基于联盟链的电子健康记录隐私保护和共享*

巫光福†, 余 攀, 陈 颖, 李江华
(江西理工大学 信息工程学院, 江西 赣州 341000)

摘 要: 医院正在逐渐采用电子健康记录(EHR)的方式去记录患者的医疗信息。然而,医疗数据的隐私性和EHR 标准的差异化阻碍了医疗数据在病人和医院之间的共享。因此,针对隐私信息泄露和难于共享的问题,提出了一个基于联盟链的隐私保护数据共享模型。此外,基于匿名算法提出了(p,α,k)匿名隐私算法,能够解决EHR 隐私信息泄露的问题。通过理论分析和实验证明,提出的基于联盟链和(p,α,k)隐私匿名算法模型能够在保护数据隐私的前提下,实现病人和医院之间的数据安全共享。对比前人的模型,该模型具有所需节点少、减少主链压力、容错性强和病人对 EHR 完全控制等优势。

关键词: 电子健康记录; 隐私保护; 共享; (p,α,k)隐私匿名算法; 联盟链
中图分类号: TP309.7 **文献标志码:** A **文章编号:** 1001-3695(2021)01-006-0033-06
doi:10.19734/j.issn.1001-3695.2019.09.0584

Privacy protection and sharing of EHR based on consortium blockchain

Wu Guangfu†, Yu Pan, Chen Ying, Li Jianghua
(*School of Information Engineering*, *Jiangxi University of Science & Technology*, *Ganzhou Jiangxi* 341000, *China*)

Abstract: Hospitals are gradually adopting EHR methods to record patients' medical information. However, the privacy of medical data and the differentiation of EHR standards hinder the sharing of medical data between patients and hospitals. Therefore, for the problem of leakage of private information and difficulty in sharing, this paper proposed a privacy protection data-sharing model based on the consortium blockchain. In addition, based on the anonymous algorithm, it proposed the (p,α,k) anonymous privacy algorithm, which could solve the problem of EHR private information leakage. Through theoretical analysis and experimental results, the proposed consortium-based chain and (p,α,k) privacy anonymity algorithm model can realize data security sharing between patients and hospitals under the premise of protecting data privacy. Compared with the previous models, the proposed model has the advantages of fewer nodes, less main chain pressure, better fault tolerance and complete control of the patient's EHR.

Key words: electronic healthcare record(EHR); privacy protection; sharing; (p,α,k) privacy anonymity algorithm; consortium blockchain

0 引言

近年来,随着病历电子化、医院上云、远程问诊等业务的开展,越来越多的个人健康信息被连入网络,病历信息逐渐被记录在电子健康记录中。虽然这在很大程度上提升了便捷性,但同时也增加了病人信息数据泄露的风险。EHR 有着很高的价值,其中包含患者的姓名、年龄、居住地址、电话、病史、银行账户等信息,蕴涵着重要的财富价值,同时也涉及大量个人隐私信息。然而,病人的 EHR 只是简单地存放在各大医院的医疗系统中,加上医院缺少专业的技术人员使得 EHR 成为黑客和医院内部工作人员窃取的对象,导致易对病人造成进一步的损失且无法明确责任人,各个医院的系统是独立存在的,且每个医院的 EHR 标准可能有所区别,所以存放在各个医疗系统中的数据无法实现实时共享,存在特别严重的信息孤岛现象。假设病人曾在医院 A 就诊,当病人更换到医院 B 就诊时,则以前在医院 A 就诊的 EHR 无法被医院 B 使用,使得医生对病人的诊断变得非常困难。不法分子对医疗系统的攻击,使得部分医院遭受了严重的医疗数据泄露。根据隐私信息安全法,所有识别的信息特别是个人信息应该有责任得到安全保密。在 EHR 中可以识别特定个人的特有信息,必须依法得到保护[1]。

为了解决 EHR 隐私信息泄露和难于共享的问题,迫切需要一个去中心化、可溯源、可实时共享和保护隐私的系统。因此,有必要设计一个安全可靠的分布式共享系统,以确保 EHR 的安全共享和隐私信息的保护。目前,区块链被认为是一种快速发展的颠覆性技术[2]。区块链是一个去中心化的账本系统,记账由不同地域的多个节点共同完成,而且每一个节点记录的都是完整的账目,因此它们都可以参与监督交易的合法性。它是一种按照时间顺序将数据区块相连组合成的一种链式数据结构。区块链使用加密技术加密账户身份信息和交易信息,只有在数据拥有者授权的情况下才能访问,从而保证了数据的安全和个人的隐私。每个节点都可以根据分布式网络中的公开密钥来验证交易数据的有效性,因此它们之间没有信任共识。比特币的系统依赖于所有节点参与称为工作量证明(PoW)的共识算法来完成对交易数据的验证。区块链还可以提供智能合约[3]技术,智能合约是一套以数字形式定义的承诺,包括合约参与方可以在上面执行这些承诺的协议,并且一旦部署,只要满足设定的条件就会自动执行和自动验证。此外,区块链使用独特的经济激励机制来吸引节点完成工作(即

收稿日期: 2019-09-26; **修回日期:** 2019-11-19 **基金项目:** 国家自然科学基金资助项目(11461031);江西省教育厅科技类重点项目(GJJ170492);江西省自然科学基金资助项目(20181BBE58018);江西省教育厅科技类一般项目(GJJ170516,GJJ180442)

作者简介: 巫光福(1977-),男(通信作者),江西玉山人,教授,硕导,博士,主要研究方向为信息论与编码、密码学与信息安全、区块链技术、人工智能、统计方法(wuguangfu@126.com);余攀(1993-),男,江西抚州人,硕士研究生,主要研究方向为密码学、区块链技术;陈颖(1995-),女,江西宜春人,硕士研究生,主要研究方向为密码学、区块链技术;李江华(1976-),男,河南新野人,教授,硕导,博士,主要研究方向为数字水印、数据工程、语义 Web.

