基于关键词的加密云数据模糊搜索策略研究

方忠进[1,2,3] 周 舒[1,3] 夏志华[1,3]
(南京信息工程大学计算机与软件学院 南京 210044)[1] (南京信息工程大学滨江学院 南京 210044)[2]
(南京信息工程大学江苏省网络监控工程中心 南京 210044)[3]

摘 要 随着个人和企业用户产生的数据量越来越多,云存储的价格便宜、存储空间使用灵活等优势也越来越明显。随着大量的数据外包到云服务器端,用户一般采用加密方法实现对敏感数据的保护和使用限制,这使得传统的基于明文的搜索方案不再适用。如何在保护隐私的基础上实现高效的数据文件搜索是首要考虑的问题。在已有的加密数据搜索方案的基础上,分析中文所特有的模糊音和多义的特点,利用中文和英文实现关键词的同义词构建,并建立关键词的模糊音词和同义词集合,提出了基于关键词的加密云数据模糊搜索方案,实现了对中文模糊音和同义关键词的搜索并利用伪随机函数对私钥进行保护。最后通过安全性分析和系统实验验证了该方案具有较高的安全性、良好的实用性和较高的搜索成功率。

关键词 云存储,模糊搜索,关键词,同义词,加密搜索
中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.3.028

Research on Fuzzy Search over Encrypted Cloud Data Based on Keywords

FANG Zhong-jin[1,2,3] ZHOU Shu[1,3] XIA Zhi-hua[1,3]
(School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)[1]
(Binjiang College, Nanjing University of Information Science & Technology, Nanjing 210044, China)[2]
(Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing 210044, China)[3]

Abstract With the increasing amount of data generated by individuals and business users, the advantages of cloud storage such as lower price and flexible use of storage space are obvious. As a large amount of data are outsourced to the cloud server, encryption methods are used to achieve protection and limitations of sensitive data generally. It makes traditional search scheme based on plaintext no longer applicable. How to achieve efficient file search on the basis of privacy protection is the primary consideration. On the basis of the existing encrypted data search schemes, Chinese characteristics of homophone and polysemy were analyzed, and synonyms of keyword were constructed using Chinese and English, and the sets were constructed respectively. Fuzzy search scheme over encrypted cloud data based on keywords was proposed. Search of Chinese fuzzy pinyin and synonyms is achieved, and the private key is protected by pseudo-random function. High security, good practicability and high searching success rate of the system were verified by security analysis and system experiment.

Keywords Cloud storage, Fuzzy search, Keywords, Synonyms, Searchable encryption

1 引言

随着信息技术的发展和普及,普通用户和企业存储在本地的数据文件数量越来越多,对本地存储造成的压力也越来越大,一旦本地硬件出现故障或者严重损害将极大地影响到用户和企业对于数据的使用,甚至将永远丢失重要的数据。因此,云存储服务以使用方便、节约成本等优势受到越来越多用户的欢迎。但是云存储服务的使用也面临着一些问题。例如,企业的一些涉及商业机密的数据必须能够保护隐私以及防止被非法使用。因此,这些数据一般都在本地加密后再外包给云存储服务器。这给数据的使用带来了很大的麻烦,因为用户不可能将所有的数据重新下载到本地后再解密使用,这也受到网络带宽和本地存储容量的限制。因此研究设计支持隐私保护和加密搜索的云数据服务是一项具有重要意义和实用价值的研究课题。

国内外研究人员在基于公钥加密的加密数据的关键词搜索方面已取得了很多研究成果,如 Song、Wagner、Perrig[1] 等人于 2000 年提出了基于对称加密的密文关键词搜索;Boneh、Crescenzo、Persiano[4] 等人在 2004 年提出基于公钥加密的密文关键词搜索。近年来取得的成果也很多,如 Jin Li、Qian

到稿日期:2014-08-22 返修日期:2014-09-22 本文受国家自然科学基金(61173141),江苏省普通高校研究生科研创新计划项目(CXZZ12_0512),江苏省网络监控工程中心开放基金课题资助项目(KJR1106),江苏高校优势学科建设工程项目资助。
方忠进(1979-),男,博士生,讲师,主要研究方向为网络信息安全、云安全,E-mail:nuist@163.com;**周 舒**(1984-),女,硕士,助理研究员,主要研究方向为信息安全;**夏志华**(1983-),男,博士,讲师,主要研究方向为信息安全、隐写分析。

