软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(4):880−895 [doi: 10.13328/j.cnki.jos.004554]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

可搜索加密机制研究与进展*

沈志荣[1], 薛 巍[1,2], 舒继武[1,2]

[1](清华大学 计算机科学与技术系,北京 100084)

[2](信息科学与技术国家实验室(清华大学),北京 100084)

通讯作者: 舒继武, E-mail: shujw@tsinghua.edu.cn

摘 要: 随着云计算的迅速发展,用户开始将数据迁移到云端服务器,以此避免繁琐的本地数据管理并获得更加便捷的服务.为了保证数据安全和用户隐私,数据一般是以密文存储在云端服务器中,但是用户将会遇到如何在密文上进行查找的难题.可搜索加密(searchable encryption,简称 SE)是近年来发展的一种支持用户在密文上进行关键字查找的密码学原语,它能够为用户节省大量的网络和计算开销,并充分利用云端服务器庞大的计算资源进行密文上的关键字查找.介绍了 SE 机制的研究背景和目前的研究进展,对比阐述了基于对称密码学和基于公钥密码学而构造的 SE 机制的不同特点,分析了 SE 机制在支持单词搜索、连接关键字搜索和复杂逻辑结构搜索语句的研究进展.最后阐述了其所适用的典型应用场景,并讨论了 SE 机制未来可能的发展趋势.

关键词: 可搜索加密;数据安全;隐私;密码学;云计算;云存储

中图法分类号: TP309 **文献标识码**: A

中文引用格式: 沈志荣,薛巍,舒继武.可搜索加密机制研究与进展.软件学报,2014,25(4):880−895. http://www.jos.org.cn/1000-9825/4554.htm

英文引用格式: Shen ZR, Xue W, Shu JW. Survey on the research and development of searchable encryption schemes. Ruan Jian Xue Bao/Journal of Software, 2014,25(4):880−895 (in Chinese). http://www.jos.org.cn/1000-9825/4554.htm

Survey on the Research and Development of Searchable Encryption Schemes

SHEN Zhi-Rong[1], XUE Wei[1,2], SHU Ji-Wu[1,2]

[1](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

[2](Tsinghua National Laboratory for Information Science and Technology (Tsinghua University), Beijing 100084, China)

Corresponding author: SHU Ji-Wu, E-mail: shujw@tsinghua.edu.cn

Abstract: With the rapid development of cloud computing, users are beginning to move their data to the cloud servers in order to avoid troublesome data management at local machines and enjoy convenient service. To protect data security and user privacy, data are usually stored in encrypted form in the cloud, but it activates the inconvenience when the user tries to retrieve the files containing some interested keywords. Searchable encryption (SE) is a recently developed cryptographic primitive that supports keyword search over encrypted data, which not only saves huge network bandwidth and computation capacity for users, but also migrates the cumbersome search operation to the cloud server to utilize its vast computational resources. This paper first introduces the research background and the current development of SE schemes and compares the different features between symmetric key cryptography based SE schemes and public key cryptography based SE schemes. The research status of the search query supported in SE schemes is then provided. The discussion includes the support of single keyword search query, conjunctive (and multi-keyword) search query and complex search query, respectively. Finally, this study presents the typical application scenario of SE schemes, and discusses the possible development tendency.

Key words: searchable encryption; data security; privacy; cryptography; cloud computing; cloud storage

* 基金项目: 国家自然科学基金(61232003); 国家科技重大专项(2013ZX03002004-003); 中美软件合作研究项目(61361120098)
收稿时间: 2012-09-08; 定稿时间: 2013-12-05; jos 在线出版时间: 2014-01-14
CNKI 网络优先出版: 2014-01-14 13:02, http://www.cnki.net/kcms/doi/10.13328/j.cnki.jos.000013.html

近年来,随着云计算技术[1−3]的迅猛发展,一些典型的云服务产品也顺势发布,并得到了人们广泛的关注,例如云网络存储工具 Dropbox[4]、亚马逊简易储存服务(Amazon simple storage service)[5]和微软的云计算平台 Windows Azure[6]等.它们在云端服务器上为用户保存数据和搭设虚拟系统环境,通过网络向用户传输对数据的操作服务,并根据用户所使用的硬件资源和服务时间进行收费.

由于其方便快捷的特性和灵活的收费方式,越来越多的用户选择将本地的数据迁移到云端服务器中,以此来节省本地的数据管理开销和系统维护开支.由于数据脱离了用户的物理控制而存储在云端,云端服务器管理员和非法用户(如黑客等不具有访问权限的用户)可以尝试通过访问数据来试图获取数据所包含的信息,这将可能造成数据信息和用户隐私的泄露.近年来,由于黑客的非法入侵和云端服务器管理员的不当操作造成了多起云安全事故的发生,直接导致了大量用户资料和私人数据的泄露.例如,Sony 公司在 2011 年由于黑客入侵导致上亿用户资料外泄事故[7]和 Google 公司在 2011 年发生的 Gmail 大规模用户数据泄露事件等,这些频繁发生的云事故,让用户开始更加审慎考虑当数据存放在云端时的安全性以及自己的个人隐私是否能够得到有效保护等问题.为了保证数据的机密性,越来越多的公司和个人用户选择对数据进行加密,并将数据以密文形式存储在云端服务器,但是当用户需要寻找包含某个关键字的相关文件时,将会遇到如何在云端服务器的密文进行搜索操作的难题.一种最简单的方法是将所有密文数据下载到本地进行解密,然后在明文上进行关键字搜索,但是这种操作不仅因为很多不需要的数据浪费了庞大的网络开销和存储开销,而且用户也需要因为解密和搜索操作付出巨大的计算开销.另外,这种方法也极不适用于低带宽的网络环境中.而另一种极端的方法是将密钥和关键字发给云端服务器,让云端服务器解密密文数据,并进行明文上的搜索操作.但是这种做法又将让用户的个人数据重新曝光于云端服务器管理员和非法用户的视线之下,严重威胁到数据的安全和用户的个人隐私.

为了更好地解决这个问题,可搜索加密(searchable encryption)便应运而生,并在近几年中得到了研究者的广泛研究和发展[8−32].用户可以首先使用 SE 机制对数据进行加密,并将密文存储在云端服务器;当用户需要搜索某个关键字时,可以将该关键字的搜索凭证(search capability)发给云端服务器;云端将接收到的搜索凭证对每个文件进行试探匹配,如果匹配成功,则说明该文件中包含该关键字;最后,云端将所有匹配成功的文件发回给用户.在收到搜索结果之后,用户只需要对返回的文件进行解密.在安全性上,云端服务器在整个搜索的过程中除了能够猜测任意两个搜索语句是否包含相同的关键字(即,搜索模式,search pattern),并知道多次搜索的结果(即,访问模式,access pattern)、文件密文、文件密文大小和一些搜索凭证之外,不会获得关于所请求搜索关键字内容以及文件的明文信息(一些工作也尝试使服务器端无法获得访问模式,例如文献[31,33,34]的工作).在访问效率上,通过以上过程可以直观发现使用 SE 机制给用户带来的方便性:首先,用户不需要为了没有包含关键字的文件浪费网络开销和存储空间;其次,对关键字进行搜索的操作交由云端来执行,充分利用了云端强大的计算能力;最后,用户不必对不符合条件的文件进行解密操作,节省了本地的计算资源.

1996 年,文献[31]首次提出了隐藏用户访问模式的密文搜索机制,但是它要求用户和服务器端进行多重对数轮交互,这种方法在实际中运用的效率不高.2000 年,文献[8]提出了一种基于对称算法下的 SE 方法的实现方法,开创了使用实用性的 SE 机制来实现在密文上进行关键字搜索的先例.在此之后,许多研究者不断推动着使用对称加密算法支持多个关键字搜索的 SE 机制[15,22].2004 年,文献[16]提出了基于公钥密码学算法上的 SE 机制,为后来的研究者通过公钥密码学实现更加多样的 SE 方案提供了指导.另外,研究者不仅在理论上设计多样的 SE 机制,并且也开始尝试将它们应用于真实应用场景中.例如,文献[9]提出了在日志数据上实现密文搜索的方法,微软公司提出的 Cryptographic Cloud Storage[35]中实现了基于对称密码学下的密文搜索的功能,以及针对云计算环境下所做的一些具有现实指导意义的工作[12,13,21]等.

在此基础上,本文主要分析和总结了当前 SE 机制的研究现状,探讨了其未来的发展趋势.从分析中可以得出:由于网络技术的迅速发展导致现今的数据获取将更加方便和快捷,在一些特殊应用场景中,例如电子病历存储、个人邮件处理、财政数据审计等,用户更加在意数据的安全性和个人隐私保护.因此,SE 机制在未来的一段时间内仍然是工业界和学术界的研究热点.特别是在数据以密文形式存储在云端的前提下,一种高效而又灵活的 SE 机制的设计和实现,都将对云计算的普及起到推动性的作用.

本文第 1 节主要从总体上介绍 SE 机制的主要研究内容.第 2 节主要对现有的 SE 机制的算法进行分类,分析它们的优劣,并介绍一些公钥密码学简单的术语.第 3 节主要介绍现有 SE 机制对搜索语句的支持情况.第 4 节针对 SE 机制的应用场景进行分析.最后,第 5 节总结当前的研究现状,并对未来的研究方向加以展望.

