软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(6):1577−1591 [doi: 10.13328/j.cnki.jos.005005]

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

对加密电子医疗记录有效的连接关键词的搜索*

张丽丽[1], 张玉清[1,2], 刘雪峰[1], 全韩彧[1]

[1](综合业务网理论及关键技术国家重点实验室(西安电子科技大学),陕西 西安 710071)

[2](国家计算机网络入侵防范中心(中国科学院大学),北京 101408)

通信作者: 刘雪峰, E-mail: liuxf@nipc.org.cn

摘 要: 随着云计算的发展,医院或医疗组织为了节省存储资源将加密的电子医疗记录的存储和管理外包给云服务器.尽管加密有助于保护用户数据的机密性,但是对加密的数据执行安全而有效的搜索是一个挑战性的问题.首先构造了被称为 MCKS_I 的简单的多域连接关键词搜索(MCKS)方案,该方案仅支持连接相等查询.为了实现更加灵活而复杂的多域关键词连接查询,例如子集查询和范围查询,又提出了被称为 MCKS_II 的提高方案.该方案利用了分层属性的矢量表示方法.这两个方案被证明能够抵抗已知明文攻击.大量的分析和实验数据表明,该方案有效实用.

关键词: 云计算;电子医疗记录;连接关键词搜索;谓词加密;非对称的向量积保持加密(ASPE)

中图法分类号: TP309

中文引用格式: 张丽丽,张玉清,刘雪峰,全韩彧.对加密电子医疗记录有效的连接关键词的搜索.软件学报,2016,27(6):1577−1591. http://www.jos.org.cn/1000-9825/5005.htm

英文引用格式: Zhang LL, Zhang YQ, Liu XF, Quan HY. Efficient conjunctive keyword search over encrypted electronic medical records. Ruan Jian Xue Bao/Journal of Software, 2016,27(6):1577−1591 (in Chinese). http://www.jos.org.cn/1000-9825/5005.htm

Efficient Conjunctive Keyword Search over Encrypted Electronic Medical Records

ZHANG Li-Li[1], ZHANG Yu-Qing[1,2], LIU Xue-Feng[1], QUAN Han-Yu[1]

[1](National Key Laboratory of Integrated Services Networks (Xidian University), Xi'an 710071, China)

[2](National Computer Network Intrusion Protection Center (University of Chinese Academy of Sciences), Beijing 101408, China)

Abstract: With advances in cloud computing, hospitals and healthcare organizations can outsource the storage and management of their encrypted electronic medical records (EMRs) to the cloud services for great flexibility and economic savings. Although encryption helps protecting user data confidentiality, designing secure and practically efficient search functions over encrypted data remains challenging problem. This paper first constructs a multi-field conjunctive keyword search (MCKS) scheme, called MCKS_I, which supports equality query. For more flexible and complex MCKS, such as subset and range query, it then proposes an improved scheme, MCKS_II, based on a novel vector representation of hierarchical attributes. The new schemes are proven to be able to resist known plaintext attack. Extensive analysis and experiments show that the proposed schemes are extremely practical.

Key words: cloud computing; electronic medical record (EMR); conjunctive keyword search; predicate encryption; asymmetric scalar-product preserving encryption (ASPE)

随着信息技术的快速发展,云计算作为一种推动信息技术实现按需供给、促进信息技术和数据资源得到充分利用的计算模式,受到越来越多人的关注.云计算等外包服务器是由第三方互联网商业公司(如亚马逊、谷歌

* 基金项目: 国家自然科学基金(61272481, 61402352, 61370220)
Foundation item: National Natural Science Foundation of China (61272481, 61402352, 61370220)
收稿时间: 2015-08-15; 修改时间: 2015-10-09; 采用时间: 2015-12-05; jos 在线出版时间: 2016-01-21
CNKI 网络优先出版: 2016-01-22 11:20:06, http://www.cnki.net/kcms/detail/11.2560.TP.20160122.1120.015.html

等)维护与管理的.尽管云计算可以提供强大的存储和计算能力,但是由于用户与第三方商业公司通常不属于同一个信任域,很多公司和组织对云计算持谨慎态度,最主要的原因是他们担心失去对数据安全和用户隐私的控制.

近几年来,随着信息医疗化的不断发展,电子医疗记录(EMR)数量越来越大,怎样对日益增大的EMR执行有效的存储和管理是很多医院和医疗组织面临的需要解决的问题.具有强大存储和计算能力的云计算[1]等外包服务提供了潜在的解决方案.通过将病人的EMR和详细的病例外包给云服务器,医院和医疗组织极大地减少了管理和存储负担.例如,医院将表 1 以及表中每条记录对应的具体的个人病例外包给云服务器.表中的每一行指一个病人的属性记录,例如年龄、性别、疾病和地区.这些属性都是多值的数值属性或非数值属性.例如,年龄是属性值为 1~100 的数值属性,性别是属性值为男和女的非数值属性.每一个属性称为关键词域,而属性值称为关键词.EMR 的属性涉及病人的隐私,为了保护病人的隐私安全,EMR 以及相应的详细的病人病例通常都是以密文形式存储在云服务器上.这给用户搜索包含某些关键词的 EMR 操作带来了极大的不便.例如,一个授权的用户为了实现信息统计的目的提交查询"(40<年龄<50) AND (性别=女) AND (疾病=糖尿病)".因此,在确保数据安全和病人隐私的情况下,开发一个针对加密的EMR的快速而且精确的连接关键词搜索方案,是一个尤其需要解决的问题.可搜索加密(searchable encryption,简称 SE)[2]是近年来发展的一种支持用户在密文上进行关键词查询的密码学原语.

Table 1 An example of EMR

表 1 一个 EMR 的例子

ID	性别	年龄	疾病	地区
1	男	60	糖尿病	北京
2	女	45	心脏病	上海
3	男	70	皮肤病	郑州
4	男	52	肠炎	北京
5	女	30	肺癌	沈阳
…	…	…	…	…

然而,在现有的 SE 方案中,服务器或者返回不精确的查询结果,或者花费巨大的计算开销.一般来说,基于用户查询中包含的关键词,SE 可以分为单关键词查询,多关键词连接查询和多关键词排序查询.

因为单关键词 SE 方案[3−10]只支持单个关键词的查询,所以他们通过一次查询不能完成多关键词的连接查询.针对单关键词 SE 查询方案的不足,简单的多关键词搜索方案已经提出[11−16].然而这些方案只支持简单的连接关键词相等查询,例如"(年龄=60) AND (疾病=糖尿病)".为了完善搜索功能,支持范围和子集查询的多域上的连接关键词搜索方案在文献[17–22]中提出,在这些方案中,一般的方法是在复杂的代数结构,例如复合阶群或者双对矢量空间(dual pair vector space,简称 DPVS)上构造矢量.由于在复杂代数结构上双线性对的计算开销比在单阶群上的计算开销要大的多,因此这类方案的计算开销成了一个最大的问题.基于内积相似分数或者欧几里德距离的多关键词排序搜索是另外一类连接关键词搜索[23–27].为了增加安全性,这些方案给相似分数或距离引入一些随机变量,从而引起了结果的不精确性.而且,这一类查询仍然是多关键词连接相等查询,只是在返回的结果中增加了排序功能.

构造支持多域上的连接关键词搜索的加密方案仍是一个开放的问题.针对该问题,本文有以下几点贡献.

(1) 我们定义了基于对称密码的谓词加密,提出了基本的和分层的两种属性和查询矢量的表示方法.

(2) 利用谓词加密模式,我们构造了两个 MCKS 方案,方案 MCKS_I基于基本的属性和查询矢量的表示方法,仅支持多关键词连接相等查询.由于采用分层属性的矢量表示方法,方案 MCKS_II支持具有复杂查询结构的连接关键词查询,能实现范围查询和子集查询,而且这个方案实现比方案 MCKS_I更低的计算开销.经证明这两个方案都能抵抗已知明文攻击.

(3) 我们对这两个方案做了详细的性能模拟,结果显示我们的方案比其他方案具有更高的实用性.

本文第 1 节讨论单关键词和多关键词搜索的相关工作.第 2 节介绍系统体系结构、设计目标及攻击模型.第 3 节详细介绍两个 MCKS 方案并进行安全分析.第 4 节对提出的两个方案进行性能模拟.第 5 节总结了全文.

1 相关工作

SE 机制按照构造算法的不同可以分为两类,即基于对称密码算法(symmetric key cryptography,简称 SKC)的 SE 机制和基于公钥密码算法(public key cryptography,简称 PKC)的 SE 机制.按照查询中包含的关键词数目,可以分为支持单关键词搜索的 SE 机制和支持多关键词搜索的 SE 机制.

1.1 单关键词搜索

Song 等人[3]首次研究了基于 SKC 的单关键词搜索方案.继文献[3]之后,文献[4–6]也研究了单关键词的搜索.后来一些研究[7–9]按照关键词出现的频率实现了具有排序功能的单关键词搜索.Boneh 等人[10]首次提出了基于 PKC 的单关键词的搜索方案.该论文中,他们研究了利用公钥加密系统解决对加密数据的搜索问题,为利用公钥密码实现更多样化的 SE 机制提供了指导意义.支持单关键词的 SE 机制不能满足多关键词搜索的要求.