Wang 和 Cong Wang[11]等人提出了使用编辑距离量化关键词之间的相似性并形成一种基于关键词模糊搜索的新技术；Bijit Hore，Sharad Mehrotra[13]等人在 2012 年提出了基于范围分组的范围搜索方法。

从上述分析可知，加密数据搜索研究已经取得了比较丰硕的成果，但是在语义搜索和容错方面仍然存在许多亟待解决的问题：以上基于关键词的搜索方法都是针对关键词的精确或模糊匹配方面进行搜索的，同时并不完全适用于中文环境。中文字除了字形外还包含有拼音和含义两部分，拼音由声母和韵母组成。因此，在正常情况下一个中文关键词有很多同义词、近义词及拼音相似的词等。目前国内外在语义与语音相近的中文关键词搜索方面的研究较少。因此，本文提出了基于关键词的加密云数据模糊搜索策略，探索在云存储环境下中文模糊音和同义关键词的模糊搜索的执行效率和方法。

2 问题描述

2.1 系统与威胁模型

云环境下的数据存储服务的体系结构如图 1 所示。图中显示该体系结构中包含了 3 个主要实体：数据拥有者、用户和云服务提供商。其中，数据拥有者可以是个人或者企业用户，他们将拥有的数据文件集合 $C=(F_1,F_2,\cdots,F_n)$ 存储在云服务器上。与文件集合 C 相关的不同关键词集合预先进行定义并且表示为 $W=\{w_1,w_2,\cdots,w_p\}$。为了保证敏感数据不被未授权的人使用，数据集合 C 在外包至云服务器之前需要进行加密处理。由于中文存在大量的近音和同义词，为了提高云数据利用效率和检索成功率，该体系结构需要提供加密数据的模糊音和同义词的模糊搜索功能。为了实现该服务，数据拥有者需要将搜索请求生成私钥 sk 并分发给其他授权用户，如团队成员或企业员工。当私钥分配完成后，对于任一个输入的关键词 w，为了能够安全地搜索出相关的文件集合，被授权用户利用私钥 sk 和单向生成函数将需要查询的关键词转换成一个搜索请求（以下称为陷门），并且提交给云服务器。云服务器在未解密数据的情况下执行搜索并且将搜索到的与关键词 w 或 w 的模糊音或同义词相关的目标文件集合（记为 FID_w）发送给数据查询者。

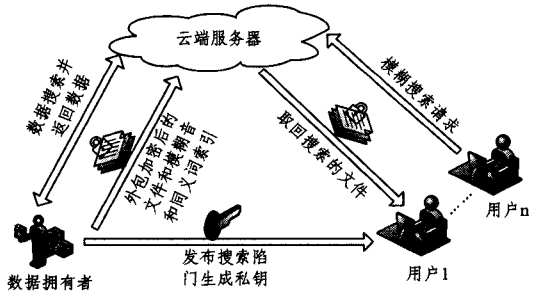

图 1 模糊搜索策略系统框架

本文认为云数据服务体系结构中所涉及的云服务器是诚实但是具有好奇心的，它可以正确执行指定的协议规范，但是又会通过用户的输入来推断和分析相关信息。因此，在设计同义关键词搜索方案时，我们仍然遵循传统对称加密中所涉及的安全定义。除了搜索结果和搜索模型以外，不应该泄露与存储的文件及索引相关的其他任何内容。

2.2 设计目标

为了能够对上述模型中的加密云数据实现安全且高效的同义关键词搜索，本文需要实现以下目标：1）模糊关键词搜索功能，即探索不同机制设计外包云数据的高效且正确的模糊关键词搜索策略；2）保证安全，即防止云服务器在搜索过程中学习到与数据文件或关键词相关的知识；3）保证效率，即在尽可能小的存储、通信和计算资源占用情况下实现上述目标。

2.3 预备知识

C：外包的文件集合，表示为 n 个数据文件的集合，即 $C=(F_1,F_2,\cdots,F_n)$。

W：提取自文件集合 C 的不同关键词集合，表示为 p 个词的集合，即 $W=(w_1,w_2,\cdots,w_p)$。

I：为可隐私保护的模糊关键词搜索而建立的索引。

Tw：陷门，即用户输入搜索关键词 w 后由单向函数所生成的搜索请求。

FID_{wi}：文件集合 C 中包含有关键词 w_i 或其近音或同义的文件 ID 集合。

$f(key,\cdot),g(key,\cdot)$：伪随机函数（$PRF$），定义为：$\{0,1\}^*\times key\rightarrow\{0,1\}^l$。

