第35卷 第22期 计 算 机 工 程 2009年11月
Vol.35 No.22 Computer Engineering November 2009
·安全技术· 文章编号：1000—3428(2009)22—0153—02 文献标识码：A 中图分类号：TP309

加密数据中连接关键词的安全搜索算法

刘星毅
(钦州学院数学与计算机科学系，钦州 535000)

摘 要：现有关键词搜索算法只能处理单个关键词，且检索复杂度高。针对该问题提出同时搜索多个连接关键词的加密数据安全搜索算法，该算法把生成“能力”的过程分为线下和线上2个部分，它对多个连接关键词的安全搜索时间比原有算法降低了80%左右，实例分析结果验证了其正确性和有效性。

关键词：数据库安全；加密数据；连接关键词搜索

Secure Search Algorithm for Conjunctive Keyword in Encrypted Data

LIU Xing-yi
(Department of Mathematics and Computer Science, Qinzhou University, Qinzhou 535000)

【Abstract】 Existing keyword search algorithms can only deal with single keyword and presenting high time complexity. Aiming at this problem, this paper proposes a secure search algorithm for encrypted data which searches several conjunctive keywords at the same time. This algorithm partitions the processes for generating “Capability” into two parts——online part and offline part. Its run time of secure search is about eighty percents less than existing algorithms. Example analysis results show that this algorithm is reasonable and effective.

【Key words】 database security; encrypted data; conjunctive keyword search

1 概述

由于事先不知道服务器是否可以信任(即使知道目前可以信任，也无法确定将来是否可以信任)，因此用户经常把加密的文件或数据存储到不可信任的服务器上。例如，将电子邮件放到邮件服务器时，用户一般不知道当前邮件服务器是否可以信任。为了保证自己的数据不被泄露，用户在存放资料时通常都对自己的资料进行加密，以保证即使服务器得到用户的资料也不能浏览资料内容。但当用户想从服务器检索或重新取出自己的资料时，由于用户加密，使得服务器不知道用户资料的内容，当服务器得到用户的检索词时，无法确定将哪份资料提供给用户。因此，为了检索方便，在存储的同时用户通常会赋予服务器一个“能力”，使服务器能查询用户加密的文件但不能浏览用户文件内容。通过该“能力”，服务器可以在用户存储的文件或数据中找到用户要检索的材料但不能趁机浏览资料内容。此时，用户资料得到了安全保障。在数据库安全领域内，通过关键词对加密文件进行的搜索行为通常被称为“加密数据中的关键词搜索”。

文献[1-4]对“加密数据中的关键词搜索”问题进行研究，但给出的方法不能实现同时搜索多个关键词。例如，用户想从服务器中检索关键词“小王并且工作”，其实质是用逻辑连接词连接多个关键词、在加密数据中的安全搜索问题，目的是找到来自小王的工作文件。文献[5]对该问题做了一定研究，但没有给出有效的解决方案。一些现有解决方法存在较多缺陷，例如，空间复杂度过大或采用非标准模型分析算法而降低了安全性等。鉴于此，本文提出一种可以安全地对保存在不安全服务器中的资料进行连接关键词检索的算法。

2 加密数据中连接关键词的安全搜索算法

2.1 问题定义和模型建立

每个加密文件对应一些关键词，用户把它们存储到一个不可信任的服务器上，希望在服务器的帮助下，通过多个关键词搜索到这些文件中用户需要的文件。在此过程中必须确保服务器不能浏览文件和关键词的内容，且不知道需要搜索哪个文件或关键词组合对应哪个文件。其关键是用户定义一个安全的“能力”给服务器。

为构建一个健全的算法体系，不失一般性地，本文根据文献[4-5]做如下假设：(1)假设每个文件对应 m 个关键词领域，例如在邮件问题中，一般可以分为来源、目的地、日期和分类4个关键词领域。(2)在2个不同的关键词领域中不存在相同的关键词。例如，关键词“来源于小王”属于“来源”关键词领域，而不属于“到小王”领域，因为后者属于“目的地”关键词领域。(3)每个关键词领域由其对应的文件决定。本文假设文件数是 n，每个文件对应一个有 m 个向量的关键词。