1.2 多关键词连接搜索

针对单关键词搜索机制的不足,Golle 等人[11]首次提出了基于 SKC 的支持连接关键词相等搜索的 SE 机制,该文中的第 1 个方案产生了大量的计算开销;第 2 个方案减小了通信开销,然而他们是基于双线性对操作,因而实用性较差.Ballard 等人[12]分别基于 SKC 和 PKC 构建了两个方案.其中基于 SKC 的方案实现了很高的效率,然而这个方案的陷门与文件的数量成线性关系.Park 等人[14]以及 Hwang 等人[15]基于 PKC 实现了连接关键词相等搜索.这些方案可以用于像邮件路由系统这样的应用场所,其中邮件内容和关键词用公钥加密.由于是基于双线性对操作的,这些系统产生了很大的计算开销.上述方案均基于数据源集中化假设,即由单一数据源集中创建可搜索索引.2015 年,Liu 等人[28]为多数据源的场景设计了基于 SKC 的连接关键词相等搜索算法,允许各数据源以分散的方式生成索引.该算法成功地保护了数据文件和检索结果的隐私,却泄露了数据源数目、数据文件数目以及访问模式和搜索模式等信息.

为了丰富查询功能,Boneh 等人[18]基于 PKC 提出了支持范围查询和子集查询的复杂查询.Shi 等人[29]也基于 PKC 提出了针对加密数据的多维的范围查询,用于解决有关网络审计日志的分享的隐私问题以及其他的一些应用.在他们的方案中,每个域上都支持范围查询.然而他们是基于双线性对的,在加解密时产生较大的计算开销.

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.Katz 等人[19]基于 PKC 提出了支持或,多项式方程以及内积的谓词加密.主要思想是在复合阶群上构造支持合取范式(conjunctive normal formulation,简称 CNF),析取范式(disjunctive normal formulation,简称 DNF)以及多项式方程的矢量.在复合阶群上构造的双线性对的计算开销是在素数阶群上的 50 倍.Attrapadung 等人[20]通过牺牲属性隐私的方法提高了文献[19]的效率.Li 等人[17]基于 PKC 利用分层的谓词加密(hierarchical predicate encryption,简称 HPE[30])解决了在加密的个人健康记录上具有隐私保护的关键词搜索问题.这个方案实现了范围查询和子集查询.文献[17]中的方案可以实现文献[30]中的方案的安全.文献[30]构建的 HPE 是基于 DPVS 的.在 DPVS 上的双线性对的计算与 n 阶素数阶群的计算量相当.基于 PKC 的谓词加密方案有一个共同的缺点就是不能保护查询隐私.也就是说,云服务器可以了解用户正在查询的内容.这是因为云服务器可以利用公钥加密各种属性值对应的信息,之后对用户提交的查询陷门发起词典攻击.

Shen 等人发现基于 PKC 的谓词加密固有的缺点后,提出了基于 SKC 的支持多关键词连接相等查询的谓词加密方案[16].尽管该内积谓词加密可以扩展为范围谓词加密,但是需要花费相当大的代价.Lu 等人[22]扩展了文献[16]中方案的功能并支持范围查询,并且提高了搜索效率.

1.3 多关键词排序搜索

Cao 等人[23]第一次提出了多关键词排序搜索方案.云服务器根据查询中的多个关键词计算与每一个文档的相似分数,返回相似分数最高的 k 个文档.然而搜索复杂度与数据库中的文档数成线性关系.Yang 等人[24]大大提高了搜索效率.以后,Sun 等人[25,26]提出了基于 MDB 树的搜索方案,这类方案的搜索速度比线性搜索大很多.

然而在这些查询方案中,在服务器返回的查询结果中引入了不精确性.Strizhov 等人[31]克服了引入不确定性的缺点,提出多关键词相似性加密可搜索方案 MKSim,该方案的检索时间与文档总数之间存在亚线性关系.然而这类查询仍然是多关键词连接相等查询,只是对返回的结果进行了排序.

1.4 其他类型的搜索加密

模糊关键词查询[32−36]已经被提出.这些方案采用了拼写检查机制,例如代替错误拼写的“wireless”而搜索“wireless”,或者数据格式可能不一样,例如“datamining”和“data-mining”.Li 等人[37]提出了一种查询方案,不仅能返回精确的匹配文档,还能返回包含语义上与关键词相关的词语的文档.这与我们研究的场景有点远.

从现有的相关工作可以看出,基于 SKC 的多关键词查询方案一般只能实现查询结构比较简单的连接关键词查询.相比基于 SKC 的 SE 方案,基于 PKC 的 SE 方案可以实现具有更加复杂查询结构的连接关键词查询,例如实现范围查询,子集查询以及“或”的查询.然而相比基于 SKC 的 SE 机制来说,他们的效率更低.因为基于 PKC 的 SE 机制中一般需要产生许多耗时的对操作.我们希望基于 SKC 构造一种既能实现高效查询,又能实现范围查询,子集查询的多关键词连接查询机制.

2 问题描述

2.1 系统体系结构

本文中,系统体系结构包括 3 个实体:数据拥有者、云服务器和数据用户,如图 1 所示.数据拥有者是一类特殊的用户,例如医院或者医疗机构.用户通常指那些可以对加密的数据执行搜索的人或组织.云服务器存储加密的数据库并根据用户的查询需求执行特定的查询.一般地,操作步骤如下:

(1) 数据拥有者加密每一个病人的电子医疗记录及对应的详细的病人病例并把它们外包给云服务器.

(2) 数据用户向数据拥有者请求搜索授权,数据拥有者根据授权规则决定是否授权,假如用户获得授权,数据拥有者执行第 3 步.

(3) 数据拥有者向数据用户分发一个搜索陷门和解密秘钥.这两项分别可用于加密的 EMRs 的搜索和匹配记录的解密.

(4) 用户提交搜索陷门给云服务器,云服务器执行匹配的查询.搜索陷门不揭露查询的任何信息.

(5) 云服务器返回满足特定搜索规则的记录,解密秘钥应该能实现匹配记录的解密.

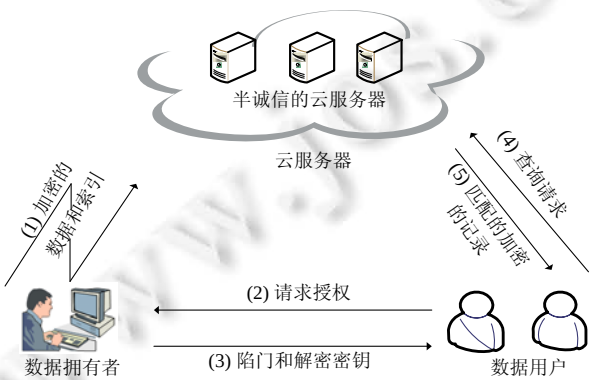


Fig.1 System architecture

图 1 系统体系结构

我们假定云服务器是半诚实的,也就是说,云服务器会按照协议规定的方式执行协议,但同时也会分析存储在它上面的信息和协议中接收到信息流,以了解更多的信息.

2.2 攻击模型

根据云服务器拥有的信息,下面定义 3 种攻击模型.

唯密文攻击(ciphertext only attack,简称 COA).在这种攻击模型中,云服务器拥有加密的病人病例文档以及这些病例所对应的 EMRs(这两项都是来自数据拥有者)以及加密的查询矢量,也即陷门(来自用户).

已知明文样本的攻击(known sample attack,简称 KSA).在这种攻击模型中,除了拥有 COA 模型中的信息外,云服务器还拥有 EMRs 中一些记录的明文信息,然而不知道他们对应的密文信息.

已知明文攻击(known plaintext attack,简称 KPA).在这种攻击模型中,除了拥有 COA 模型中的信息外,云服务器还知道 EMRs 中一些记录的明文信息及其所对应的密文信息.

这 3 种攻击级别是依次升高的,一个更高级别的攻击比更低级别的攻击更强大,假如一个加密方案可以抵御更高级别的攻击,那么它也能抵御更低级别的攻击.在这 3 种攻击模型中,在实际中,第 2 级别,即 KSA 较常见.注意 KPA 模型在实际中比较少,因为对没有加密秘钥的人来说,很难将一些记录的明文与相应的密文联系起来.

2.3 设计目标

我们的 MCKS 方案应该能实现下列主要的安全及性能目标.

(1) 多域上的连接关键词查询.设计的系统应该能支持针对加密的 EMRs 的多域上的连接关键词搜索.在实际应用中有很多这种类型的查询,如病人匹配,关联规则的挖掘等.

(2) 属性和查询隐私安全.最基本的安全目标是,除了从搜索结果中获得信息外,云服务器不能得到有关 EMR 属性以及用户查询的关键词信息.

