HIS 中授信额度访问控制隐私保护模型

摘 要： 随着医疗数据规模日益增长，医疗数据复杂的情况下，大部分医院都引进使用医疗信息管理系统（Hospital Information System，HIS），无纸化的病历系统提供了便利的同时患者隐私泄露的风险也随之提高，医生的访问权限控制很有必要。为了降低诊疗过程中医疗数据被访问时患者隐私泄露的危险，本文提出一种基于信任的动态权限访问控制模型。根据医生诊疗过程中给出的目标和实际访问记录的相关程度刻画访问的合理性，使用熵值法计算医生行为相关度。考虑信任的历史性和时间，通过HIS中医生的历史访问记录计算得到授信额度，并通过访问控制策略授权匹配来达到限制医生访问权限的目的。最后使用合作医院提供的真实数据的完成相关实验，证明本文提出的基于信任的访问控制模型能够实现在HIS背景下对医生诊疗过程隐私泄露的有效控制。
关键词： HIS；数据隐私；授信额度；访问控制

The privacy protection model of credit line access control in HIS

Abstract: With the increasing scale of medical data and the complexity of medical data, most hospitals have introduced and used the Hospital Information System (HIS). The paperless medical record System provides convenience while the risk of patient privacy disclosure also increases, so it is necessary to control the access rights of doctors.In order to reduce the risk of patient privacy disclosure when medical data is accessed in the process of diagnosis and treatment, this paper proposes a trust-based dynamic access control model.The rationality of the visit was described according to the correlation degree between the target given by doctors and the actual visit records, and the correlation degree of doctors' behavior was calculated by using the entropy method.Considering the history and time of trust, the credit limit is calculated by the historical visit records of doctors in HIS, and the access control policy authorization matching is used to limit the access rights of doctors.Finally, the actual data provided by cooperative hospitals are used to complete relevant experiments, proving that the trust-based access control model proposed in this paper can effectively control the privacy leakage of doctors in the process of diagnosis and treatment in the context of HIS.
Key words: HIS; Data privacy; Line of credit; Access control

在医疗数据方面个人用户成为了数据的重要来源，医疗隐私信息对个体往往意味着难以启齿的隐痛、痛苦不堪的经历这类消极信息。不法分子一旦窃取医疗数据就能够轻而易举的得知患者的名字，家庭住址，联系方式、检验报告、诊断结果甚至医保等重要信息，以此来伪造资料诈骗或者购买医疗设备，因此医疗行业数据盗窃后果非常严重[1]。

随着中国加入 WTO 和社会信息化进程的加快，是否拥有功能完整的 HIS 已经成为衡量一个医院综合实力的重要标志[2]。但由于 HIS 的使用医生通过电子病历能轻易接触到大量的医疗信息，由此衍生的医疗隐私泄露风险也令人担忧。与此同时市面上普遍使用的 HIS 并没有针对医院内部人员可能造成的信息泄露采取的措施，相关研究也比较缺乏。如果

医生接触患者的医疗信息没有相应的规则约束，医生的信息泄露成本非常低。为了使得医生的访问尽可能的满足 need-to-know 原则[3]，本文针对目前 HIS 中对于医生访问权限自适应控制的空白，且现有的访问控制模型[4, 5]无法应对医疗数据爆炸且访问细粒度问题，提出了一种基于授信额度、面向 HIS 中医生诊疗过程的访问控制模型。