$Enc(key,\cdot),Dec(key,\cdot)$：基于语义安全的对称密钥加密/解密函数。

编辑距离：编辑距离是字符串的相似性的一种描述方法。对于两个单词 w_1 和 w_2，编辑距离 $ed(w_1,w_2)$ 表示在二者实现转换所需的最小数量的操作，这些操作可以是增加、修改和删除字符的操作。对于给定的单词 w 和整数 d，用 $S_{w,d}$ 表示与其相似的单词 w'，满足 $ed(w,w')\leqslant d$。

$SP_{w,d}$：关键词 w 的模糊拼音所对应的关键词集合。对于给定的中文关键词 w 和整数 d，其拼音 PY_w 对应的模糊音集合 $SP_{w,d}=\{w_1',w_2',\cdots\}$，满足与关键词 w 的拼音的编辑距离小于 d 的所有类似拼音的关键词集合，表示为 $ed(PY_w,PY_{w'})\leqslant d$。即对任意 $w_i'\in SP_{w,d}$，有 w_i' 的拼音 $PY_{w_i'}\in S_{PY_{w'd}}$。

SY_w：与关键词 w 同义的关键词集合。在同一种语言中描述同一事物的不同关键词集合 $SY_w=(w_1',w_2',\cdots)$，转换为另一种语言时，w_i' 一般会对应相同的关键词 w_e，则称集合 $SY_w=(w_1',w_2',\cdots)$ 为关键词 w 的同义词集合，其中，$syn(w)=syn(w_i')$，$syn()$ 是同义转化函数。本文使用中文和英文实现同义转换。

模糊关键词搜索：给定由 n 个加密后的数据文件所构成的集合 $C=(F_1,F_2,\cdots,F_n)$，预定义的不同中文关键词集合 $W=(w_1,w_2,\cdots,w_p)$，输入的搜索关键词 w 和 d，执行同义关键词搜索后将返回文件 ID 集合 $\{FID_{w_i}\}$，其中 $w_i=w,w_i\in SY_w$ 或者 $w_i\in SP_{w,d}$。

3 系统框架

3.1 系统框架图

根据以上目标分析，系统总体框架设计如图 1 所示。

3.2 设计概述

模糊搜索的目标是根据不同用户输入的关键词，尽可能返回所有与该关键词相似（包括同义和模糊音）的结果。然而，这种基于关键词同义和模糊音词的模糊搜索对于云端数据的匹配具有很大的挑战性。任意两个中文单词在明文状态

下很容易获取其所具有的模糊音词或者同义词，但是在经过单向加密函数加密（如伪随机函数或其他加密算法）后就很难发现其中相似的规律。传统的加密搜索策略是通过用户提交的搜索陷门和可搜索加密索引之间的相等比较进行搜索的，但在这里的模糊搜索中是无法使用的。

为了解决这个问题，本文提出了分步方案来降低与云端加密数据模糊匹配的难度。第一步，数据拥有者在客户端构建模糊关键词集合，该集合主要包含 3 部分：关键词、关键词的拼音及模糊音、关键词对应的英文单词，相应的索引信息就包含 3 张表：中文关键词及文件 ID 表、中英文关键词对照表和拼音及模糊音对照表。第二步，基于该模糊关键词集合，设计安全高效的模糊搜索方法，在后续章节中将进行详细阐述。

对于外包于云端的数据，除了安全问题，用户在使用中最为关心的就是操作的效率。因此，本文采用对称加密作为可搜索加密框架。

4 方案实现

本文在设计中文模糊关键词搜索系统框架时首先考虑模糊关键词集的建立，然后再分析如何生成搜索请求，最后是如何实现安全且高效的加密数据搜索。

4.1 建立模糊关键词集

建立关键词集合是进行高效模糊搜索的前提。给出关键词 w 和相似约束条件 d，则要生成 $SP_{w,d}$ 和 SY_w，其中 $w' \in SP_{w,d}$ 满足 $ed(PY_w, PY_{w'}) \leqslant d$；$w' \in SY_w$ 满足 $syn(w) = syn(w')$。具体实现如下：

