

基于数据处理的数据挖掘隐私保护技术分析

李玲娟, 郑少飞

(南京邮电大学 计算机学院, 江苏 南京 210003)

摘 要:随着数据挖掘技术的发展与应用,如何在得到准确的挖掘结果的同时保护隐私信息不被泄露,已经成为必须解决的问题。基于数据处理的数据挖掘隐私保护是一种有效的途径,通过采用不同的数据处理技术,出现了基于数据匿名、数据变换、数据加密、数据清洗、数据阻塞等技术的隐私保护算法。文中对基于数据处理的数据挖掘隐私保护技术进行了总结,对各类算法的基本原理、特点进行了探讨。在对已有技术和算法深入对比分析的基础上,给出了数据挖掘隐私保护算法的评价标准。

关键词:数据挖掘;隐私保护;数据处理

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2011)03-0094-04

Analysis of Data Mining Privacy Preserving Technology Based on Data Processing

LI Ling-juan, ZHENG Shao-fei

(College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: As the development and application of data mining, it is a problem which must be resolved that how to protect privacy from leaking when obtaining accurate result. Data mining privacy preserving based on data processing should be an effective way to resolve the problem. Based on different data processing technologies, various privacy preserving algorithms, such as data anonymity, data distortion, data encryption, data purification and data obstructing, have been developed. In this paper, the technologies of data mining privacy preserving based on data processing are surveyed; the mechanisms and characteristics of various algorithms are discussed. Following a comprehensive comparison and analysis of the existing technologies as well as the algorithms, the criteria of evaluating data mining privacy preserving algorithms are given.

Key words: data mining; privacy preserving; data processing

0 引言

数据挖掘能从大量的数据中挖掘出隐含的、未知的、用户可能感兴趣的和对决策有潜在价值的知识和规则^[1]。这些规则蕴含了数据库中一组对象之间的特定关系,揭示出一些有用的信息,可以为经营决策、市场策划和金融预测等方面提供依据。但是,数据挖掘在产生价值的同时也产生了隐私泄露的问题。

数据挖掘领域中的隐私包括两类。一类是被挖掘的原始数据中所含有的敏感数据,如信用卡帐号、电话号码、病情等;另一类是这些数据中蕴含的敏感知识,如优质客户的行为习惯等可能影响公司本身利益的知

识。

随着数据量的不断增多及数据挖掘应用领域的不断扩展,隐私信息泄露的现象在诸如医学、军事及经济等数据挖掘的应用领域时有发生。因此,如何在用于数据挖掘的数据集被发布或共享之前,对数据集中含有的隐私信息进行处理,防止敏感信息被其他方获取,就成为数据挖掘和信息安全领域的一个很有意义的研究课题。

1 相关技术与算法

数据挖掘中的隐私保护主要关注两个方面^[2]:一方面是原始数据集中敏感数据的处理,另一方面是对敏感知识的保护。由于数据挖掘的过程可以划分为数据预处理、应用数据挖掘算法、知识表示三个阶段,因此,可以在数据预处理阶段对原始数据集进行处理以防止隐私在数据挖掘后续阶段的泄露,达到隐私保护的

收稿日期:2010-08-29;修回日期:2010-11-28

基金项目:国家重点基础研究发展计划(973计划)资助项目(2011CB302903);江苏省高校自然科学基金基础研究项目(08KJB620002);南京邮电大学校科研基金(NY207051)

作者简介:李玲娟(1963-),女,辽宁辽阳人,教授,CCF会员,研究方向为数据挖掘、分布式计算等。

主要方法包括数据清洗、数据阻塞、数据变换、数据匿名等,统称为基于数据处理的数据挖掘隐私保护技术。

1.1 基于数据清洗的隐私保护技术与算法

Oliveira 等人基于数据清洗思想,提出了一系列隐私保护关联规则挖掘算法。其中的 SWA^[3] 算法通过删除部分敏感数据来实现隐私保护的。具体来讲该算法解决的问题可以描述为: D 为一个交易数据集,给定最小支持度阈值 σ , D 上所有频繁模式的集合为 F 。设 F_s 和 F_r 为两个不相交的子集。其中 F_s 是由数据拥有者所确定的 F 中敏感模式的集合,称为敏感模式集; F_r 是数据拥有者希望在 D' 中仍然出现的频繁模式的集合,称为发布模式集。数据集 D' 是对 D 做数据变化后的结果,要求在 D' 中,任意的敏感模式 $q \in F_s$ 不被泄漏,同时尽可能地减少对任意发布模式 $p \in F_r$ 的支持度的影响。