1 SE 机制

SE 机制的正确性和安全性、支持搜索语句的效果、密钥和密文长度以及算法的性能,是现今 SE 的主要研究内容.现今 SE 的研究内容主要有以下几个方面(如图 1 所示):

1) 灵活、高效的搜索语句的设计

灵活的搜索语句不仅能够让用户可以更加精确地定位到所需要的数据文件,同时也可以让用户能够更加灵活地表述搜索需求.SE 机制从研究初期的支持单词搜索,到后来逐渐发展为支持连接关键字搜索,再到支持区间搜索和子集搜索等复杂的逻辑语句.其研究难点在于:如何达到支持复杂的搜索请求的效果以及如何寻找到适合的困难假设来证明其安全性,同时使得该机制又具有可以接受的性能.随着云计算的发展,在海量用户和海量数据的应用场景下,提供安全、灵活、高效的 SE 机制将是研究者所极力追求的目标之一.

2) 模糊搜索和基于相似度排序的模糊搜索

由于用户表述不精确或者输入错误等各种原因,造成精确匹配搜索将有可能无法找到用户所真正需要的文件,因此,模糊搜索的引入能够智能地寻找与用户所输入的搜索词相关的文件.由于数据存储形式为密文且基于安全需求的考虑,目前在明文上所广泛应用的模糊搜索方法无法直接运用到密文上的搜索中,因此,设计支持模糊搜索和基于相似度排序的模糊搜索的可搜索加密机制,是一个亟待研究的内容之一.

3) 在不同现实场景中对 SE 机制的应用

自从 SE 机制提出后,针对其所能部署的应用场景得到了研究者的关注.从 SE 机制提出初期的数据所有者独享数据[8,10,15,26],到后来数据所有者将搜索的能力共享给其他用户[12,13,21,22,28],以及云存储环境下的一些特殊场景中用户私密数据的管理[12,14,21]等.针对不同的应用场景,需要相应的 SE 机制来支持,因此,设计适合目的应用场景的 SE 机制,是应用密码学领域的研究方向之一.

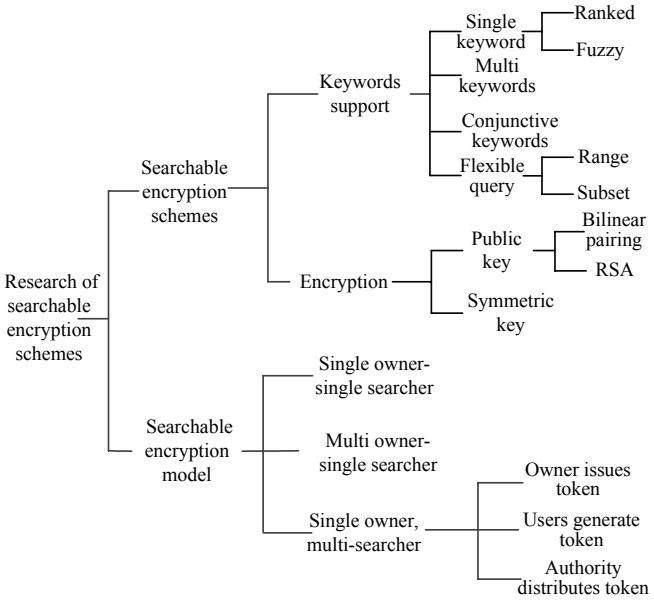


Fig.1 Research of SE schemes

图 1 SE 机制的研究内容

2 SE 机制的构造算法

现今众多的 SE 机制的设计可以根据其构造算法的不同而分为两类,即:基于对称密码学算法(symmetric key cryptography based)的 SE 机制[8−10,12,13,15,21,22,26,32]以及基于公钥密码学算法(public key cryptography based)的 SE 机制[9,11,14,16,20,23,24,29,30].前者主要是使用一些伪随机函数生成器(pseudorandom function generator)、伪随机数生成器(pseudorandom number generator)、哈希算法和对称加密算法构建而成;而后者主要是使用双线性映射等代数工具,并将安全性建立在一些复杂性问题的难解性之上.这两种方法的区别是:首先,由于大部分基于公钥密码学算法的 SE 机制都是基于双线性映射之上来构造的,在搜索的过程中需要进行群元素之间和双线性对的计算,因此,它们的开销要远高于基于对称密码学算法的 SE 机制;其次,基于对称密码学的 SE 机制更加适用于单用户创造数据并与多用户共享的应用场景,而基于公钥密码学的 SE 机制则允许除数据所有者之外的用户使用可搜索加密机制来产生数据密文并生成加密后的索引表.

2.1 SE机制的主要算法

由于现今 SE 机制的构造方法众多,因此其形式化描述方法各不相同.基本的 SE 机制主要包括 4 种算法[20],分别是 Setup,GenToken,BuildIndex 和 Query:

(1) Setup:该算法主要由权威机构或者数据所有者进行并生成密钥.在基于公钥密码学的 SE 机制中,该算法会根据输入的安全参数(security parameter)来产生公钥和私钥;在基于对称密码学的 SE 机制中,运行该算法后会产生一些私钥,例如伪随机函数的密钥等;

(2) GenToken:该算法以根据用户需要搜索的关键字为输入,产生相应的搜索凭证.算法的执行者主要由应用场景决定,可以由数据所有者、用户或者权威机构来执行(具体场景部分将在第 4 节中进行讨论);

(3) BuildIndex:该算法由数据所有者执行.在这种算法中,数据所有者将根据文件内容,选出相应的关键字集合,并使用可搜索加密机制建立索引表.在基于公钥密码学的 SE 机制中,数据所有者会使用公钥对每个文件的关键字集进行加密;在基于对称密码学的 SE 机制中,数据所有者会使用对称密钥或者使用基于密钥的哈希算法对关键字集进行加密.而文件内容主体将会使用对称加密算法进行加密;

(4) Query:该算法是由服务器端进行.服务器将以接收到的搜索凭证和每个文件中的索引表为输入,进行协议所预设的计算,最后通过输出结果是否与协议预设的结果相同来判断该文件是否满足搜索请求.服务器最后将搜索结果返回.

最后,用户在获得返回的文件密文之后,再使用相应的对称密钥对数据密文进行解密.

2.2 基于对称密码学的SE机制

对称密码学算法指的是:加密的密钥和解密的密钥都是衍生于同一个密钥,它们或者相等或者之间需要一些简单的转换[36].对称密码学算法的优点是计算开销小,适用于大块数据的加密,缺点则是加密方和解密方需要在事先实现密钥的协商,而密钥则需要通过安全信道传输.

而基于对称密码学的 SE 机制是采用伪随机函数、伪随机置换以及哈希算法(或者散列算法)对关键字按照一些步骤进行处理,当需要对关键字进行搜索时,首先将关键字随机化处理,然后让服务器端根据协议所预设的计算方式进行关键字的匹配,如果最后的结果是某种特定的格式,则说明匹配成功.例如:在文献[8]的工作中,当服务器端将关键字的密文 $E(w)$和服务器端的密文 C 进行异或运算后,需要判断结果的前$(n-m)$位的哈希值是否是后 m 位;在文献[9]的工作中,需要判断最后的结果是否是$(flag|K)$格式.基于对称密码学的 SE 机制在运算性能上较基于公钥密码学的 SE 机制更高效,但在支持搜索语句的灵活性上,现今的基于对称密码学的 SE 机制只能支持单个关键字或者是连接关键字的搜索,而且搜索凭证的大小需要与所搜索的关键字数目呈线性关系.

2.3 基于公钥密码学的SE机制

现今应用比较普遍的基于公钥密码学的 SE 机制大部分构建于双线性对(bilinear pairing)之上,其安全性都是基于不同的安全假设.下面首先给出关于双线性对的定义以及一些应用较为普遍的困难问题.

定义 1(双线性对定义)[16]. 对于双线性映射 $e:G_1\times G_1\to G_2$,需要满足以下条件:

(1) 双线性(bilinear):$\forall a,b\in\mathbb{Z}_q,\forall g,h\in G_1, e(g^a,h^b)=e(g,h)^{ab}$;

(2) 非退化性(non-degenerate):$\exists g\in G_1$,使得 $e(g,g)\neq 1$;

(3) 可计算性(computable):群 G_1,G_2 中的运算以及双线性映射 e 运算在多项式时间内可解.

定义 2(DDH,离散 Diffie-Hellman 问题)[11]. 假设 G 是一个素数阶 p 的群,其中,g 是 G 的生成元,随机地从 $\{0,\ldots,p-1\}$中选择元素 a,b,c,给定元组(g,g^a,g^b,g^c),判断 g^c 是否等于 g^{ab}.

定义 3(BDH,双线性 Diffie-Hellman 问题)[16,20]. 对于群 G 及其生成元 g,给定 g^a,g^b,g^c,计算 $e(g,g)^{abc}$.

基于公钥密码学的 SE 机制由于涉及到群元素之间的运算,开销较大.同时,正是由于双线性对的特性和复杂性假设的存在,使得一些支持更加复杂的搜索语句的工作得以发展.另外,基于公钥密码学的 SE 机制更加适用于一些不安全的网络中,它不需要加密方和解密方事先协商密钥,用户可以直接使用对外公开的公钥对关键字集合进行加密,而数据所有者可以使用私钥产生搜索凭证进行密文上的关键字搜索.