挖掘),从而促使节点提供计算能力和资源[4]。激励机制激励节点交互以改善系统的活动,这使其能够稳步发展。EHR 产业类似于区块链中网络节点的分布式结构。因此,该技术可以为其提供解决方案。薛腾飞等人[5]利用改进的委托权益证明(delegated proof of stake,DPOS)共识机制提出了一种医疗机构联盟服务器群(medical institution federate servers,MIFS)和审计联盟服务器群(auditing federate servers,AFS)相结合的医疗区块链系统 MDSM,但是该方案所需要的节点数目达到 121 个且病人无法实现对自己的 EHR 完全掌控。Azaria 等人利用以太坊[6]区块链实现了一个医疗区块链与大数据相结合的医疗信息共享平台 MedRec[7],但是该方案增加了对主链的压力,并且存在医疗信息泄露的威胁。Ivan[8]分析了将区块链作为保护医疗健康数据存储的新颖方法、实施障碍以及从当前技术向区块链解决方案逐步过渡的计划。Shrier 等人[9]采用美国麻省理工学院的 OPAL/Enigma 加密平台与区块链技术相结合的方式,为医疗保健信息的存储和分析创造了一个安全环境。Kuo 等人[10]采用了隐私保护在线机器学习与私有区块链技术相结合的模式,但是该方案增加了主链的压力,且系统中使用 PoW 共识算法,容错性较弱。Witchey[11]介绍了医疗交易单(transaction)验证系统和方法,但是也存在隐私泄露的问题。

因此,针对目前所提出的医疗共享模型存在的隐私泄露和数据共享难的问题,本文提出基于联盟链和(p,α,k)隐私匿名算法的 EHR 隐私保护和共享模型。由于公有链存放的数据是透明的,且共识算法的效率低和无容错性,所以系统选择联盟链为底层平台。联盟链中只有部分节点才能访问智能合约(链码)和数据交易,有效地保护了数据的隐私和安全;在公有链中区块容量有限,且交易信息不断存储,严重制约着交易的速度,故本文在区块中只存放交易信息的哈希值;EHR 中含有人口统计学等涉及隐私的数据,故提出(p,α,k)隐私匿名算法去处理 EHR 中的敏感信息,通过匿名算法处理的 EHR 被存储在星际文件系统(IPFS)[12], IPFS 是一种点对点分布式文件系统,旨在将所有计算设备与同一系统的文件连接起来。当病人或授权的第三方需要使用 EHR,首先需要通过区块链中存储的 hash 值进行身份验证,只有验证通过的用户才能获取 EHR 的使用权。验证通过后,用户可以下载自己的 EHR,医生根据病人所提供的完整 EHR 对病人进行准确的诊断。

本文通过匿名算法处理 EHR 中的隐私信息,处理后存放在 IPFS 中,大大地保护病人的隐私和减少主链的压力,并在一定程度下提高区块的交易处理速度;可以一定程度打破 EHR 存在的不能共享的局限性,有效地缓解信息孤岛的现象;采用 PBFT 共识算法,可以在很大程度上减少算力开销,并在一定程度上减少网络上的数据传输量[13]。

1 密码学基础

1.1 数字签名

椭圆加密算法(ECC)[14]是基于椭圆曲线数学的一种公钥加密体制,最初由 Koblitz 和 Miller 两人于 1985 年提出,其安全性依赖于椭圆离散对数的计算困难性。随着椭圆曲线的发展,其被应用到各个领域。区块链系统交易的签名和验证过程中,同样需要使用椭圆曲线。为保证存储于区块链中信息的安全和完整,区块链的定义和构造中使用了椭圆曲线公钥密码技术对交易的数据进行签名,算法 1 描述了签名的详细过程。签名过程中涉及的参数如表 1 所示。

算法 1 数字签名算法

输入:产生随机数 d,$1\leqslant d\leqslant n-1$。

输出:(r,s)数字签名结果。

a)计算 $dG=(x_1,y_1)$ $x_1\rightarrow\bar{x}_1$, $r=\bar{x}_1 \bmod n$。

b)如果 $r=0$ 满足条件后转步骤 a),不满足条件则执行步骤 c)。

c)计算 $H(m)\rightarrow e$,$d^{-1}\bmod n$ 和 $s=d^{-1}(e+kr)\bmod n$

d)如果 $s=0$,满足条件后转步骤 a),不满足条件程序执行完毕。

当验证签名时,矿工节点利用椭圆曲线的参数和公钥 Q 对签名进行验证,算法 2 展示了验证的过程。

算法 2 数字签名验证算法

输入:m,r,s,G,Q。

输出:验证成功或验证失败。

a)$H(m)\rightarrow e$,计算 $w=s^{-1}\bmod n$, $r=\bar{x}_1 \bmod n$。

b)计算 $u_1=ew \bmod n$, $u_2=rw \bmod n$。

c)计算 $X=u_1G+u_2Q$,其中 X 是椭圆曲线上的一点。

d)如果 $X=0$ 直接 break,不满足条件则执行步骤 c)。

e)计算 $x_1\rightarrow\bar{x}_1$ 和 $v=\bar{x}_1 \bmod n$。

f)如果 $v=r$,验证通过,否则验证失败。

表 1 数字签名参数

Tab.1 Digital signature parameters

符号	描述	符号	描述
m	交易信息	G	群
p	大素数	n	G 的阶
a,b	参数 $a,b\in F_p$,F_p 是有限域	k,Q	公私钥

只有当区块链中的矿工节点通过签名的验证,才能成为区块链中的区块。区块链中不仅大量使用数字签名,而且使用 Merkle 树作为快速归纳和校验完整性的数据结构。Merkle 哈希树是一类基于哈希值的二叉树或多叉树,其叶子节点上的值通常为数据块的哈希值,而非叶子节点上的值,是将该节点的所有的子节点组合结果的哈希值。如图 1 所示,Merkle root 就是通过叶子的哈希值组合而成。区块链中的 Merkle 树用于存储交易信息,每个交易两两配对,构成 Merkle 树的叶子节点,进而生成整个 Merkle 树。

