基于 P2P 的半分布式网络的隐私保护 *

熊 博,周 娅

(桂林电子科技大学 计算机与控制学院,广西 桂林 541004)

摘 要:分析了现有的隐私保护技术和对等网拓扑结构,将随机化方法引入到 P2P 网络中,设计了一个基于 P2P 的隐私保护模型,介绍了普通节点及超级节点数据处理方法的运用,验证了基于 P2P 的学生成绩管理数据库模拟的有效性。

关键词:隐私保护;P2P;普通节点;超级节点;随机化

中图分类号:TP391 **文献标识码**:A

Privacy preserving based on semi-distributed P2P network

XIONG Bo, ZHOU Ya

(Computer & Controll Collenge, GuiLin University of Electronical Technology, Guilin 541004, China)

Abstract: Analyzing both the existing privacy preserving and the characteristics of the P2P topology, the randomized process method is introduced to P2P network and a P2P-based privacy preserving model is designed in this paper. The method of processing privacy data is used by super-peer and common-peer. Simulation experiments which are established at a P2P-based student performance management database show that the schemes are effective.

Key words: privacy preserving; P2P; common-peer; super-peer; randomization

对等网 P2P(Peer to Peer)技术是目前国际计算机网络技术领域研究的一个热点,被《财富》杂志誉为将改变 Internet 未来的 4 大新技术之一,包括微软、Sun、IBM 等很多著名的企业和公司都投入到对其技术的研究之中。P2P 是指网络中没有专用的服务器,每一台计算机的地位都是平等的网络[1]。隐私保护技术是针对数据挖掘领域存在隐私泄露的问题而提出的。在数据挖掘领域,隐私被划分为两类:一类隐私是原始数据本身具有的;另一类隐私是原始数据所隐含的知识,如某公司优质客户的行为特征等规则。目前,对于数据集中分布的隐私保护技术主要以 C/S 模式和用户进行交互为主。本文通过引入隐私保护技术到 P2P 网络模型中,研究如何构建基于 P2P 网络的隐私保护模型。

1 P2P 网络中的隐私保护技术

1.1 P2P 网络的数据挖掘

P2P 网络的隐私保护和数据挖掘是一个相对较新的、相关文献较少的领域。研究人员已经发现了一些不同的计算 P2P 网络基本操作的方法(例如求平均值、总和、最大值、随机样品)。

1.2 隐私保护数据挖掘

隐私保护数据挖掘可以分成两组:数据隐藏和规则隐藏。数据隐藏的主要难点在于转移数据或者设计新的计算协议,使得私有数据在数据挖掘操作后仍然保持私有性,而基础数据的模式或模型仍然可以发现。附加扰乱、乘法扰乱[2]、安全多方计算都属于这一类;另一方面,规则隐藏试图转移这种数据库的敏感规则是伪装的,以及所有其他的基本模式仍可以被发现。

2 基于 P2P 的隐私保护模型构造

2.1 基于 P2P 的隐私保护模型

半分布式拓扑结构 P2P 网络[3]有利于网络资源的快速检索,不会由于某个节点失败导致整个系统瘫痪,具有更强的容错性。半分布式拓扑结构 P2P 网络是由提供查询服务的超级节点和其他的普通客户节点组成,并且在资源共享方面,所有节点的地位相同。在提出隐私保护的模型之前,先给出本文所采用的 P2P 网络模型。

* 基金项目:广西青年科学基金(桂科青 0832101)

基于P2P的网络模型如图1所示。

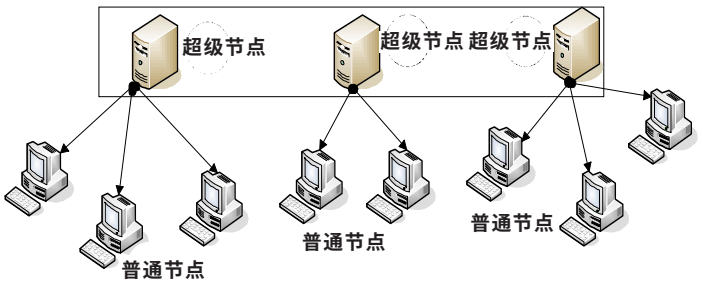

图1 基于P2P的网络模型

此模型中涉及到的对象：

(1)超级节点SP(Super-Peer)：超级节点的功能要大大强于普通节点。在P2P网络中，SP可以为其他的对等点提供一个网络位置，并可以定位其他节点和资源。所有普通节点都向它发送请求，超级节点提供所知道的普通节点的信息。