数据清洗方法在该算法中具体表现为,通过从原始数据集 D 中的某些交易记录中删除一些项来得到结果数据集 D' 。为了防止一个敏感模式 q 的泄漏,该方法只需要考虑那些包含 q 的交易记录(敏感交易记录)。然后,从这些交易记录中删除若干包含在模式 q 中的项(敏感项)。这样,就可以减少 q 在 D' 中的支持度,达到在 D' 中防止泄漏 q 的目的。数据清洗方法虽然会使 D 中的一些非敏感模式在 D' 中的支持度有所降低,但是它不会在 D' 中带来虚假频繁模式(即原来在 D 上不是频繁的,甚至没有出现过的模式)。

Oliveira 等人还提出了类似的数据处理方法,文献[4]中介绍了一种通过数据清理来隐藏敏感模式的方法。该方法设计了一个基于倒排文件索引和布尔查询的检索引擎来实现对数据集的清理,以此减少支持敏感规则的那些敏感数据项,降低支持度,实现对敏感模式的隐藏。

1.2 基于数据变换的隐私保护技术与算法

数据变换是指从整体上对数据集进行几何变换,从而对数据集中的敏感信息进行隐藏的技术,可用于关联规则、分类、聚类等挖掘的隐私保护。

(1) 基于数据变换的关联规则隐私保护。

文献[5]介绍了一种利用随机正交变换法对原始数据集进行处理以保护隐私信息的方法。其基本思想是数据提供方随机从矩阵库中选择符合要求的正交矩阵对原始数据集中任意配对的两两独立属性向量组进行变换,然后输出变换后的数据提供给使用方。在使用方得到的信息中,虽然原始的细节信息被数据提供方做了变换干扰和扭曲,但由于关联规则挖掘是基于数据集合的聚集信息值而不是一个详细的数据项,因此,随机正交法可很好地用于隐私保护的关联规则挖

掘中。

该方法有三个关键步骤:一是建立合适的矩阵库,这部分工作可以在变换之前进行,一般情况下建立符合最低隐私保护度要求的任意维的正交矩阵库即可,通常只需建立二维和三维的矩阵库。二是选择属性向量对,变换的属性向量是两两独立的,为了计算简单,一般把矩阵库中每两个属性向量分成一个属性向量对。在属性向量的个数是奇数的情况下,最后的三个属性向量将组成一对。三是变换属性向量对,从矩阵库中随机地选择正交矩阵进行变换。

(2) 基于数据变换的分类隐私保护。

数据挖掘中的分类是把待分类的数据集中的数据映射到预先给定的类中的过程;通常针对训练数据集运用一定的分类算法得到分类规则,再基于分类规则对被分类的数据进行类别划分。文献[6]中提出了一种用于分类的隐私保护数据挖掘算法。该算法先在原始数据上加随机偏移量使之被变换,再利用贝叶斯公式推导出原始数据的密度函数,重建判定树。该算法也被归为随机扰动方法。这里算法的基本思想是通过数据集进行整体值变换来实现对数据集中敏感数据的隐藏。

举例来讲,数据集中的一组原始数据值 (x_1, x_2, \dots, x_n) 可以统一加上或是减去一组具有均匀分布或是高斯分布的数据 (y_1, y_2, \dots, y_n) ,从而得到处理后的结果数据集。而后对结果数据集进行分类,由于已知结果数据集同原始数据集的关系,因此可以通过结果数据集的分类结果推算出原始数据集的分类结果。

(3) 基于数据变换的聚类隐私保护。

文献[7]介绍了一种用于聚类的隐私保护算法,该算法通过对数据进行整体旋转变换(RBT)来达到对敏感数据的隐藏。

算法的具体步骤为:首先对数据进行统一的规范化处理,如某属性列上的数据可以与该属性上的最大值、最小值或是与该属性列上取值的均值与方差进行运算,得出变换后的新值;在对数据进行规范化处理的过程中,可以按照规定好的准则去除一些不影响聚类结果的敏感数据(数据记录或是属性列)。然后,根据规定的阈值计算变换矩阵,计算准则为:随机选出的两组列向量与变换矩阵相乘后产生结果列向量,结果列向量与原始列向量相减产生的差向量的方差应该大于等于规定的阈值。再通过该准则选出合适的变换矩阵。所有属性列与该变换矩阵运算后得到的数据集即为最终的结果数据集。