3 SE 机制搜索效果分析

搜索语句是用户搜索兴趣的表现,灵活的搜索语句不仅可以让用户能够更加准确地描述自己的搜索意愿,同时也能够更加精确地定位到用户需要的文件.在本节中,根据 SE 机制支持的搜索语句,分为单词字搜索、连接关键字搜索和复杂逻辑结构语句这 3 类.

3.1 支持单词搜索的SE机制

支持单词搜索指的是用户一次只能对一个关键字进行搜索,云端服务器通过该凭证搜索到包含该词的文件,然后将最后满足条件的结果返回给用户.早期的 SE 机制研究主要考虑的应用场景是:数据所有者为了节省本地存储空间,将数据以密文形式存储在远端服务器,并在带宽受限的环境下进行关键字搜索操作.在该情况下所衍生出来的 SE 机制,主要是以基于对称密码学算法为主.

文献[8]于 2000 年提出了一种基于对称密码学算法的实用 SE 方案,具体如图 2 所示.

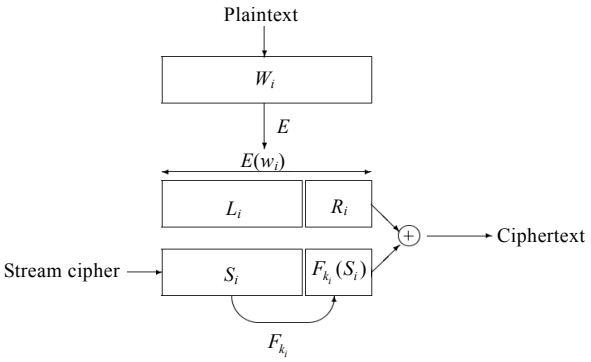


Fig.2 Final scheme in Ref.[8]

图 2 文献[8]中的最终机制

在这个方案中,伪随机函数 F 和 f 公开.在对数据加密时,文件内容被分割为固定大小的文件块,经过加密后的文件块 $E(w_i)$被分为两个部分,分别是 L_i 和 R_i.同时,数据所有者利用伪随机位(pseudorandom bits)S_i 生成 $(S_i,F_{k_i}(S_i))$,并与(L_i,R_i)经过异或运算生成密文 C_i,其中,L_i 和 S_i(或 R_i 和 $F_{k_i}(S_i)$)的长度分别为$(n-m)$位(或 m 位),并且满足 $k_i=f_{k'}(L_i)$.当用户需要搜索关键字 w_i 时,数据所有者将 $E(w_i)$和 k_i 发给服务器端.服务器端将密文 C_i 与 $E(w_j)$进行异或运算,然后判断得到的结果是否满足$(s,F_{k_i}(s))$的形式,如果满足,则说明匹配成功,并将该文件返回.这种方法虽然能够基本上实现单词搜索,但却存在着一些缺陷:云端服务器需要对每个文件的内容进行扫

描,看密文内容是否存在给定的关键字的密文形式相匹配的内容,造成的计算开销将与文件大小呈线性关系,在海量数据环境下,该方法效率不佳.同时,服务器端可以通过统计攻击的方法获得一些额外的用户隐私信息,例如,通过得到的搜索凭证来判断用户前后搜索的关键字是否相同等.

针对文献[8]中所提 SE 方案搜索效率低下等缺点,文献[10]提出了使用安全索引的方法来快速实现对海量密文数据的搜索.这种搜索机制建立在 Bloom Filter[37]之上,即,每个文件都有对应的一些独立的哈希函数和 Bloom Filter 数据结构.在文件加密之前,需要对文件中的关键字使用私钥加密,再使用哈希函数映射到 Filter 之上并记录,最后,将映射后的 Filter 和文件的密文上传到服务器中.当用户需要进行密文搜索时,需要将关键字的密文发送给云端服务器,再由云端服务器使用每个文件的哈希函数进行关键字到 Filter 的映射.如果映射到的位置之前都有记录的痕迹,则说明这个关键字有很大的概率是在该文件中,最后,云端服务器将得到的匹配文件发给用户.这种方法的优点是能够利用哈希函数计算快速的特点,快速地查找关键字所在的密文文件;但是它也继承了 Bloom Filter 存在错误率的特点,有可能导致一些文件本来并不包含关键字,最后却能够通过哈希函数的检测,而被云端作为结果返回给用户,给用户带来一些额外的带宽和计算开销.

由于之前的 SE 机制都是基于对称密码学算法,文献[16]给出了基于公钥密码下的可搜索加密机制 PEKS 的构造方法,该文的作者们主要考虑的应用场景是:用户 Alice 掌握着私钥(private key),并将相对应的公钥(public key)公开,为了让电子邮件网关(E-mail gateway)分拣接收到的邮件,Alice 会事先将一些特定关键字的陷门(trapdoor) T_w 发送给电子邮件网关,使得它能够通过判断邮件中是否包含关键字 W 来选择接受设备.与此同时,电子邮件网关在判断的过程中无法获得关于关键字和邮件内容的有效信息.在 PEKS 机制的加密算法中使用到了两个哈希算法 H_1 和 H_2,用户使用哈希算法 H_1 将每个关键字映射到群 G_1 中,然后选取随机数 r 和双线性映射将关键字 W 随机化映射到群 G_2 中.这样,即使是相同的关键字所生成的密文也将有所不同,并将 r 的信息以 g^r 形式保存,最后生成密文(g^r,$H_2(t)$).当需要生成搜索凭证时,使用 H_1 将关键字映射到 G_1 中,并选取随机数 a,将 $H_1(W)^a$ 发给服务器端,使得即使是相同的关键字,由于随机数 a 的作用,所产生的搜索凭证也将有所不同.最后,服务器端利用双线性映射的性质进行关键字的匹配判断.该方法的优点是支持数据接收者对多个发送者所加密的密文中进行搜索的应用场景,而且由于随机数的作用,系统的加密效果为非确定性加密,导致了服务器端无法通过密文是否相同来判断索引表(或搜索凭证)中是否具有相同的关键字.其缺点是计算开销因为双线性对的引进而加大,特别是对操作(pairing operation)的计算开销较大,使得该方法在海量数据处理场景中的应用性受到一定的限制.另外,PEKS 的安全性在随机语言机模型(random oracle model)下成立,并不适合现实应用.

文献[9]分别基于对称密码学和基于身份加密(identity based encryption,简称 IBE)给出了两种 SE 机制的设计方法以实现在加密后的审计日志上进行关键字的搜索.在基于对称密码学的设计中,审计日志服务器(audit log server)首先从数据中萃取出具有代表性的关键字 $\{w_i\}_{i=1}^n$,然后使用带有密钥的伪随机函数和一个随机位串 r,计算得到密文.用户需要向审计代理(audit escrow agent)请求搜索关键字 w 操作,审计代理根据每个审计日志服务器的密钥 $\{S_j\}_{j=1}^t$,使用伪随机函数生成搜索凭证(search capability)并发给用户.用户持有该搜索凭证并利用 r 对密文进行 XOR 运算,如果结果满足某种特定的格式,则说明搜索成功.而在基于 IBE 的设计中,审计日志服务器以关键字 w_i 为公钥,对($flag|K$)使用 IBE 的加密算法进行加密.当收到搜索关键字 w 的请求后,审计代理通过计算得到 w 的私钥 d_w 作为搜索凭证发给用户.用户使用 d_w 对密文使用 IBE 中的解密算法进行解密,若得到的结果是($flag|K$)形式,则说明解密成功.与其他 SE 机制的应用场景不同,该方法需要审计日志服务器对日志进行加密,同时要求用户访问服务器进行搜索,需要用户自己承担搜索开销.

文献[15]则考虑了两种应用场景,分别是当用户拥有足够的空间存储生成关键字的字典以及当用户所拥有的存储空间无法存放字典的情况.所谓的字典就是用来表示每个关键字的、长度为 d 的二元组合(i,w_i),其中,$i\in 2^d$.在第 1 种情况中,用户需要为每个文件 j 建立二进制位串索引 I_j,并使用伪随机函数将该文件包含的每个关键字的 id 使用伪随机置换函数进行置换,并将置换结果所在的位置 1,并使用另一个随机位串对 I_j 进行异或运算,隐藏 I_j 中每一位的值.在第 2 种情况中,用户将字典用伪随机置换函数随机化后存储在服务器端.当需要进行关键字的搜索时,则需要与服务器端进行两轮的交互,第 1 轮是从服务器中取出想要搜索关键字的二进制串表

示,第 2 轮中则是计算出关键字的陷门并发给服务器端.

针对之前的 SE 机制只是假设攻击者(adversary)并不会考虑搜索的陷门(trapdoor)和以前搜索结果的特点,文献[22]提出了一种更强的自适应(adaptive)攻击者模型,在该模型中,攻击者会将之前搜索过的陷门和搜索结果作为参照来决定下一次的查询语句.在文献[22]所做的工作中,提出了两种设计方案:第 1 种设计方案是在假设攻击者是 non-adaptive 的情况下保证安全,并使用了连接表(linked list)、数组和查找表等数据结构将不同文件中的相同关键字连接起来,提高了搜索的性能;而在第 2 种方案中,则保证了自适应的语义安全(adaptive semantic security).同时,针对以前基于公钥密码学算法中只有掌握私钥的用户才能对用公钥加密的数据进行搜索的限制,文献[22]采用了广播加密(broadcast encryption)的方法,在共享用户群中共享密钥 r,使得该共享用户群也能够对这些密文数据进行搜索.