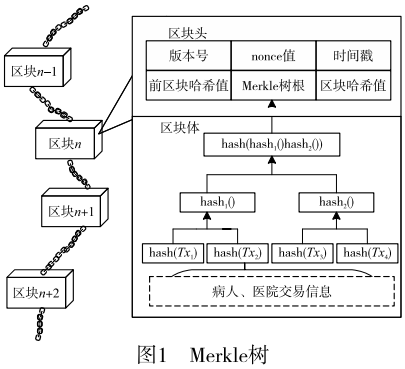

图1 Merkle树

Fig.1 Merkle tree

1.2 k-匿名

对于原始的数据集的属性,将其分成三类:a)标志符属性,可以唯一地确定一条用户记录,如姓名和地址;b)准标志符(QI)属性,可以以较高的概率结合一定的外部信息去确定一条用户记录;c)敏感属性,就是需要保护的用户数据,如健康状况。

对于标记符属性,在入库之前可能都会删除或者经过加密处理,所以对于入侵者无法根据这个属性去确定用户身份。但是,入侵者可能根据准标志符属性结合外部的一些信息使用记录链接技术[15]去搜集个人身份信息。

定义 1 QI-组。QI 是原始数据表中所有含有相同的 QI 属性值记录的集合,则由所有的 QI 构成的集合称为 QI-组。在不同的文献中可能有不同的术语,如等价类[16]和 QI-簇[17]。

定义 2 k-匿名[18]。每个 QI-组在原始数据表中至少出现 k 次,且在修改后的数据表中同样满足。

1.3 纠错码

纠错码(error correcting code)[19]是一种在传输过程中发生错误后能在接收端自行发现或纠正的编码。纠错码的基本思路是在所有的由发送符号组成的序列中,仅挑出其中一部分作为信息的代表向信道发送,并使得所挑出的这些序列之间有

尽可能多的差异。每个被挑出的允许发送的序列被称为一个码字,而所有码字就称为码。在发送端把信息变换成码字的过程称为编码;在接收端从接收到的信号判定所发码字、从而恢复信息的过程称为解码(或译码)。在解码时,若收到的信号不是码中的一个码字,则可以肯定在传输中出现了差错,从而着手对差错进行纠正。下面对纠错码的基本构成要素进行说明。

定义 3　线性分组码。有限域 F_q 内,信息被划分成若干个组,每一组由 k 个码元组成,然后再通过编码器使一组变成 n 个码元,这 n 个码元构成分组码的一个码字。在编码的过程中,每个码元的取值有 q 种可能性,如果所有的码字集合构成一个 k 维的线性空间,则称它是一个(n,k)线性分组码。

定义 4　汉明距离与汉明重量。对于 $\boldsymbol{v}=(v_1,v_2,\cdots,v_n)$,$\boldsymbol{u}=(u_1,u_2,\cdots,u_n)\in F_q^n$,用 $W_H(\boldsymbol{v})$ 表示非零向量 $\boldsymbol{v}_i(1\leqslant i\leqslant n)$ 的个数,叫做向量 $\boldsymbol{v}$ 的汉明重量。$d_H(\boldsymbol{u},\boldsymbol{v})=W_H(\boldsymbol{u}-\boldsymbol{v})$ 叫做向量 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 之间的汉明距离。

汉明距离的一些性质:a)$d(\boldsymbol{u},\boldsymbol{v})\geqslant 0$ 并且 $d(\boldsymbol{u},\boldsymbol{v})=0\Leftrightarrow\boldsymbol{u}=\boldsymbol{v}$;b)$d(\boldsymbol{u},\boldsymbol{v})=d(\boldsymbol{v},\boldsymbol{u})$;c)$d(\boldsymbol{u},\boldsymbol{v})\leqslant d(\boldsymbol{u},w)+d(w,\boldsymbol{v})$。

定义 5　纠错能力。纠错能力一般用 d_{min}(最小汉明距离)来表示,d_{min} 越大,纠错能力越强。一般来说,线性分组码的理论纠错能力 t 满足 $t=(d_{min}-1)/2$。

1.4　PBFT

PBFT(practical Byzantine fault tolerance)[20] 是实用拜占庭容错算法。该算法是 Castro 等人在 1999 年提出来的,解决了原始拜占庭容错算法效率不高的问题,将算法复杂度从指数数量级降低到多项式数量级,使得拜占庭容错算法在实际系统应用中变得可行。基于消息传递执行 PBFT 算法,显著提高了事务确认速度和事务吞吐量。该算法解决了原始拜占庭容错算法的低效问题,并将算法的复杂度从指数数量级降低到平方数量级。PBFT 一致性算法可以解决数据丢失和数据延迟问题。该算法在保证活性和安全性的前提下提供了$(n-1)/3$ 的容错性。图 2 展示了在没有发生主节点失效的情况下算法的正常执行流程,其中副本 0 是主节点,副本 3 是失效节点,而 C 是客户端。

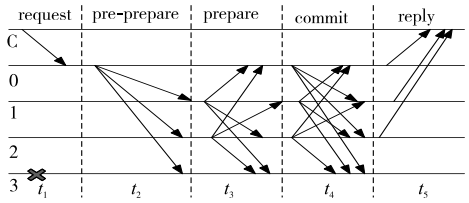

图2　PBFT执行流程

Fig.2　PBFT implementation process

2　隐私保护和共享的网络架构

本文提出的基于联盟链和 (p,α,k)隐私匿名算法的 EHR 存储和共享系统具体的网络框架如图 3 所示。区块链技术底层由 P2P 网络支撑,所有节点共同维护和验证区块链网络,非常适合应用于医疗领域需要多方共同参与的业务场景中。利用区块链技术的去中心化、数据防窜改、数据可溯源三大特性,从底层优化医疗行业信息难于共享和信息孤岛的难题。由于区块链的不可窜改性和可溯源等特性,当病人信息出现泄露时,可以马上发现事件的源头。使得数据的溯源、验证、查询流程摆脱传统的人工审计,提高效率的同时大大降低了成本。