由于该算法处理后的结果数据集并未改变原始数据集中数据间的相似性,所以在对结果数据集进行聚类时仍然能保持原始数据集的聚类效果。

1.3 基于数据阻塞的隐私保护技术与算法

数据阻塞法通过向原始数据集引入不确定的值来隐藏数据库中的敏感规则^[8],比如把敏感事务中的“1”变成“?”。

文献[9]提出了一种针对关联规则挖掘的隐私保护的数据阻塞法。与数据清洗方法不同,该方法通过将原始数据集中的部分数据修改为阻塞量来使敏感知识的支持度或置信度降低至一定区间,从而达到隐藏原始数据集中敏感知识的目的。

该算法实现的具体步骤为:首先把原始数据集表示为0-1矩阵(若原始数据集记录较多,可以分多次处理),然后对支持敏感知识的数据记录中项数最少的记录进行修改,直至该敏感知识的支持度或置信度降至阈值以下。循环执行上述步骤直至对所有的敏感知识处理完毕。算法执行完后含有阻塞量的0-1矩阵可以看作是算法处理后的结果数据集。

1.4 基于数据加密的隐私保护技术与算法

基于密码学的隐私保持技术通常针对分布式数据对象^[10]。文献[11]针对医学数据的挖掘,提出了一种对隐私信息进行保护的数据集加密算法,具有一定的代表性。医学数据大致可以分为数字型数据、字符型数据、时间型数据、图像型数据四类,因此可以定义一组可逆的转换规则 $F = \{f_1, f_2, f_3, f_4\}$ 来对不同类型的数据进行加密,其中 f_1 表示数字型数据的转换函数; f_2 表示字符型数据的转换函数; f_3 表示时间型数据的转换函数; f_4 表示图像型数据的转换函数。利用这组转换规则按照旧表结构建立对应的新表,然后按照转换规则把旧表的内容转换为新表的内容,从而产生新的数据集。医学科研机构只将新数据集送给数据分析专业人员,当医学科研机构得到分析结果后只需按 F 的逆变换将结果转换回原始状态即可。

1.5 基于数据匿名的隐私保护技术与算法

数据匿名即隐藏数据或数据来源,对于集中式数据集来说,数据匿名主要用来对数据集本身进行处理,对于分布式数据集来说,数据匿名主要用来对数据来源进行隐藏。数据匿名一般采用两种基本操作:抑制和泛化。抑制指对某数据项、某记录或某属性组进行隐藏,亦即不发布上述信息;泛化是对数据进行更概括、更抽象的描述。譬如,把一个整数模糊描述为一个整数区间。

为了更好地度量匿名化效果,出现了若干匿名化原则,如 k -匿名、 l -diversity、 t -Closeness。 k -匿名原则要求同一个类的 k 条记录中的任何一条记录都不能区分于其他记录,其中不能区分指满足 k -匿名原则的数据中敏感属性可以区分同一类的不同记录而非敏感属性不能区分; k 值越大,则同一类中含有的记录数越

多,隐私被隐藏的效果越好,但在数据处理过程中丢失的信息越多,所以针对不同数据集选择恰当的 k 值尤为重要。文献[12]中提出了一种 k -匿名算法 Incognito,它是一种广为应用的匿名化算法,它首先建立包含所有泛化操作的图(泛化图),在选取最优泛化操作前,预先对图进行修剪以减少不必要的搜索,从而提升搜索最优泛化操作的效率,然后应用泛化操作处理数据直至数据满足匿名原则。 l -diversity 则在 k -匿名原则基础上保证同一类中敏感属性至少有 l 个不同的取值,使得攻击者得到敏感信息的概率降至 $1/l$ 以下。与 l -diversity 相比, t -closeness 又进一步考虑了同一类中不同记录敏感数据的分布问题,它要求同一类中敏感数据的分布尽量同该属性列的取值分布保持一致。

