 计算机学报,2009,32(11):2130-2136

(上接第 100 页)

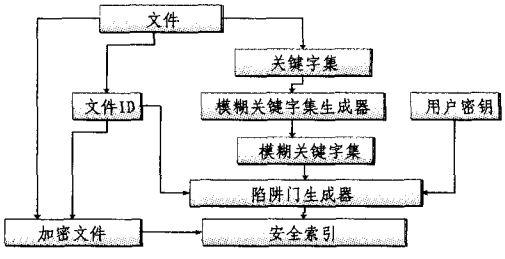

图 2 用户段的文件索引生成

2. 当被授权的用户检索他所感兴趣的文件时,首先根据用户输入的关键字生成模糊关键字集合,然后使用密钥 k_s 加密生成关键字陷阱门,将这个陷阱门上传到云服务器进行检索。

3. 云服务器接受到授权用户上传的陷阱门后,通过和数据拥有者上传的索引表进行对比(此时对比的是数据拥有者上传的数据模糊关键字集索引表和由检索用户输入关键字生成的模糊关键字集合陷阱门),返回所有相关联的加密文件 id,然后发送加密文件给授权的用户。用户使用密钥 k_s 进行解密后得到期望的文件。对于用户的检索关键字信息,检索到的文件信息都是经过加密的,对服务器是不可见的。

结束语 本文提出了一种在加密的云计算环境下进行高效模糊检索的方案。将生成的模糊关键字集与加密文档 ID 一起生成安全的索引。在保证数据私密性的情况下,可以进行高效的数据模糊检索。

在接下来的工作里面,会从服务器检索的方面来考虑提升系统的效率和安全性。安全性方面我们考虑将存储以及查询分离;效率方面我们将设计一种特殊的新的索引来辅助检索。

参 考 文 献

[1] Li Jin, Wang Qian, Wang Cong, et al. Fuzzy Keyword Search over Encrypted Data in Cloud Computing[C]// Proceedings of INFOCOM'2010. 2010: 441-445

[2] An Introduction to Information Retrieval[M]. Cambridge University Press, 2008

[3] Goh E-J. Secure Indexes[Z]. Stanford University, March 2004

[4] Kamara S, Kistin S, Lauter K. Cryptographic Cloud Storage[Z]. January 2010

[5] Shay A, Adam K, Calvin N, et al. Encrypted Keyword Search in a Distributed Storage System