之前的 SE 机制只是简单地将满足关键字的搜索结果返回但却缺乏对搜索结果的整理,而用户需要在本地对结果进行整理并找出最相关的结果.文献[12]针对搜索结果进行了一些优化,提出了一种在云数据中实现安全排序的方法,这种方法不仅能够使云端服务器进行关键字的密文搜索,同时还能够对搜索的结果对于关键字的相关性程度进行排序,最后,将前 k 个最相关的文件返回给用户.这种操作更加有利于用户更快地找出自己所感兴趣的数据,节省了用户的开销.为了实现排序的效果,首先需要文件所有者在上传文件之前,针对每个关键字在文件中出现的次数,根据相关性分数模型计算该文件对于每个关键字的相关性得分,然后在本地构建一个关键字和文件的倒排索引表,每个表格的数据则是该关键字和文件的相关性分数,然后再将索引表加密.同时,使用保序加密(order-preserving encryption[38])对相关性分数加密,当用户需要搜索密文关键字时,云端服务器再根据每个文件对该关键字相关性分数的密文结果进行排序,并将经过排序后的前 k 个文件返回给用户.这种方法的优点是,在每次搜索的过程中,用户可以得到在相关性分数模型下对该关键字最为相关的文件,提高了用户体验,并节省了一定的带宽和因为解密带来的计算开销,但是云端服务器也掌握了每个文件对于某个关键字相关性程度的信息.

针对有可能发生的印刷错误以及用户由于疏忽导致精确匹配无法发生作用的情景,文献[13]则考虑了在云计算中实现密文上的模糊关键字搜索的应用场景.在文献[13]的设计中,主要考虑了 3 种情况下的模糊搜索,分别是插入、删除和替换操作.该文作者使用了编辑距离(edit distance)的概念,对于某个单词,他们将在某个编辑距离之内的所有单词都加密,然后上传到云端服务器.当用户需要搜索某个单词时,可以定位到与该单词具有某个编辑距离的所有单词,然后将包含这些单词的文件返回给用户.该方法虽然可以实现对于关键字的模糊搜索,但是带来的代价是需要存储的可能数据量过于庞大;而且用户虽然可以得到近似的结果,但是结果的不精确性也将给用户带来极大的网络和计算开销.

3.2 支持连接关键字搜索(conjunctive keyword search)的可搜索加密机制

由于支持单词的 SE 机制只允许用户一次只能发送一个单词的搜索凭证,这极不符合现实生活中多词搜索的应用需求,特别是当单词无法精确定位到用户所想要的文件时,单词搜索的限制可能需要用户使用不同关键字多轮搜索,或者是经过一轮密文搜索后,对返回结果解密,通过在明文上进行搜索来寻找目标文件,而这样的结果将给用户带来极差的操作体验.针对这些不足,支持连接关键字搜索的可搜索加密机制开始得到研究者广泛的关注和研究.

文献[11]针对之前的搜索机制中只能使用单词搜索的不足,提出了两种支持连接关键字搜索的 SE 机制.在这一方案中,每个文件都有固定数量的关键字域,每个域中都有特定的关键字来表征这些文件的特性.例如:在邮件中具有关键字域“主题、发送方、接收方”,而在“主题”域中可能具有关键值“会议”等.第 1 种机制能够达到固定的在线网络开销,所谓固定指的是用户数据所有者进行在线交互的网络开销是依赖于每个文件中的关键字域数量,在这个方案中,用户需要发送两个部分的搜索凭证:第 1 个部分可以在高速网络中离线发送到服务器端,称为“原型凭证(proto-capability)”,其大小与存储在服务器端的文件数量线性相关;第2个部分称为“查询部分(query part)”,需要用户与数据所有者进行在线交互而得到.当用户将查询部分发给服务器端时,服务器端会将其与原型凭证整合成完整的搜索凭证,并进行搜索.该机制的安全性建立在 DDH(decisional Diffie-Hellman)问题

的复杂性之上.针对第 1 种机制可能导致网络开销过大的情况,在第 2 种机制中使用了固定网络开销的搜索凭证,即,搜索凭证对于文件数量而言是固定的,但是依然与关键字域数量线性相关.

文献[26]分别基于对称密码学和公钥密码学提出了两种实现连接关键字搜索的高效 SE 机制,但是需要保证没有重复的关键字.在基于对称密码学的 SE 机制中,首先对关键字采用伪随机函数进行加密,当用户需要搜索 m 个关键字时,就使用秘密共享(secret sharing)中的 recover 算法,将这 m 个关键字的密文作为输入,并将结果作为陷门发给服务器端.服务器端也调用 secret sharing 中的 recover 算法,对于每个文件进行判断,如果结果等于陷门,则说明该文件包含了所有的关键词.但是该方案要求陷门的大小和文件数量呈线性关系.针对这个问题,文献[26]提出了基于公钥密码学的 SE 机制,通过使用双线性映射使得陷门的大小固定,但其的安全性则是建立在 DDH,XDH 和 MXDH 的复杂性之上.

之前的基于公钥密码学机制要么只能支持“多个发送者-单个接受者”的应用场景,要么会出现密文过于庞大的情况,文献[24]针对这种情况提出了一种基于公钥密码学的 SE 机制,其核心思想是:通过使用多接受者公钥加密方法(multi-receiver public key encryption)将所有接受者的公钥对关键字集合进行加密,通过支持连接关键字的公钥加密方法,实现了支持每个接受者只需使用自己的私钥就能对连接关键字进行搜索以及在“单个发送者-多个接受者”场景下支持密文搜索的效果.

作为文献[12]工作的拓展,文献[21]的工作能够让服务器对用户所请求搜索的多个关键字,根据每个文件对于所请求关键字的得分排序,并将排名最高的 k 个文件返回给用户,服务器端将无法获得用户搜索的关键字信息、文件是否包含某个关键字信息以及最后每个文件的得分信息.其核心思想是:采用 kNN[39]的思想,首先生成两个二进制位串,分别称为文件的数据向量(data vector)和用户的查询向量(query vector).这两个向量中的每个位都分别与关键字进行一一对应,并以该位的值来表示该文件以及用户的查询请求是否包含某个关键字,然后使用两个互逆的矩阵分别对这两个位串加密,保证文件包含关键字的信息和用户查询语句对云端服务器不可见.在计算得分的时候,还需要对两个位串的乘积通过加入随机数来进行随机化处理.这种方法与前面的连接关键字搜索的不同之处在于:普通的连接关键字搜索是返回的文件需要保证包含每个域上的关键字;而在这里的多词搜索中,即使某个文件没有全部包含所请求的关键字,但只要其得分位列于前 k 中,依然可以被返回.另外,随机数的引入也导致了最终得分的不精确性,根据文献[21]中的描述,当引入随机变量的正态分布标准差 σ=1 时,最后结果的不准确度最高可达到 20%.

该机制的安全性建立在文献[39]之上,能够抵御已知明文攻击(known-plaintext attack).

除了在普通数据上进行搜索之外,文献[32]也尝试着在对加密后的图结构数据(graph-structure data)上进行搜索的机制,该机制遵循“过滤和确认(filtering-and-verification)”原则,即:在过滤步骤里,预先为每个图建立一个基于特征的索引(feature-based index),其中,特征(feature)是原图的子图;而在确认阶段,则需要判断每个图像的特征是否与请求同构.由于特征集合的大小和整个图像集合相比更小,因此这种操作减轻了比较的工作量.与文献[21]不同,为了让云端服务器无法通过相关性得分来判断在一次搜索中所匹配的关键字数量这一信息,该文放弃了在每个图形索引中使用二进制位串中的每一位的值来表征是否包含特征的方法,而是分别从一个随机数串 S 和一个随机数矩阵 M 中取值来表征,其中,$M[i][j]<S[j]$.如果图形 G_i 包含特征 F_j,则相应的索引表取值为 $S[j]$;否则,取值为 $M[i][j]$.为了保证索引表对云端服务器不可见,使用了一对互逆的矩阵分别对图形索引表和查询语句中的数据向量和查询向量进行隐藏.最后,在搜索时,数据所有者会将搜索预期值作为搜索凭证的一部分发给用户,而云端服务器在计算相关性得分时,如果某个图形的相关性得分等于该预期值,则说明该图形包含所有的特征;反之,如果得分小于预期值,则说明该图不满足所有的特征,但是云端服务器无法在不满足的图形中获取它们包含特征数量多少的信息.该机制能够抵御已知明文攻击.

3.3 支持复杂逻辑结构的可搜索加密机制

另外,近年来逐渐发展的谓词加密(predicate encryption)是一种涵盖面比较广的密码学原语(cryptographic primitive),它涵盖了基于属性的加密机制(attribute-based encryption scheme[40,41])和基于身份的加密机制(identity-based encryption scheme[42]).在谓词加密中,每个密文和刻画其性质的属性 $I\in\Sigma$相联系(其中,Σ是所有属

性的集合),而对应着谓词$f\in F$的私钥(secret key)记作SK_f.如果SK_f和I满足$f(I)=1$,则SK_f能够解密与I相联系的密文.谓词加密可以被运用于密文搜索中,其中每个密文的属性可以用该文件含有的典型关键字来刻画,谓词可被认为是用户的查询语句,私钥可看作是根据用户查询语句生成的搜索凭证,而$f(I)=1$可认为密文的关键字满足用户的查询语句.