从总体上来看,匿名化算法大多根据通用匿名原则除去数据集中的敏感信息,并不针对特定的应用。这些算法的步骤一般包括:列出各种不同的泛化操作,并根据设定好的准则选取出最优的泛化操作;然后利用最优泛化操作对数据进行处理,进而产生处理后的结果数据集。近年来,出现了针对特定应用的匿名化算法,如面向聚类的 r -gather 和 r -cellular 算法^[13]。这两种算法根据 k -匿名原则,要求每个聚类中至少包含 k 个数据点且聚类的最大半径尽量小,以达到隐藏数据中敏感信息的目的。

2 各种隐私保护技术与算法的分析比较

上述的基于数据清洗技术的各种隐私保护算法,操作简单,只需根据定义好的规则选择敏感事务和项加以删除即可,一般可实现所有敏感规则的隐藏。但是这些算法大都针对基于交易数据集的关联规则挖掘,且算法扫描数据的次数多依赖于敏感规则的数量,而且这类算法在保护隐私的同时,会付出挖掘结果精度损失的代价。

基于数据变换技术的各种隐私保护算法中,用于关联规则挖掘的“随机正交变换法”,由于对数据变换后得到的数据的记录结构与原始数据库中的记录结构是相同的,因此,使用该方法进行隐私保护的关联规则挖掘时,不仅可以使使用 Apriori 算法,也可以使用 FP-Growth 算法,而且还可以使用基于图的关联规则挖掘算法,即随机正交变换法对不同的关联规则挖掘算法具有较好的适用性。应用正交变换方法可以处理集中式和分布式(垂直分布式和水平分布式)数据集的数据,因此,该方法对不同分布形式的数据集也具有较好的适用性。

由于在用于分类的“随机扰动方法”中,基于扰动数据重构出来的数据的分布相对于原始数据的分布几乎不变,因此利用基于重构数据的分布进行分类器训

练而得到的决策树能对数据进行准确的分类。

由于用于聚类的 RBT 算法是基于旋转变换的,属于等距变换,因此基于距离的聚类算法在原始数据集上的挖掘结果将等同于在变换后的数据集上的挖掘结果。

基于数据阻塞技术的隐私保护方法将原来规则的支持度和置信度从一个确定值转换为不确定的支持度区间和置信度区间,对于有些数据集并不适用,如病人诊断记录等,如采用此方法将会产生不可预料的结果,可能导致医生的诊断建立在错误的数据挖掘结果上。

采用增加噪音法或数据清洗法,会降低结果的精确性,因此这两种方法对于医学方面的数据集并不适用,因为医学对数据的安全性和结果的准确性要求都非常高,而采用数据加密方法恰好可以避免这个问题的产生。基于数据加密技术的隐私保护方法使转换后的数据集跟转换前的数据集结构相同,语义也相同,因此能保证数据挖掘结果的准确性。但其算法效率依赖于数据集的大小以及可逆转换规则的转换效率,因此如何在加密前对数据集进行清洗以及如何定义高效的不可逆转换规则是此类算法需要进一步研究的内容。

采用数据匿名技术的数据处理算法大多可以对面向不同应用的多种类型的数据集进行处理,具有良好的通用性。但是,面向聚类的匿名化算法仍面临着如何确定原始数据集中各个属性的权值,如何对不同含义的属性进行统一度量等问题。

3 数据挖掘隐私保护算法的评价标准

通过对上述典型算法的总结与分析可以看出,对数据挖掘隐私保护算法做出恰当的评价是非常重要的。文献[2,8,10,14]分别提出了一些评价因素。笔者认为,对基于数据处理的数据挖掘隐私保护算法可以从以下几个方面进行较为全面的评价:

有效性:有效的算法应该能够最大限度地防止非法获取隐私信息,对隐私信息进行保护;同时能够准确地对数据进行处理,在对敏感信息进行保护的同时,不影响非敏感数据的使用以及非敏感知识的产生。

复杂性:时间复杂度和空间复杂度是衡量算法效率的重要标准。在分布式环境下,通讯复杂性也是一个必须考虑的主要因素。复杂性尽可能低应该是算法设计的重要目标之一。

通用性:一个通用性好的算法应该能够适用于不同类型的数据集,即能对多种类型的数据集进行处理,清除其中的敏感信息。

扩展性:一个扩展性好的算法不会因为数据量的增大而在效率上呈现出明显和快速的变化。不难看出扩展性与复杂性是相关的。

4 结束语

从数据处理角度对目前比较典型的数据挖掘隐私保护算法进行了总结和分析,将之归纳为基于数据匿名、数据变换、数据加密、数据清洗、数据阻塞技术的几类数据挖掘隐私保护算法。其中对数据集的数据清洗可以看作是从原始数据集中删除某些信息;数据变换是对原始数据集进行整体的几何变换;数据阻塞、数据加密是对原始数据集中的数据项进行修改;数据匿名可以看作是对原始数据集进行删除与修改等综合操作。此外,还提出了综合评价数据挖掘隐私保护算法的基本准则。

