开放数据集中的安全规则建立及查询校验

张 圆 王 梅 乐嘉锦
(东华大学计算机科学与技术学院 上海 201620)

摘 要 数据开放为海量数据中的价值得以最大化利用提供了可能,然而数据的安全性问题却成开放共享的最大阻碍。针对开放数据集查询分析结果包含的隐私信息,基于其与数据访问行为的直接联系,设计一种安全规则存储结构。提出面向自然语言的安全需求描述接口,对待保护的隐私信息进行灵活、方便的描述,进一步提出自然语言隐私需求描述到安全规则的自动转换方法。在此基础上建立数据安全自动审核模型,该模型根据数据拥有者的安全需求对访问者的数据行为进行审核,实现在数据隐私不泄露的前提下保证数据开放的最大化。通过真实数据集上的实验表明:安全规则能够准确地捕获数据提供者的隐私保护需求,并能够有效地保证数据的安全性。

关键词 数据开放 隐私数据 安全规则 自然语言 查询审核

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2019.01.009

ESTABLISHMENT OF SECURITY RULES AND QUERY VERIFICATION FOR OPEN DATASET

Zhang Yuan Wang Mei Le Jiajin
(*College of Computer Science and Technology, Donghua University, Shanghai 201620, China*)

Abstract Data openness provides the great opportunities to maximize the value in massive data, but the data security issue becomes the big obstacle to open sharing. In view of the privacy information contained in the results of open data set query analysis, a storage structure of the security rules was designed based on its direct connection with data access behavior. Natural language-oriented security requirements description interface was proposed to describe the protected privacy information flexibly and conveniently. Furthermore, an automatic conversion method from privacy requirement description of natural language to security rules was proposed. On this basis, an automatic data query verification model was established, which checked the data behavior of users according to the security needs of data owners. It realized the maximization of data openness without revealing data privacy. Experiments on real data sets show that security rules can accurately capture the privacy protection needs of data providers and effectively guarantee the security of data.

Keywords Data openness Privacy data Security rules Natural language Query verification

0 引 言

数据开放共享是指政府和公共数据资源应该开放给公众查看和使用。在大数据时代,公共数据资源的开放共享为充分释放海量数据中的价值提供了可能,可带来巨大的社会与经济效益[1]。因此,各国政府对数据资源的开放利用给予了高度的重视[2-3]。然而由于数据开放所带来的数据安全和隐私泄露的风险[4-5],使得数据资源的开放共享变得十分困难。如何在数据开放的过程中保证数据的隐私和安全性成为了目前亟待解决的关键问题。

开放数据集中的隐私数据可大致分为两类,一类是本身为敏感信息的数据,如医疗数据中的医疗卡号、姓名等个人隐私数据;另一类是对发布后的数据进行分析获得的隐私数据,如医生的手术成功率、检查/检

收稿日期:2018-09-04。上海市科技创新行动计划项目(16JC1400803)。**张圆**,硕士生,主研领域:数据库,自然语言,安全规则,数据溯源。**王梅**,教授。**乐嘉锦**,教授。

验结果的一致性等。对上述信息进行保护,对于前者可采用访问控制、数据加密等技术。基于 BLP 安全访问控制模型[6]的分级开放[7],根据数据来源、数据重要性等指定数据的安全等级,当数据使用者的安全等级高于数据安全级别时方可访问数据。这种基于安全等级的访问方式保证了用户不能越级访问受限数据,具有较高的保密性。基于数据加密的方式,通过 AES[8-9]、ECC[10]和 RSA[11-12]等加密算法以及它们的混合加密算法[13],对原始数据进行加密,实现信息隐蔽,保证隐私数据不被其他用户访问。然而,上述方法主要关注于保护原始数据的安全性,很难应对后续数据分析带来的隐私泄露。更多的隐私保护方法被提出以应对数据发布和数据分析带来的隐私威胁问题,如 k-匿名[14]隐私保护、差分隐私[15]等。差分隐私通过对原始数据、原始数据的压缩数据、原始数据的统计信息添加噪声扰动来达到隐私保护效果,是目前公认较为严格和强健的保护方法。围绕如何添加噪声满足差分隐私,现已存在多种差分隐私算法[16-18]及相应的基于差分隐私的查询处理技术[19]。然而现有方法仍存在关键两个问题有待解决:一是如何减少噪声带来的误差,以提高数据的可用性[20]。尤其是对于以下的场景:医生的手术成功率为医院的敏感信息,然而手术成功的患者的用药信息则是可查询的信息。此时,若采用差分隐私方法对手术结果字段进行扰动,则可能导致后一条查询的查询结果无法准确获知,极大地影响数据的可用性。二是对于如何表示隐私数据,现有方法均缺乏有效考虑。因此,现有数据开放模型,往往在用户的查询结束后,需通过人工审核数据处理流程和查询结果进行审核。如何准确地表达、描述多种安全需求,自动识别隐私数据的数据访问数据行为方面研究不足。

针对上述问题,本文提出了开放数据集中的安全规则建立及查询校验方法。本文首先提出了面向自然语言的安全需求描述接口,对不同行业背景的数据提供者提供了灵活、方便的隐私数据描述途径。进一步,本文基于隐私数据与其数据访问行为的直接关系,借助于数据库自然语言接口技术,本文实现了自然语言安全需求到数据访问安全规则的自动转换。在此基础上,本文提出了基于安全规则的数据访问行为自动审核方法。对数据使用者的查询需求进行细粒度审核,自动过滤违反安全约束的非法请求,实现在数据隐私不泄露的前提下保证数据开放的最大化。最后,基于实际医疗数据,对本文提出的安全规则创建和审核方法进行测试,验证了本文安全规则建立及查询校验的有效性。

本文主要贡献如下:

(1) 针对开放数据的提供者,本文提出了面向自然语言的安全需求描述接口。从数据的访问行为出发,对用户的安全需求进行解析,提取该安全需求对应的数据访问核心元素,构成相应的安全规则,在此基础上设计了面向数据行为的安全规则存储结构及自然语言安全需求描述到安全规则转换的详细流程。