区块链由于区块容量和交易速度的问题,故系统将(p,α,k)隐私匿名算法处理后的 EHR 存放在 IPFS,这样有利于减少区块的压力、提高交易的速度,最重要的是保护病人的隐私数据;当病人和医院进行交互时,产生的交易信息通过改进的 hash 函数进行计算,产生不可改变的 hash 值,存储在联盟链区块中,并且节点账户保留此 hash 值。当需要访问数据时,病人或授权医生需要通过 hash 值验证其身份,验证通过后,则授权给需要访问 EHR 的病人或医生,其可以通过 IPFS 下载病人的 EHR。系统中只有验证通过的病人或医生才能下载 EHR,这极大地增强了系统的安全性。

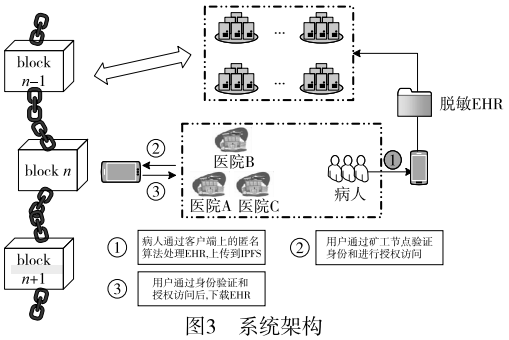

图3　系统架构

Fig.3　System structure

2.1　数据隐私保护存储过程

医疗信息是一笔重要的财富,医生结合以前的 EHR 对病人的病情进行分析,准确地对病人作出诊断并给出相应的治疗方案。在这个过程中,病人和医院会产生一系列的交易数据和 EHR,EHR 中主要包括:病人的一些个人信息,如姓名、性别和家庭地址等;医生的诊断信息,如病症、疾病的种类等;临床表现和治疗方案等信息。不法分子可以根据 EHR 非法获取病人的隐私信息,故需要对病人的 EHR 中敏感信息进行处理,故本文提出(p,α,k)隐私匿名算法对敏感信息进行保护处理。

定义 6　敏感属性分组。假设 A 是原始数据表 T 中敏感属性值的集合,根据集合 A 中敏感属性的隐私保护程度,A 被分为 m 个分组,则$(A_1,A_2,\cdots,A_m)$是敏感属性集合 A 的值域。

定义 7　隐私泄露率。病人隐私信息泄露的百分比。隐私泄露率的定义为

$$\alpha=\mathrm{count}(B,A_i)/\mathrm{count}(N_B)\tag{1}$$

其中:$\mathrm{count}(B,A_i)$是等价类 B 中敏感属性分组的隐私保护度 A_i 的属性值的数量;$\mathrm{count}(N_B)$是等价类 B 中敏感属性值的总数。对于敏感属性集合 $A=(A_1,A_2,\cdots,A_m)$,每个隐私保护度 A_i 都有对应的隐私泄露率,用 $\alpha_{A_1},\alpha_{A_2}\cdots,\alpha_{A_m}$ 表示,其中 $\alpha_{A_1}<\alpha_{A_2}<\cdots<\alpha_{A_m}$。

在隐私保护过程中设置约束值非常重要,通常由专家给出。通常专家会根据敏感属性值隐私保护程度的分组为不同的敏感属性分组提供不同的 α,该值的大小由等价类中敏感属性分组属性值的泄露严重性确定。为了给敏感属性分组的值提供更高级别的保护,属性分组可以设置为较小的 α,较低的保护级别可以设置为较大的 α。α 不仅可以直观地反映数据隐私保护的水平,而且可以避免更高的敏感属性值出现在同一等价类中。

定义 8　(p,α,k)隐私匿名算法。对于原始的数据 D 被匿名算法处理后的数据表 D'满足以下条件:k-匿名的性质;至少包含 p 个不同的敏感属性分组;任何敏感属性分组都满足 α。

算法 3　(p,α,k)隐私匿名算法

输入:数据集合 D,p,k,专家给的 α_{A_i}值,其中 $0<\alpha_{A_i}\leqslant 1$。

数据集 D 敏感属性的分组 $A=(A_1,A_2,\cdots,A_m)$。

输出:脱敏后的数据。

judge function / * 判断函数,对原始数据进行判断是否满足 k-匿名算法 * /

a)判断数据集 D 是否满足 k-匿名,首先令 $flag=$ true。

b)循环遍历数据集

```
for each QI-group in D:
    if count(QI 记录) > = k - 1:
        flag = true
    else
        flag = false
```

translate function / * 当不满足 k-匿名算法,需要对数据集进行转换 * /

c) 对原始数据集进行转换,产生等价类 D'。

d) judge(D', p,k),返回 D'。/ * 利用判断函数对产生的等价类进行判断,如果满足则返回等价类,如果不满足则执行步骤 e) * /

e) while QI-group$\geq p$ where $2 \leq p \leq k$

对 EHR 原始记录作匿名处理并将匿名处理后的 EHR 记录归并到 QI-groups,然后返回 D'。

(p,α,k)隐私匿名算法处理后 EHR 通过客户端上传到 IPFS,文件先存储在医疗记录文件夹地址 fileAddress 下以便上传到 IPFS 服务器。上传的 EHR 文件地址会记录在 folderAddress 文件中,folderAddress 是 IPNS 协议经过哈希加密后的医疗记录文件夹地址。当需要使用 EHR 文件的时候,可以通过 folderAddress 快速找到对应的文件,极大减少文件的访问时间。对 EHR 进行隐私处理,极大增强其安全性,保护病人的隐私数据。