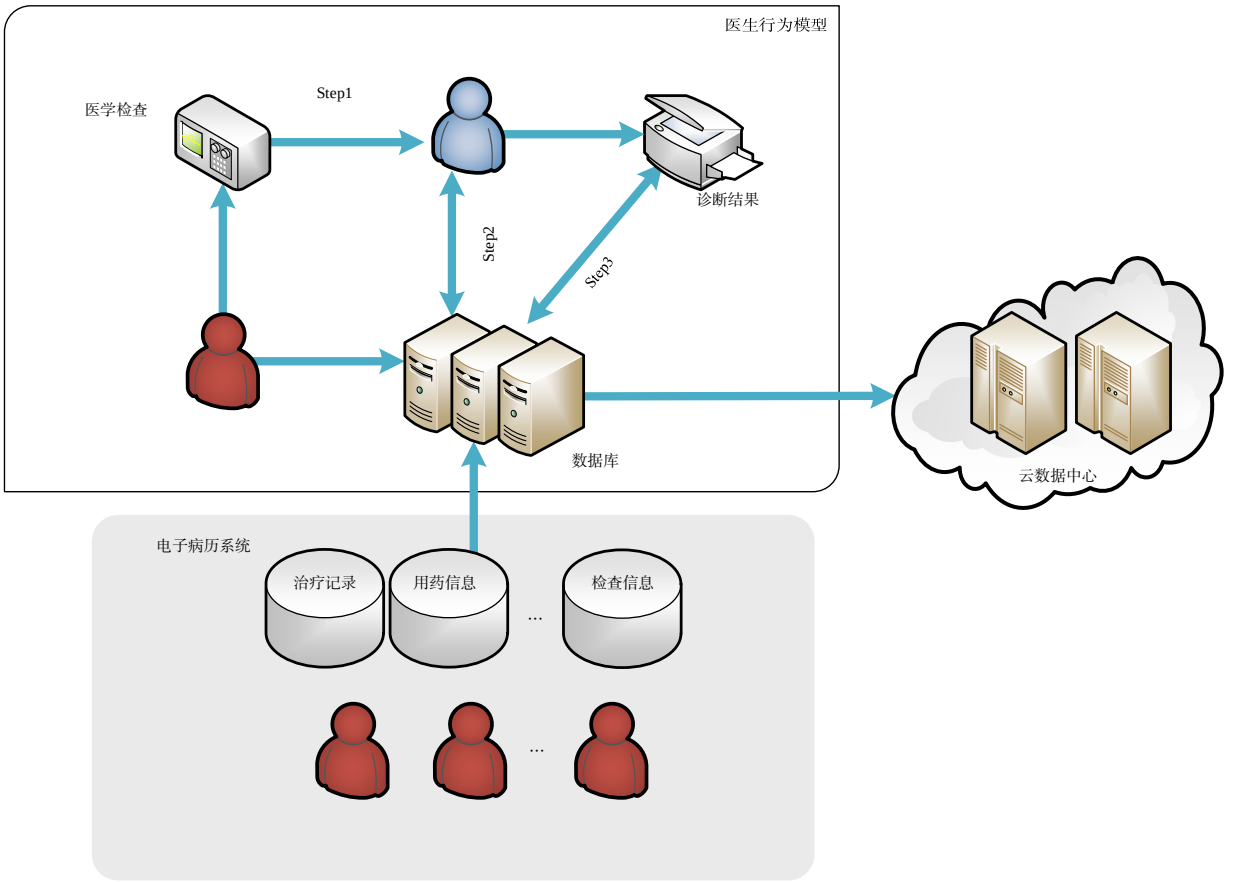


Fig.1　Doctor behavior model

图 1　HIS 医生行为模型

HIS 中个子系统之间的高度集成提高了医院整体运行效率，改善了患者就医环境，同时为医院的管理、临床等提供了数据化支持。电子病历系统[6]主要功能就是以电子的方式将患者的就诊记录保存下来，不但包括患者的用药信息，还包括了患者的治疗记录，化验及检查记录等信息。医生在给患者的诊疗期间就可以调取患者以往的病历和病史，能够更精准的分析患者病情对患者进行治疗。图 1 抽象表示了医生诊疗过程中的行为模型。简化 HIS 诊疗过程的用户都为医生。针对每个患者的诊疗过程我们称之为一次任务，在任务中医生的诊治步骤一般是：首先浏览患者的基本信息，例如姓名、年龄、过往病史。如果患者在该医院以往有就诊史，医生能通过 HIS 的数据库看见患者过去的检查项目和结果。随后患者根据医生的安排将获得新的检查结果，同样储存在 HIS 数据库中。医生根据查看患者的化验结果，以及某些相关医疗记录（比如数据库中其他相似诊断患者的医学影像资料等）来为该患者得出最终诊断。医生可查看的 HIS 数据库中的医疗数据涉及到病人的敏感信息，但考虑道德因素和泄密成本，假设模型中的医生不会泄露自己主治患者的任何信息。则基于以上行为模型，好奇医生隐私泄露的行为将会发生在以下三个步骤中：