(2) 针对开放数据的使用者,本文设计了基于安全规则的数据访问行为自动审核方法。对用户查询语言进行细粒度审核,自动过滤违反安全约束的请求,实现在数据隐私不泄露的前提下保证数据开放的最大化。

(3) 最后,基于实际医疗数据,对本文提出的安全规则创建和审核方法进行测试,验证了本文安全规则的有效性和创建的准确性。

1 基于安全规则的查询审核模型

本文提出的基于数据访问的安全规则接口主要由三部分构成,依次为规则生成、规则存储和数据访问行为审核。图 1 描述了系统的完整框架。首先,数据提供者可以根据隐私数据保护需求提交规则创建请求,经规则生成模块构建安全规则后,由存储模块保存并入库。当数据使用者发起查询请求后,数据行为自动审核模块便会主动查询规则库,判断该查询是否违反安全规则限定,以确定是否进行下一步数据库查询操作。下面分别对安全规则的结构设计和构建,以及基于规则的查询审核进行详细介绍。

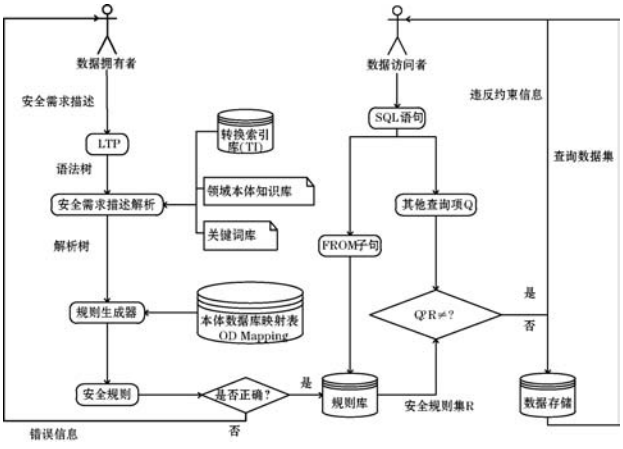

图 1 数据安全访问模型框架

2 安全规则结构设计

假定某数据集存在安全需求“普通用户不能查询

医生的用药习惯”,很明显该数据集上待保护的隐私数据为“用药习惯”,即医生根据诊断结果经常开的药。将数据集上待保护的某给定数据记录(集)记为 S,将用户运行数据库查询语言后的结果集记为 S'。很明显,若 $S \subseteq S'$,则说明存在隐私泄露和安全问题。为此,对于用户提出的给定安全需求,需要明确其中的待保护信息,进一步明确如何获得上述待保护信息,对该数据访问行为进行限制,从而避免信息的泄露。

目前数据开放的大量数据均为结构化数据,以二维表的形式进行存储。数据访问行为本质上都是作用在不同二维表连接形成的数据源,再在数据源上进行选择和投影等操作实现不同复杂度的数据获取。如图 2 所示,图中外围矩形 DATA SOURCE 表示用户数据访问请求所涉及的数据源,即表连接后得到的结果集。与此对应的,水平线表示对数据进行条件筛选,列表示对数据进行属性投影,对阴影矩形进行分组、聚集等操作即可得到该数据访问行为结果。开放数据集中的待保护隐私数据同样通过这样的方式获得。由此可见,数据的安全需求对应一个受限的数据访问行为。本文把该数据访问行为定义为一条安全规则。很明显,该数据访问行为不同于数据库查询语句,因此查询语句的形式多样,同一数据访问结果可有不同形式的查询语句获得,然而其最终的查询结果一致,获取查询结果的关键因素一致。

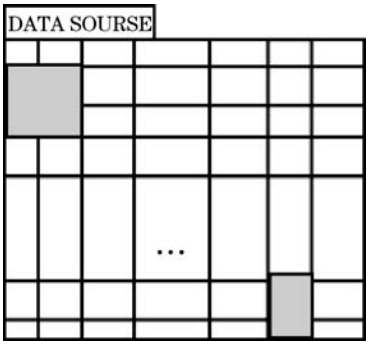

图 2　安全访问设计原理

基于以上的分析,给出本文安全规则(记作 r)的形式化定义:

$$r = \{DS, SL, CD, GB\} \tag{1}$$

式中:DS 表示数据源,由不同数据表的连接产生,SL 为隐私信息涉及的属性集合。条件项 CD 以条件的逻辑组合形式对隐私记录的取值范围进行了规定。考虑到分组操作对结果集的影响(如聚合操作通过不同字段分组得到的结果不同),本文在安全规则中使用分组项 GB 对分组字段进行记录。为了后续描述方便,本文将数据源、属性项、条件项和分组项统称为限定项,每一限定项又可由若干限定子项组成。以条件项为例,这里的选择条件中可能是多个条件的组合,即 $CD = \{c_1\theta_1c_2\theta_2\cdots c_{n-1}\theta_{n-1}c_n\}$。其中,$n$ 为 CD 中的限定子项数,c_i 为第 i 个条件子项,θ_i 为第 i 个条件子项与第 $i+1$ 个条件子项之间的逻辑关系。式(1)表示各条件限定子项通过一定的逻辑关系共同作用得到一个完整的条件限定项。

规则库 R 由所有的安全规则构成。假设规则库中存在 n 条规则,则规则库 R 的形式化定义为:

$$R = \{U_{i=1}^{n} r_i\} \tag{2}$$

表 1 展示了医疗数据集上的安全规则实例。从表中可以发现,规则“诊断一致性”(即“患者接受的超声诊断与穿刺诊断中诊断结果相同的概率”)中聚合操作“COUNT(诊断结果)”被保存在属性限定项 SL 中,表示该聚合操作的统计结果将作为该数据源的一个新生成的属性列,因此加入到 SL 中。与“患者”的“ID”共同构成了该规则的属性限定项。

表 1　安全规则结构

名称(RN)	数据源(DS)	属性项(SL)	条件(CD)	分组(GB)
诊断一致性	超声诊断表,穿刺诊断表,患者表	患者.ID,COUNT(诊断结果)	超声诊断.诊断结果 = 穿刺诊断.诊断结果	患者.ID,诊断结果