(2)普通节点CP(Common-Peer)：存储和提供检索资源。

(3)聚簇：1个超级节点与受其管辖的普通节点组成1个聚簇。聚簇的大小就是在该聚簇中节点的数量，包括超级节点本身。

(4)父节点：SP是它本身所在聚簇的所有CP的父节点。

(5)子节点：在同一个聚簇中，所有的CP都是该簇SP的子节点。

2.2 基于P2P网络隐私保护模型

在此模型中，选取P2P网络任意一个聚簇为样板，来构造基于P2P网络的隐私保护模型[4]，如图2所示。

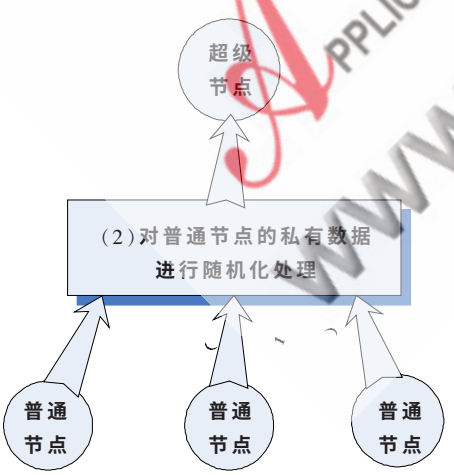

图2 基于P2P的隐私保护模型

隐私保护的过程主要分成3个阶段：

(1)普通节点提交数据；

(2)随机化处理数据，对数据进行伪装；

(3)超级节点对数据进行重构和建模。

3 关键问题分析

很多公司都希望建立顾客的信息集合模型，比如，一家名牌服装店想知道最有可能买D&G的顾客年龄和收入；一个广告公司需要知道客人的个人爱好，以便更好地、有目标地做广告；一家网上零售公司希望了解顾客的个人喜好，以便更好地对网页进行排版以适应顾客的口味。以上这些情况均包括1台服务器(公司)和多个客户端(顾客)。服务器需要建立1个数据聚集模型来应用关联规则挖掘算法或分类算法。通常，结果模型不再包含个人的可识别信息，只包括建立在大量数据之上的平均值。一般情况下，在建立数据聚集模型之前，客户已经将私有信息发布到服务器上了，但越来越多的人们开始关注自己的隐私保护，不想把和此次交易不相关的数据上传。但是，公司仍需要数据聚集模型。可能的方案是，在用户提交数据之前，每个客户端将数据打乱，一些真实数据被取走，而用一些假的随机产生的数据取代它，这种方法叫做随机化方法(在本文的隐私保护模型中，普通节点为客户端，超级节点为服务器)。

针对隐私保护模型通过以下3个步骤进行分析：

3.1 普通节点提交数据

在此假设有N个普通节点，分别用$c_i(i=1,2,\cdots,N)$表示，每个普通节点均有要发布的属性$x_i,i=1,2,\cdots,N$，设其中每个x_i都是一个随机变量x_i的实例，x_i是独立的、并且可以被无差别分发的。累积的分布函数(对每个Xi都一样)定义为F_x，超级节点需要知道函数F_x或者它的近似模拟，这些就是允许超级节点事先知道的聚集模型。这样，超级节点就可以通过模型获取普通节点的信息，但同时要限制超级节点知道真正的x_i。

3.2 数据的随机化处理

每个普通节点均加一个随机产生的偏移量y_i到x_i上，偏移量y_i是累积分布函数F_y无差别产生的独立的随机变量。F_y是事先选定的，并且SP通过普通节点C_i将随机产生的值$z_i=x_i+y_i$发送到SP，SP的任务就是用预知的F_y和$z_1,\cdots,z_n$的值来模拟F_x[2]。