2.2 身份验证授权和 EHR 共享调用过程

考虑到区块的容量和交易速度的局限,在系统中将病人和医院之间的交易数据通过改进的 hash 函数进行运算,存储在区块链区块中。各类 hash 算法在运行速度和 CPU 占有率上有很大不同的原因是迭代轮数、hash 值长度和区块大小等的不同[21]。区块链技术运行速度慢和存储空间占用大的问题可通过优化设计 hash 算法中的迭代轮数、hash 值长度得到改善[21]。故在系统中采用基于纠错码的 hash 函数对交易数据进行 hash 处理。算法处理流程如下:

a)将每个比特消息块中 512 位的信息分成 16 个子分组,每个子分组是 32 位。每个分组要进行 4 轮运算,1 轮 16 步,共需 64 步。每一步的运算通用形式如式(2)所示。

$$a = b + ((a + g(b,c,d) + X[k] + T[i]) <<< s) \tag{2}$$

其中:a、b、c、d 为 MD 缓冲区中寄存器 A、B、C、D 中的 4 个字;$g()$表示基本逻辑函数,基本逻辑函数如表 2 所示;$<<< s$ 表示对 32 位字循环左移 s 位;$X[k]$表示第 k 个 32 位字;$T(i)$表示加法常数表 T 中的第 i 个元素,加法常数表 T 中的元素是用于混淆的常数。

b)算法中通过纠错码生成矩阵构造,以达到更好的混淆效果,从而具备更好的随机性。根据纠错码的性质,纠错码的选择需要基于以下两点:(a)码字中 0 和 1 的个数尽可能一样多;(b)码间的最小汉明距离尽可能大。故本文根据其特性选择线性分组码(32,6,16)[22],根据线性分组码(32,6,16)生成矩阵。

c)为了使加法常数的随机性最大化,消除数据中的任何规律性,还可以对生成矩阵进行循环移位变化,将生成矩阵循环左移 6 位再进行运算,这样得到有效常数 62 个。取 MD5 算法中第一个和最后一个加法常数,结合得到的 62 个有效常数,重新构造纠错码的加法常数表。

d)把重新构造的加法常数表中的加法常数嵌入算法进行完整数据运算,最后输出 hash 值。

表 2 基本逻辑函数

Tab. 2 Basic logic function

轮数	轮函数 $RF()$	基本逻辑函数 $g()$
1	$RF_F(b,c,d)$	$(b \wedge c) \vee (b \wedge d)$
2	$RF_G(b,c,d)$	$(b \wedge d) \vee (c \wedge \overline{d})$
3	$RF_H(b,c,d)$	$(b \oplus c \oplus d)$
4	$RF_I(b,c,d)$	$(c(b \wedge \overline{d}))$

表 2 中,b、c、d 是取值为 0、1 的比特值。

算法 4 改进的 hash 函数算法

输入:交易信息 T。

输出:hash 值。

a)将 512 bit 交易信息分割成 16 组,

对 16 个分组进行 4 轮运算,每一轮 16 次。

b) for i in range(4)

for j in range(16)

$$a = b + ((a + g(b,c,d) + X[j] + T[m]) <<< s)$$

c)计算 $H(m) \rightarrow e$ 和 $d^{-1} \bmod n$,计算 $s = d^{-1}(e + kr) \bmod n$。利用线性分组码(32,6,16)生成矩阵 $\boldsymbol{G}_x$,然后在对矩阵左移 6 位。

$$\boldsymbol{G}_x = \begin{bmatrix} g_{00} & \cdots & g_{031} \\ \vdots & \ddots & \vdots \\ g_{50} & \cdots & g_{531} \end{bmatrix}$$

d)左移的过程产生 62 个新的常数,利用新的常数进行 hash 运算。

病人和医院之间产生的交易信息通过改进后的 hash 算法处理后,通过客户端上传到区块中,且节点账户保留此 hash 值。当病人或者医院需要病人的 EHR 时,通过客户端上传前面处理过交易的哈希值到区块链,节点通过上传的哈希值与其区块中的哈希值进行验证匹配,验证申请者的身份。如果匹配成功后,则验证通过,否则验证失败。由于系统中使用联盟链,联盟链的信息非授权节点无法看到区块的信息,故可以保证区块存放数据的安全性。当系统对访问者的身份验证通过后,访问者则可以通过 IPFS 下载相应的 EHR。医疗信息共享调用的过程,必须对申请者进行身份的验证,验证通过后,才能对其授权,极大地保证了 EHR 的安全性。

3 对系统的评价分析

3.1 安全性分析

本模型特点是采用 P2P 结构,避免了单点攻击,通过所有节点共同维护系统,可以很好地保证系统稳定性。系统中使用的基于纠错码的 hash 函数对交易数据进行 hash 运算,满足修改明文中的任意一个或多个比特都将导致输出比特串的近一半发生变化。故满足雪崩效应,加强了算法的安全性。利用(p,α,k)隐私匿名算法对 EHR 中疾病、手机号等信息进行敏感信息处理,处理后的数据上传到 IPFS 进行存储,故进一步加强了 EHR 隐私的安全性。

3.2 可靠性分析

该模型借助 Fabric 联盟链进行 EHR 的共享。对于系统的网络,记账节点和服务节点具有高可用性的特点,当网络出现抖动时,不会影响系统的服务;对于共享节点,节点具有高可用、支持 failover 同步和备份恢复的特性,并且只有授权的节点才能加入区块链;对于节点的账户,不同节点下的账户信息具有高可用性。根据上诉对系统网络、共享节点和节点账户分析得出,该系统具有高可靠性。