3　安全规则的构建

自然语言作为人类的基本技能,若能向数据提供者提供基于自然语言的安全需求接口,必然会大大降低用户的学习成本,提升规则创建的效率和友好度。因此,本文设计了一种面向自然语言的安全规则创建方法,数据拥有者以自然语言的方式提交安全需求描述,系统对其进行语义解析自动转换为安全规则四元组结构并保存。考虑到用户自然语言查询描述的简洁性、模糊性,我们借助本体知识库提高转换的准确度。图 3 展示了本文基于临床医疗数据构建的本体结构,其中包括概念(如医生、患者),属性(如姓名)以及它们之间的关系(如接受、治疗等)。在该图所示的例子中,“诊断”被称为父概念,而与其以“is-a”连接的各种诊断被称作子概念,“医生”和“患者”与“用户”概念之间以“unionOf”连接,分别被称作成员和集合概念。在本体构建过程中,同时建立本体-数据库映射表(OD Mapping),将数据库与本体知识库中元素的映射关系进行保存。

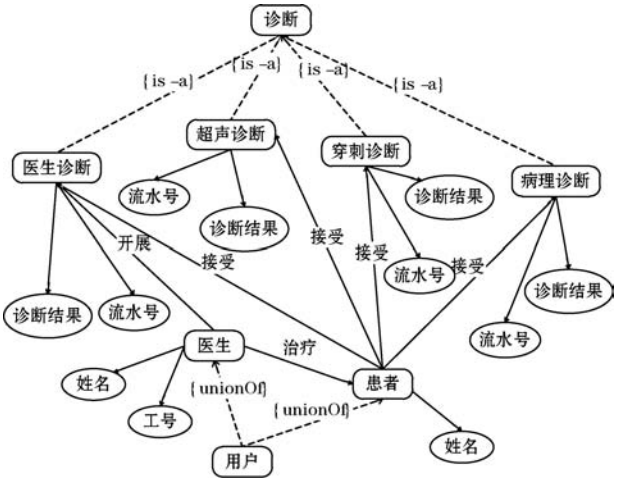

图 3 涉及到诊断的本体结构

基于本体的安全规则创建过程主要包括以下五步：

步骤一 安全需求描述分词。系统接受用户安全需求的自然语言描述后通过 LTP 对其进行分词处理[21]，生成一棵初步的语法树，记为 GT。

步骤二 分词预处理。由于数据提供者上传的数据往往是某特定领域的，因此易存在分词错误的情况。本文将利用本体结构对分词结果进行预处理，纠正错误分词，生成新的语法树，记为 GT^1。

步骤三 节点映射。遍历步骤二生成的语法树 GT^1 的各个节点，在本体知识库中查找可能存在的对应节点，建立语法树节点到本体结构的映射，得到语法解析树。

步骤四 规则生成，建立解析树节点与安全规则的对应关系，生成规则结构，最后通过数据库-本体映射表将该结构转化为面向数据库的最终存储结构并入库。

下面分别对安全需求的预处理，解析树的生成和规则转换过程做详细的介绍。

3.1 解析树预处理

本文采用 LTP 对用户的安全需求描述进行分词，图 4(a)展示了语句“患者接受的医生诊断、超声诊断、穿刺诊断结果相同的概率”经分词后生成了语法树结构，可以发现医生、超声、穿刺与诊断被分成两个词，这显然不符合数据拥有者的描述意图。本文利用本体知识对语法树进行预处理，对其中存在的错误分词进行纠正，图 4(b)为调整后的树结构，可以发现，除关键字节点外，经调整后的树节点与本体结构中的元素基本能够一一对应(包括同义词)。

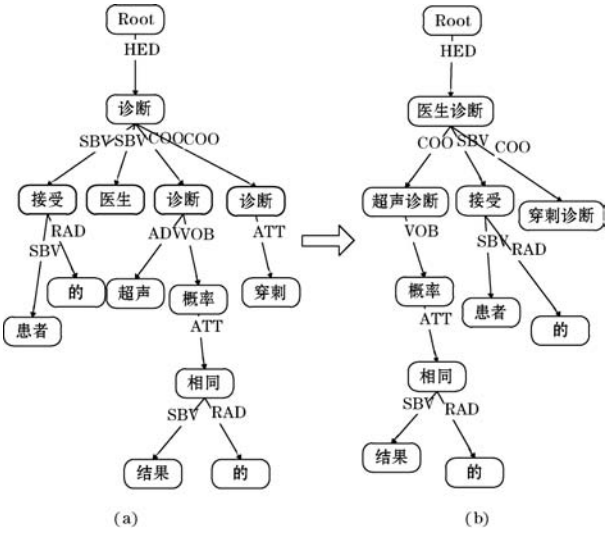

图 4 语法树结构调整

分词预处理包括分词纠正和语法树结构调整。观察 LTP 分词后得到的语法树结构可以发现，错误分词一般出现在相邻两个或多个节点上，于是，本文提出了一种递归的节点组合方法来解决错误分词的问题，即将语法树中相邻两个或多个节点进行组合，找出可能的分词方式。进一步利用语法树中的词性标记确定节点词性，确定分节点是否应该组合成一个新的节点。最后根据判断结果调整语法树结构，生成新的语法树。

分词预处理的详细过程如算法 1 所示。其中算法 3 - 8 行递归生成所有组合节点，若组合节点在本体知识库中存在对应元素时，则利用语法树中的语义标记确定分节点的词性(算法第 14 行)，当分节点不全为名词时，说明该分节点与本体知识库的映射不满足语义要求，直接使用组合节点替换 GT1 中对应分节点即可(代码 15 - 16 行)。

算法 1 安全需求描述预处理

1 输入：语法树 $GT = \{ < vi, ei, vi + 1 > \mid 0 < i < n \}$，本体知识库 O

2 输出：语法树 GT^1

3 let node set CT = {} //保存临时组合节点集合

4 foreach vertex v_i in GT //遍历语法树中各节点

5 let $v_i v_{i+1} = v_i$ concatenate v_{i+1} //相邻节点组合形成组合节点

6 while $v_i v_{i+1} \in O$ //若组合节点在本体知识库中存在对应元素

7 add $v_i v_{i+1}$ to CT //将组合节点加入临时组合节点集合 CT 中

8 let $v_i v_{i+1} = v_i v_{i+1}$ concatenate v_{i+2} //递归组合相邻节点，直到组合节点在本体知识库中没有对应元素