(3) 查询无连接性(query unlinkability).除了授权的用户,其他人(包括云服务器)不能从以前的搜索陷门产生新的搜索陷门.而且两个同样的查询每次应该生成不同的搜索陷门.并且,云服务器不能推断出陷门之间的关系.

(4) 搜索模式隐藏.根据文献[6]的定义,一个用户的搜索模式是指云服务器获得了两个搜索是否是由同样的关键词生成从而获得的信息.具体地说,一个用户的搜索模式是指,云服务器在知道“两个搜索是由相同的关键词产生”或者“两个搜索不是由相同的关键词产生”的前提下而获得的任何知识.显然,实现搜索模式隐藏的基本方法是在陷门产生过程中引入不确定性.

(5) 效率.上面提到的功能以及安全目标均应该以较低的计算和通信开销完成.

2.4 定义和注释

我们非正式的定义基于 SKC 的谓词加密,采用文献[17]中定义的一般框架.假定Σ表示二进制矢量的有限集合,$\mathcal{F}$表示谓词集合.Σ和$\mathcal{F}$可能依赖安全参数.一个基于 SKC 的谓词加密方案由 4 种 PPT 算法 Setup, GenKey,Enc 和 Dec 组成.

Setup: 以安全参数 1^n 作为输入,输出私钥 SK.

Enc: 以私钥 SK,属性 $I\in\Sigma$和消息空间中的消息 M 作为输入,输出密文 C,记做 $C\leftarrow Enc_{sk}(I,M)$.

GenKey: 以私钥 SK 和一个谓词 $f\in\mathcal{F}$作为输入,输出一个秘钥,也就是陷门,简记为 TD_f.

Dec: 以秘钥 TD_f和密文 C 作为输入,输出消息 M 或者是不可辨识的符号$\perp$.对于解密的正确性,对所有的 n,由 **Setup**(1^n)产生的 $SK,f\in\mathcal{F},TD_f\leftarrow Genkey_{sk}(f)$ 及所有的 $I\in\Sigma$,**Dec** 都需要满足:

假如 $f(I)$=1,那么 $Dec_{TD_f}(Enc_{SK}(M,I))=M$;

假如 $f(I)$=0,那么 $Des_{TD_f}(Enc_{SK}(M,I))=\perp$.

本文中,我们考虑上述定义的变体,称为仅谓词(predicate only)方案.仅谓词方案加密时只将属性 I 作为输入,而消息 M 不作为输入.解密时只输出一个标识,标识被查询的文档与查询是否匹配.正确性需要满足: $Dec_{TD_f}(Enc_{SK}(I))=f(I)$. 我们考虑一类谓词是函数 $\mathcal{F}:\{f_x\mid x\in\Sigma\}$, 当且仅当$\langle x,y\rangle$=0 时,$f_x(y)$=1($\langle x,y\rangle$表示两个 n 位二进制矢量 x 和 y 的内积 $\sum_{i=1}^{n}x_i.y_i$; Σ是一个根据特定的矢量表示方法产生的二进制矢量集合(见第 3.1.1 节)).

为了方便表达,我们在 MCKS_I方案中引入了一些主要的注释.

R:各个病人病例的集合.假定有 m 个病人,记为 $R=\{R_1,R_2,...,R_m\}$.这里是指 EMR 中每条记录对应的病人的病例.

W:从各个病人的病例中抽取的关键词的集合,假定有 k 个关键词,记为 $W=\{W_1,W_2,...,W_k\}$.

$\overrightarrow{I_i}$:根据 EMRs 的第 i 行(第 i 个记录)生成的二进制属性矢量(相应于病例 R_i),也称索引矢量.

I_i':是 $\overrightarrow{I_i}$ 的密文矢量(i=1,…,m).

q:一个包含查询关键词的查询表达式.

$\vec{q}$:根据 q 生成的 n 位二进制查询矢量.

$\overrightarrow{q'}$:一个 n 位全部设置为 1 的二进制矢量,称为查询矢量 q 的参考二进制矢量.

$\vec{Q}$:一个 n 位的二进制矢量,也称为 $\overrightarrow{q'}$ 和 $\vec{q}$ 的差分矢量,$\vec{Q}=\overrightarrow{q'}-\vec{q}$.

3 具有隐私保护的高效 MCKS 方案

本节将描述属性和查询矢量的表示方法,之后详细描述 MCKS 方案的构造.

3.1 MCKS_I方案

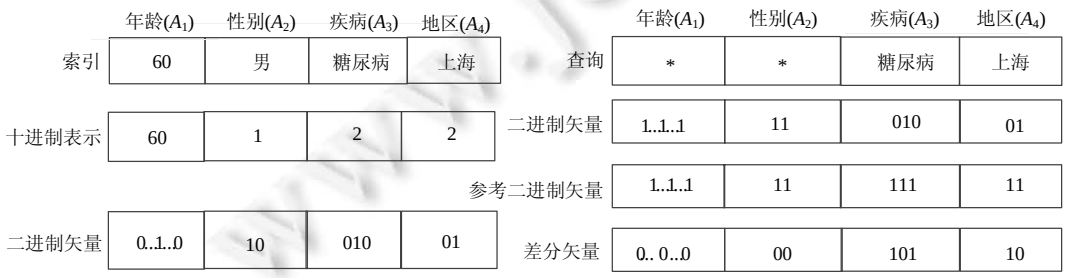
3.1.1 基本的属性矢量和查询矢量表示

本文考虑的场景中,EMR 中每条记录都有固定的属性,每个属性都有固定的属性值.这些属性需要用矢量表示出来,下面介绍属性的矢量表示.

假定某一个属性有 m 个属性值.属性矢量按下述方法表示:每一个属性值被按顺序分配[1,m]范围内一个整数值,那么这 m 个整数中的每一个都被表示成 m 位的二进制矢量.对于整数 i,二进制矢量是第 i 位为 1,其余 m–1 位都为 0.假如整个 EMR 中总共有 n 个属性值(也称关键词),那么 EMR 的每一条记录都可以表示成 n 位的二进制矢量,1 表示相应的关键词存在.

根据查询请求中包含的关键词,可以按下述方法产生查询矢量,对用户不关注的属性,相应位全部设置为 1,其余位按照属性矢量表示方法表示.

我们通过一个例子详细说明基本的矢量的表示方法.假定 EMR 总共有 4 个属性{年龄,性别,疾病,地区},年龄有 100 个属性值{1,2,…,100},性别有两个属性值{男,女},疾病有 3 个属性值{心脏病,糖尿病,皮肤病},地区有两个属性值{北京,上海}.那么 EMRs 中共有 107 个属性值,因此每个索引和查询矢量都可以表示为 107 位的二进制矢量,每一位依次对应于关键词{1,2,…,100,男,女,心脏病,糖尿病,皮肤病,北京,上海},对于 EMR 中的一条具有属性值{60,男,糖尿病,上海}的记录,属性矢量被表示为 0…1…01001001,其中对应年龄的前 100 位中只有第 60 位是 1,其余位都是 0.如图 2(a)表示.



(a) 索引的矢量表示 (b) 查询的矢量表示

Fig.2 Representations of index and query vector. Note: “ * ” stands for “don’t care”

图 2 索引和查询矢量的表示.注释“*”表示不关心

假定一个授权的医疗研究者提交查询请求“(疾病=糖尿病) AND (地区=上海)”,在这个查询请求中,该用户

不关心年龄和性别,因此将该查询矢量的相应于年龄和性别的 102 位全部设置为 1,那么查询矢量 $\vec{q}$ 将被表示为 1…1…101001.设定一个 107 位全部为 1 的 $\vec{q}$ 的参考矢量 $\vec{q'}$, 即 $\vec{q'}=1...1...1$, 我们可以很容易的获得 $\vec{q}$ 和 $\vec{q'}$, 的差分矢量 $\vec{Q}=\vec{q'}-\vec{q}=0...0...010110$. 该查询的矢量如图 2(b)表示.

3.1.2 支持 MCKS 的谓词加密方案

根据图 2(a),任何包含属性值{糖尿病,上海}的记录都可以表示为后面 5 位是 01001 的矢量,那么可以获得一个重要的结论:任何包含属性值{糖尿病,上海}的记录的属性矢量和图 2(b)的差分矢量 $\vec{Q}$ 的内积为 0.在实际中,EMRs 中有更多的属性和属性值.由于采用相同的矢量表示方法,上述结论成立.

根据这个发现,构造支持 MCKS 的加密方案的思想是当属性矢量和差分矢量内积为 0 时利用内积保持的加密.文献[38]中提出的 ASPE 方案可以满足内积保持要求.现在我们对 ASPE 做个简要的描述:$E_T()$和 $E_Q()$分别是属性矢量和查询矢量的加密算法.假定 I_i' 和 q'分别是 $\vec{I_i}$ 和 $\vec{q}$ 的加密矢量.即 $I_i'=E_T(\vec{I_i},k)=M^T\vec{I_i}, q'=E_Q(\vec{q},k)=M^{-1}\vec{q}$, 其中,$M$ 是 $n\times n$ 的可逆矩阵.ASPE 保持了 $\vec{I_i}$ 和 $\vec{q}$ 的内积.即 $I_i'.q'=\vec{I_i}^T MM^{-1}\vec{q}=\vec{I_i}^T\vec{q}=\vec{I_i}.\vec{q}$. 其中,“.”表示内积.假如差分矢量 $\vec{Q}$ 采用查询矢量 $\vec{q}$ 的加密方式,那么 ASPE 将保持属性矢量 $\vec{I_i}$ 和差分矢量 $\vec{Q}$ 的内积.