F_y的选择应该尽量遵循2个原则：(1)SP能够合理地模拟F_x；(2)z_i的值尽可能无法揭示x_i的值，这种揭示的量度由置信区间决定，给定一个区间$[z-w_1,z+w_2]$，对所有的未随机化的量x：

$$p[z-w_1\leqslant x\leqslant z+w_2|z=x+y,y\sim F_y]\geqslant c\%$$

一般针对置信区间，用它在c%置信度级别上的最短宽度$w=w_1+w_2$来衡量隐私的数量。

3.3 超级节点对数据进行重构和建模

一旦分发函数F_y已定义并且数据经过随机化处理之后，SP将面对重构问题。给定F_y及$z_1,\cdots,z_n,z_i=x_i+y_i$，求$F_x$。用基于贝叶斯定律的循环算法来处理，定义$x_i(F_x$

的引申值)的分布密度为 f_x,算法描述如下:

(1) f_x^0=统一分布;

(2) $j:=0$;//计数器

(3) loop

$$①f_x^{j+}.(a):=\frac{1}{N}\sum_{i=1}^{N}\frac{f_Y(z_i-a)f_x^j(a)}{\int_{-\infty}^{+\infty}f_Y(z_i-z)f_x^j(x)dz}$$

②$j:=j+1$;

until(不满足规则)

为了让算法更实用,可以把属性域分成 k 个区间,分别为 $l_1,\cdots,I_k$,用分段的常数函数来代替密度函数 f_x^j,其中 $m(I_t)$是 I_t 的中点,则上面的公式变为:

$$f_x^{j+1}(I_p):=\frac{1}{N}\sum_{i=1}^{N}\frac{f_Y(m(z_i)-m(I_p))f_x^j(I_p)}{\sum_{i=1}^{K}f_Y(m(z_i)-m(I_t))f_x^j(I_t)|I_t|}$$

4 实例说明

基于 P2P 的学生成绩管理数据库应用随机化处理数据,进行重构后的结果分布对照如图 3 所示。重构方法较好地模拟了原始的成绩分数函数,基本上达到了对于隐私数据保护的目的。

尽管现在提出了一些隐私保护技术,但是大多数都是存在于数据集中式分布的数据库中,即仍然局限于 C/S 模式,这种模式有着固定的网络瓶颈、扩展性较差等缺陷。对此,本文提出了基于随机化处理数据的隐私保护构想,设计了一个基于 P2P 的隐私保护模型,介绍了构建该模型所采用的数学方法,并结合实际的数据库实例进行模拟,证明了所采用的隐私保护策略是有效的。

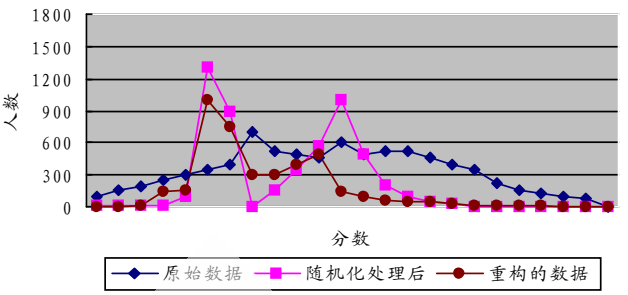

图 3 学生成绩管理数据库随机化处理后的重构结果

参考文献

[1] ABERER K, DESPOTOVIC Z. Managing trust in a peer-to-peer information system [C]. In proc. of 10th Int'l. Conf. on information and knowledge management (CIKM), 2001: 310-317.

[2] LIU K, KARGUPTA H. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining [J]. IEEE Transaction on Knowledge and Data Engineering(TKDE), 18(1): 92-106, January 2006.

[3] 罗杰文.Peer to peer (P2P)综述[A/OL].[2005-11-3]. http://www.intsci.ac.cn/users/luojw/papers/p2p.htm.

[4] LINDELL Y, PINKAS B. Privacy preserving data mining[A].CRYPTO[C].2000.

(收稿日期:2009-07-01)