文献[30]提出了支持析取子句、多项式方程和内积形式的谓词加密构造方法.通过在复合阶群(composite order group)上构建向量(vector)来表示析取范式、合取范式和多项式方程等复杂逻辑结构.虽然该方法支持多种逻辑结构,但是由于建立在复合阶群之上的双线性对的计算开销大约是素数阶群的50倍[43,44],因此该方法的性能不是很理想.在该方法的实现中,虽然可以使用一些预处理来加速双线性对计算[45],但是总的效率依然远低于其他建立在素数阶群之上的SE机制.另外,支持内积的谓词加密也可以适用于子集和区间的关键字搜索,可以将判断是否属于子集和区间的关系式转换为以"或"关系连接的子句.例如,对于判断语句"$x\in[1,3]$",可以将其转换为"$(x=1)\vee(x=2)\vee(x=3)$",然后再将该子句转换为内积形式.针对谓词加密的代理,文献[12]考虑了文献[20]中的代理问题,并提出了实现的方案.另外,文献[25]还考虑了谓词的隐私性问题,构造了一个实现谓词隐私保护的密码学原语,该构造建立在一个复合群之上,其中,该复合群可以表示成4个子群的直积.该方法能够达到更高强度的属性隐藏安全(attribute-hiding security),意味着除了用户通过密钥获得的额外信息之外,每个密文所关联的属性信息都将被隐藏.

文献[29]构造出了基于DPVS(dual pairing vector space)的分级谓词加密方案(hierarchical predicate encryption,简称HPE),其中,DPVS是n个素数阶群的直积所展成的空间,并且DPVS上的双线性运算相当于在n个素数阶群上进行双线性运算.通过在DPVS上构造出单位正交向量$(\boldsymbol{b},\boldsymbol{b}^*)$,将向量$\boldsymbol{b}$用于构造密文,将向量$\boldsymbol{b}^*$用于生成密钥,使得用户可以根据自己的私钥将部分权限代理给其他用户.假设用户Alice能够解密的文件集合为F_A,而当Alice将部分权限代理给Bob后,Bob所能解密的文件集合F_B应该满足$F_B\subseteq F_A$.该方案能够在RDSP和IDSP假设下,对选择性明文攻击(chosen plaintext attack)达到选择性属性隐藏(selective attribute-hiding)安全.

文献[17]考虑到网络入侵检测的问题而提出了MRQED机制,在不可信的远端服务器中存放网络审计日志的密文形式并进行搜索,当网络入侵发生时,审计者(auditor)可以通过被授权的*key*来查找在某个特定区间中的网络流信息.在他们的工作中,提出多维区间查询(multi-dimensional range query)机制,其中,在每个维度上支持区间查询,并定义了两种安全模型,分别是匹配保护安全(match-concealing security)和匹配泄露安全(match-revealing security),最后证明了所提出的方案在这两种安全模型下能够达到选择性安全(selective secure).另外,他们的工作可以防止串谋攻击(collusion attack).

文献[14]考虑了在加密后的个人健康记录(personal health record,简称PHR)上进行授权搜索的应用场景,并利用分级的谓词加密(hierarchical predicate encryption,简称HPE)支持灵活的搜索语句.在他们的设计中,用户首先在本地产生合取范式形式(conjunctive normal form,简称CNF)的查询语句,需要先向本地的LTA申请,本地的LTA认证用户的请求并将合取范式的查询语句转换为谓词向量(predicate vector)和属性向量(attribute vector)来表示,最后,LTA将转换成的谓词向量作为输入,使用分级谓词加密方法中的代理算法(delegation algorithm)产生相应的搜索凭证(search capability)并颁发给用户.为了保证查询语句的隐私,文献[14]引入了一个代理服务器(proxy server)对密文进行再处理,使得云端服务器无法达到通过公钥遍历生成密文来试探搜索语句的内容的效果.由于该机制建立在文献[29]之上,因此其安全性能够达到文献[29]中所提方案的安全性.

3.4 小 结

从以上的分析可以看出:随着研究者们对SE机制的研究和发展以及为了满足一些特定的应用需求,SE机制在搜索效果上得到迅速的发展:首先,从对搜索语句的支持上,SE机制的搜索效果从简单的单词搜索,逐步发展成灵活、复杂的查询语句,不仅支持合取范式和析取范式等查询语句,而且在每个维度上支持相同匹配(equality)、子集(subset)和区间(range)查询,这将更有利于用户描述自己的搜索意愿;其次,从对搜索结果的优化处理上看,SE机制的搜索结果从只满足用户搜索请求,逐步发展成支持返回与用户搜索关键字相关的前k个文件,以及支持容错的密文搜索等,这将更有利于用户能够更快地找到自己所需要的文件.表1对一些SE机制在所

基于的算法、查询语句支持等方面进行总结,其中,对于一些基于公钥密码学的 SE 机制,还归纳了它们所基于的安全性假设.

Table 1 Comparison of SE schemes

表 1 SE 机制对比

Algorithm based	Query support	Schemes	Assumptions
Symmetric key cryptography based SE schemes	Single keyword	SWP00[8]	\
		WBDS04[9]	\
		G03[10]	\
		CGKO06[22]	\
		WCLRL10[12]	\
		LWWCRL10[13]	\
		CM05[15]	\
	Multi-Keyword	CWLRL11[21]	\
	Conjunctive eyword query	CYWRL11[32]	\
		BKM05[26]	\
Public key cryptography based SE schemes	Single keyword	WBDS04[9]	BDH
		BCOP04[16]	BDH
		DRD10[23]	RSA assumptions
	Conjunctive keywords	HL07[24]	DLDH
		GSW04[11]	DDH, BDDH
		BKM05[26]	DDH, XDH, MXDH
	Complex search query	OT09[29]	RDSP, IDSP
		KSW08[30]	Large number factorization
		LYCL11[14]	RDSP, IDSP
		SBCSP07[17]	DBDH, DLA
		BW07[20]	BDH, C3DH

另外,针对一些具有典型意义的基于公钥密码学的 SE 机制,我们给出了它们在公钥长度、索引密文长度、搜索凭证大小、索引表的加密以及每次匹配搜索时间开销的比较,具体见表 2.从表 1 中可以看出:基于公钥密码学的 SE 机制大部分都使用了双线性对工具,然后使用了一些安全假设来证明该机制的安全性.同时,从表 2 中可以看出:在一些支持连接关键字的 SE 机制中,搜索凭证的大小和关键字数目呈线性关系,例如 DRD10[23], BKM05[26];有的则和文件的数量呈线性关系,例如 GSW04[11].而对于支持内积的谓词加密机制而言,所需要的公钥大小通常与向量长度 n 有关,例如在 OT09[29]中,公钥大小是 $O(n^2)$,这是由于其所需要的公钥为 $O(n)$个长度为 n 的向量.另外,对于支持“单个发送者-多个接受者”的情况,公钥大小还与接受者的数量有关,例如 HL07[24].对于支持子集和区间查找的 SE 机制,密文和搜索凭证的长度都与维度的数目以及每个维度中关键字的数量有关,如 SBCSP07[17],BW07[20].

Table 2 Comparison of some public-key cryptography based SE schemes

表 2 一些基于公钥密码学的 SE 机制的对比

Scheme	PK size	CT size	Search cap size	Encryption time	Search time
GSW04[11]	$O(1)$	$O(W)$	$O(F)$ & $O(KW)$	$O(W)G$	$O(KW)G$ & $O(KW)P$
BCOP04[16]	$O(1)$	$O(W)$	$O(1)$	$O(W)G+O(W)H_1$	$O(1)P+O(1)H_2$
LYCL11[14]	$O(n^2)$	$O(n)$	$O(n)$	$O(\sum_{i=1}^{I} u_i \cdot n)\, G$	$O(\sum_{i=1}^{I} u_i \cdot n)\, P$
SBCSP07[17]	$O(D\log T)$	$O(D\log T)$	$O(D\log T)$	$O(D\log T)G+O(1)G'$	$O(\log T^D)P+O(\log T^D)G'$
BW07[20]	$O(DT)$	$O(DT)$	$O(D)$	$O(W)G$	$O(KW)P$
DRD10[23]	$O(1)$	$O(W)$	$O(KW)$	$O(W)G$	$O(1)G)+O(1)H$
HL07[24]	$O(1)$	$O(W+U)$	$O(KW+U)$	$O(U)G_1+O(W)G_1$	$O(KW)G_1+O(1)G_1+O(1)P$
BKM05[26]	$O(1)$	$O(W)$	$O(KW)$	$O(W)G_1$	$O(KW)G_1+O(1)G_T+O(1)P$
OT09[29]	$O(n^2)$	$O(n)$	$O(n)$	$O(\sum_{i=1}^{I} u_i \cdot n)\, G$	$O(\sum_{i=1}^{I} u_i \cdot n)\, P$
KSW08[30]	$O(n)$	$O(n)$	$O(n)$	$O(n)G$	$O(n)P$

KW: The number of keywords, W: The number of words, F: The number of documents, D: The number of dimensions, T: The number of keywords over each dimension, U: The number of users, n: The length of the vector, P: The pairing operation, $G/G'/G_1/G_T$: Operations in group $G/G'/G_1/G_2$, H_1/H_2: The hash operation

4 SE 机制的应用模型

近年来,SE 机制已被广泛地应用于实际场景中.根据应用场景,SE 机制可以分为 3 类:第 1 类是数据所有者并不将数据共享给其他用户,而是独自拥有对数据的搜索的权利;第 2 类是数据所有者允许其他经过认证后的用户对其数据进行搜索;第 3 类是多个数据所有者允许某个特定的用户对数据进行搜索,例如邮件处理场景.