在上述假设下，多个连接关键词在加密文件中的搜索方法 CKSE(Conjunctive Keywords Security Encrypt)一般包含以下 5 个步骤：(1)参数生成 $Param\left(1^k\right)$；(2)关键词生成 $KeyGen(\rho)$；(3)加密 $Enc(\rho,K,D_i)$；(4)能力生成；(5)验证。本文算法主要对第(4)步进行改进。

基金项目：广西自然科学基金资助项目(桂科自0899018)；广西教育厅科研基金资助项目(200808MS062)

作者简介：刘星毅(1972－)，男，副教授、硕士，主研方向：计算机网络，数据库技术

收稿日期：2009-05-23 **E-mail：**qznc@163.com

2.2 多连接关键词的安全搜索算法

如何定义用户对服务器赋予的“能力”是 CKSE 中最关键的问题[4]。在本算法中，为了降低算法复杂度，把“能力”的定义分成 2 个部分，即初始化部分和查询部分。在初始化部分，先把复杂度高的初始化流程通过线下完成，可以在无网络下完成，然后把结果通过宽带网络上传，不占用在线查询的时间。查询部分的“能力”定义必须在线完成。根据文献[5]，“能力”初始化部分占整个“能力”部分(即初始化部分和查询部分)的 60%左右，却占用了算法至少 95%以上的运行时间[4]。且它们与查询部分“能力”的定义无关，即线下部分对“能力”的定义对算法最终结果没有影响。因此，在线下事先定义部分“能力”是可行的，并能有效提高算法效率。

初始化部分描述如下：加密好的文件数据被储存在服务器中，一些与查询部分独立的“能力”被事先放在此处。该部分的复杂度很大，但由于与下一部分相独立，可以线下完成，因此能在一定程度上有效降低算法复杂度。

查询部分描述如下：一个常数长度的连接查询的“能力”被送到服务器中，该部分必须在线完成。

图 1 描述了本文算法中用户赋予服务器“能力”的过程。一个不可信任的服务器通常具有以下特点：(1)高的存储能力，至少能存储用户的所有“能力”。例如，现实中的邮件服务器具有极大的存储能力。(2)可信赖的网络连接，此网络连接必须安全，否则在传输过程中数据就不安全。例如，邮件服务器的网络连接通常很安全。(3)具有随时传输服务的能力。用户要事先把部分“能力”和所有加密文件送到服务器中，服务器必须随时提供此要求。现实中，通常可以通过邮件服务器 24 h 不间断地传输数据。用户在平时通过个人电脑把自己的加密文件和部分与查询独立的“能力”送到不可信任的服务器上储存，此时，用户可以使用宽带宽的网络加速传输。而服务器接收到这些信息后，把接收到的“能力”和所有接收到的加密文件放在一起，当用户查询时，方便及时地将其传送给用户。当用户使用小带宽网络(如手机上网)时，要向服务器传输一个加密查询，服务器结合原来得到的“能力”和当前从用户得到的“能力”作出判断，立即根据查询送一个结果给用户。此时，服务器除了知道有一个查询和应答了一个查询外，不知道用户的任何信息，即实现了安全多连接关键词查询。

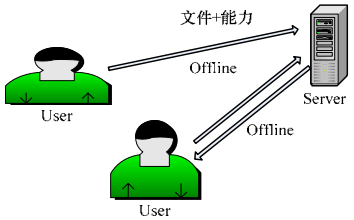

图 1 用户赋予服务器“能力”的过程

对本文算法具体阐述如下：

(1)系统参数算法

在本算法中，系统自动生成参数 $\rho = Param\left(1^k\right)$ 并返回参数 ρ。ρ 与服务器无关，只有用户知道，且即使服务器知道，对安全仍然没有影响。$\rho=\left(G, g, f(\cdot,\cdot), h(\cdot)\right)$，其中，$g$ 是 G 的一个生成算子；$f:\{0,1\}^k\times\{0,1\}^*\to Z_q^*$ 是一个产生关键词的函数；h 为哈希函数，被当作一个随机的预测，即一个黑盒子。安全参数 α 模糊地表示 G, f 和 h 的选择。