Step$_1$：检查信息和初步诊断的匹配率低，即患者做的检查结果可以直接表明患者并没有传染病的可能，但是医生依旧给出了可能患有感染病的初步诊断，随后访问传染病的患者相关医疗记录。

Step$_2$：假设医生在 *Step*$_1$ 中给出了与检查信息相符合的初步诊断，但在根据初步诊断访问医疗记录时查询了不必要的医疗记录。

Step$_3$:假设医生在 *Step*$_1$、*Step*$_2$ 中正常操作，但是最终诊断与查询的医疗记录相关度偏低，怀疑医生访问了不必要医疗记录。

给出符号的形式化描述如下，并且把过程抽象表示：

M：检查信息的集合；

I：初步诊断的集合；

F：最终诊断的集合；

R：医疗记录的集合；

$S_1: M, \mathrm{I} \to [0,1]$:定义检查信息和初步诊断的相关性函数，其中 $m \in M, i \in I$，函数返回值反应了某次诊疗过程二者的相关程度。

$S_2: \mathrm{I}, \mathrm{R} \to [0,1]$:定义初步诊断和医疗记录的相关性函数，其中 $i \in I, \mathrm{r} \in R$，函数返回值反应了某次诊疗过程二者的相关程度。

$S_3: R, F \to [0,1]$:定义医疗记录和最终诊断的相关性函数，其中 $r \in R, \mathrm{f} \in F$ 函数返回值反应了某次诊疗过程二者的相关程度。

1　基于信任评估的隐私保护

随着时代的发展，信任的含义有所拓展。用户基于其在整个网络中的表现以及与其他用户关系互动和信息互动而获得的可信任程度，可作用于风险规避。复杂的网络环境中信任问题需要通过建立可行和合理的信任评估模型来解决，因此信任评估模型在数据安全隐私保护研究领域中发挥着基础性作用，它需使用一种信任评估方法对复杂的网络环境中的用户信息进行信任评估[7]。