然而上述 ASPE 不能抵御已知明文攻击.假定一个攻击者知道明文 $I=\{\vec{I_1},\vec{I_2},...,\vec{I_n}\}$ 和相应的密文 I_i' ($i\in[1,n]$).这里 $\vec{I_i}$ 是属性矢量.攻击者可以通过建立下列方程组求出 M^T.假如 A,B 是 $n\times n$ 的矩阵满足 $A=(\vec{I_1},\vec{I_2},...,\vec{I_n})$ 和 $B=(I_1',I_2',...,I_n')$. 攻击者可以得到下列式子 $M^T A=B$.这里,A 是可逆矩阵,因为各个 $\vec{I_i}$ 是线性独立的.因此攻击者可以计算出 $M^T=BA^{-1}$,从而能恢复出 EMRs 中任何记录的属性矢量.

ASPE 的脆弱之处在于,给定 EMRs 中 d 个属性矢量($d\geqslant n$),已知明文攻击者可以建立 n 个方程从而求出 M^T 中的各个未知数.针对这个不足,我们利用 p 划分(splitting)和 q 划分[38].属性矢量 $\vec{I_i}$ 划分为 $\vec{I_{ia}}$ 和 $\vec{I_{ib}}$, 同时差分矢量 $\vec{Q}$ 被划分为 $\vec{Q_a}$ 和 $\vec{Q_b}$, 利用两个可逆矩阵 M_1 和 M_2,我们加密 $\vec{I_{ia}}$ 和 $\vec{I_{ib}}$ 得到 I_{ia}' 和 I_{ib}', 即 $I_{ia}'=M_1^T\vec{I_{ia}}, I_{ib}'=M_2^T\vec{I_{ib}}$, 同时加密 $\vec{Q_a}$ 和 $\vec{Q_b}$ 得到 Q_a' 和 Q_b', 即, $Q_a'=M_1^{-1}\vec{Q_a}, Q_b'=M_2^{-1}\vec{Q_b}$. 下列等式成立:

$$I_i'.Q'=(M_1^T\vec{I_{ia}},M_2^T\vec{I_{ib}}).(M_1^{-1}\vec{Q_a},M_2^{-1}\vec{Q_b})=\vec{I_i}.\vec{Q} \quad (1)$$

很显然,服务器可以利用上述方法在加密的数据库中执行 MCKS,满足等式(1)的属性矢量即是满足查询请求的属性矢量.然而等式(1)无论 EMR 的属性矢量和查询差分矢量的内积是否为 0 时都满足内积保持的性质.所以利用等式(1)云服务器可以得到很多信息.

因为 $\vec{I_i}.\vec{Q}$ 等于关键词不匹配的数目,所以云服务器很容易地从 $I_i'.Q'$ 的结果中推断关键词不匹配的数目,为了解决这个问题,在加密 $\vec{I_i}$ 之前,乘以用一个不等于 0 的随机值 ε_i.为了实现搜索模式的隐藏,在产生陷门时,即对差分矢量加密时,采用不确定性加密.也就是,在加密之前,我们用一个随机变量 β_i 乘以差分矢量 $\vec{Q_i}$. 通过引入更多的不确定性到加密的属性矢量和查询矢量的内积使云服务器不能获取额外的信息.基于以上分析,我们利用 Predicate only 加密构建一个 MCKS 方案.下面我们详细描述这个方案.

Setup:在初始化阶段,数据拥有者随机产生了 n 位二进制矢量 S 和两个 $n\times n$ 的可逆矩阵$\{M_1,M_2\}$,私钥 SK 是一个三元组$\{S,M_1,M_2\}$.

Enc(sk, $\vec{I_i}$):数据拥有者首先根据属性矢量产生方法为 EMR 中的每条记录产生一个属性矢量 $\vec{I_i}$,这里 $\vec{I_i}[j]\in\{0,1\}$表示相应位的关键词是否存在.加密过程如下:

(1) 用一个随机值ε_i 乘以矢量 $\vec{I_i}$,得到 $\hat{I}_i$,即 $\hat{I}_i=\varepsilon_i.\vec{I_i}$, 这里$\varepsilon_i\neq 0$.

(2) 对 $\hat{I}_i$ 执行 p 划分.对于 j=1 到 n,假如 $s[j]$=1,把 $\hat{I}_i[j]$ 随机划分成 $\widehat{I_{ia}}[j]$ 和 $\widehat{I_{ib}}[j]$ 满足 $\widehat{I_{ia}}[j]+\widehat{I_{ib}}[j]=\hat{I}_i[j]$; 假如 $s[j]$=0,设置 $\widehat{I_{ia}}[j]$ 和 $\widehat{I_{ib}}[j]$ 都等于 $\hat{I}_i[j]$.

(3) 产生加密的属性矢量 I_i', $I_i'=\{M_1^T\widehat{I_{ia}},M_2^T\widehat{I_{ib}}\}$.

Genkey$_{SK}$($\vec{Q}$):其中, $\vec{Q}=\vec{q'}-\vec{q}$, 其中, $\vec{q'}$ 和 $\vec{q}$ 按照第 3.1.1 节描述的规则生成.算法输出 $TD_{\vec{Q}}$.

(1) 用一个随机数β乘以差分矢量 $\vec{Q}$ 得到 $\hat{Q}$,即 $\hat{Q}=\beta.\vec{Q}$,这里,$\beta\neq 0$.

(2) 对 $\hat{Q}$ 执行 q 划分.对于 j=1 到 n,假如 $s[j]$=0,将 $\hat{Q}[j]$ 随机分成 $\widehat{Q_a}[j]$ 和 $\widehat{Q_b}[j]$,满足 $\widehat{Q_a}[j]+\widehat{Q_b}[j]=\hat{Q}[j]$; 假如 $s[j]$=1,设置 $\widehat{Q_a}[j]$ 和 $\widehat{Q_b}[j]$ 都等于 $\hat{Q}[j]$.

(3) 产生陷门 $TD_{\vec{Q}}$. $TD_{\vec{Q}}=\{M_1^{-1}\widehat{Q_a},M_2^{-1}\widehat{Q_b}\}$.

$\boldsymbol{Dec}(\boldsymbol{I_i'},\boldsymbol{TD_{\vec{Q}}})$:假定 $I_i'=\{M_1^T\widehat{I_{ia}},M_2^T\widehat{I_{ib}}\},TD_{\vec{Q}}=\{M_1^{-1}\widehat{Q_a},M_2^{-1}\widehat{Q_b}\}$, 当且仅当 $\{M_1^T\widehat{I_{ia}},M_2^T\widehat{I_{ib}}\}.\{M_1^{-1}\widehat{Q_a},M_2^{-1}\widehat{Q_b}\}=0$ 时, $Dec(I_i',TD_{\vec{Q}})$ 输出 1.

此方案利用了基本的属性和查询矢量的表示方法,并且支持多关键词连接查询,我们称其为 MCKS_I方案.

3.2 正确性和安全性分析

3.2.1 正确性

$$\begin{aligned}&\{M_1^T\widehat{I_{ia}},M_2^T\widehat{I_{ib}}\}\cdot\{M_1^{-1}\widehat{Q_a},M_2^{-1}\widehat{Q_b}\}\\&=(M_1^T\widehat{I_{ia}})\cdot(M_1^{-1}\widehat{Q_a})+(M_2^T\widehat{I_{ib}})\cdot(M_2^{-1}\widehat{Q_b})\\&=\widehat{I_{ia}}\cdot\widehat{Q_a}+\widehat{I_{ib}}\cdot\widehat{Q_b}\\&=\widehat{I_i}\cdot\hat{Q}\\&=\varepsilon_i\cdot\beta\cdot\vec{I_i}\cdot\vec{Q},\end{aligned}$$

其中,$\varepsilon_i\neq0,\beta\neq0$.

假如 $\vec{I_i}\cdot\vec{Q}=0$, 那么 $\{M_1^T\widehat{I_{ia}},M_2^T\widehat{I_{ib}}\}\cdot\{M_1^{-1}\widehat{Q_a},M_2^{-1}\widehat{Q_b}\}=\varepsilon_i\cdot\beta\cdot\vec{I_i}\cdot\vec{Q}=0$.

在其他情况下,$\{M_1^T\widehat{I_{ia}},M_2^T\widehat{I_{ib}}\}\cdot\{M_1^{-1}\widehat{Q_a},M_2^{-1}\widehat{Q_b}\}=\varepsilon_i\cdot\beta\cdot\vec{I_i}\cdot\vec{Q}\neq0$.