9 end while

10 end for

11 let GT^1 = GT //结果语法树初始化

12 foreach $v_i v_{i+1}$ in CT　　//遍历组合节点
13 if $v_i v_{i+1}$'s separate node v_i and $v_j \in$ elements in O then
　　//判断分节点在本体知识库中是否存在对应元素
14 get the POS of each node according to [20]
　　//利用语义标注确定分节点词性
15 if v_i and v_j are not both Noun. then //若分节点不全为名词
16 replace v_i and v_j in GT^1 with $v_i v_{i+1}$
　　//用 $v_i v_{i+1}$ 替换对应分节点
17 end for
18 return GT^1

3.2　节点映射

语法树的预处理过程解决了错误分词的问题,使得语法树能够更加准确地描述安全需求的语义信息。进一步,将语法树中的各节点映射至本体结构中对应元素。在节点映射过程中,由于用户的描述与本体结构中的元素并非完全对等,可能存在近义或形近词的情况,因此本文设计了一种相似度算法为不同映射进行评分,并仅将评分高于一定阈值的映射加入候选节点中。算法综合应用 JACCARD 相似度和知网相似度计算词元映射的相似性。下面对该相似度算法做形式化描述:

将语法树中的词元记为 t,待映射的本体元素词元记为 e,词形相似度的比较以汉字为单位,即计算两个词中相同汉字所占比例。t 与 e 的 JACCARD 相似度表示为 $Jac_\delta(t,e)$,知网相似度表示为 $sim(t,e)$,于是本文相似度 $rsim(t,e)$ 的计算方法可以表示为:

$$rsim(t,e)=\begin{cases} stm(t,e) & sim(t,e)>\sigma \\ Jac_\delta(t,e) & Jac_\delta(t,e)>\varepsilon \\ \dfrac{(t,o)+Jac_\delta(t,e)}{2} & sim(t,e)<\sigma \text{ and } Jac_\delta(t,e)<\varepsilon \end{cases} \tag{3}$$

式中:σ 为同义词阈值,ε 为形近阈值。由于同义词在日常生活中使用较为频繁,本文在计算相似度时首先对同义词的相似性进行判断,若 $sim(t,e)$ 达到阈值 σ,则认为它们是同义词,令 $rsim(t,e)=sim(t,e)$。否则判断 JACCARD 相似度,若高于阈值 ε,则相似度取 $Jac_\delta(t,e)$。反之,则将知网相似度和 JACCARD 相似度的平均值作为两个词的相似度。再根据整体的相似度阈值 γ 过滤相似度过低的映射。于是得到一个完整的映射集合,记为 $V=\{U_{i=1}^{n} t_i \rightarrow E_i\}$,其中 t_i 为语法树中第 i 节点的词元,E_i 为 t_i 在本体知识库中所有映射元素的集合,映射集合 V 表示语法树中各节点与本体元素的映射关系。不同映射元素对应的本体对象、属性及关系的最小覆盖构成了候选解析图,进一步调用最小生成树算法可去掉不必要的连接关系,以生成最终的最小覆盖树。

得注意的是,在自然语言描述的语法树中,除包含与本体结点对应的结点外,还可能包含诸如“最多”和“张医生”等关键词和数据值类型的节点,此类节点将通过关键字库和转换索引库实现映射,其中关键词库的形式为{keyword:function},表示 keyword 关键词对应的功能操作为 function,如{“最多”: MAX}。转换索引库中保存典型数据值与本体中对应属性的关系,如{“甲状腺肿”:“诊断结果”,“张 *”:“医生姓名,患者姓名”}。同时,某一数据值可能同时存在于不同属性中,在对此类数据值解析时,需要根据语法树中的语义标记选择合适的所属属性。我们将关键字结点映射后的结点称为功能结点,根据关键字描述的功能不同,功能结点分类如表 2 所示。

表 2　结点映射中功能节点及其作用

功能节点	节点作用
比较符	指定该节点所在属性的范围
逻辑符	指定该节点与同一级节点间的逻辑关系(未指定默认为与)
查询	指定该节点所在属性为待查询属性
分组	指定按该节点所在属性分组
概率	分别对分子和分母指定安全规则

3.3　安全规则转换

上一节中生成的最小覆盖树中,各节点与安全规则中各限定项之间存在着一定的对应的关系。根据该对应关系,规则生成器将解析树的各个节点转换为面向本体知识库的形式化结构,最后根据本体-数据库映射表将本体元素转换为数据库中对应字段的表示方式,生成安全规则存储结构。图 5 展示了规则“患者诊断一致性”的解析树与规则结构之间的对应关系。

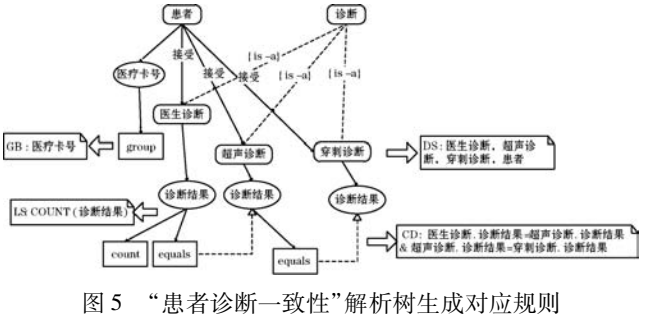

图 5　“患者诊断一致性”解析树生成对应规则

下面对各限定项的转换规则进行详细阐述:

数据源(*DS*):数据源对应解析树中的概念节点,因此只需取出相关概念节点,以“,”连接即可。

属性项(*SL*):属性项中包含概念的属性和聚合函数等,在解析树中通过功能节点指明该属性为属性限定项。转换过程中只需找出所有标记为属性限定项的节点,通过“,”连接即可构成完整属性项。