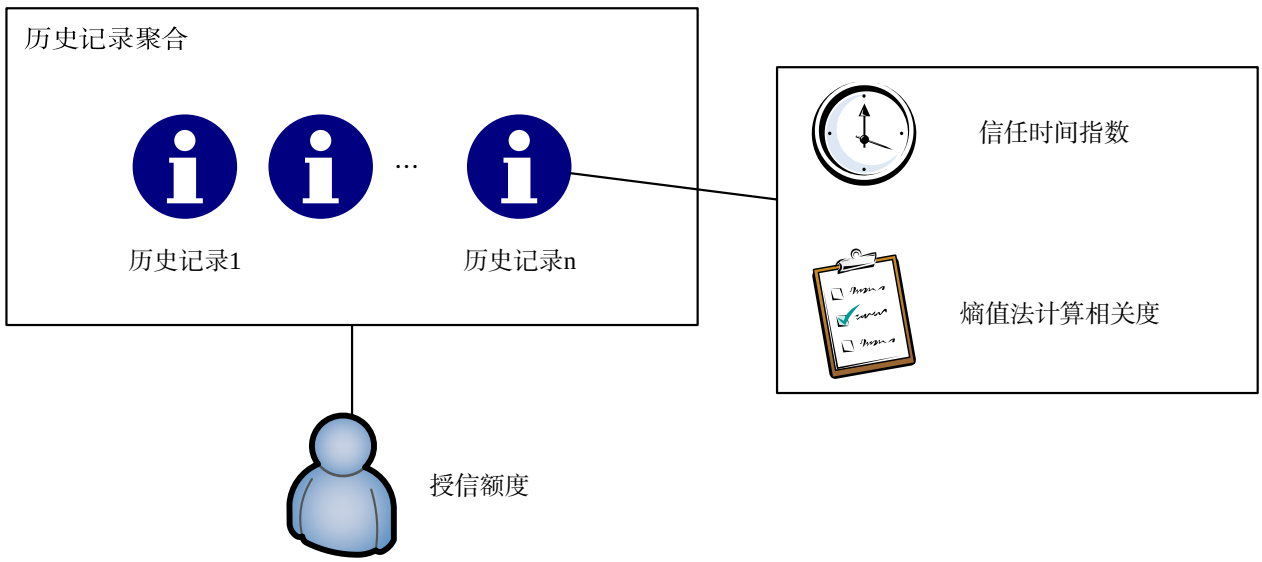


Fig.2　Trust calculation process

图 2　信任计算过程

2.1 信任属性体系

现有的信任体系相对单一，通常分为直接信任和间接信任[7]，在医生诊疗的背景下，这种信任模型无法准确评估医生行为的可信度。医生在 HIS 中是大量接触患者和医疗记录的角色，但是每个医生之间，或者患者和其他医生（未为自己诊断）之间缺少有效的直接信任联系。因此本文根据医生诊疗的行为特点和医疗资源体系的结构的特殊性，在信任属性中不考虑间接信任，只考虑医生的历史访问记录会直接影响其授信额度。

2.2 授信额度思想

授信的概念源自于金融领域，授信额度[8]意为银行在授信期内经过计算评估给予用户的最高授信值，也代表了银行对该用户最高风险承担能力的量化。

定义 1 授信额度（credit line）：根据 HIS 医生用户历史记录、访问行为因素综合评估，授予医生用户的可以透支使用的信任额度。

通过 HIS 读取医生的历史记录计算得出授信额度，医生持续的诚信行为记录有助于授信额度的增加，相关度低的访问则会导致额度下降。本文引入国际疾病分类编码 ICD（International Classification of Diseases）[9]作为医生诊断时电子病历使用的代码。历史记录包括信任时间指数和操作相关度两个子属性。

2.3 信任计算

医生诊断过程中涉及三个步骤都有不同大小的隐私泄露风险，通过相关性函数来计算风险值。度量相似性的方法有很多[10]。大量使用到的度量相似性的方法例如余弦或者相关相似性在医疗数据类别基数比较大的背景之下，由于数据密集容易到相似度很高的错误结论，故将交叉熵[11]引入计算两个随机变量之间的相似度。

定义 2 熵（Entropy）：信息量的期望值，假设有一个随机变量 x，取值范围为集合 X，其概率分布函数可以表示为 $\mathrm{p}(x)=\Pr(X=x), x\in X$ 信息量为 $I(x_1)=-\lg(p(x_1))$ 当一件事发生的概率越大，即 $p(x_1)$ 越大，那么它所携带的信息量就越小[12]。极端情况下 $p(x_1)=1$，信息量等于零，意为当一件事发生的概率是百分之百时，那这件事发生不会引入太多的信息量。当我们已知信息量来衡量一件事发生的不确定度，可以通过对所有可能得到的结果带来的额外信息量求期望 $\left(E\left[I(x)\right]\right)$ 即得到了熵。

$$H(X)=Ep\left[\lg p(x)\right]=-\sum x\in Xp(x)\lg p(x) \tag{1}$$

定义 3 交叉熵（Cross entropy）：设初步诊断服从一随机分布 p，某医生某次诊疗过程中的访问记录服从随机分布 q，交叉熵计算 p、q 的相似程度。按照分布 p 得到的期望为 $H(p)$，对于医生的诊断过程，访问记录为离散变量，用 q 分布来表示 p 分布得到 $H(p,q)$。根据 ICD 编码统计涵盖的疾病计算得出每个初步诊断（疾病）需要的平均的信息量为阈值。