3.2.2 安全性分析

很明显,MCKS_I方案在唯密文模式下是安全的.在外包 EMR 之前,数据拥有者对 EMR 记录的进行加密,用户进行查询服务时,通常采用将查询关键词加密成陷门的方式保护查询关键词,因此,只要私钥没有泄露,加密的属性矢量和查询矢量所隐含的明文信息就能得到很好的保护,这在文献[38]中给出了证明.非零随机数εi 和β 引入更多的混淆到加密的属性矢量和查询矢量的内积,因此云服务器很难获得额外的信息,例如不匹配的关键词数和是否两个查询是相同的查询,那么也完成了搜索模式的隐藏.在陷门产生过程中由于采用了矢量的 q 拆分,以及引入了随机数β,从而使每次生成的陷门都不同,即使是两个对相同的查询也会产生两个不同的陷门,这保证了查询的无连接性.

MCKS_I方案可以抵抗已知明文攻击.

证明:假定一个攻击者知道数据库 EMR 一组明文记录 I 以及它们相应的密文.不失一般性,我们假定没有随机数引入(注意随机数的引入将加强方案的安全性)对任何矢量 $\vec{I_i}\in I$, 根据方案,一个明文攻击者知道加密的值 (I_{ia}',I_{ib}'). 假如攻击者不知道用于矢量拆分的矢量 s,他将把 $\vec{I_{ia}}$ 和 $\vec{I_{ib}}$ 模型化为两个随机的 n 位矢量.解转化矩阵的方程是 $I_{ia}'=M_1^T\vec{I_{ia}}$ 和 $I_{ib}'=M_2^T\vec{I_{ib}}$. 其中,$M_1$ 和 M_2 是两个 $n\times n$ 的未知矩阵.方程的个数是 $2n|P|$,然而在 $\vec{I_{ia}}$ 和 $\vec{I_{ib}}$ 有 $2n|P|$个未知数,在 M_1 和 M_2 中有 $2n^2$ 个未知数.很显然,攻击者没有足够的信息解出置换矩阵.因此该方案可以抵御已知明文攻击. □

本文方案基于文献[38].Yao 等人[39]提出对文献[38]中方案的安全性的攻击.攻击是基于等式

$$p\cdot q=p_a'\cdot q_a'+p_b'\cdot q_b' \tag{2}$$

的,其中,p 和 q 指明文的属性矢量和查询矢量, p_a' 和 p_b' 是 p 执行 p 划分后的加密矢量. q_a' 和 q_b' 是 q 执行 q 划分后加密矢量.然而等式(2)在我们的方案中只是 $p\cdot q$=0 时才成立.所以这种攻击不适用于我们的方案.

3.3 MCKS_II方案

方案 MCKS_I中属性矢量和查询矢量的维数都很大,这从第 3.1.1 节中属性和查询的矢量表示方法可以看出,而且,该方案只能支持的多关键词连接相等查询.例如,通过一次查询,方案 MCKS_I不能实现类似“(61≤年龄

≤100) AND(疾病=糖尿病) AND (年龄=61 OR 年龄=66) AND (疾病=糖尿病)”的查询.这类查询在一个或者多个域上是范围或者是子集.为了实现上述类型的查询,我们提出分层属性的矢量表示方法.若采用这种矢量表示方法后,方案 MCKS_I的功能和效率都大幅度的提高.我们称提高后的方案为 MCKS_II方案.

3.3.1 数字属性的分层

假定 A 是一个属性值集$\mathcal{W}_A$为[1,N]的数字属性,在$\mathcal{W}_A$上定义一个最深层为 k 的平衡树 $Tr(A)$,它的每一个节点都有一个预先分配的唯一的 ID.我们定义 $Tr(A)$是树中所有节点 ID 的集合.每一个中间节点表示一个范围而叶子节点代表$\mathcal{W}_A$ 中的一个元素.假如 int(ID)表示一个间隔,那么身份标识符为 ID 的中间节点表示范围(间隔) $int(ID)=int(ID_1)\cup...\cup int(ID_j)$, 其中, $ID_1,...,ID_j$ 表示 ID 各个子节点的身份标识符.当 ID 是第 i 个叶子节点时,int(ID)=i.数字属性“年龄”的分层结构如图 3 所示.

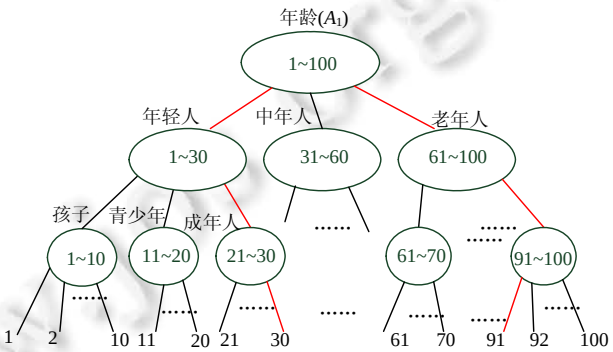


Fig.3 Hierarchy of numerical attribute “age”

图 3 数字属性“年龄”的分层结构

3.3.2 非数字属性的分层

非数字属性 A 的分层用数字属性分层相似的方法定义,不同的是每一个中间节点 ID 语义上包含所有的子节点.节点 ID 的语义范围包含了所有能从 ID 到达的叶子节点.假如 Z 和 $\{Z_1,Z_2,...,Z_u\}$ 语义上是等价的,我们说 Z_i 语义上包含于 Z.表示为 $Z_i\subset Z(i=1,\ldots,u)$.与数字属性“年龄”相似,“地区”和“疾病”也都可以用分层的方式定义.为了避免重复,省略该部分内容.

不论是哪种类型的属性分层 $Tr(A)$,我们称第 d 层的节点集是 d 层属性,用 $Tr_d(A)$标记,其中每一个节点称为 d 层关键词.最顶层的节点称为根节点.从叶子节点 x 到根节点的路径记为$\mathbb{P}(x)$,对路径$\mathbb{P}(x)$上的每一个节点 ID,都有 $x\in int$(ID)或者 $x\subset$ID.

3.3.3 分层属性的矢量表示

对于最大层为 k 的任何分层属性 A_i,我们都可以将其扩展为 k 个子属性(也称子域): $A_{i,1},...,A_{i,k}$, 其中,k 被称为扩展因子.子属性 $A_{i,d}$ 值的集合记为 $Tr_d(A_i)$,对于叶子节点 x,它的 $A_{i,d}$ 的属性值是沿路径$\mathbb{P}(x)$从根节点开始的第 d 个元素.在最大层为 k 的平衡树 $Tr(A_i)$中,叶子节点 x 的矢量表示过程如下:

(1) 找到路径$\mathbb{P}(x)=(N_1,\ldots,N_k)$,其中,$N_d(d=1,\ldots,k)$是路径$\mathbb{P}(x)$上从根节点开始的第 d 个节点.

(2) 让 $A_{i,j}=N_j(j=1,2,...,k)$.

(3) $A_{i,1}$用二进制 1 表示,因为第 1 层属性只有一个属性值.

(4) 假如 $Tr_d(A_i)(1\leqslant d<k)$的每一个节点都有 n 个子节点,那么 $A_{i,d+1}$ 被表示成 n 位的二进制矢量.若 $A_{i,d+1}$ 是 $A_{i,d}$ 的第 j 个子节点,那么只有第 j 位为 1,其他 n–1 位都为 0.

(5) 假如 $Tr_d(A_i)(1\leqslant d<k)$中的节点 N 比同层其他节点有更多的子节点,那么 $A_{i,d+1}$ 被表示成 m 位的二进制矢量,其中 m 是节点 N 的子节点的个数.

以图 3 中“年龄=30”为例,我们详细描述分层年龄(A_1)的矢量表示.

(1) 找到路径$\mathbb{P}$(30)={“1~100”,“1~30”,“21~30”,“30”}.

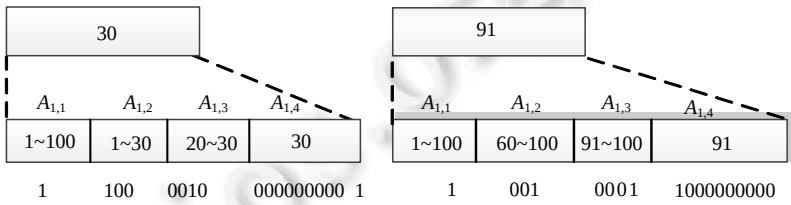
(2) 让($A_{1,1}$,$A_{1,2}$,$A_{1,3}$,$A_{1,4}$)=(“1~100”,“1~30”,“21~30”,“30”).

(3) $A_{1,1}$(“1~100”)用二进制 1 表示.

(4) $A_{1,1}$(“1~100”)有 3 个子节点并且 $A_{1,2}$(“1~30”)是它的第 1 个子节点,因此 $A_{1,2}$ 被表示成 100.