目前已有的算法大都针对特定的数据集,各种面向特定应用的数据挖掘隐私保护算法的研究仍将是热点课题;同时如何综合各种数据处理技术,提出具有一定的通用性的数据挖掘隐私保护算法,也将是值得研究的课题。

参考文献:

- [1] 陈安,陈宁,周龙骧. 数据挖掘技术及应用[M]. 北京: 科学出版社,2006.
- [2] 郭宇红,童云海,唐世渭,等. 数据库中的知识隐藏[J]. 软件学报,2007,18(11): 2782-2799.
- [3] Oliveira S R M, Zaiane O R. Protecting sensitive knowledge by data sanitization [C]//Proc. of the 3rd IEEE International Conference on Data Mining (ICDM'03). Melbourne: [s. n.],2003:613-616.
- [4] Oliveira S R M, Zaiane O R. Privacy preserving frequent itemset mining [C]// Proc. of the IEEE international conference on Privacy, security and data mining. Maebashi: Australian Computer Society,2002:43-54.
- [5] 许焕霞,邵良杉,褚丽丽. 随机正交变换法在隐私保持关联规则挖掘中的应用[J]. 科技和产业,2010,1(10): 75-79.
- [6] Agrwal R,Srikant R. Privacy- Preserving Data Mining [C]// In Proc. of the ACM SIGMOD Conference on Management of Data. Dallas,Texas: [s. n.],2000:439-450.
- [7] Oliveira S R M,Zaiane O R. Achieving Privacy Preservation When Sharing Data For Clustering [C]//In Proc. of the Workshop on Secure Data Management in a Connected World (SDM'04) in conjunction with VLDB'2004. Toronto,Ontario, Canada: [s. n.],2004:67-82.
- [8] 李学明,刘志军,秦东霞. 隐私保护数据挖掘[J]. 计算机应用研究,2008,25(12):3550-3555.
- [9] Saygin Y,Verykios V S,Elmagarmid A. Privacy Preserving Association Rule Mining [C]// In Proc. of 12th RIDE. [s. l.]: [s. n.],2002:151-158.
- [10] 吕品,陈年生,董武世. 面向隐私保护的数据挖掘技术研究[J]. 计算机技术与发展,2006,16(7):147-149.