条件项(*CD*):条件字段与规则结构间的转换关系相对复杂,主要集中在各条件子项之间逻辑关系的确定和不同类型功能节点的识别。如“患者诊断一致

性”中“医生诊断. 诊断结果”的功能节点为“ = ”,该功能节点通过引用指向“超声诊断”的子节点“诊断结果”,据此添加条件限定子项“医生诊断。诊断结果 = 超声诊断. 诊断结果”。

分组项(*GB*):若解析树中包含“分组”功能节点,则将该节点的父节点添加至分组限定项中,以“,”连接,表示该规则同时以多个字段进行分组。

4 查询审核

本节总体介绍安全规则查询审核的过程,并对条件和聚合操作审核的过程进行详细介绍。

4.1 基于安全规则的查询校验

本文安全规则依次从数据源、属性项、条件项和分组字段上对安全规则和数据使用者的查询行为进行限定。系统接受用户的查询请求后,首先对查询语句进行解析,获得各查询子项。然后依次对比查询子项和安全规则集合中对应的限定项,获得各查询子句违反的安全规则集合。若最终获得的安全规则集合非空,说明有访问隐私数据的风险,给出提示返回。下面给出查询审核的完整步骤:

步骤 1 查询解析,解析查询语句的各查询子句,以二叉树结构保存,记为 STree。

步骤 2 数据源审核,从 STree 中取出 FROM 子句,检索规则库中数据源为 FROM 子句子集的规则,形成规则集,记为 *RD*。

步骤 3 属性项审核,取出 STree 中 SELECT 子句,将其记为 *SS*。判断规则集 *RD* 中各规则的 *SL* 限定项是否为 *SS* 的子集,取出满足条件的规则集,将其记为 *RS*。

步骤 4 条件项审核,从 STree 中取得条件子句(WHERE 子句),将其记为 *W*。通过算法 2 获得其与规则集 *RS* 的交集,记作 *RW*。若 *RW* 为空集,则说明查询语句满足安全需求,返回安全标志。反之则继续执行步骤 5。

步骤 5 分组字段审核,在 STree 中取出分组字段,记作 *SG*。判断 *SG* 与 *RW* 中规则的 *GB* 限定项是否一致,是则获取聚合操作,判断字段和函数名是否一致,若非一致,由于其中两个操作能够推导出另一个操作的结果,进一步判断是否存在此类安全隐患,是则返回违反安全约束标志。

4.2 条件审核

由于聚合函数等对条件和分组字段的依赖性强(不同条件或按不同字段进行分组都将会导致结果不同),算法会在选择规则集的过程中首先判断查询字段是否存在聚合函数,若存在,则判断查询语句的 WHERE 子句中各字段的作用区间是否与 RS 中给定规则的条件限定项的作用区间相同。判断之前首先对用户查询中条件子句与安全规则中条件限定项相同的字段分别进行合并,之后对两者对应字段进行区间对比,各区间均相同则认为条件子句违反该规则的条件限定项,并直接将规则添加到 *RW* 中。

算法 2 描述了上述查询语句条件子句的判断过程,其输入是前一步骤得到的可能违反的规则集 *RS* 及查询语句的 WHERE 子句,其输出是经过进一步判断的规则集。若规则集为空,则该查询安全。若规则集不为空,继续下一步骤的判断。

算法 2 条件项审核

1 输入: 规则集 $RS = \{r_1 \cup r_2 \cdots \cup r_m\}$, 条件子句 $W = \{w1\&1w2\&2\cdots\&l-1wl\}$

2 输出: 结果集合 *RW* or NULL

3 let set *RW* = { } //用以保存可能违反的规则集合

4 for each rule r_i in RS

5 get $SL_i = \{ l_1 \cup l_2 \cdots l_{k-1} \cup l_k \}$, $CDi = \{c_1\theta_1 c_2\theta_2\cdots c_{n-1}\theta_1 c_n\}$ from r_i

6 if j in *SLi* is an aggregate function

//若属性限定项中存在聚合函数,则判断条件语句

//与条件限定项区间是否一致

7 if *CDi* = *W* then

//条件限定项和条件子句对应字段作用区间均相同

8 add r_i to *RW* //直接将规则添加到规则集 *RW* 中

9 else

//若不存在聚合函数,则比较条件子句与对应限定条件项

10 for each w_j and $\&_j$ in *W*

//遍历条件子句中各条件和相关逻辑操作符

11 let $RW_{ij} \leftarrow \{(c_1 \wedge w_j)\theta_1(c_2 \wedge w_j)\theta_2\cdots\theta_{n-1}(c_n \wedge w_l)\}$

//取规则 r_i 中 *CDi* 的各限定子项与对应字段的 w_j 的交集,

//并按规则的逻辑关系合并,生成交集集合 RW_{ij}

12 end for

13 let $RW_i \leftarrow \{ RW_{i1}\&_1 RW_{i2}\&_2\cdots\&_{l-1} RW_{il}\}$

//对各条件子句的集合按查询条件中的

//逻辑关系联合得到该规则与条件语句的交集 RW_i

14 end

15 $RW \leftarrow U_{i=1}^{m} RW_i$

//对各规则的交集取并集得到最终结果集

16 return *RW*

若属性项中不包含聚合函数,则将条件子句与条件限定项的对应限定子项进行对比,判断两者作用区间是否存在交集,若存在交集,则判定该条件子句违反规则的条件项约束。判断交集的过程如算法第10 - 13 行所示。首先分别对条件子句和规则的条件限定项中表示同一字段的限定子项按逻辑关系进行合并,然后比较两者相同字段的作用区间是否存在交集,最后通