3.3 性能分析

采用对比分析法,从能否减轻主链压力、需要节点数、是否依赖可信的第三方、共识机制、隐私保护、安全存储和控制病历这七个方面的性能,对近几年提出的主流医疗模型进行分析。其中薛腾飞等人提出的 MDSM 使用改进的 DPOS 共识机制,能够减轻主链压力,但在网络的稳定性方面、病人对病历的控制和运算量有一定劣势,且需要依赖可信的第三方。ModelChain 在隐私保护和安全存储方面做得很好,但是在减轻主链压力方面有所欠缺。本文方案通过改进的 PBFT 增强网络的稳定性,降低了资源利用率,通过联盟链和改进的 hash 算法对身份进行验证,在对 EHR 进行存储前,通过匿名隐私算法对敏感信息进行处理。本文方案与具有代表性的模型进行对比的具体情况如表 3 所示。

3.4 效率分析

方案中数据块的交易确认时间设置为 10 min,而传统区块链(例如比特币系统)的交易块确认时间是 60 min。在获得数据块的共识方面,PBFT 需要节点之间的对等通信,因此该通信机制所需的共识节点的数量不需要过大,并且共识过程仅在病人和医院之间执行,而不是所有的网络节点。在此模式下,节点同意的速度更快,延迟更低,大大提高了区块共识的速度,也

从另一方面提高了交易的速度。与传统的区块链相比,方案将确认数据块所需的时间缩短了近 6 倍,并将传输效率提高了 83.33%,随着所需要确任的区块数量不断增加,两者之间所需的时间变化,如图 4 所示。由于节点数量的控制,本文方法不会像区块链系统那样消耗太多的算力,并且大大提高了整个网络的吞吐量。特别是数据块共识,仅需要实现相应节点的共识过程,而不是所有连接的节点。

表 3　性能对比分析

Tab. 3　Performance comparison analysis

性能指标	MDSM[5]	MedRec[7]	ModelChain[10]	MediBchain[23]	本文方案
减轻主链压力	是	否	否	否	是
需要节点数	121 个	多	多	多	少
是否依赖可信的第三方	是	是	是	是	否
共识机制	改进 DPOS	PoW	POI	PoW	改进 PBFT
隐私保护	是	是	是	是	是
安全存储	是	否	是	否	是
控制病历	控制不完整	控制不完整	控制完整	控制完整	完全控制

全网采用改进的 PBFT 共识机制,保证数据的真实有效。通过实验收集的数据得出,改进的 PBFT 共识机制比原以太坊使用的 PoW 共识机制更适用于本模型。如图 5 所示,数据线性趋势对比可以看出,改进后的 PBFT 比原来的 PoW 共识机制占用的时间明显减少,对请求可作出快速响应。

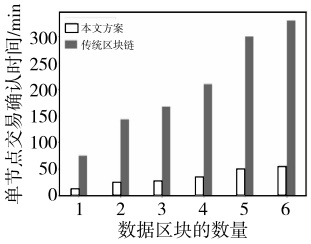

图4　区块确认时间

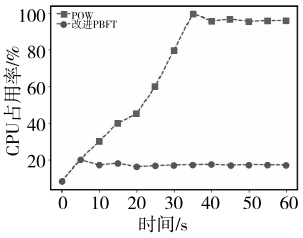

图5　共识机制 CPU占用率

Fig.4　Block confirmation time　Fig.5　Consensus mechanism CPU usage

3.5　隐私算法分析

本文算法使用 http://archive. ics. uci. edu/ml/datasets. php? format = &task = &att = &area = &numAtt = &numIns = &type = &sort = nameUp&view = table 网站和 http://people. dbmi. columbia. edu/ ~ friedma/Projects/DiseaseSymptomKB/index. html 网站中的公开医疗数据集。数据集中包括病人电话号码、种族、性别、年龄和所患疾病等信息。表 4 所列的是部分原始数据集。

表 4　原始医疗数据

Tab. 4　Original medical data

phone	race	gender	age	illness
86240259	Caucasian	female	[70 – 80)	fibroid tumor
114298353	African American	female	[40 – 50)	HIV
80367957	Caucasian	male	[70 – 80)	cancer
86209290	Caucasian	female	[70 – 80)	Alzheimer's disease
82844478	Caucasian	male	[70 – 80)	cancer
89709309	Caucasian	male	[70 – 80)	migraine disorders
80499960	Caucasian	male	[80 – 90)	hyperglycemia
82198143	Caucasian	male	[70 – 80)	upper respiratory infection
89869032	African American	female	[40 – 50)	bronchitis
114960726	Caucasian	female	[50 – 60)	hepatitis B
89193870	Caucasian	female	[40 – 50)	anemia
82894275	Caucasian	male	[60 – 70)	Alzheimer's disease
80177094	Caucasian	female	[80 – 90)	gout
114751503	Caucasian	female	[40 – 50)	influenza
80156115	Caucasian	female	[70 – 80)	epilepsy
86797872	Caucasian	female	[70 – 80)	malignant tumor of colon
82339443	Caucasian	male	[70 – 80)	fibroid tumor
89076051	African American	male	[50 – 60)	fibroid tumor
86880384	Caucasian	female	[70 – 80)	hyperglycemia
114194385	African American	female	[50 – 60)	neuropathy

由于不法分子可以根据 EHR 非法获取病人的隐私信息,故需要对病人的 EHR 中敏感信息进行处理。对原始数据中的疾病根据敏感程度对其进行分组,如表 5 所示。

对于原始数据集观察可知,对于疾病和电话号码等敏感属性没有作出相应的处理,不法分子可以利用信息之间的关联性,可以确认 EHR 信息对应到具体某个人。在上传 EHR 到系统前,需要对敏感信息进行脱敏处理。假设程序中 $p=3, k=4$,根据敏感隐私分组,利用$(3,\alpha,4)$隐私匿名算法对原始数据集进行匿名处理。部分原始数据处理的结果如表 6 所示。

表 5　敏感属性分组

Tab. 5　Sensitive attribute grouping

grouping ID	sensitive property value	grouping privacy
1	lymphoma, fibroid tumor, malignant tumor of colon, malignant neoplasms, cancer, HIV	A1
2	hepatitis B, Alzheimer's disease, neuropathy, asthma, epilepsy	A2
3	hypoglycemia, gout, hyperglycemia, anemia, hemiparesis	A3
4	influenza, migraine disorders, upper respiratory infection, bronchitis	A4