1）模糊拼音关键词集合建立

中文字的拼音主要由声母和韵母组成，且声母与韵母的组合符合特定规律，以下列出了具体集合。

声母集合：{b,p,m,f,d,t,n,l,g,k,h,j,q,x,zh,ch,sh,r,z,c,s,y,w}

单韵母集合：{a,o,e,i,u,ü}

复韵母集合：{ai,ei,ui,ao,ou,iu,ie,üe,er}

前鼻韵母：{an,en,in,un,ün}

后鼻韵母：{ang,eng,ing,ong}

中文拼音不会出现类似英文单词那样的任意组合的情况，如声母是 b 时，韵母所构成的集合中只能包含固定数量：{a,o,i,u,ai,ei,ao,ie,an,en,in,ang,eng,ing,ian,iao}。而且在中文拼音拼写中最容易出错的就是模糊音，如声母中的平舌音和翘舌音、l 和 n 等，韵母中的前鼻音和后鼻音、ian 和 iang 以及 uan 和 uang。因此，建立模糊音关键词集合最简单的方法就是可以枚举出可能的拼音组合，进而找出与这些组合相同拼音的关键词集合。举例如下：

假设用户给定的 $d=2$，输入关键词 w 的拼音是 lin，则根据拼音构成规则生成对应的模糊音关键词组合 $SP_{w,d} = \{w_1', w_2', \cdots\}$，其中 w_i' 的拼音应该包含于集合{lin,nin,ling,ning,*in}。

2）同义词集合建立

中文中存在很多含义相同或相近的词，而这些在近义词词典中却无法体现出来，如“计算机”、“电脑”和“微机”。这 3 个中文词含义一致，用户在执行云端数据搜索时，与这 3 个词相关的文件 ID 都应该返回给用户，但如何去实现类似的同义词比较却一直没有很好的办法。通过对比中文和英文在描述同一个事物时的表达方式的类似性，提出了利用语言间的表达差异来实现同义词转换的方法，如上述 3 个词的英文单词均是“computer”，这样如果中文关键词 w_i 对应的英文翻译一致，那么就说明这些词是同义词。执行过程如下：

假设用户输入的关键词 w，执行函数 $syn(w)$，将 w 翻译为英文单词 w_e，然后在中英文对照表中查找英文翻译为 w_e 的中文关键词并返回关键词集合 SY_w。

生成相应的模糊音和同义词集合后，调用加密函数 $Enc(key, \cdot)$ 将得到的 $SP_{w,d}$ 和 SY_w 进行加密，并连同加密后的文件一起发送到云端进行保存。

4.2 生成搜索请求

当用户输入关键词 w 后，系统执行模糊搜索，返回相应的文件 ID 集合 $\{FID_{w_i}\}$，其中 $w_i = w, w_i \in SY_w$ 或者 $w_i \in SP_{w,d}$。搜索请求的生成过程与关键词索引的生成方式类似，即根据输入的 w 和 d，调用模糊拼音和同义词生成函数得到模糊拼音关键词集合 $SP_{w,d}$ 和同义词集合 SY_w，将 $w, SP_{w,d}$ 和 SY_w 加密后生成搜索陷门，提交给云服务器，即完成搜索请求生成工作。

4.3 模糊搜索方案

在云服务系统中，为了避免云端获取敏感信息，部分工作需要在客户端执行，如搜索索引的建立、陷门的生成等；而在大量数据中执行搜索是非常消耗资源的工作，这些则交由云服务器去完成。基于关键词的加密云数据模糊搜索方案执行流程如下。

系统预处理阶段：

1）数据拥有者随机选择两个数 a 和 b 作为私钥 sk。

2）构建索引。索引 $I_1 = \{f(a, w_i'), Enc(sk_{w_i'}, FID_{w_i})\}, w_i' \in SP_{w,d}$，$I_2 = \{f(a, w_i'), Enc(sk_{w_i'}, FID_{w_i})\}, w_i' \in SY_w$，其中 $1 \leqslant i \leqslant p$，密钥 $sk_{w_i'} = g(b, w_i')$。

3）将索引表 I_1 和 I_2 以及加密数据文件外包到云服务器存储。

搜索阶段：

1）用户输入 sk、w 和 d，客户端系统生成 $SP_{w,d}$ 和 SY_w，同时生成陷门 $T_{1w'} = (f(a, w'), g(b, w')), w' \in SP_{w,d}$；$T_{2w'} = (f(a, w'), g(b, w')), w' \in SY_w$。将陷门集 $T_{1w'}$ 和 $T_{2w'}$ 发送到云端。

2）云服务器将接收到的陷门 $T_{1w'} = f(a, w'), w' \in SP_{w,d}$；$T_{2w'} = f(a, w'), w' \in SY_w$ 分别与索引 $I1$ 和 $I2$ 进行比较得到符合条件的文件 ID 集合 $\{FID_{w_i}\}$，其中 $w_i = w, w_i \in SY_w$ 或者 $w_i \in SP_{w,d}$，并将结果发送给用户端。