过一定逻辑关系的组合得到条件子句与规则集 *RS* 的条件限定项的交集 *RW*(算法 11 – 15 行)。

在条件子句中,比较关系包括“ =, < >, >, <, >=, <=, BETWEEN, IN, LIKE”等,所比较的类型可能包括字段、数值、字符等,不同类型的条件求交集的方式有所不同,表 3 展示了条件子句与限定子项存在交集的情况(操作符两端等效)。表中 A、C 为属性,b、d 为具体的属性值或表达式。为更形象地说明该过程,假设规则库中存在规则“甲状腺手术中成功率小于 50% 的医生信息”,该规则中的条件子句包括“手术名称 LIKE ‘甲状腺’”、“手术成功率 < 50%”和“手术. 医生工号 = 医生. 医生工号”等。这时当用户输入查询语句中包含条件子句 “手术名称 = ‘甲状腺结节’”,则判断交集的过程按表中第三行“A = C & b LIKE d”进行,即 A = C = “手术名称”,d = “% 甲状腺%”,b = “甲状腺结节”,于是判断“b LIKE d”为真,最后取“甲状腺结节”作为相交的结果集,因此该交集的返回值为“手术名称 = ‘甲状腺结节’”。同样地,若用户的查询中包含条件“手术成功率 <60%”,通过第五行条件“A = C & b∩d≠∅”判断交集是否为空,并最终得到它们的交集 “A ∈ (– ∞, 50%) ∩ (– ∞, 60%) = (– ∞, 50%)”。结果集不为空说明该查询将受限。此外,表中第 7 行中规则和条件子句均为模糊匹配,下式用于取出它们的交集:

$$SM(b,d) = sub(b) \Delta sub(d) \tag{4}$$

式中:$sub(b)$ 函数用于取出模糊字符串 b 的子串列表,Δ 操作用于将两个子串列表中的子串不重复地组合成新的模糊字符串。值得注意的是,新生成的字符串为以不同顺序组合子串得到的并集,以避免字符顺序对结果集合造成的影响。

表 3　不同类型条件子句与限定项取交集的操作

编号	条件子句(用户查询)	限定子项(安全规则 ri)	存在交集的情况	结果集
1	A = b	C = d	A = C & b = d	A = b
2		< >, <, >, <=, >=, BETWEEN, IN (C ∈ d)	A = C & b ∈ d	
3		C LIKE d	A = C & b LIKE d	
4	< >, <, >, <=, >=, BETWEEN, IN (A ∈ b)	C = d	A = C & d ∈ b	A = d
5		< >, <, >, <=, >=, BETWEEN, IN (C ∈ d)	A = C & b∩d≠∅	A ∈ b∩d
6	A LIKE b	C = D	A = C & d LIKE b	A = d
7		C LIKE d	A = C & b LIKE d & d LIKE b	A = SM(b,d)

4.3　聚合操作审核

步骤 5 对查询中包含聚合函数的情况做了进一步的安全审核,算法 3 展示了审核的详细过程。算法以步骤 6 的输出规则集 *RS*、选择子句 *S* 和分组子句 *QG* 为输入。经过分组限定项和聚合函数的审核得到最终规则集 *RA*。

算法 3　聚合操作审核

1　输入: 规则集 $RW = \{r_1 \cup r_2 \cdots \cup r_m\}$, 分组子句 *QG*, 选择子句 S

2　输出: 规则集合 *RA* 或 NULL

3　let set *RA* = { }　　//用以保存最终结果集合

4　foreach GB_i of r_i in *RW*

　　//取出规则集 RW 中各规则的分组项 GB_i

5　if $GB_i = QG$　　//若分组字段与分组项 GB_i 相同

6　add r_i to *RA*

　//将满足条件的分组字段所属的安全规则添加到 RA 中

7　end for

8　foreach SL_i of r_i in *RA*　//获取规则集 *RA* 中的各查询项 *SL*

9　if SL_i has aggregation function and S has aggregation function　　//若查询项 SL_i 和查询子句 *S* 中均存在聚合操作

10　if agg(SL_i) ∪ agg(*S*) ≠ { COUNT, AVG, SUM } and agg(SL_i) ≠ agg(*S*) //agg( ) 函数为获取聚合操作,这里判断聚　　//合操作是否会导致隐私泄露

11　remove r_i from *RA*

　　//从 *RA* 中移除聚合操作不会导致隐私泄露的规则

12　end for

13　return *RA*

首先,算法判断分组字段是否一致(若存在),将一致的规则添加到结果集中。由于 COUNT、AVG 和 SUM 操作能够通过其中两者得到另一者的结果,因此在前面步骤中未对聚合函数进行处理。算法第 9 – 12 行判断存在聚合函数的情况下,当聚合字段和函数都相同,或字段相同,但由查询的函数能够推导出规则限定的结果时,则可判定查询违反该条规则。算法第 11 行将安全的查询由违反安全规则的集合中移除,最终得到查询所有违反的安全规则集合。

5　实验与结果

本节通过实际医疗数据和安全需求验证本文基于安全规则的审核模型的有效性。

5.1　数据集

医疗数据作为一类典型的敏感数据,存在较多的数据安全和隐私威胁问题,本文实验采用某三甲医院的医疗数据集,其中包含患者基本信息、各类诊断信

息、手术及用药等在内的 16 个数据表。

5.2 安全规则转换效果

本实验验证本文提出的规则转换的有效性。

本文前期对医院的安全需求进行了相关调研，了解到其安全需求可被划分为患者、医生和医院三个维度。其中，患者维度存在的信息泄露主要涉及患者个人信息、患者既往用药情况、患者既定就医规则等。医生维度数据访问安全需求主要包括医生个人基本信息、就诊病人信息、医生 ICD 编码诊断习惯、检查习惯、用药习惯、医生临床科研的实验数据和结果数据泄露等。医院维度信息泄露主要包括独特的、具有优势的院内制剂配方；药品、耗材等使用状况数据；经营财务状况数据；院内诊疗方案、诊疗规范等。

通过对上述安全需求进行分析，并结合本文实际数据集，本文选择了如表 4 所示的 9 条与数据访问相关的安全需求及其对应的自然语言描述。观察表中各安全需求描述语句可以发现，患者和医生维度的安全需求与具体患者或医生关联才会导致隐私的泄露，而由于该数据集的来源已经确定，因此医院维度不用指定具体医院。

表 4 医疗安全需求及对应自然语言描述

安全维度	安全需求	需求自然语言描述
患者	患者个人数据	患者个人基本信息
	既往用药情况	患者经常用药的名称
	既定就医规则	患者接受诊断的时间
医生	就诊病人个人信息	医生治疗病人的个人信息
	医生用药习惯	医生经常开药的药品名称
	医生手术成功率	医生开展手术结果为成功的概率
	医生诊断错误记录	医生开展诊断与超声诊断结果不同的记录
医院	手术失败的记录	手术结果为失败的记录
	药品、耗材等使用状况	药品使用量