(5) $A_{1,2}$(“1~30”)有 3 个子节点,但同是第 2 层的“61~100”有 4 个子节点(最多),因此,$A_{1,3}$ 被表示成 4 位的二进制矢量.由于 $A_{1,3}$(“21~30”)是 $A_{1,2}$(“1~30”)的第 3 个子节点,因此,$A_{1,3}$ 被表示成 0010.

(6) 因为第 3 层的每一个节点有 10 个子节点,而且 $A_{1,4}$(“30”)是 $A_{1,3}$(“21~30”)的第 10 个子节点,因此 $A_{1,4}$ 被表示成 0000000001.综上,“年龄=30”被表示成 110000100000000001,如图 Fig.4(a)所示.类似地,“年龄=91”被表示成 100100011000000000,如图 4(b)所示.



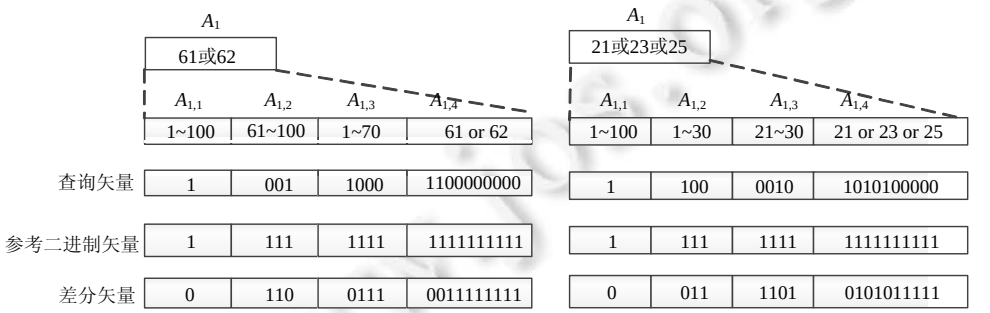
(a) “30”的矢量表示 (b) “91”的矢量表示

Fig.4 Vector representation of hierarchical age

图 4 分层“年龄”的矢量表示

采用属性分层表示能实现多关键词连接相等查询(很显然),而且也能实现复杂的多关键词连接查询,如范围查询和子集查询.结合图 3 中的年龄分层,我们举例详细阐述在范围查询和子集查询中的矢量表示.不失一般性,我们只考虑分层年龄的查询矢量的表示方法.

(1) 子集查询.由于采用属性分层,我们可以实现子集查询.例如“年龄=61 or 年龄=62”属于 age(A_1)的子集,查询矢量被表示为 100110001100000000,如图 5(a)所示.对于另一个例子,“年龄=21 or 年龄=23 or 年龄=25”的矢量表示为 110000101010100000,如图 5(b)所示.注意,子集中的关键词应该是同一个中间节点的子节点.



(a) “年龄=61 或年龄=62”的矢量表示 (b) “年龄=21 或年龄=23 或年龄=25”的矢量表示

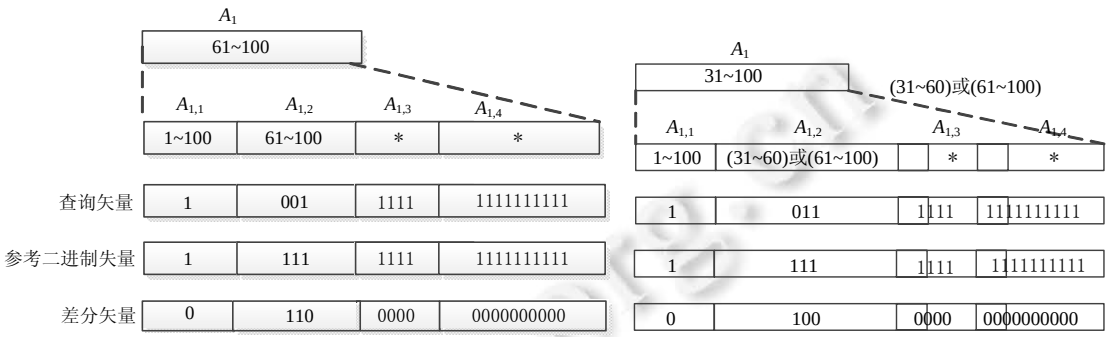
Fig.5 Vector representation of subset

图 5 子集的查询矢量表示

(2) 范围查询.采用属性分层,可以实现范围查询.例如“61≤年龄≤100”是 age 属性上第 2 层的简单范围,它的矢量表示为 10011111111111111,如图 6(a)所示.另外,“31≤年龄≤100”是年龄属性上第 2 层的连续范围,它的矢量表示为 10111111111111111,如图 6(b)所示.假定一个用户希望在第 i 个属性上查找一个范围,他应该在该属性同一层上选择一个简单范围或者语义简单范围.这里,简单范围意思是只能用同一层的节点表示.他可以选择一个或多个简单范围组成一个连续范围或者是不连续的范围.本文我们不实现任意范围的查询,这一点与文献

[17]是一致的.

像“年龄”一样,其他分层属性,例如“疾病”和“地区”也采用分层属性的矢量表示.其他不可以分层的属性,如“性别”仍用基本的矢量表示方法表示.



(a) “61≤年龄≤100”的矢量表示 (b) “31≤年龄≤100”的矢量表示

Fig.6 Vector representation of range

图 6 范围的查询矢量表示

4 性能分析

首先,总结主要的 SE 方案,见表 2.根据基于的算法,SE 机制可以分为两大类:基于 PKC 的和基于 SKC 的.我们只关注支持连接关键词查询的方案.在基于 SKC 的 SE 机制中,文献[11,13,16,28]中的方案与文献[12]中的第 1 个方案只支持简单的连接关键词搜索,也就是相等查询.多关键词排序方案[23−27]考虑文档和关键词之间的相似程度,而且他们引入了不精确性到服务器返回的结果.

Table 2 Comparison of main SE schemes

表 2 主要的 SE 方案的比较

基于的算法	支持的查询	方案
基于对称密码算法的	单关键词查询	文献[3–6]
	连接关键词相等查询	文献[11,12]1(文献[12]的第 1 个方案),文献[13,16,28]
	排序查询	单关键词排序[7−9],多关键词排序[23−27]
基于公钥密码算法的	单关键词查询	文献[10]
	连接关键词相等查询	文献[12]2(文献[12]的第 2 个方案),文献[14,15]
	具有复杂结构的查询	文献[17−20,29,30]

从表中可以看出,以前的基于 SKC 的多关键词查询方案一般只都支持查询结构简单的多关键词连接查询.相比基于 SKC 的 SE 方案,基于 PKC 的 SE 方案可以实现具有复杂查询结构的多关键词查询,例如实现范围查询,子集查询以及多域关键词“或”的查询.然而相比基于 SKC 的 SE 机制来说,他们的效率更低.因为基于 PKC 的 SE 机制中一般需要产生许多耗时的对操作.在基于 PKC 的 SE 方案中,文献[17]与本文考虑同样的场景,完成相同的查询功能.下面我们估计我们方案的计算开销并与文献[17]相比较.因为还没有用于研究的目的 EMRs 数据库可以公开利用.用于实验的数据库是我们用程序随机产生的类似表 1 的数据库,数据库有:年龄,性别,疾病和地区 4 个属性.我们在 CPU1.77Ghz,32 位 Ubuntu Linux 的服务器上运行程序的,采用的语言是 C 语言.

4.1 初始化

在 MCKS(MCKS_I,MCKS_II)方案中,除了产生两个 $n\times n$ 可逆矩阵 M_1 和 M_2 以及 n 位的二进制矢量 S 外,初始化阶段的开销还包括预计算 $M_1^T,M_2^T,M_1^{-1},M_2^{-1}$ 的时间,初始化阶段的主要开销是矩阵求逆,时间复杂度是 $O(n^3)$.为了模拟文献[17]中的方案,我们利用了文献[17]中的实验时采用的参数.在文献[17]的方案中,初始化过程中的主要开销是产生基 $\hat{B}$ 和 B^*,其计算复杂度为是 $O(n_0^2)=O(n^2)(n_0=n+3)$,n_0 是 HPE 中 ECC 矢量空

间的维度,n 是查询和属性矢量的长度.图 7(a)和图 7(b)分别表示在 MCKS 方案和文献[17]方案中初始化时间与 n 之间的关系.当 n=46 时,文献[17]方案中的初始化时间是 45s,而我们方案的初始化时间是 4.2ms.当 n=400 时,文献[17]方案中的初始化时间是 2170s,而我们方案的初始化时间是 1 770ms.从图 7 可以明显地观察到我们方案的初始化时间比文献[17]的少得多.