3）用户端使用相应的 $g(b, w')$ 解密文件 ID，并取回需要的文件，调用 $Dec(key, \cdot)$ 解密后使用。

5 安全性分析

本文所设计的加密搜索方案中，当用户输入相同的搜索请求时，云端总是会返回同样的搜索结果。虽然云服务器看不到底层明文是什么，但在与用户的交互中，它仍然可以建立访问模式和搜索模式。因此系统的安全方面应确保除了模式和搜索请求以外的内容被泄露。本节将证明本文设计的模糊搜索方案是符合非自适应语义的安全保障要求的。非自适应攻击模型只考虑敌手（如云服务器），该敌手不能选择基于陷门的搜索请求以及以前的搜索结果[16]，这是可以接受的。因为只有拥有授权私钥的用户可以生成搜索陷门。下面介绍在

文献[17]中所用的一些概念来分析模糊搜索方案的安全性。

历史记录(History):用户和云服务器之间的交互,其由文件集合 C 和一组由用户搜索的关键词组成,表示为 $Hq=(C,w_1,w_2,\cdots,w_p)$。

视图(View):根据密钥 K,给出历史记录 Hq,云服务器1)只能看到加密过的历史记录,即视图 $Vk(Hq)$,包括:文件集合 C 的索引 I;查询关键词的陷门 $T_{1w'}$,其中 $w'\in SP_{w,d}$ 和2)$T_{2w'}$,其中 $w'\in SY_w$;加密文件集合 C,3)记为 $\{e_1,\cdots,e_n\}$。

追踪:给定一个历史记录 Hq 和一个加密文件集合 C,3)$Tr(H_q)$ 捕捉到云服务器所学习的精确信息,包括:加密文件的大小($|F_1|,\cdots,|F_n|$);每个搜索的输出结果 $\{FID_{w_i}\}$,其中 $w_i=w,w_i\in SY_w$ 或者 $w_i\in SY_{w,d}$;每个搜索的模式为 Π_q。Π_q 是一个对称矩阵,$\Pi_q[i,j]$存储了两个集合的交集,Π_{q_1} 和 Π_{q_2} 分别是模糊音记录的交集和同义记录的交集:$\Pi_{q_1}=\{T_{w'}\}_{w'\in SP_{w_i,d}}\cap\{T_{w'}\}_{w'\in SP_{w_j,d}}$,$\Pi_{q_2}=\{T_{w'}\}_{w'\in SY_{w_i}}\cap\{T_{w'}\}_{w'\in SY_{w_j}}$。

一般来说,本方案的安全强度体现在:对于具有相同轨迹的两个历史记录,云服务器无法区分它们的视图。换句话说,云服务器无法根据查询中泄漏的信息(即轨迹)提取更多的信息内容,因此,本方案是安全的。本文的模糊搜索方案的安全性结论在以下定理中予以说明。因为在以上内容中已经说明了模糊搜索方案,那么以下结论同样适用于实例化的模糊搜索。

定理 本模糊关键词搜索方案满足非自适应语义安全性。

证明:为了证明语义安全性,我们构造了一个仿真器 S,它具有轨迹 $Tr(H_q)$,可以模拟一个视图 V_q^*,其可以完全模拟云服务器的视图 $VK(H_q)$,对于任意 $q\in N$、任意 H_q 以及随机选择的 K。伪随机函数 $f(key,\cdot)$ 的安全参数 l,$T=\max\{|SP_{w_i,d}|,|SY_{w_i}|\}(w_i\in W)$,$|W|$ 以及 FID_{w_i} 的大小对 S 都是已知的。

对于 $q=0$,S 生成 $V_q^*=\{e_1^*,e_2^*,\cdots,e_n^*,I^*\}$,$e_i^*$ 从 $\{0,1\}^{|F_i|}(1\leqslant i\leqslant n)$ 中随机选取。令 $I^*=(T^*,C^*)$,分别用 $T^*[i]$和 $C^*[i]$来表示 T^* 和 C^* 中第 i 行记录,生成如下:

生成 T^*:对于 $1\leqslant i\leqslant\tau|W|$,$S$ 选取一个随机的 $t_i^*\in\{0,1\}^l$,并使 $T^*[i]=t_i^*$。