$$H(\mathrm{p})=\sum_i p(i)*\lg\frac{1}{p(i)} \tag{2}$$

$$H(p,q)=\sum_i p(i)*\lg\frac{1}{q(i)} \tag{3}$$

根据以上给出的定义假设一疾病的 p 可以表示为 $[1,0,0][1,0,0]$，一名医生 A 访问历史记录得到的 q 为 $[0.5,0.4,0.1][0.5,0.4,0.1]$，那么根据（3）交叉熵的计算方法可以得出医生该次诊断过程中的访问行为和医生 A 给出的初次诊断之间的交叉熵为：

$$H(p=[1,0,0],q=[0.5,0.4,0.1])=-(1*\lg 0.5+0*\lg 0.4+0*\lg 0.1)\approx$$
$$0.3H(p=[1,0,0],q=[0.5,0.4,0.1])=-(1*\lg 0.5+0*\lg 0.4+0*\lg 0.1)\approx 0.3$$

如果对于给出的是同样初步诊断结果的医生 B 访问记录 q 是 $[0.8,0.1,0.1][0.8,0.1,0.1]$，那么医生 B 该次诊断过程中访问和初步判断之间的交叉熵为：

$$H\left(p=[1,0,0],q=[0.8,0.1,0.1]\right)=$$
$$-(1*\lg 0.8+0*\lg 0.1+0*\lg 0.1)\approx$$
$$0.1H\left(p=[1,0,0],q=[0.8,0.1,0.1]\right)=$$
$$-(1*\lg 0.8+0*\lg 0.1+0*\lg 0.1)\approx 0.1$$

从计算结果可以明显看出医生 B 的交叉熵数值较小，即操作相关度要高。若已知该疾病的阈值为 0.2，那么可以得出结论医生 B 是在安全的访问医疗数据，医生 A 有较大隐私泄露的嫌疑。

定义 4 信任时间指数（Trust time function）：在授信额度计算的过程中，历史记录的时间即医生每个诊疗发生的时间。早期的历史记录将对授信的影响会随着时间渐弱，相反如果近期发生了不诚信访问行为，授信额度会被更大程度的被影响，在一段时期内保持低数值，以此作为惩罚。将每个历史访问记录的时间映射在$[0,1]$值域内。映射方法如下：设医生诊疗历史任务集 $HT=\{T_{hk}(1\le k\le q)\}(q=|HT|)$ 对应的诊疗发生时间 $DT=\{dt_k|1\le k\le q\}$，则对应某任务 $T_{hk}(1\le k\le q)$，信任时间指数为：

$$D_k=\frac{1-dt_k/\sum_{k=1}^{q}dt_k}{\sum_{k=1}^{q}\left(1-dt_k/\sum_{k=1}^{q}dt_k\right)} \tag{4}$$

历史记录越久远，信任时间指数越小，符合现实生活中越早时期的历史记录将对当前授信产生越小影响的情况[13]。

2 访问控制模型

访问控制技术，指防止对任何资源进行未授权的访问[14]，从而使计算机系统在合法的范围内使用。基于信任的访问控制通过用户身份及其所归属的信任评估结果来限制用户对某些信息项的访问，或限制对某些控制功能的使用[15]。传统访问控制带来的工作量显然不能适应大数据背景下 HIS 中海量数据的情况，为了细化云环境中数据访问控制的粒度，构建灵活的动态访问控制模型[16]，要先解决权限需预先设定分配等问题，以此提高权限控制的兼容性自动化，从而弥补静态策略的不足。