与安全需求对应的转换结果如表 5 所示。从转换结果中可以发现，借助基于领域的本体知识库，本文所有的安全需求描述均能解析出正确的安全规则。从解析结果可以发现，诸如“基本信息”在关键字库中保存有对应的解析方式，本文默认为患者的所有个人隐私信息，以“*”表示。语句“患者经常用药的名称”所生成的规则是由选择产生的，查看该语句生成的解析树的权重可以发现，默认的解析结果与所选择的规则的权重相等（在解析树中体现为“名称”与用药表中“科室名称”和“药品名称”的相似度相同）。这是因为当候选解析树权重相同时，系统会随机选择其中一个作为默认解析树，并将转换的安全规则返回给用户。

表 5 医疗数据中安全需求与对应的安全规则

安全需求描述	转换的安全规则			
	数据源	属性项	条件	分组
患者个人基本信息	患者	*		
患者经常用药的名称	患者，用药	患者 ID，用药.药品名称		
患者接受诊断的时间	患者，诊断	患者. ID，诊断. 时间		
医生治疗病人的个人信息	医生，患者	医生. 医生编号，患者.*		
医生经常开药的药品名称	医生，医生诊断	医生. 医生编号，医生诊断. 诊断结果		医生. 医生编号，医生诊断. 诊断结果
医生开展手术结果为成功的概率	医生，手术	医生. 医生编号，COUNT（手术. 结果）	手术. 结果 = ‘成功’	医生. 医生编号，手术. 结果
医生开展诊断与超声诊断结果不同的记录	医生诊断，超声诊断	诊断结果	医生诊断. 诊断结果 < > 超声诊断. 诊断结果	
手术结果为失败的记录	手术	*	结果 = ‘失败’	
药品使用量	药品	药品名称，COUNT（药品名称）		药品名称

本文同时列举除了本文方法无法正常解析的安全需求描述，如表 6 所示。分析转换过程可以发现，医生维度中“医生使用利多卡因针的频率”语句中由于“利多卡因针”在本体中没有对应元素，同时其并未指定所属属性是药品名称，导致在本体知识库中无法找到对应项。而在医院维度中，“医院经营财务状况”的“经营状况”在本文所使用的数据中没有对应数据。最终这两条语句会返回解析出错的提示。

表 6　不能被正确解析的安全需求描述

维度	安全需求描述
医生	医生使用利多卡因针的频率
医院	医院经营财务状况

5.3　安全审核准确性测试

安全审核的准确性验证包括两个方面:一是对于违反安全规则的数据访问需求进行拒绝访问;二是对不违反安全规则的数据访问需求不会返回错误结果。由于用户在数据的访问过程中输入的查询访问需求是可变、多样的,因此对安全审核的准确性验证十分困难。对于验证一,虽然查询 SQL 的形式具有不确定性,但在查询审核过程中解析出用于查询校验的数据源、条件限定项等子项是确定的,因此验证一相对容易。而验证二,需要对所有 SQL 进行判定,不会产生拒绝访问的结果较难实现。为此,本文提出如下的解决方法。在测试安全规则审核的准确性实验中,针对每一条安全规则,本文从上一节生成的安全规则的各限定项出发,制定可能违反或符合查询约束的查询子句,并对各子句进行组合形成多个查询语句,并逐一判断语句是否违反安全规则限定。

以“患者诊断一致性”为例,从数据源出发制定查询子句有“FROM 医生诊断,超声诊断,穿刺诊断,患者”、“FROM 超声诊断,穿刺诊断,患者”和“FROM 医生诊断,超声诊断,穿刺诊断,患者,医生”;而从条件限定项出发指定的条件子句则包括“WHERE 医生诊断. 诊断结果 = 超声诊断. 诊断结果 AND 超声诊断. 诊断结果 = 穿刺诊断. 诊断结果”、“WHERE 医生诊断. 诊断结果 = 超声诊断. 诊断结果 AND 超声诊断. 诊断结果 < > 穿刺诊断. 诊断结果”、“WHERE 医生诊断. 诊断结果 = 超声诊断. 诊断结果”和“WHERE 医生诊断. 诊断结果 = 超声诊断. 诊断结果 AND 超声诊断. 诊断类型 = ‘门诊’”等。再对以上各子句进行组合形成新的查询语句。

表 7 展示了几个典型的违反和符合安全规则例子,语句二和三分别因为分组字段和条件字段与安全规则中限定项不同,因此被认为是安全的查询语句。而语句一由于与规则完全吻合而被拒绝访问。从结果可以发现,查询语句都能够按预期完成审核。从表 7 可以看出,对于相同的属性列,通过基于规则的查询审核,若访问该属性列的最终结果无隐私泄露风险,则允许执行,从而保证了数据的可用性。

表 7　“患者诊断一致性”查询审核测试

查询语句	是否安全
SELECT COUNT(诊断结果) FROM 医生诊断,超声诊断,穿刺诊断,患者 WHERE 医生诊断. 诊断结果 = 超声诊断. 诊断结果 AND 超声诊断. 诊断结果 = 穿刺诊断. 诊断结果 GROUP BY 患者. ID,诊断结果	违反约束
SELECT COUNT(诊断结果) FROM 医生诊断,超声诊断,穿刺诊断,患者 WHERE 医生诊断. 诊断结果 = 超声诊断. 诊断结果 AND 超声诊断. 诊断结果 = 穿刺诊断. 诊断结果 GROUP BY 患者. ID	安全
SELECT COUNT(诊断结果) FROM 医生诊断,超声诊断,穿刺诊断,患者 WHERE 医生诊断. 诊断结果 = 超声诊断. 诊断结果 OR 超声诊断. 诊断结果 < > 穿刺诊断. 诊断结果 GROUP BY 患者. ID,诊断结果	安全

6　结　语

为了解决数据开放共享中的安全问题,本文提出了一种基于安全规则的数据行为审核模型。模型利用本体知识库的语义优势将用户的安全需求描述转换为安全规则存储结构,并根据安全规则自动对用户的数据访问行为进行审核。本文通过实际医疗数据验证了本文安全规则的有效性。

参　考　文　献

[1] 张兰廷. 大数据的社会价值与战略选择[D]. 北京:中共中央党校, 2014.