表 6　匿名处理结果

Tab. 6　Anonymous processing results

phone	race	gender	age	grouping ID
86 ******	Caucasian	female	[70 – 80)	1
86 ******	Caucasian	female	[70 – 80)	1
86 ******	Caucasian	female	[70 – 80)	3
86 ******	Caucasian	female	[70 – 80)	2
82 ******	Caucasian	male	[60 – 80)	1
82 ******	Caucasian	male	[60 – 80)	1
82 ******	Caucasian	male	[60 – 70)	2
82 ******	Caucasian	male	[60 – 80)	4
114 ******	*	female	[40 – 60)	2
114 ******	*	female	[40 – 60)	2
114 ******	*	female	[40 – 60)	4
114 ******	*	female	[40 – 60)	1
80 ******	Caucasian	*	[70 – 90)	3
80 ******	Caucasian	*	[70 – 90)	3
80 ******	Caucasian	*	[70 – 90)	2
80 ******	Caucasian	*	[70 – 90)	1
89 ******	*	*	[40 – 60)	3
89 ******	*	*	[40 – 60)	4
89 ******	*	*	[40 – 60)	1
89 ******	*	*	[40 – 60)	4

4　结束语

本文分析了联盟链的特性,探索了医疗信息交易主体交易结构,并提出了基于联盟链的 EHR 隐私保护和共享的系统,结合现阶段提出的医疗信息系统对比分析,发现本文研究对于其他医疗信息共享平台体系的构建具有一定的参考价值。

目前,由于区块链处理速度、操作壁垒的限制问题,再加上各大医院的设备条件问题和专业人员的缺乏,基于联盟链的 EHR 隐私保护和共享的系统在现实中推广使用还是不现实的;区块链技术需要不断完善,建立全球统一的交易体系架构及协议,提高操作的便利性、界面的友好性以及处理速度,才能推进其在医疗平台的应用。

参考文献:

[1] Thompson L A, Black E, Duff W P, *et al*. Protected health information on social networking sites: ethical and legal considerations[J]. Journal of Medical Internet Research, 2011, 13(1): 1-11.

[2] Sharma P K, Moon S Y, Park J H. Block-VN: a distributed blockchain based vehicular network architecture in smart city[J]. Journal of Information Processing Systems, 2017, 13(1): 184-195.

[3] Luu L, Chu D H, Olickel H, *et al*. Making smart contracts smarter [C]//Proc of ACM SIGSAC Conference on Computer and Communi-

cations Security. New York:ACM Press,2016:254-269.
[4] Yli-Huumo J, Ko D, Choi S, *et al*. Where is current research on blockchain technology? A systematic review[J]. PLoS One,2016,11(10):e0163477.
[5] 薛腾飞,傅群超,王枞,等.基于区块链的医疗数据共享模型研究[J].自动化学报,2017,43(9):1555-1562.(Xue Tengfei, Fu Qunchao, Wang Cong, *et al*. A medical data sharing model via blockchain[J]. Acta Automatica Sinica,2017,43(9):1555-1562.)
[6] Wood G. Ethereum:a secure decentralised generalised transaction ledger[EB/OL].(2018)[2018-08-19]. http://gavwood.com/paper.pdf.
[7] Azaria A, Ekblaw A, Vieira T, *et al*. Medrec:using blockchain for medical data access and permission management[C]//Proc of the 2nd International Conference on Open and Big Data. Piscataway,NJ:IEEE Press,2016:25-30.
[8] Ivan D. Moving toward a blockchain-based method for the secure storage of patient records[EB/OL].(2018)[2018-08-19]. https://www.healthit.gov/sites/default/files/9-16-drew_ivan_20160804_blockchain_for_healthcare_final.pdf.
[9] Shrier A A, Chang A, Diakun-thibault N, *et al*. Blockchain and health IT: algorithms, privacy, and data[EB/OL].(2018)[2018-08-19]. http://www.truevaluemetrics.org/DBpdfs/Technology/Blockchain/1-78-blockchainandhealthitalgorithmsprivacydata whitepaper.pdf.
[10] Kuo T T, Ohnomachado L. ModelChain:decentralized privacy-preserving healthcare predictive modeling framework on private blockchain networks[EB/OL].(2018)[2018-08-19]. https://www.healthit.gov/sites/default/files/10-30-ucsd-dbmi-onc-blockchain-challenge.pdf.
[11] Witchey N. Healthcare transaction validation via blockchain proof-of-work, systems and methods: U. S, US20150332283A1[P/OL].(2015-11-19). https://patentimages.storage.googleapis.com/72/da/6f/95585e5e8709c3/US20150332283A1.pdf
[12] Benet J. IPFS-content addressed, versioned, P2P file system[EB/OL].(2014-07-14). https://arxiv.org/abs/1407.3561.
[13] 黄秋波,安庆文,苏厚勤.一种改进PBFT算法作为以太坊共识机制的研究与实现[J].计算机应用与软件,2017,34(10):288-293,297.(Huang Qiubo, An Qingwen, Su Houqin. Study and realization of an improved PBFT algorithm as an Ethereum consensus mechanism[J]. Computer Applications and Software,2017,34(10):288-293,297.)
[14] Koblitz N. Elliptic curve cryptosystems[J]. Mathematics of Computation,1987,48(177):203-209.
[15] 霍峥,孟小峰,黄毅. PrivateCheckIn:一种移动社交网络中的轨迹隐私保护方法[J].计算机学报,2013,36(4):716-726.(Huo Zheng, Meng Xiaofeng, Huang Yi. PrivateCheckIn:trajectory privacy-preserving for check-in services in MSNS[J]. Chinese Journal of Computers,2013,36(4):716-726.)
[16] Wong R C W, Li Jiuyong, Fu A W C, *et al*. (α,k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing[C]//Proc of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM Press,2006:754-759.
[17] Truta T M, Campan A, Meyer P. Generating microdata with p-sensitive k-anonymity[J]. Lecture Notes in Computer Science,2007,4721(1):124-141.
[18] Sweeney L. k-ANONYMITY:a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,2002,10(05):557-570.
[19] MacWilliams F J, Sloane N J A. The theory of error-correcting codes[M].1977.
[20] Castro M, Liskov B. Practical Byzantine fault tolerance[C]//Proc of the 3rd Symposium on Operating Systems Design and Implementation. 1999:173-186.
[21] 巫光福,曾宪文,刘娟,等.基于纠错码的hash函数的设计与分析[J].信息网络安全,2018,18(1):67-72.(Wu Guangfu, Zeng Xianwen, Liu Juan, *et al*. Design and analysis of hash function based on error correcting code[J]. Netinfo Security,2018,18(1):67-72.)
[22] 巫光福.基于拟阵理论的二进制线性分组码的构造的研究[D].厦门:厦门大学,2012.(Wu Guangfu. Research on the construction of binary linear block codes based on matroid theory[D]. Xiamen: Xiamen University,2012.)
[23] Omar A A, Rahman M S, Basu A, *et al*. MediBchain: a blockchain based privacy preserving platform for healthcare data[C]//Proc of International Conference on Security, Privacy, and Anonymity in Computation, Communication, and Storage. Cham: Springer, 2017: 534-543.
[24] Nakamoto S. Bitcoin: a peer-to-peer electronic cash system[EB/OL].(2008). https://bitcoin.org/bitcoin.pdf.
[25] 张小红,郭焰辉.基于椭圆曲线密码的RFID系统安全认证协议研究[J].信息网络安全,2018,214(10):57-67.(Zhang Xiaohong, Guo Yanhui. Research on RFID system security authentication protocol based on elliptic curve cryptography[J]. Netinfo Network Security,2018,214(10):57-67.)
[26] Zhang Fangguo, Kim K. Efficient ID-based blind signature and proxy signature from bilinear pairings[C]//Proc of the 8th Australasian Conference on Information Security and Privacy. Berlin: Springer, 2003:312-323.
[27] Bellare M, Ran C, Krawczyk H. Keying hash functions for message authentication[C]//Advances in Cryptology-CRYPTO. 1996:1-15.
[28] Jean-Sebastien C, Dodis Y, Cecile M, *et al*. Merkle-Damgard revisited: how to construct a hash function[C]//Proc of the 25th Annual International Cryptology Conference. 2005:430-448.
[29] 陈兰香,邱林冰.基于Merkle哈希树的可验证密文检索方案[J].信息网络安全,2017(4):1-8.(Chen Lanxiang, Qiu Linbing. A verifiable ciphertext retrieval scheme based on Merkle hash tree[J]. Netinfo Security,2017(4):1-8.)