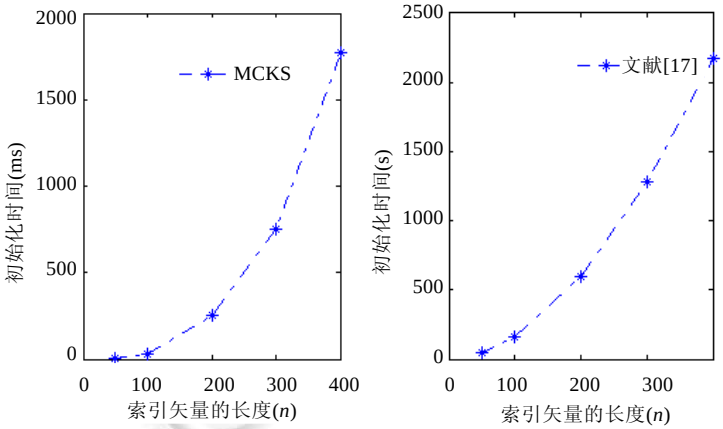


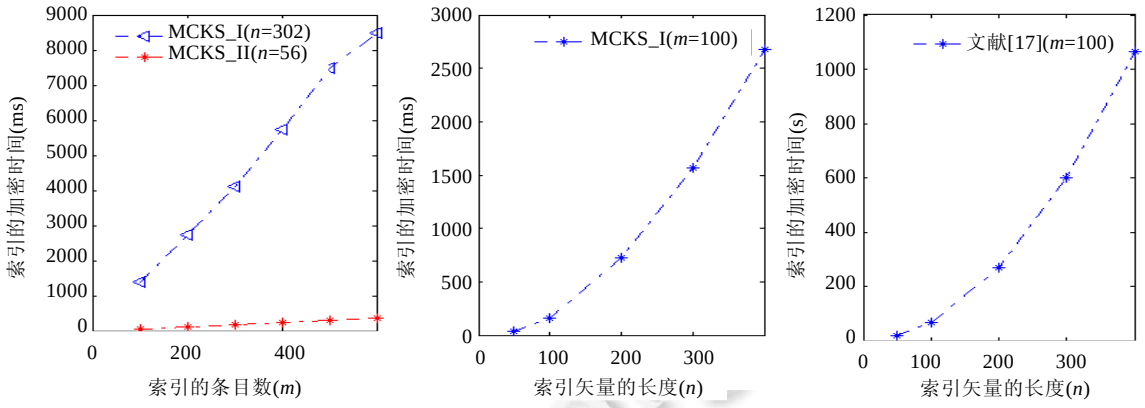
Fig.7 Relation of setup time and the length of index vectors

图 7 初始化时间与索引矢量长度的关系

4.2 属性矢量的加密

为了加密属性矢量(索引),第 1 步是从每一个病人的各个属性值(关键词)映射为一个二进制矢量,接着加密每个二进制矢量.这个阶段主要消耗的时间主要依赖于数据库中记录条目(即行数)以及索引矢量的长度,索引矢量的长度主要由数据库关键词的个数以及属性结构决定的.我们实验中忽略第 1 步的时间.假定“年龄”,“疾病”和“地区”各有 100 个属性值.加上两个“性别”属性值{男,女},相应于整个 EMRs 数据集总共有 302 个属性值(关键词).MCKS_1中索引矢量和查询矢量都是采用基本的表示方法,所以在 MCKS_1方案中,索引矢量和查询矢量的长度是 302bit.即 n=302;在 MCKS_II方案中,假定“年龄”采用图 3 的结构分层,“疾病”和 “地区”采用与年龄同样的方式进行分层,那么“年龄”,“疾病”和“地区”都将表示成 18bit 的二进制矢量,加上不分层年龄属性的两个属性值,在 MCKS_II方案中索引矢量和查询矢量长度是 56bit.即 n=56.图 8(a)显示了索引矢量的加密时间几乎与 EMR 中记录的数成线性关系.MCKS_II方案加密索引的构建时间大约是 MCKS_I方案的 1/29.为了获得属性矢量的加密时间与属性矢量长度(n)之间的关系,我们从数据集中选择前 100 行(即 m=100)作为输入数据计算加密时间.图 8(b)显示属性矢量加密时间复杂度为 $O(n^2)$.当 n=50 时,文献[17]中的属性矢量的加密时间大约是 17s,我们方案的加密时间是 40ms.当 n=400 时,文献[17]中的属性矢量的加密时间大约是 1 066s,我们方案的加密时间是 2 680ms.相比文献[17]的方案,我们方案的效率提高大约 425 倍.

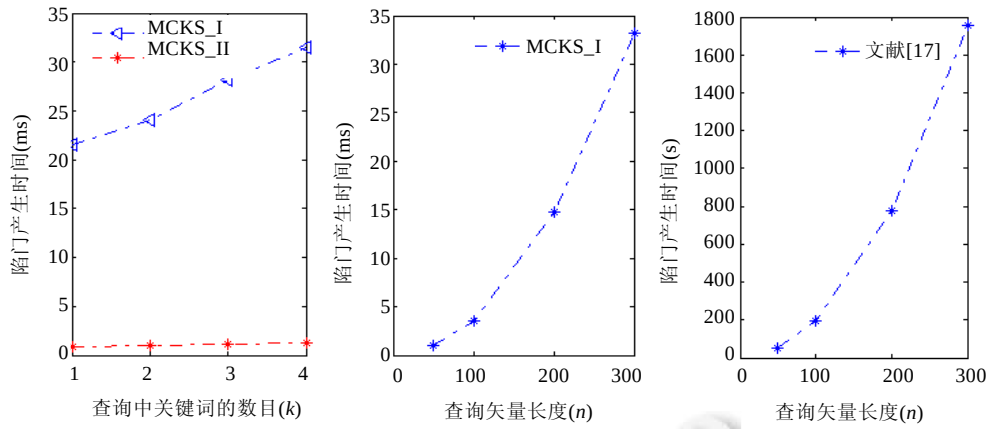
查询时间(query time). 测量查询时间所用的数据集是测试属性矢量加密时间所用的数据集.当测试 MCKS_I方案时,属性矢量的长度是 302bit,而当测试 MCKS_II方案时,属性矢量的长度是 56bit.从图 10(a)说明了,查询时间主要是 EMRs 中索引的数量(即数据库的行)决定,对于行数相同的数据集,MCKS_I方案的查询时间大约是 MCKS_II方案的 2.5 倍.图 10(b)和图 10(c)说明了我们方案的查询时间比[17]有效得多.



(a) 索引的数目是变化的 (b) 索引的矢量长度是变化的 (c) 索引的矢量长度是变化的

Fig.8 Comparison of the time of encrypting index vectors

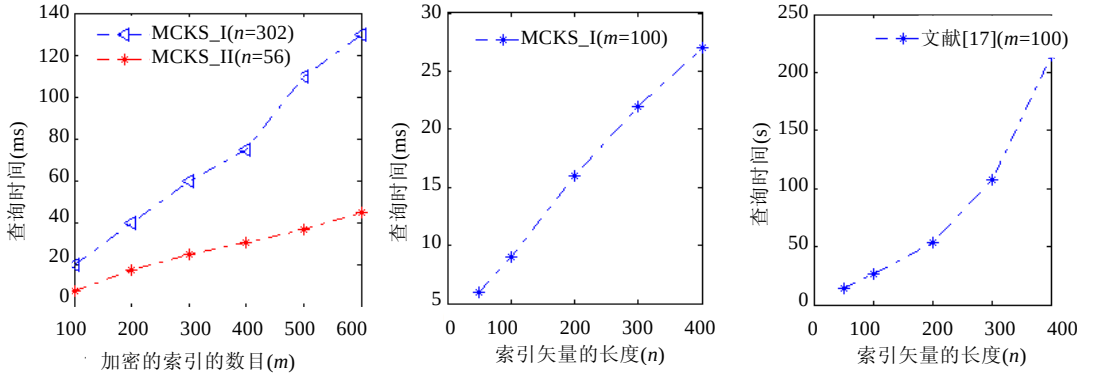
图 8 加密索引矢量所需时间的比较



(a) 查询中包含的关键词数是变化的 (b) 索引矢量长度是变化的 (c) 索引矢量长度是变化的

Fig.9 Trapdoor generation time

图 9 陷门产生时间的比较



(a) 索引数目在变化 (b) 索引数恒定,索引矢量长度在变化 (c) 索引数恒定,索引矢量长度在变化

Fig.10 Query time

图 10 查询时间的比较

5 结 论

本文我们主要研究对于加密的 EMR 的安全查询.我们定义了基于 SKC 的谓词加密,并且提出了新颖的索引和查询矢量的表示方法.我们基于 ASPE 构造了两个 MCKS 方案.这两个方案被证明能抵抗已知明文攻击.MCKS_I方案只能支持多域上的连接关键词相等查询,利用分层属性的矢量表示方法,MCKS_II方案不仅能支持灵活的连接关键词查询,如范围查询和子集查询.而且该方案实现了比 MCKS_I方案更高的效率.大量的分析和实验数据表明,本文提出的方案更加有效.然而本文提出方案的最大不足是搜索时间与数据库文档数成线性关系,这将影响在大型数据库中的使用性.我们将构造一个既能实现具有复杂查询结构的,又能实现比线性搜索更快的多关键词连接搜索方案作为我们下一阶段的研究目标.