(2)关键生成算法

用户先根据上述系统参数算法得到的系统参数 ρ，得到 CKSE 算法第(2)步中的 $G_1=<g_1>$和 $G_2=<g_2>$。服务器可以知道 G_1，但 G_2 只有用户知道，然后返回一个加密的 $\alpha\in\{0,1\}^k$，该加密的关键词只有用户知道。

(3)加密文本函数生成

用户通过上述关键生成算法对各个文件的关键词进行加密后，根据加密后的关键词对文本进行加密，已有很多文本加密方法，本文采用对称加密方法。假设 $h_{i,j}=H\left(W_{i,j}\right)$，$j$=1,2,⋯,$m$，$r_i$ 从 Z_q^* 中均匀随机地获取，则输出结果为

$$Enc(\rho, K, D_i)=\left(e(\alpha, g_2^{r_i}), g_2^{r_i}, (h_{i,1})^{r_i}, \cdots, (h_{i,m})^{r_i}\right)$$

其中，$D_i=\left(W_{i,1}, W_{i,2}, \cdots, W_{i,m}\right)$。

(4)“能力”生成

$Cap=GenCap\left(\rho, K, j_1, j_2, \cdots, j_t, W_{j_1}, W_{j_2}, \cdots, W_{j_t}\right)$ 是本算法的核心，。“能力”由 2 个部分组成：(1)实现用户线下输入的加密文件和部分“能力”；(2)用户查询时给出的“能力”，统一记为 $Cap=(\alpha\prod_{t=1}^{l}(H(W_{j_t}))^s, g_2^s)$，假设 s 均匀随机地从集合 Z_q^* 中获取。

(5)验证函数

服务器得到了用户的查询和查询关键词的“能力”(已加密)，服务器结合用户线下存储的“能力”，对服务器里加密的文本计算 $Cap=(Cap_1, Cap_2)$，$Enc(\rho,K,D_i)=(V_i, Enc(\rho,K,D_{i,0}), Enc(\rho,K,D_{i,1}),\cdots,Enc(\rho,K,D_{i,m}))$。如果得到 $e(Enc(\rho,K,D_{i,0}),Cap_1)/e(\prod_{t=1}^{l}Enc(\rho,K,D_{i,j_t}),Cap_2)=V_i$，则返回“真”，否则输出“假”。

2.3 算法安全性分析

“能力”的定义使服务器可以把用户的文件分成 2 个部分，一部分满足“能力”函数，另一部分不满足“能力”函数。如果服务器从加密的文件和“能力”中没有学到其他信息，则称该连接关键词查询是安全的，否则是不安全的，本文对安全性定义采用文献[4]的 3 个安全性定义原则，即 security Game ICC(Indistinguishability of Ciphertext from Ciphertext), security Game ICR(Indistinguishability of Ciphertexts from Random), and security Game ICLR (Indistinguishability of Ciphertexts from Limited Random)。根据文献[4-5]的结论可知，如果一个对手(想取得用户资料的人或服务器)在没有得到用户授予“能力”的情况下，想获取用户资料是不可能的，因为其条件一定不可能同时满足上述 3 个安全性原则。由此可知，本文设计的算法是安全的。

2.4 实例分析

表 1 描述了一个经典的机密数据中的关键词搜索实例，它是有关邮件的例子，从第 2 行起的每行代表一个事件。例如，*Alice* 在 08/01/2007 写了一封“*TopSecret*”邮件给 *Bob*。假设用户要查找关键词“*Alice*”和“*Bob*”，下文将解释服务器如何给用户答案。

表 1 机密数据中的关键词搜索实例

来源(*From*)	目的地(*To*)	日期(*Date*)	分类(*Subject*)
Alice	*Bob*	08/01/2007	*TopSecret*
Alice	*Charles*	09/07/2007	*Secret*
…	…	…	…
Dave	*Alice*	11/25/2007	*Unclassified*

上述问题的邮件领域是 {*From*,*To*}，在此邮件领域下的关键词领域为 $\{Alice_{From}, Bob_{To}\}$，其处理过程如下：

(1)用户通过系统参数得到一个 ρ，这只有用户知道，对整个过程没有影响，其出现只是为了得到一个加密的关键词。

(下转第 158 页)