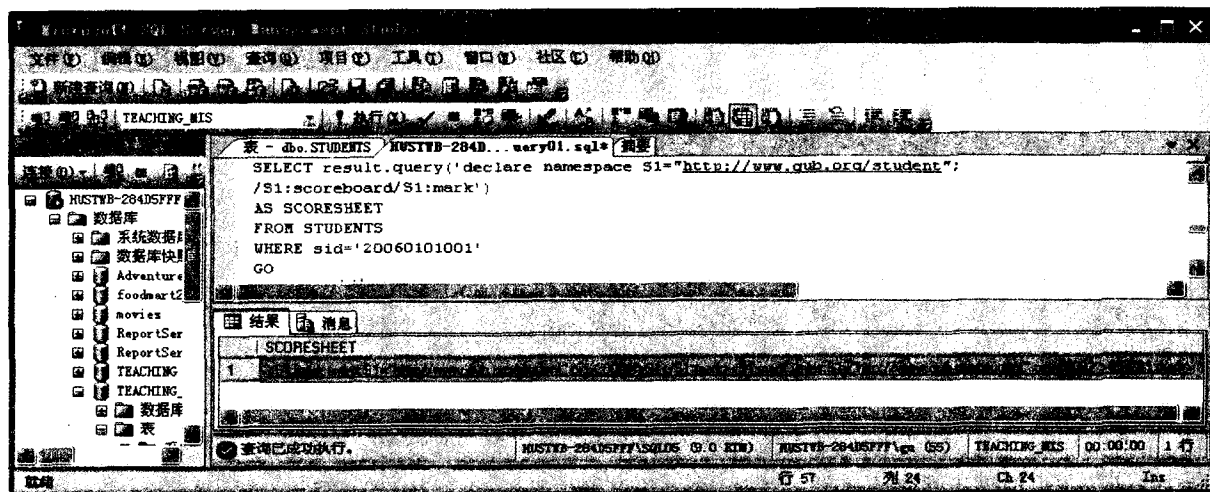


图4 在 SQL Server 2005 中通过 XQuery 表达式修改 XML 数据值阶段

```
SET result.modify('replace value of (/S1:scoreboard/S1:mark[1]/text())[1] with "91"')
WHERE SID = '20060101001'
GO
```

上述语句执行结果,是将学号为"20060101001"的学生所修第1门课程成绩修改为91分。其语句执行情况,可通过XQuery查询得到,如图4所示。

4 结束语

XQuery 被设计用来查询以任何 XML 形式呈现的数据^[10]。作为 W3C 所发布的标准,是各大厂商和 XML 应用开发者要遵循的规范,但由于目前浏览器、XML 处理器对其的支持等原因,其易用性尚显不足。SQL Server 2005 是当前商用大型数据库应用的主流产品,其对 XML 技术提供了全面支持,特别是对于 XQuery 语言的支持,为 XML 数据管理及使用提供了极大便利。但其在 XQuery 结构及其使用上都进行了扩展^[11],文中通过对比 W3C 的 XQuery 标准和 SQL Server 2005 中 XQuery 表达式的特点指出了其异同,而其最大的不同是在 XQuery 表达式中,扩展了对于 XML 数据流的操作^[12]。这一功能上的扩充,不仅使得使用 SQL Server 2005 进行 XML 数据管理、操作与查询更为简便,同时,也与数据库管理与操作统一起来。

参考文献:

- [1] W3C XQuery 1.0: An XML Query Language[EB/OL]. 2007. <http://WWW.W3.org/TR/xquery>.
- [2] 牛杰,黄东. SQL Server 2000 XML 技术及应用[J]. 计算机技术与发展,2006,16(7):242-244.
- [3] 李元韬,曹志宇. XML 查询语言 XQuery 的分析与研究[J]. 太原科技,2010(1):90-92.
- [4] 孙鑫. XML、XML Schema、XSLT2.0 和 XQuery 开发详解[M]. 北京:电子工业出版社,2009.
- [5] 吴君. XQuery 语言查询优化策略研究[J]. 计算机与数字工程,2009,37(10):182-185.
- [6] 施振掇,曹渠江. 基于 XQuery 查询优化的研究[J]. 计算机应用与软件,2008,25(11):86-88.
- [7] Homer A. SQL Server 2005 XQuery and XML-DML - Part 1 [EB/OL]. 2005. <http://www.15Seconds.com/issue/050803.htm>.
- [8] Klein S. SQL Server 2005 XML 高级编程[M]. 北京:清华大学出版社,2007.
- [9] 王国仁,于戈,杨晓春,等. XML 数据管理技术[M]. 北京:电子工业出版社,2007.
- [10] 华珊珊,谢铨洋. XML 查询语言 XQuery 的研究与实现[J]. 计算机技术与发展,2009,19(4):48-49.
- [11] 刘建民,赵政. 一种有效的 XQuery 更新操作[J]. 微处理机,2008(7):113-115.
- [12] 杨卫东,施伯乐. XML 流管理研究综述[J]. 计算机研究与发展,2009,46(10):1721-1728.

(上接第97页)

- [11] Brumen B, Welzer T. Protecting Medical Data for Analyses [C]//Proceedings of the 15th IEEE Symposium on Computer-based Medical Systems (CBMS 2002). [s. l.]: [s. n.], 2002:102-107.
- [12] Le Fevre K, Dewitt D J, Ramakrishnan R. Incognito Efficient full domain k-anonymity [C]//Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD). Balti-

more, Maryland: [s. n.], 2005:49-60.

- [13] Aggarwal G, Feder T, Kenthapadi T, et al. Achieving anonymity via clustering [C]//Proceedings of the Symposium on Principles of Database Systems (PODS). Chicago, Illinois, USA: [s. n.], 2006: 153-162.
- [14] 陈晓明,李军怀,张璟. 隐私保护数据挖掘算法综述[J]. 计算机科学,2007,34(6):183-186.