[2] Bright J, Margetts H Z, Wang N, et al. Explaining Usage Patterns in Open Government Data: The Case of Data. Gov. UK[EB]. Social Science Electronic Publishing, 2015.

[3] Yang T M, Jin L, Jing S. To open or not to open? Determinants of open government data[J]. Journal of Information Science, 2015, 41(5): 596 - 612.

[4] Chaudhuri S. What next?: a half-dozen data management research goals for big data and the cloud[C]//Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems. ACM, 2012:1 - 4.

[5] Green M, Hohenberger S, Waters B, et al. Outsourcing the decryption of ABE ciphertexts[C]. Proceedings of the 20th USENIX conference on Security, 2011: 34 - 34.

[6] Bell D E, Padula L J L. Secure Computer System: Unified Exposition and Multics Interpretation[J]. Secure Computer System Unified Exposition & Multics Interpretation, 1976: 161 - 161.

(下转第 77 页)

[5] 黄曾阳. HNC 理论与自然语言语句的理解[J]. 中国基础科学,1999(S1):85-90.
[6] 黄曾阳. HNC(概念层次网络)理论:计算机理解语言研究的新思路[M]. 北京:清华大学出版社,1998.
[7] 司贝贝,杨进才. 基于依存关系的复句关系词搭配库建设[J]. Software Engineering & sapplications,2015,4(4):81-87.
[8] 李艳翠,孙静,周国栋. 汉语篇章连接词识别与分类[J]. 北京大学学报(自然科学版),2015,51(2):307-314.
[9] 杨进才,郭凯凯,沈显君,等. 基于贝叶斯模型的复句关系词自动识别与规则挖掘[J]. 计算机科学,2015,42(7):291-294.
[10] 刘挺,车万翔,李正华. 语言技术平台[J]. 中文信息学报,2011,25(6):53-62.
[11] Xue N. Chinese word segmentation as character tagging[J]. 中文计算语言学期刊,2003, 8(1):29-47.
[12] 杨宪泽. 21 世纪高校特色教材,人工智能与机器翻译[M]. 成都:西南交通大学出版社,2006.
[13] Zhang H. The optimality of naive bayes. [C]//Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, Usa. 2004.
[14] Schwenker F. Hierarchical support vector machines for multi-class pattern recognition[C]//International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies, 2000. Proceedings. IEEE, 2000:561-565 vol.2.
[15] Osuna E, Freund R, Girosi F. Training support vector machines: an application to face detection[C]//Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 1997:130-136.
[16] 忻栋,杨莹春,吴朝晖. 基于 SVM—HMM 混合模型的说话人确认[J]. 计算机辅助设计与图形学学报,2002, 14(11):1080-1082.
[17] Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning [C]//International Joint Conference on Artificial Intelligence. AAAI Press, 2016:2873-2879.
[18] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[EB]. eprint arXiv:1404.2188v1, 2014.
[19] Kim Y. Convolutional neural networks for sentence classification[EB]. eprint arXiv:1408.5882, 2014.
[20] Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification [J]. eprint arXiv: 1607.01759, 2016:427-431.
[21] Shamma D A, Shamma D A, Friedland G, et al. YFCC100M: the new data in multimedia research[J]. Communications of the Acm, 2016, 59(2):64-73.
[22] Luhn H P. The automatic creation of literature abstracts [M]. IBM Journal of Research and Development, 1958, 2(2):159-165.

(上接第 53 页)

[7] Youman C E, Sandhu R S, Feinstein H L, et al. Role based access control models[J]. Information Security Technical Report, 1996, 6(2):21-29.
[8] National Institute of Standards and Technology. Federal Information Processing Standard(FIPS). US, Department of Commerce, Advanced Encryption Standard[OL]. 2001. http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf.
[9] 闫乐乐,李辉. 基于复合混沌序列的动态密钥 AES 加密算法[J]. 计算机科学,2017,44(6):133-138,160.
[10] 张险峰, 秦志光, 刘锦德. 椭圆曲线加密体制的性能分析[J]. 电子科技大学学报, 2001, 30(2):144-147.
[11] Gentry C. Fully homomorphic encryption using ideal lattices [C]//Proceedings of the Annual ACM Symposium on Theory of Computing, 2009:169-178.
[12] 韩天悦,谢静. RSA 加密解密算法及相关攻击方法[J]. 电脑与信息技术,2018,26(1):53-55.
[13] 缪昌照,徐俊武. AES 与 ECC 混合加密算法研究[J]. 软件导刊,2016,15(11):63-64.
[14] Sweeney L. k-Anonymity: A Model for Protecting Privacy [J]. International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems, 2002, 10(5):557-570.
[15] Dwork C. Differential Privacy: A Survey of Results[C]//International Conference on Theory and Applications of MODELS of Computation. Springer-Verlag, 2008:1-19.
[16] 熊平, 朱天清, 王晓峰. 差分隐私保护及其应用[J]. 计算机学报, 2014, 37(1):101-122.
[17] Machanavajjhala A, He X, Hay M. Differential Privacy in the Wild: A Tutorial on Current Practices & Open Challenges[J]. Proceedings of the Vldb Endowment, 2016, 9(13):1611-1614.
[18] Zhang J, Xiao X, Xie X. PrivTree: A differentially private algorithm for hierarchical decompositions[C]//Proceedings of the 2016 International Conference on Management of Data. ACM,2016:155-170.
[19] 朱作玉. 支持隐私保护的极限学习机研究[D]. 沈阳:东北大学, 2014.
[20] 张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报, 2014(4):927-949.
[21] Zhang M, Deng Z, Che W, et al. Combining Statistical Model and Dictionary for Domain Adaption of Chinese Word Segmentation[J]. Journal of Chinese Information Processing, 2012, 26(2):8-12.