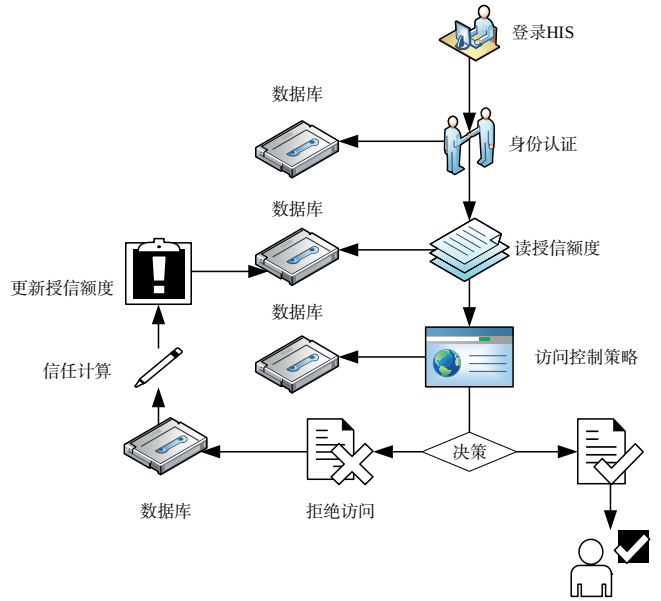


Fig.3 Access control flowchart

图 3 访问控制流程图

3 授信额度的访问控制

本文引入授信额度的概念针对现有的 HIS 中医生诊疗过程访问控制进行改进。医生根据身份信息（医生登录设备、时间、地点）等进行登录 HIS，每个用户根据系统计算得出相应的授信额度，用于根据访问控制策略匹配合理的权限。

医生每次诊断结束后，访问记录等数据 HIS 会保存在历史记录中。通过信任计算可以得到一段期间内用户的授信额度。

所在的信任区间对应了权限的开放程度。当某医生的授信额度为 t，根据访问控制策略，该医生只允许在诊疗过程中访问授信额度 t 所在区间对应的相关度，如果医生访问了过多无关内容，操作相关度的下降会导致该医生授信额度降低。当额度不足时用户的访问要求则会被拒绝。访问控制方案总体流程图如图 3 所示。

4 实验结果及分析

4.1 数据来源

本文依托国家自然科学基金项目，根据项目合作单位昆明市某三甲医院提供的医疗数据集合完成了相关研究实验。数据集包含丰富的文本数据、图像数据及影像数据，数据总共五个数据库，大小1200G，包含了 1360 张数据表，共计 2139373 条记录。本实验抽取部分医疗数据模拟医生诊疗过程中的访问情况。

4.2 实验设置

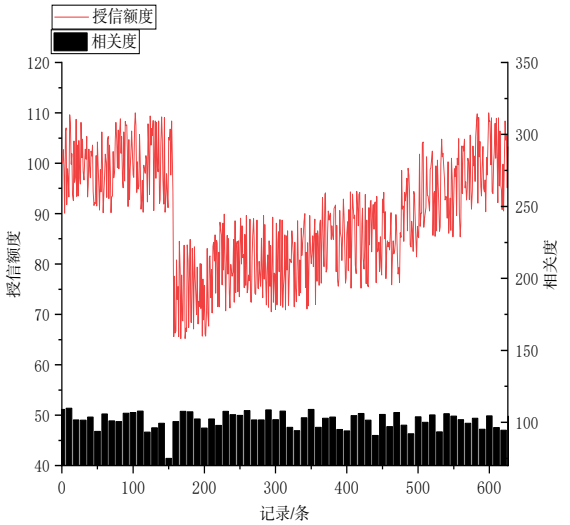
实验的目的是验证本文提出的基于 HIS 访问控制模型是否能够通过医生的历史行为记录计算出授信额度，并且通过授信额度的数值来很好的控制医生的访问权限。抽取合作医院提供的某科室的 3 名医生 HIS 账户访问记录数据计算，其中包含 1 名模拟好奇医生的行为和 1 位模拟特殊访问情形的诚实医生作为实验组。