4.1 数据独享场景

在早期的SE机制中,数据所有者独享搜索权限是主要考虑的应用场景,如图3所示.用户将数据上传到服务器端,并不是出于与其他用户共享的目的,而是为了节省本地存储空间和管理开销,同时希望能够在低带宽的网络环境下对数据进行访问.由用户独享搜索权限的数据可以是用户的私密数据,例如电子病历、邮件等涉及到个人隐私和公司内部事务等的数据.在文献[8,10,26]的工作中,主要考虑了当用户 Alice 在保留搜索能力的前提下,将个人数据的密文存放在不可信的服务器上,在带宽受限的环境中,需要对其数据进行搜索的应用场景.在文献[15]的工作中,主要考虑两种应用场景:第 1 种是当用户在使用家中的电脑存储的字典产生搜索凭证,并发给远端服务器的密文搜索;第 2 种应用场景是当用户使用存储空间较小的手机时,需要将字典存放在远端服务器,然后经过与服务器端进行两轮交互,对服务器端的密文数据进行搜索.

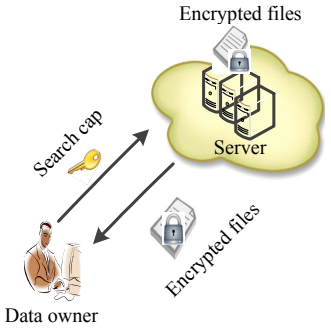


Fig.3 Scenario that data are searched by the owner only

图 3 数据只由数据所有者独享

4.2 数据共享场景

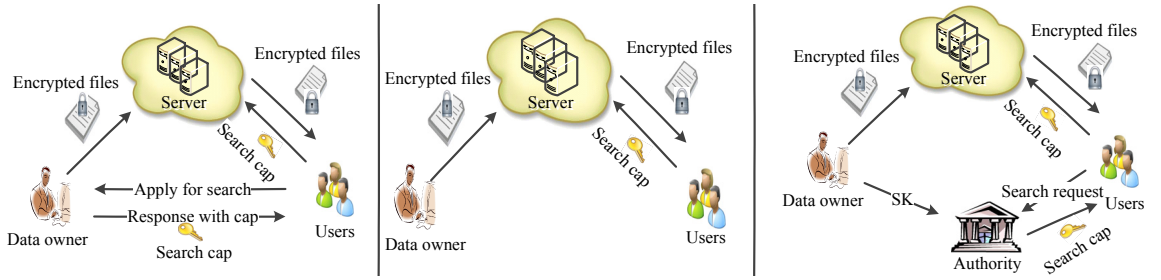
数据共享场景主要适用于数据所有者将自己的一些数据与其他用户共享的场景.当数据所有者允许经过认证后的用户对其数据进行搜索时,则有 3 种方式使得其他用户能够根据自己感兴趣的关键字获取到搜索凭证,分别是:由数据所有者生成搜索凭证;或者数据所有者将一部分密钥信息发布给授权的用户,由授权用户在本地产生搜索凭证;或者将分发搜索凭证的责任交由信任的第三方权威机构来执行.

这 3 种方法各有所适用的应用场景.

4.2.1 由数据所有者生成搜索凭证

由数据所有者生成搜索凭证可以保证数据实时地由数据所有者控制,使得非法用户在没有获得搜索凭证的情况下无法对数据进行搜索操作,具体情况如图 4(a)所示.但是这种方法也要求数据所有者必须时刻在线处理用户的搜索请求,并在本地为每个请求计算生成搜索凭证,这极有可能导致数据所有者的计算能力成为整个系统的瓶颈,从而降低了系统的可扩展性.

在文献[21]的工作中,用户需要将所请求的关键字集合 $\tilde{W}$ 以及一个二元向量 Q 发给数据所有者,其中,$Q[j]$ 的作用是表征 W_j 是否在 $\tilde{W}$ 中.根据用户的请求,数据所有者使用密钥矩阵计算得到 $\{M_1^{-1}\bar{Q}', M_1^{-1}\bar{Q}''\}$,并将其返回.由于是矩阵之间的乘积,因此对数据所有者而言,开销不是很大.



(a) 数据所有者产生搜索凭证并发给用户 (b) 用户在本地产生搜索凭证 (c) 可信审计机构处理用户请求并分发搜索凭证

Fig.4 Scenario that data are shared between multiple users

图 4 数据在多用户之间共享的场景

4.2.2 由用户在本地自己生成搜索凭证

由用户自己生成搜索凭证不仅可以有效地减轻数据所有者的计算负担,还避免了与数据所有者进行交互操作所带来的网络开销和时间延迟,增强了系统的可扩展性.但是数据所有者丧失了对用户搜索关键字的认证,同时,共享给用户密钥也加大了数据泄露的可能性.该应用场景如图 4(b)所示.

在文献[12,22]的工作中,数据所有者需要将密钥 x,y 共享给认证用户,用户可以根据自己感兴趣的关键字 w 生成陷门 $T=(\pi_x(w),f_y(w))$,其中,f 和 π 分别是伪随机函数(pseudo-random function)和哈希函数.

在文献[13]的工作中,数据所有者需要和认证用户(the authorized user)共享一个密钥 sk,当认证用户需要对某个关键字 w 进行搜索时,则利用单向函数 f 产生凭证 $f(sk,w)$ 并发给云端服务器.

文献[28]考虑了一种比较新颖的系统模型,每个用户都拥有自己单独的私钥,并且都可以自己向中立的数据库服务器端上传密文.与此同时,每个用户都有权利针对某个关键字使用自己的私钥来生成搜索凭证,并对整个数据库进行搜索.

4.2.3 由可信审计机构生成搜索凭证

在多数据所有者和多用户的环境中,由于服务器端的数据归属复杂,数据所有者可以将分发搜索凭证的责任交由可信审计机构执行,具体情况如图 4(c)所示.由可信审计机构生成搜索凭证可以不要求数据所有者时刻在线,并利用可信审计机构强大的计算能力来承担产生搜索凭证的计算负担,同时达到对用户的搜索请求进行授权的目的.但是这也要求数据所有者对其完全信任,并将数据的控制权限交给该机构.其中,基于公钥密码学的 SE 机制较为适用于这种应用场景.

文献[14]考虑了用户在加密的电子病历上进行搜索的应用场景,其中,电子病历由可信的第三方机构 TA 进行保存管理.该文提出了使用 TA/LTA 对用户的搜索权限进行认证并颁发搜索凭证的系统框架,每个 LTA 都被赋予了一些属性密钥,这些属性密钥决定了该 LTA 对数据的访问权限.当认证用户的搜索请求时,LTA 使用自己的属性密钥对用户的请求进行代理操作(delegation operation),即,经过认证后的用户请求能够检索到的文件不能超过该 LTA 的权限.这种方法能够缓解 TA 成为整个系统性能瓶颈的情况,并提高了系统的可扩展性.但在该系统框架中,由于 TA 拥有了最高的权限,造成数据的明文信息将完全曝光在 TA 的视线之中,而数据所有者丧失了对数据的完全控制能力.

在文献[9]的工作中构造了两种机制:在第 1 种机制中,每个调查者(investigator)在搜索之前都需要将所要搜索的词 w 发给审计代理机构,以此获取搜索凭证(search capability) $d_w:=\langle H_{S_1}(w),...,H_{S_t}(w)\rangle$,其中,$H_{S_i}(w)$ 是以 S_i 为密钥的哈希函数;在第 2 种机制中,审计代理使用 IBE 中的哈希函数以及随机数,调用 IBE 中密钥生成算法得到 $d_w:=sH_1(w)$.

在文献[17]的工作中,主要考虑审计者在网络入侵发生后,对网络流量进行审计并找出入侵者的应用场景.审计者需要向权威机构进行申请,并获得搜索凭证.

4.3 邮件处理场景

这种情况主要使用于基于公钥密码学的SE机制.邮件发送者使用公钥对邮件进行加密,邮件接收者利用私钥对一些感兴趣的关键字生成搜索凭证并发给网关,让网关根据感兴趣的关键字对邮件进行分拣.

在文献[16]的工作中,主要考虑邮件接收者对其他用户利用其公钥加密的邮件进行判断的场景.邮件接受者针对感兴趣的关键字使用私钥生成搜索凭证并发给网关,网关根据搜索凭证来判断邮件中是否有接受者所制定的关键字,并以此来判断邮件的紧急程度,从而选择其最终的接受设备.在文献[20]的另一项工作中,主要考虑的是邮件接受者根据感兴趣的谓词条件生成搜索凭证 P 并发给网关,使其根据搜索凭证来判断邮件是否满足谓词条件 P,以此来决定接受的设备.该谓词条件包括某个关键字是否属于某个区间,以及判断该关键字集合是否是某个集合的子集等.