(上接第32页)

[9] Zheng Zibin, Xie Shaoan, Dai Hongning, *et al*. An overview of blockchain technology: architecture, consensus, and future trends[C]//Proc of IEEE International Congress on Big Data. Piscataway, NJ:IEEE Press,2017:557-564.
[10] 于戈,聂铁铮,李晓华,等.区块链系统中的分布式数据管理技术——挑战与展望[J/OL].计算机学报.[2020-09-29]. http://kns.cnki.net/kcms/detail/11.1826.tp.20191029.1604.004.html.(Yu Ge, Nie Tiezheng, Li Xiaohua, *et al*. The challenge and prospect of distributed data management techniques in blockchain systems[J/OL]. Chinese Journal of Computers,[2020-09-29]. http://kns.cnki.net/kcms/detail/11.1826.tp.20191029.1604.004.html.)
[11] 范吉立,李晓华,聂铁铮,等.区块链系统中智能合约技术综述[J].计算机科学,2019,46(11):1-10.(Fan Jili, Li Xiaohua, Nie Tiezheng, *et al*. Survey on smart contract based on blockchain system[J]. Computer Science,2019,46(11):1-10.)
[12] 武岳,李军祥.区块链共识算法演进过程[J].计算机应用研究,2020,37(7):2097-2103.(Wu Yue, Li Junxiang. Evolution process of blockchain consensus algorithm[J]. Application Research of Computers,2020,37(7):2097-2103.)
[13] 王煜,朱明,夏演.非对称加密算法在身份认证中的应用研究[J].计算机技术与发展,2020,30(1):94-98.(Wang Yu, Zhu Ming, Xia Yan. Application research of asymmetric encryption algorithm in identity authentication[J]. Computer Technology and Development,2020,30(1):94-98.)
[14] Magrahi H, Omrane N, Senot O, *et al*. NFB: a protocol for notarizing files over the blockchain[C]//Proc of the 9th IFIP International Conference on New Technologies, Mobility and Security. Piscataway, NJ:IEEE Press,2018:1-4.
[15] Renner T, Myuller J, Kao O. Endolith: a blockchain-based framework to enhance data retention in cloud storages[C]//Proc of the 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing. Piscataway,NJ:IEEE Press,2018:627-634.
[16] 黄凯峰,张胜利,金石.区块链智能合约安全研究[J].信息安全研究,2019,5(3):10-24.(Huang Kaifeng, Zhang Shengli, Jin Shi. The security research of blockchain smart contract[J]. Journal of Information Security Research,2019,5(3):10-24.)
[17] Nguyen V, Trang H, Nguyen Q. Building mathematical models applied to UTXOs selection for objective transactions[C]//Proc of the 5th NAFOSTED Conference on Information and Computer Science. Piscataway,NJ:IEEE Press,2018:160-164.
[18] Kang Jiawen, Xiong Zehui, Niyato D, *et al*. Incentivizing consensus propagation in proof-of-stake based consortium blockchain networks[J]. IEEE Wireless Communications Letters,2019,8(1):157-160.