生成 C^*:对于 $1\leqslant i\leqslant\tau|W|$,$S$ 选取一个随机的 $c_i^*\in\{0,1\}^{|FID_{w_i}|}$,并使 $C^*[i]=c_i^*$。

由于对称加密的语义安全性,对于敌手而言,不可能出现可以区分 e_i 和 e_i^* 或者 $Enc(sk_{w_i'},FID_{w_i})$ 和 c_i^* 的情况。由于陷门生成函数的伪随机性,同样不存在可以区分 $f(x,w_i')$ 和一个随机串 t_i^* 的情况。所以,$V_K(H_0)$ 和 $V_0{}^*$ 是无法区分的。

对于 $q\geqslant 1$,S 构造集合 $V_q^*=\{e_1^*,e_2^*,\cdots,e_n^*,I^*,\{T_{1w'}\}_{w'\in SP_{w,d}},\{T_{2w'}\}_{w'\in SY_w}\}$,其中 e_i^* 是由 $\{0,1\}^{|F_i|}(1\leqslant i\leqslant n)$ 随机生成的。令 $I^*=(T^*,C^*)$,第 j 次搜索请求的返回结果为 $\{FID_{w_i}\}$,其中 $w_i\in SY_{w_j}$ 或者 $w_i\in SY_{w_{j,d}}$,是与索引匹配的记录对应文件 ID 的集合。我们可以把它改写为 $\bigcup_{k=1}^{\alpha_j}FID_{w_{j,k}}$,每个 $w_{j,k}$ 满足 $w_{j,k}\in SY_{w_j}$ 或者 $w_{j,k}\in SY_{w_{j,d}}$,α_j 表示第 j 次搜索匹配的数目。对于搜索输入 w^1,构造第一个陷门集 $\{T_{w'}^*\}=\{\{T_{1w'}\}_{w'\in SP_{w,d}},\{T_{2w'}\}_{w'\in SY_w}\}$。仿真器 S 执行以下操作:

1)选择 α_1 个随机串 $t_{1,1}^*,\cdots,t_{\alpha,1}^*\in\{0,1\}^l$,并将其设置为 $T^*[i_{1,1}],\cdots,T^*[i_{1,\alpha_1}]$。

2)选择 α_1 个随机串 $\rho_{1,1}^*,\cdots,\rho_{\alpha,1}^*\in\{0,1\}^l$,并设置 $C^*[i_{1,k}]=Enc(\rho_{1,k}^*,FID_{w_{i,k}})$,其中 $1\leqslant k\leqslant\alpha_1$。

3)设置其他的 $\tau-\alpha_1$ 个陷门为随机值对 $(t_{1,k}^*,\rho_{1,k}^*)\in\{0,1\}^l\times\{0,1\}^l$,其中 $\alpha_1\leqslant k\leqslant\tau$。

对于陷门仿真 $w^j(2\leqslant j\leqslant q)$,所有的 $i<j$,假如 $|\Pi_q[i,j]|=0$,陷门仿真将重复和陷门 w^1 相同的操作;否则,用 β_1 表示文件 ID 列表 $\{FID_{w_{j,k}}\}_{1\leqslant k\leqslant\alpha_j}$ 的文件数,这些文件 ID 已经在列表 $\{FID_{w_{i,k}}\}_{1\leqslant k\leqslant\alpha_i,i<j}$ 中匹配成功。接下来,S 执行:

1)从已经存在的陷门集 $\{T_{w'}^*\}_{w'\in\{SP_{w^i,d},SY_{w^i}\},i<j}$ 中选择 β_j 个陷门,其与 β_j 个文件 ID 列表 $\{FID_{w_{j,k}}\}_{1\leqslant k\leqslant\alpha_j}\cap(\bigcup_{i=1}^{j-1}\{FID_{w_{j,k}}\}_{1\leqslant k\leqslant\alpha_i})$ 相对应,并将它们分配给陷门仿真 w^j。

2)假如 $\alpha_j>\beta_j$,S 将用仿真陷门 w^1 相同的过程在索引 I^* 建立 $\alpha_j-\beta_j$ 个记录。

3)S 进一步检查 $\Pi_q[i,j]$,其中 $i<j$,从已经存在的陷门集 $\{T_{w'}^*\}_{w'\in\{SP_{w^i,d},SY_{w^i}\},i<j}$ 中查找 γ_j 个陷门(在索引 I^* 中没有匹配项),并将它们分配给当前的陷门仿真 w^j。