4.3 访问控制实验

实验设置授信额度计算的历史记录时间周期 N 为一个月，每位医生每周坐诊三天，日均就诊记录在 50 左右。根据实验结果表明，医生发生恶意访问行为后在保持原有平均相似度水平的情况下，授信额度完全稳定在原先水平，需要大概 650 条记录接近一个月的时间。医生发生恶意访问行为后若有意的积极提高自己的平均访问相关度，如下图实验结果所示，授信额度想要完全稳定原先水平，需要大概 250 条记录，接近半个月。实验证明，在恶意访问发生时，授信额度的数值会即刻受到影响，作为隐私风险惩罚在较长的一段时间内授信额度都会处于低值以此来警示用户的不良行为，并同时达到访问控制的效果。

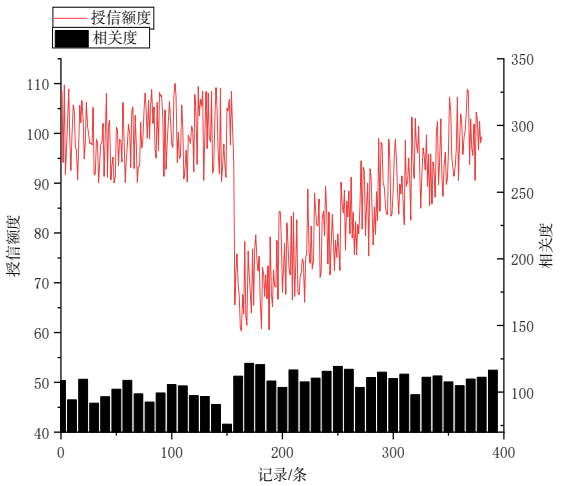
4.4 对比实验

经了解知项目合作医院目前使用的为不带访问控制的传统 HIS，对比实验随机选取医院内 100 名医生进行黑盒测试实验，医生被分成两组分别使用传统 HIS 和本文提出的信任访问控制模型的 HIS 进行周期为一个月的对比试验。在医生都未知实验内容的情况下，对两组医生的历史访问记录进行分析。



(a) General behavior after malicious access

(a) 恶意访问后一般行为



(b) Positive behavior after malicious access

(b) 恶意访问后积极行为

Fig.4 Credit line and correlation change

图 4 授信额度及相关度变化

根据图表可知实验一周时间内，两种 HIS 的访问记录相关度并没有很大的差距，实验整个周期内使用传统 HIS 的总访问相关度没有明显变化，而信任 HIS 模型则有显著的提升，这说明授信额度的提出可以一定程度上规范用户的行为。

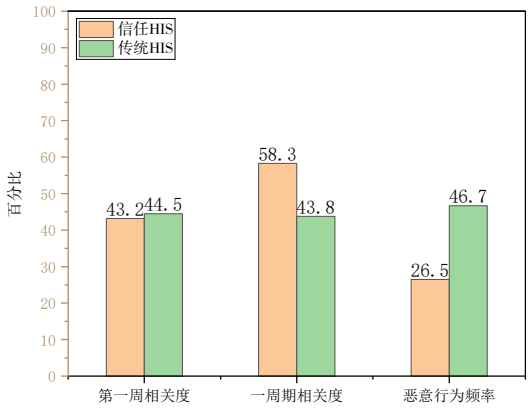


Fig.5 Comparison of correlation degree and behavior constraint ability of HIS

图 5 HIS 相关度及行为约束能力比较

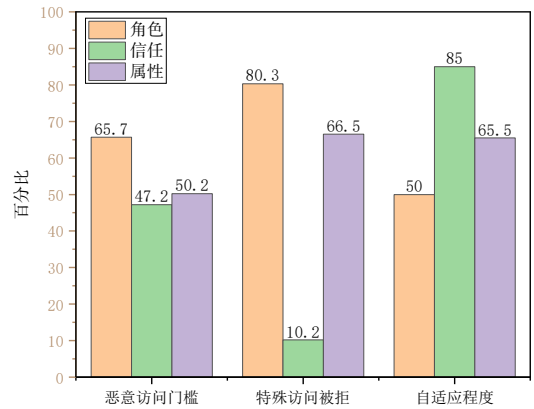


Fig.6 Comparison between traditional HIS and trust-based access control HIS

图 6 传统 HIS 与基于信任的访问控制 HIS

本文提出的授信额度模型与基于用户的角色判定访问范围的访问控制模型（以下简称角色）和基于属性根据科室划分的访问控制模型（简称为属性）对比，根据全部医生 256 用户的反馈如下表。