针对文献[16,20]所考虑的应用场景是多个用户发送给某个邮件接收者的情况,文献[24]考虑了其对偶问题,即,一个用户需要将一封邮件发送给多个接受者的情况.他们考虑的是使用多接受者的公钥加密方法对邮件进行处理,从而避免因为分别使用每个接受者的公钥对邮件加密而带来的重复加密开销,以及支持邮件接受者单独使用私钥生成搜索凭证对邮件进行搜索.

4.4 小 结

本节主要对SE机制的应用场景和系统模型进行了介绍,分析了每种应用场景的特点.从以上的分析可以得出以下几个结论:

- 首先,SE 机制的研究与现实的需求紧密相关,具体体现在各个 SE 机制都能在一些特定的现实场景中为用户节省大量的开销和繁琐的操作.另外,对于每个应用的场景都有各自的威胁模型和系统模型的定义,而它们所关注和改善的目标都极为不同,而这些不同之处也产生了额外的问题,进而带来新的设计,例如,如果需要减轻数据所有者的负担而给予用户生成搜索凭证的能力,则需要将相关的密钥分发给用户,同时保证授权用户不会将该密钥泄露;
- 其次,现实中的一些特定的应用需求,也极大地凸显出其与现有SE机制的不合之处,从而让研究者们开始分析和设计适应该需求下的 SE 机制,进而达到提高安全性的同时又在极大程度上保证其效率的效果.

5 总结与展望

本文针对可搜索加密机制的研究现状进行了较为全面的介绍和讨论:首先,针对SE机制的研究内容进行介绍;其次,对 SE 机制的构造算法、对搜索语句的支持程度以及其被考虑的应用场景进行了分析和讨论.从以上介绍中可以看出:SE 机制的研究逐渐成熟化,将逐渐成为云计算环境下用户对数据密文进行操作的有利工具.未来的一段时间,SE 机制依然将被视为解决云计算中的安全问题的研究热点之一.随着越来越多的数据存储在云端服务器中,以及用户对数据安全和个人隐私的敏感程度越来越强,如何高效、精确且安全地对存储在云端服务器中的密文进行搜索,将是研究者不断研究的方向.我们认为,进一步的研究应重点解决以下问题:

第一,高效率且支持灵活查询语句的 SE 机制是未来重点的研究方向之一.

随着越来越多的用户接受云的概念,同时在云端服务器存储的数据逐渐向海量级别发展,在这种条件下,如何给用户良好的搜索体验,对 SE 机制的性能将是极大的考验.基于对称密码学的 SE 机制在性能上较为优越,而一些基于公钥密码学的 SE 机制虽然能够支持灵活的查询语句,但是由于其要么构造于复合群之上(例如文献[20,30]的工作),要么是构建于多个素数群的直积之上(例如文献[29]的工作),由于公钥密码学的计算开销约是对称密码学算法的 1 000 倍[46],导致了它们的性能效果将难以适用于海量用户和海量数据的应用场景中.虽然一些能够提高群元素间运算速度的硬件[47]已经推出,以及一些并行化技术也能够在一定程度上减轻搜索时所带来的时间等待,但是只有设计出高效的 SE 机制,才是在算法角度上根本性地成为加快现今搜索效率的有效方法.

第二,支持模糊搜索(fuzzy search)和支持按相关性排序的可搜索加密机制依然是未来需要解决的问题.

由于现今的大部分可搜索加密机制都是基于匹配搜索(equality search),因此在密文上实现快速、高效的支持按相关性排序搜索和模糊搜索之上的工作依然处于初始研究阶段.一些工作[13,21,48,49]虽然在一定程度上实现了基于相关性排序和模糊搜索,但是依然存在着一些不足,例如,文献[13]中的方法其存储开销较大且仅支持单词搜索等.

第三,支持关系运算(>,<,==等)的可搜索加密机制依然是未来需要研究的内容.

现有的一些工作虽然能够实现区间查询(range query)和子集查询(subset query),但在支持关系运算的效果上依然不够理想.在未来的一段时间里,支持关系运算的可搜索加密机制的研究依然是研究的一个热点,它的研究也将为加密数据库上的查询提供一些借鉴.

第四,保留语义的 SE 机制依然是研究难点.

加密虽然能够将明文信息隐藏,但同时也破坏了明文中的语义关系,使得用户无法像在明文上通过机器学习等方法使得返回的密文能够越来越符合用户的需求.未来的SE机制研究成果或许能够达到像明文那样,不仅能够实现精确的关键字匹配,还能够洞悉用户的真正搜索请求,然后将最符合的结果返回给用户.

最后,应用于实际场景中的 SE 机制得到业界的更多关注.

SE 机制起源于理论研究,但是由于其实现复杂度高等原因,一直无法在业界得到广泛应用.随着一些应用密码学工作的展开[9,12,13,21],SE 机制渐渐步入实际中的应用.由于云计算的逐步推广,相应的配套技术也将得到研究和部署.相信在未来的一段时间内,SE 机制不仅仅只停留在理论研究框架之内,也将逐步在实际中接受用户的检验.

References:

[1] Chen K, Zheng WM. Cloud computing: System instances and current research. Ruan Jian Xue Bao/Journal of Software, 2009,20(5): 1337−1348 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3493.htm [doi: 10.3724/SP.J.1001.2009.03493]

[2] Feng DG, Zhang M, Zhang Y, Xu Z. Study on cloud computing security. Ruan Jian Xue Bao/Journal of Software, 2011,22(1): 71−83 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3958.htm [doi: 10.3724/SP.J.1001.2011.03958]

[3] Su JS, Cao D, Wang XF, Sun YM, Hu QL. Attribute-Based encryption schemes. Ruan Jian Xue Bao/Journal of Software, 2011, 22(6):1299−1315 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3993.htm [doi: 10.3724/SP.J.1001.2011.03993]

[4] Dropbox. hhtp://www.dropbox.com/

[5] Amazon. Amazon S3. http://aws.amazon.com/s3/

[6] Windows azure. http://www.microsoft.com/windowsazure/

[7] Weber T. Cloud computing after Amazon and Sony: Ready for primetime? 2011. http://www.bbc.co.uk/news/business-13451990

[8] Song D, Wagner D, Perrig A. Practical techniques for searches on encrypted data. In: Proc. of the 2000 IEEE Symp. on Security and Privacy. Berkeley: IEEE Computer Society, 2000. 44−55. [doi: 10.1109/SECPRI.2000.848445]

[9] Waters B, Balfanz D, Durfee G, Smetters D. Building an encrypted and searchable audit log. In: Proc. of the 11th Annual Network and Distributed System Security Symp. San Diego: The Internet Society, 2004. http://www.isoc.org/isoc/conferences/ndss/04/proceedings/

[10] Goh E. Secure Indexes. In: Cryptology ePrint Archive. 2003. http://eprint.iacr.org/2003/216.pdf

[11] Golle P, Staddon J, Waters B. Secure conjunctive keyword search over encrypted data. In: Proc. of the 2nd Int'l Conf. on Applied Cryptography and Network Security (ACNS). Berlin, Heidelberg: Springer-Verlag, 2004. 31−45. [doi: 10.1007/978-3-540-24852-1_3]

[12] Wang C, Cao N, Li J, Ren K, Lou WJ. Secure ranked keyword search over encrypted cloud data. In: Proc. of the IEEE 30th Int'l Conf. on Distributed Computing Systems (ICDCS). Genoa: IEEE Computer Society, 2010. 253−262. [doi: 10.1109/ICDCS.2010.34]

[13] Li J, Wang Q, Wang C, Cao M, Ren K, Lou WJ. Fuzzy keyword search over encrypted data in cloud computing. In: Proc. of the IEEE INFOCOM Mini-Conf. San Diego: IEEE Computer Society, 2010. 1−5. [doi: 10.1109/INFCOM.2010.5462196]

[14] Li M, Yu S, Cao N Lou W. Authorized private keyword search over encrypted data in cloud computing. In: Proc. of the IEEE Int'l Conf. on Distributed Computing Systems (ICDCS). Minneapolis: IEEE Computer Society, 2011. 383−392. [doi: 10.1109/ICDCS.2011.55]

[15] Chang YC, Mitzenmacher M. Privacy preserving keyword searches on remote encrypted data. In: Proc. of the 3rd Int'l Conf. on Applied Cryptography and Network Security (ACNS). Berlin, Heidelberg: Springer-Verlag, 2005. 442–455. [doi: 10.1007/11496137_30]

[16] Boneh D, Crescenzo G, Ostrovsky R, Persiano G. Public key encryption with keyword search. In: Proc. of the EUROCRYPT. Berlin, Heidelberg: Springer-Verlag, 2004. 506–522. [doi: 10.1007/978-3-540-24676-3_30]

[17] Shi E, Bethencourt J, Chan T, Song D, Perrig A. Multi-Dimensional range query over encrypted data. In: Proc. of the IEEE Symp. on Security and Privacy. Berkeley: IEEE Computer Society, 2007. 350–364. [doi: 10.1109/SP.2007.29]

[18] Shi E, Waters B. Delegating capabilities in predicate encryption systems. In: Proc. of the 35th Int'l Colloquium on Automata, Languages and Programming (ICALP). Berlin, Heidelberg: Springer-Verlag, 2008. 560–578. [doi: 10.1007/978-3-540-70583-3_46]