4)从 $\{0,1\}^l\times\{0,1\}^l$ 中随机地选择数值对为剩下的 $\tau-\alpha_j-\gamma_j$ 个陷门赋值。

所构建的视图的正确性很容易通过在索引中搜索构建的陷门得以证明。本方案中不存在攻击者能够区分 $V_K(H_q)$ 和 $V_q{}^*$ 的情况;而且,模拟的密文采用的是对称加密方案,由其语义安全性决定密文是不可区分的。基于伪随机函数的特性,索引和陷门也是不可区分的。因此,证明定理是正确的。

6 实验分析

6.1 时间与空间消耗

本文提出的中文关键词的模糊搜索方案是在已有的搜索方案基础之上引入了中文关键词的同义词和模糊音词,因此在系统预处理阶段和搜索阶段,由于增加了关键词同义词和模糊音词而需要额外消耗时间和空间来处理这些内容,但在时间复杂度和空间消耗上与原有的方案处于同一级别。预处理阶段索引构建耗时的时间复杂度只与关键词的数量有关,即 $O(\tau|W|)$;索引的大小同样也只与关键词的数量有关,即 $O(\tau|W|)$。在搜索时,由于云服务器端系统对于多线程的支持,可以同时实现对于同义词、模糊音词的检索,因此并不会增加太多的时间消耗,搜索的时间复杂度为 $O(\tau|W|)$。

6.2 实验对比

本文以亚马逊提供的免费云平台作为实验平台,使用期刊杂志文献作为搜索对象进行实验和验证对象,对本方案的有效性和实用性进行验证。构建索引时,随着关键词数量的增多,对于 CPU 和内存的占用逐渐增大,所消耗的时间也相应增加。

构建索引的时间消耗如图2所示。

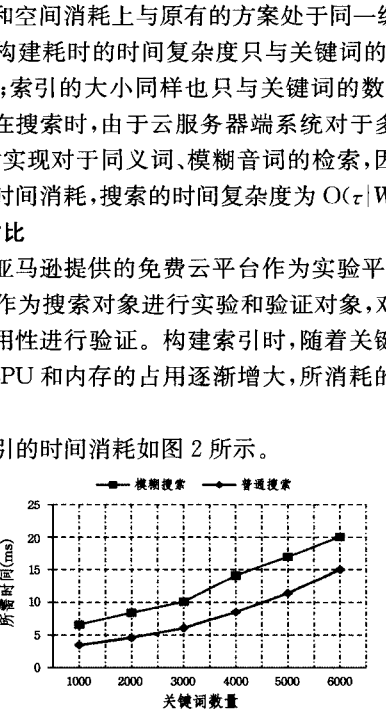

图2 索引构建时间消耗图

(下转第173页)

Workshop on Quality of Protection. 2007:15-20

[12] Tsai H Y, Huang Y L, Wagner D. A graph approach to quantitative analysis of control flow obfuscating[J]. IEEE Transactions on Information Forensics and Security, 2009, 4(2): 257-267

[13] Huang Y L, Tsai H Y. A framework for quantitative evaluation of parallel control-flow obfuscation[J]. Computers & Security, 2012, 31(8): 886-896

[14] 付剑晶,王珂. 软件迷惑变换的鲁棒性量化评价[J]. 软件学报, 2013, 24(4): 730-748

[15] Ogiso T, Sakabe Y, Soshi M, et al. Software obfuscation on a theoretical basis and its implementation[J]. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2003, 86(1): 176-186

[16] Ceccato M, Di P M, Nagra J, et al. Towards experimental evaluation of code obfuscation techniques[C]// Proceedings of the 4th ACM Workshop on Quality of Protection. Alexandria, VA, USA, 2008: 39-46

[17] Ceccato M, Di Penta M, Nagra J, et al. Towards experimental evaluation of code obfuscation techniques[C]// Proceedings of the 4th ACM Workshop on Quality of Protection. 2008: 39-46

[18] 赵玉洁,汤战勇,王妮,等. 代码混淆有效性评估[J]. 软件学报, 2012: 700-711

[19] Satty T L. The Analytic Hierarchy Process [M]. New York: McGraw-Hill, 1980

(上接第 139 页)