References:

[1] 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]

[2] 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 with English abstract). http://www.jos.org.cn/1000-9825/4554.htm [doi: 10.13328/j.cnki.jos.004554]

[3] Song D, Wagner D, Perrig A. Practical techniques for searches on encrypted data. In: Proc. of the S&P. Berkeley, 2000. 3−10. [doi: 10.1109/SECPRI.2000.848445]

[4] Waters B, Balfanz D, Durfee G, Smetters DK. Building an encrypted and searchable audit log. In: Proc. of the NDSS. SanDiego, 2004.

[5] Yan CC, Mitzenmacher M. Privacy preserving keyword searches on remote encrypted data. In: Proc. of the ACNS. New York, 2005. 442−455. [doi: 10.1007/11496137_30]

[6] Curtmola R, Garay J, Kamara S, Ostrovsky R. Searchable symmetric encryption: Improved definitions and efficient constructions. In: Proc. of the ACM CCS. New York, 2006. 79−88. [doi: 10.1145/1180405.1180417]

[7] Zerr S, Olmedilla D, Nejdl W, Siberski W. Zerber+r: Top-*k* retrieval from a confidential index. In: Proc. of the EDBT. 2009. 439−449. [doi: 10.1145/1516360.1516412]

[8] Wang C, Cao N, Li J, Lou WJ. Secure ranked keyword search over encrypted cloud data. In: Proc. of the ICDCS 2010. Genoa, 2010. 253−262. [doi: 10.1109/ICDCS.2010.34]

[9] Wang C, Cao N, Ren K, Lou W. Enabling secure and efficient ranked keyword search overoutsourced cloud data. IEEE Trans. on Parallel and Distributed Systems, 2012,23(8):1467−1479. [doi: 10.1109/TPDS.2011.282]

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

[11] Golle P, Staddon J, Waters B. Secure conjunctive keyword search over encrypted data. In: Proc. of the ACNS. Heidelberg, 2004. 31−45. [doi: 10.1007/978-3-540-24852-1_3]

[12] Ballard L, Kamara S, Monrose F. Achieving efficient conjunctive keyword searches over encrypted data. In: Proc. of the ICICS. Beijing, 2005. 414−426. [doi: 10.1007/11602897_35]

[13] Bosch C, Brinkman R, Hartel P, Jonker W. Conjunctive Wildcard Search over Encrypted Data. Secure Data Management, Berlin, Heidelberg: Springer-Verlag, 2011,6933:114−127. [doi: 10.1007/978-3-642-23556-6_8]

[14] Park DJ, Kim K, Lee PJ. Public key encryption with conjunctive field keyword search. In: Lim CH, Yung M, eds. Proc. of the WISA 2004. LNCS 3325, 2004. 73−86. [doi: 10.1007/978-3-540-31815-6_7]

[15] Hwang YH, Lee PJ. Public key encryption with conjunctive keywordsearch and its extension to a multi-user system.In: Proc. of the Intl Conf. on Pairing-Based Cryptography. Berlin, Heidelberg: Springer-Verlag, 2007. 2-22 [doi: 10.1007/978-3-540-73489-5_2]

[16] Shen E, Shi E, Waters B. Predicate privacy in encryption systems. In: Proc. of the 6th Theory of Cryptography Conf.on Theory of Cryptography. Heidelberg, 2009. 457−473. [doi: 10.1007/978-3-642-00457-5_27]

[17] Li M, Yu S, Cao N, Lou WJ. Authorized Private Keyword Search over Encrypted Data in Cloud Computing, In: Proc.of the IEEE Int'l Conf. on Distributed Computing Systems (ICDCS). Minneapolis, 2011. 383−392. [doi: 10.1109/ICDCS.2011.55]

[18] Boneh D, Waters B. Conjunctive, subset, and range queries onencrypted data. In: Proc. of the TCC. Berlin, Heidelberg: Springer-Verlag, 2007. 535−554. [doi: 10.1007/978-3-540-70936-7_29]

[19] 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]

[20] Attrapadung N, Libert B. Functional encryption for inner product: Achieving constant size ciphertext with adaptive security or support for negation. In: Nguyen P, Pointcheval D, eds. Proc. of the Public Key Cryptography (PKC 2010). LNCS 6056, 2010. 384−402. [doi: 10.1007/978-3-642-13013-7_23]

[21] Lewko A, Okamoto T, Sahai A, Takashima K. Fully secure functional encryption: Attribute-based encryption and (hierarchical) inner product encryption. In: Proc. of the EUROCRYPT. 2010. 62−91. [doi: 10.1007/978-3-642-13190-5_4]

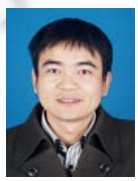
[22] Lu YB. Privacy-Preserving logarithmic-time search on encrypted data in cloud. In: Proc. of the NDSS. 2012.

[23] Cao N, Wang C, Li M, Ren K, Lou WJ. Privacy-Preserving multi-keyword ranked search over encrypted cloud data. In: Proc. of the IEEE INFOCOM. Shanghai, 2011. 829−837. [doi: 10.1109/INFCOM.2011.5935306]
[24] Yang C, Zhang W, Xu J, Yu N. A fast privacy-preserving multi-keyword search scheme on cloud data. In: Proc. of the CSC. Shanghai, 2012. 104−110. [doi: 10.1109/CSC.2012.23]
[25] Sun W, Wang B, Cao N. Verifiable privacy-preserving multi-keyword text search in the cloud supporting similaritybased ranking. IEEE Trans. on Parallel Distributed Systems, 2013,25:3025−3035. [doi: 10.1109/TPDS.2013.282]
[26] Sun WH, Wang B, Cao N. Privacy-Preserving multi-keyword text search in the cloud supporting similarity-based ranking. In: Proc. of the ASIACCS. 2013. 71−82.
[27] Sun WH, Lou WJ, Hou YT, Li H. Privacy-Preserving keyword search over encrypted data in cloud computing. In: Secure Cloud Computing. New York: Springer-Verlag, 2014. 189−212. [doi: 10.1007/978-1-4614-9278-8_9]
[28] Liu C, Zhu LH, Chen JJ. Efficient searchable symmetric encryption for storing multiple source data on cloud. Technical Report, No.20150423:023649, Sydney: University of Technology, 2015.
[29] 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]
[30] Okamoto T, Takashima K. Hierarchical predicate encryptionfor inner-products. In: Matsui M, ed. In: Proc. of the ASIACRYPT. Berlin, Heidelberg: Springer-Verlag, 2009. 214−231. [doi: 10.1007/978-3-642-10366-7_13]
[31] Strizhov M, Ray I. Multi-Keyword similarity search over encrypted cloud data. In: Cuppens-Boulahia N, Cuppens F, Jajodia S, Kalam AAE, Sans T, eds. Proc. of the IFIP SEC 2014. Heidelberg: Springer-Verlag, 2014. 52−65. [doi: 10.1007/978-3-642-55415-5_5]
[32] Li J, Wang Q, Wang C, Cao N, Ren K. 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]
[33] Chuah M, Hu W. Privacy-Aware bedtree based solution for fuzzy multi-keyword search over encrypted data. In: Proc. of the ICDCSW. Minneapolis, 2011. 273−281. [doi: 10.1109/ICDCSW.2011.11]
[34] Liu C, Zhu L, Li L, Tan Y. Fuzzy keyword search on encrypted cloud storage data with small index. In: Proc. of the ICCCIS. 2011. 269−273. [doi: 10.1109/CCIS.2011.6045073]
[35] Wang C, Ren K, Yu SC . Achieving usable and privacy-assured similarity search over outsourced cloud data. In: Proc. of the IEEE INFOCOM. Orlando, 2012. 451−459. [doi: 10.1109/INFCOM.2012.6195784]
[36] Wang B, Yu S, Lou W, Hou YT. Privacy-Preserving multi-keyword fuzzy search over encrypted data in the cloud. In: Proc. of the IEEE INFOCOM. Toronto, 2014. 2112−2120. [doi: 10.1109/INFOCOM.2014.6848153]
[37] Li C, Sun X M, Xia Z H, Liu Q. An efficient and privacy-preserving semantic multi-keyword ranked search over encrypted cloud Data. Int'l Journal of Security and Its Applications, 2014,8:323−332. [doi: 10.14257/ijsia.2014.8.2.33]
[38] Wong KK, Cheung DW, Kao B, Mamoulis N. Secure kNN 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]
[39] Yao B, Li FF, Xiao XK. Secure nearest neighbor revisited. In: Proc. of the 29th IEEE Int'l Conf. on Data Engineer. 2013. 733−744. [doi: 10.1109/ICDE.2013.6544870]

附中文参考文献:

[1] 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]
[2] 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 [doi: 10.13328/j.cnki.jos.004554]

张丽丽(1979－),女,河南商丘人,博士,主要研究领域为安全协议,云计算.



刘雪峰(1985－),男,博士,主要研究领域为云计算安全,应用密码学.

张玉清(1966－),男,博士,教授,博士生导师,主要研究领域为网络与信息系统安全.

全韩彧(1989－),男,博士生,主要研究领域为大数据,云计算安全.