Table 1 Doctor User Rating Form

表 1 医生用户评价表

	恶意访问的困难程度			特殊访问请求的被拒绝风险			系统自动化程度		
级别	1	2	3	1	2	3	1	2	3
属性	73	128	55	42	189	25	16	127	13
角色	156	56	44	206	34	16	116	25	15
信任	80	84	92	10	34	212	3	64	189

根据表格数据得到如下图，本文提出的基于信任的 HIS 访问控制模型在访问控制的灵活性、阻止恶意访问行为发生以及在系统自动化的程度上都有较好的表现。

5. 结束语

针对医疗大数据背景下的 HIS 对访问控制的需求，本文提出了一种基于授信额度的医生诊疗过程的动态访问控制模型。实验证实本文设计的模型不仅能自动的计算医生的授信额度，并且根据其值准确的甄别好奇医生，并抑制其的访问。

参考文献：

[1] 郭子菁. 罗玉川. 蔡志平, 等 医疗健康大数据隐私保护综述[J].计算机科学与探索, 15 (2021) 389-402.

[2] 王强芬. 大数据时代背景下医疗隐私保护的伦理困境及实现途径[J]. 中国医学伦理学, 29 (2016) 685-689.

[3] R.S. Sandhu.P. Samarati.Access-control - principles and practice[J] Ieee Communications Magazine, 32 (1994) 40-48.

[4] 高鹏祥. 基于信任和角色的动态访问控制模型的研究和设计[D]. 天津：天津大学, 2014.

[5] 李昊. 张敏. 冯登国,等. 大数据访问控制研究[J]. 计算机学报, 40 (2017) 72-91.

[6] S.H. Bardach, K. Real, D.R. Bardach, Perspectives of healthcare practitioners: An exploration of interprofessional communication using electronic medical records[J]. Journal of interprofessional care, 31 (2017) 300-306.

[7] 张仕斌. 许春香. 基于云模型的信任评估方法研究[J]. 计算机学报, 36 (2013) 422-431.

[8] 应千伟. 罗党论. 授信额度与投资效率[J]. 金融研究, (2012) 151-163.

[9] P. Wu, A. Gifford, X. Meng, X. Li, H. Campbell, et al .Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation[J].JMIR medical informatics, 7 (2019) e14325.
[10] 张晓琳. 付英姿. 褚培肖. 杰卡德相似系数在推荐系统中的应用[J]. 计算机技术与发展, 25 (2015) 158-161+165.
[11] A. Jamin, A. Humeau-Heurtier, (Multiscale) Cross-Entropy Methods: A Review[J]. Entropy, 22 (2020) 15.
[12] Z. Zhang, M.R. Sabuncu, Generalized cross entropy loss for training deep neural networks with noisy labels[J]. arXiv preprint arXiv:1805.07836, DOI (2018).
[13] J. Caverlee, L. Liu, S. Webb, The SocialTrust framework for trusted social information management: Architecture and algorithms[J]. Information Sciences, 180 (2010) 95-112.
[14] G.M.H. Narayanan H A, Ensuring access control in cloud provisioned healthcare systems[J]. consumer communications and networking conference, DOI (2011) 247-251.
[15] X. Wang, L. Wang, Y. Li, K. Gai, Privacy-aware efficient fine-grained data access control in Internet of medical things based fog computing[J]. IEEE Access, 6 (2018) 47657-47665.
[16] 张怡婷. 傅煜川. 杨明, 等 基于 PBAC 模型和 IBE 的医疗数据访问控制方案通信学报[J]. 36 (2015) 200-211.