[19] Yang Z, Zhong S, Wright R. Privacy-Preserving queries on encrypted data. In: Proc. of the 11th European Conf. on Research in Computer Security. Berlin, Heidelberg: Springer-Verlag, 2006. 479–495. [doi: 10.1007/11863908_29]

[20] Boneh D, Waters B. Conjunctive, subset, and range queries on encrypted data. In: Proc. of the 4th Conf. on Theory of Cryptography. Berlin, Heidelberg: Springer-Verlag, 2007. 535–554. [doi: 10.1007/978-3-540-70936-7_29]

[21] Cao N, Wang C, Li M, Ren K, Lou W. Privacy-Preserving multi-keyword ranked search over encrypted cloud data. In: Proc. of the IEEE INFOCOM. Shanghai: IEEE Computer Society, 2011. 829–837. [doi: 10.1109/INFCOM.2011.5935306]

[22] Curtmola R, Garay J, Kamara S, Ostrovsky R. Searchable symmetric encryption: Improved definitions and efficient constructions. In: Proc. of the 13th ACM Conf. on Computer and Communications Security (CCS). New York: ACM Press, 2006. 79–88. [doi: 10.1145/1180405.1180417]

[23] Dong C, Russello G, Dulay N. Shared and searchable encrypted data for untrusted servers. In: Proc. of the 22nd Annual IFIP WG 11.3 Working Conf. on Data and Applications Security. Berlin, Heidelberg: Springer-Verlag, 2008. 127–143. [doi: 10.1007/978-3-540-70567-3_10]

[24] Hwang Y, Lee P. Public key encryption with conjunctive keyword search and its extension to a multi-user system. In: Proc. of the Int'l Conf. on Pairing-Based Cryptography. Berlin, Heidelberg: Springer-Verlag, 2007. 2–22. [doi: 10.1007/978-3-540-73489-5_]

[25] Shen E, Shi E, Waters B. Predicate privacy in encryption systems. In: Proc. of the 6th Theory of Cryptography Conf. on Theory of Cryptography. Berlin, Heidelberg: Springer-Verlag, 2009. 4570–473. [doi: 10.1007/978-3-642-00457-5_27]

[26] Ballard J, Kamara S, Monrose F. Achieving efficient conjunctive keyword searches over encrypted data. In: Proc. of the 7th Int'l Conf. on Information and Communications Security. Berlin, Heidelberg: Springer-Verlag, 2005. 414–426. [doi: 10.1007/11602897_35]

[27] Baek J, Safavi-Naini R, Susilo W. Public key encryption with keyword search revisited. In: Proc. of the Int'l Conf. on Computational Science and Its Applications. Berlin, Heidelberg: Springer-Verlag, 2008. 1249–1259. [doi: 10.1007/978-3-540-69839-5_96]

[28] Bao F, Deng R, Ding X, Yang Y. Private query on encrypted data in multi-user settings. In: Proc of the 4th Int'l Conf. on Information Security Practice and Experience. Berlin, Heidelberg: Springer-Verlag, 2008. 71–85. [doi: 10.1007/978-3-540-79104-1_6]

[29] Okamoto T, Takashima W. Hierarchical predicate encryption for inner-products. In: Proc. of the ASIACRYPT. Berlin, Heidelberg: Springer-Verlag, 2009. 214–231. [doi: 10.1007/978-3-642-10366-7_13]

[30] Katz J, Sahai A, Waters B. Predicate encryption supporting disjunctions, polynomial equations, and inner products. In: Proc. of the EUROCRYPT. Berlin, Heidelberg: Springer-Verlag, 2008. 146–162. [doi: 10.1007/978-3-540-78967-3_9]

[31] Goldreich O, Ostrovsky R. Software protection and simulation on oblivious RAMs. Journal of the ACM, 1996,43(3):431–473. [doi: 10.1145/233551.233553]

[32] Cao N, Yang Z, Wang C, Ren K, Lou W. Privacy-Preserving query over encrypted graph-structured data in cloud computing. In: Proc. of the IEEE Int'l Conf. on Distributed Computing Systems (ICDCS). Minneapolis: IEEE Computer Society, 2011. 393–402. [doi: 10.1109/SP.2007.11]

[33] Goldreich O, Ostrovsky R. Software protection and simulations on oblivious RAMs [Ph.D. Thesis]. MIT, 1992.

[34] Boneh D, Kushilevitz E, Ostrovsky R, Skeith W. Public key encryption that allows PIR queries. In: Proc. of the 27th Annual Int'l Cryptology Conf. on Advances in Cryptology. Berlin, Heidelberg: Springer-Verlag, 2007. 50–67. [doi: 10.1007/978-3-540-74143-5_4]

[35] Kamara S, Lauter K. Cryptographic cloud storage. In: Proc. of the 14th Int'l Conf. on Financial Cryptograpy and Data Security. Berlin, Heidelberg: Springer-Verlag, 2010. 136–149. [doi: 10.1007/978-3-642-14992-4_13]

[36] WIKIPEDIA. http://en.wikipedia.org/wiki/Symmetric-key_algorithm

[37] Bloom BH. Space/Time trade-offs in hash coding with allowable errors. Communications of the ACM, 1970,13(7):422−426. [doi: 10.1145/362686.362692]

[38] Agrawal R, Kiernan J, Srikant R, Xu Y. Order preserving encryption for numeric data. In: Proc. of the ACM SIGMOD. New York: ACM Press, 2004. 563−574. [doi: 10.1145/1007568.1007632]

[39] Wong KK, Cheung DW, Kao B, Mamoulis N. Secure *k*NN computation on encrypted databases. In: Proc. of the 35th SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2009. 139−152. [doi: 10.1145/1559845.1559862]

[40] Bethencourt J, Sahai A, Waters B. Ciphertext-Policy attribute-based encryption. In: Proc. of the IEEE Symp. on Security and Privacy. Berkeley: IEEE Computer Society, 2007. 321−334. [doi: 10.1109/SP.2007.11]

[41] Goyal V, Pandey O, Sahai A, Waters B. Attribute-Based encryption for fine-grained access control of encrypted data. In: Proc. of the ACM Conf. on Computer and Communications Security. New York: ACM Press, 2006. 89−98. [doi: 10.1145/1180405.1180418]

[42] Boneh D, Franklin M. Identity-Based encryption from the weil pairing. In: Proc. of the Advances in Cryptology-CRYPTO. Berlin, Heidelberg: Springer-Verlag, 2001. 213−229. [doi: 10.1007/3-540-44647-8_13]

[43] Reeman D. Converting pairing-based cryptosystems from composite-order groups to prime-order groups. In: Proc. of the EUROCRYPT. Berlin, Heidelberg: Springer-Verlag, 2010. 44−61. [doi: 10.1007/978-3-642-13190-5_3]

[44] Zhang Y, Xue C, Wong D, Mamoulis N, Yiu S. Acceleration of composite order bilinear pairing on graphics hardware. IACR Cryptology ePrint Archive. 2011. http://eprint.iacr.org/2011/196.pdf

[45] The Java pairing based cryptography library (jPBC). 2011. http://gas.dia.unisa.it/projects/jpbc/

[46] Salama D, Minaam A, Abdual-Kader H, Hadhoud M. Evaluating the effects of symmetric cryptography algorithms on power consumption for different data types. Int'l Journal of Network Security, 2010,11(2):78−87.

[47] The Elliptic Semiconductor (clp-17). High performance elliptic curvecryptography point multiplier core. http://www.internetsociety.org/privacy-preserving-logarithmic-time-search-encrypted-data-cloud

[48] Wang C, Ren K, Yu S, Urs K.Achieving usable and privacy-assured similarity search over outsource cloud data. In: Proc. of the IEEE INFOCOM. Orlando: IEEE Computer Society, 2012. 451−459. [doi: 10.1109/INFCOM.2012.6195784]

[49] Shen Z, Xue W, Shu J. Preferred keyword search over encrypted data in cloud computing. In: Proc. of the ACM/IEEE IWQoS. Montreal: IEEE Computer Society, 2013. 1−6. [doi: 10.1109/IWQoS.2013.6550283]

附中文参考文献:

[1] 陈康,郑纬民.云计算:系统实例与研究现状.软件学报,2009,20(5):1337−1348. http://www.jos.org.cn/1000-9825/3493.htm [doi: 10.3724/SP.J.1001.2009.03493]

[2] 冯登国,张敏,张妍,徐震.云计算安全研究.软件学报,2011,22(1):71−83. http://www.jos.org.cn/1000-9825/3958.htm [doi: 10.3724/SP.J.1001.2011.03958]

[3] 苏金树,曹丹,王小峰,孙一品,胡乔林.属性基加密机制.软件学报,2011,22(6):1299−1315. http://www.jos.org.cn/1000-9825/3993.htm [doi: 10.3724/SP.J.1001.2011.03993]

沈志荣(1987－),男,福建三明人,博士生,主要研究领域为云存储中的数据可靠性,数据安全和隐私.

E-mail: czr10@mails.tsinghua.edu.cn

薛巍(1974－),男,博士,副教授,CCF 高级会员,主要研究领域为电力系统分析模拟,并行算法设计,集群计算.

E-mail: xuewei@tsinghua.edu.cn

舒继武(1968－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络/云存储,存储安全与可靠性,大数据存储,并行处理技术.

E-mail: shujw@tsinghua.edu.cn