关键词查询的时间消耗如图 3 所示。

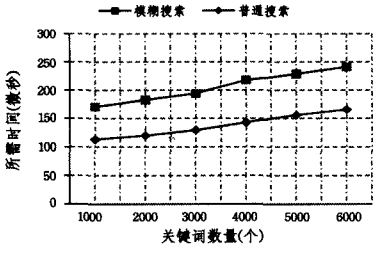

图 3 关键词查询时间消耗图

关键词查询的成功率如图 4 所示。

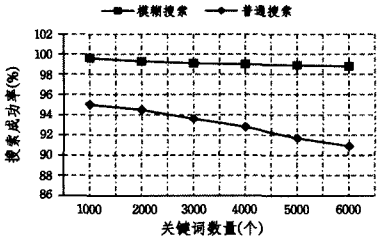

图 4 搜索成功率对比图

从以上实验结果可以看出,本文所提出的方案通过适当增加存储来实现对关键词的同义词和模糊音词的存储,虽然系统预处理和搜索时间有所增加,但却大大提高了搜索的成功率。

结束语 本文根据云存储环境下用户中文搜索的实际需求,给出了一种基于关键词的加密云数据模糊搜索策略。通过在方案中构建模糊音与同义词集,很好地解决了中文环境下搜索容易出现的输入的文字和用户想找的词存在的模糊音和同义问题,同时通过使用伪随机函数有效避免查询过程中的信息泄露问题。因此,本方案具有较高的安全性、良好的实用性和较高的搜索成功率。

参 考 文 献

[1] Song D, Wagner D, Perrig A. Practical Techniques for Searches on Encrypted Data[C]// Proceedings of IEEE Symposium on Security and Privacy. 2000: 44-55

[2] Goh E J. Secure Indexes. Cryptology ePrint Archive, Report 2003/216[OL]. http://eprint.iacr.org/, 2003

[3] Chow R, Golle P, Jakobsson M, et al. Controlling Data in the Cloud: Outsourcing Computation without Outsourcing Control [C]// Proceedings of ACM Workshop on Cloud Computing Security (CCSW'09). New York, 2009: 85-90

[4] Boneh D, Crescenzo G D, Ostrovsky R, et al. Public Key Encryption with Keyword Search [C]// Proceedings of International Conference on Theory and Applications of Cryptographic Techniques (Eurocrypt'04). Interlaken, Switzerland, Vol. 3027, 2004: 506-522

[5] 李春艳,张学杰. 基于基准测试的高性能计算云研究[J]. 计算机科学, 2013, 40(12): 23-30

[6] Li J, Wang Q, Wang C, et al. Fuzzy keyword search over encrypted data in cloud computing[C]// Proc. of IEEE INFOCOM, Mini-Conference. 2010: 441-445

[7] Hwang Y H, Lee P J. Public Key Encryption with Conjunctive Keyword Search and Its Extension to a Multi-user System[C]// Proceedings of International Conference on Pairing-Based Cryptography (Pairing'07). 2007: 2-22

[8] Ballard L, Kamara S, Monrose F. Achieving Efficient Conjunctive Keyword Searches over Encrypted Data[C]// Proceedings of International Conference on Information and Communications Security (ICICS'05). Vol. 3783, 2005: 414-426

[9] Boneh D, Waters B. Conjunctive, Subset, and Range Queries on Encrypted Data[C]// Proceedings of TCC'07. 2007: 535-554

[10] 林闯,苏文博,孟坤,等. 云计算安全:架构、机制与模型评价[J]. 计算机学报, 2013, 36(9): 1765-1784

[11] Li J, Wang Q, Wang C, et al. Fuzzy Keyword Search over Encrypted Data in Cloud Computing[C]// IEEE INFOCOM 2010 Mini-Conference. San Diego, CA, March 2010

[12] Cao N, Wang C, Li M, et al. Privacy-Preserving Multi-keyword Ranked Search over Encrypted Cloud Data[C]// Proceedings of IEEE INFOCOM 2011. 2011: 829-837

[13] Hore B, Mehrotra S, Canim M, et al. Secure multidimensional range queries over outsourced data[J]. The VLDB Journal, 2012 (21): 333-358

[14] 冯登国,张敏,张妍,等. 云计算安全研究[J]. 软件学报, 2011, 22 (1): 71-83

[15] Wang Peng, Ravishankar C V. Secure and Efficient Range Queries on Outsourced Databases Using R-trees[C]// ICDE Conference. 2013: 314-325

[16] Curtmola R, Garay J A, Kamara S, et al. Searchable symmetric encryption: improved definitions and efficient constructions[C]// Proc. of ACM CCS. 2006: 79-88

[17] Golle P, Staddon J, Waters B. Secure Conjunctive Keyword Search over Encrypted Data[C]// Proceedings of Applied Cryptography and Network Security Conference (ACNS'04). 2004: 31-45