

基于 K-匿名的电子商务匿名方法研究

解争龙, 弋改珍

(咸阳师范学院 信息工程学院, 陕西 咸阳 712000)

摘 要: 根据 K-匿名的基本模型, 为了实现在电子商务销售商以营销为目的数据分析发布过程中不泄漏消费者的消费隐私信息, 提出了电子商务隐私信息匿名方法, 将电子商务消费者个体信息分为 3 种类, 设计出了个体身份标识属性的重编码和个体身份准标识属性的泛化算法. 利用 Adult data 作为数据源, 分析了本匿名方法泛化层、信息丢失率和 K 之间的关系, 为电子商务对数据的处理决策提供一个选择方案.

关键词: 电子商务; K-匿名; 泛化; 信息丢失率

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2013)02-0136-04

Search of the Anonymity Method Based-on K-anonymity for Electronic Commerce

XIE Zheng-long, YI Gai-zhen

(School of Information & Engineering, Xianyang Normal University, Xianyang 712000, China)

Abstract: According to the model of K-anonymity, the attributes of a information tables in Electronic Commerce have been divided into three parts: identifier, Quasi Identifier, sensitive information. To conceal privacy information for consumer, the paper adopted a recode method to protect the name and cell-phone number against revealing, used the generalization algorithm to conceal the quasi identifier for consumer. The data resource of paper derived from the Adult database. It has been analyzed on the correlations among of the generalization hierarchy, the ratio of information loss and K. The result provided an alternative scheme of data processing decision for data holder.

Key words: electronic commerce; K-anonymity; generalization; ratio of information loss

1 引言

随着 Internet 技术的飞速发展, 基于互联网技术的电子商务应运而生并得到了快速发展. 电子商务正在作为一种现代商务模式, 逐步取代传统的商业运营模式. 而消费者不希望别人知道自己的消费, 买了什么, 买了多少, 消费了多少等等, 所以电子商务安全交易也逐步成为电子商务发展的核心和关键问题. 特别是随着互联网用户的不断增多和技术的不断提高, 人们以正当或不正当的手段在网上以各种形式获取信息, 特别是别人的隐私信息, 这给电子商务的消费者带来极大的威胁. 再者, 商业机构以营销为目的收集并分析消费者的个人信息, 甚至有些

商家将涉及到个人隐私的信息对外发布. 对于商家而言, 对外发布信息时, 一方面应该保证消费者的隐私信息不被泄露; 另一方面, 发布的数据应该含有足够的具有分析价值的信息. 这就成为电子商务末端环节急需解决的一个问题. 本文以 K-匿名模型为基础研究电子商务的末端消费者个体信息匿名的实现.

2 K-匿名定义

为了使得个人隐私信息在发布时不被泄露, 1998 年 Samarati P 和 Sweeney L^[1-3] 提出 K-匿名技术模型, 该模型要求在商家待发布的数据表中, 应该包含一定数量的若干属性值相同的个体. 这样, 即

收稿日期: 2012-05-18; 修回日期: 2012-06-16

基金项目: 陕西省科技厅计划项目(SJ08ZT14-8)

使攻击者通过其他手段获得有关消费者的其他数据表,也不能根据它推断出隐私信息所属的实际个体. 以此防止个人隐私的泄露. 在K-匿名模型中,把商家待发布的消费者信息表中的属性分为三类:(1)个体识别属性(IID: individually identifying attribute):能够唯一标识一个具体个体的属性,例如身份证号、手机号等;(2)准标识属性(QI: quasi identifier attribute):可以与外部表连接的属性,即待发布表与外表的公有属性,如性别、年龄、邮政编码等;(3)敏感属性(ST: sensitive attribute):表示个体隐私信息,例如在电子商务中他买了什么商品、商品的数量、支出的费用是多少等. 在K-匿名模型中,商家若以数据表的形式发布个体隐私信息时,为了使个体隐私信息不被泄露的同时,还要使得这些信息具有分析价值. 由于恶意攻击者能通过准标识属性将待发布表与通过其他手段获得的外表进行连接,进而推断出实际个体身份,并找到个体与敏感属性之间的关系. 据统计,美国约87%的人口可通过邮编、性别、生日等非显式标识符来唯一确定其个体身份^[4]. 于是,K-匿名模型把个体识别属性简单地删除,对准标识属性进行泛化,而保留隐私信息以利于数据分析.

定义1 (准标识符): 假设有实体集 E 、特定实体表 $T(F_1, F_2, \dots, F_n)$, 映射 $f_c: E \rightarrow T$ 以及 $f_g: T \rightarrow E'$, 其中 $E \subseteq E'$. 表 T 的准标识符 Q_T 为属性组 $\{F_i, \dots, F_j\}$, $\{F_i, \dots, F_j\} \subseteq \{F_1, F_2, \dots, F_n\}$, 其中 $\exists e_i \in E$ 使得 $f_g(f_c(e_i)[QI]) = e_i$.

也就是,准标识符同时存在于待发布隐私表与外表中,不能唯一标识个体信息的一组属性称为标识符. 准标识符的取值依赖于攻击者所拥有的外表信息,即隐私表与外表的关联属性,同一个隐私表与多个不同的外表关联时,可能存在不同的准标识符.

定义2 (K-匿名): 假设有数据表 $T(F_1, F_2, \dots, F_n)$, QI是与 T 相关联的准标识符, $T[QI]$ 表示 T 表的元组在QI上的投影. 我们说 T 满足K-匿名, 当且仅当在表 $T[QI]$ 中出现的每个值序列至少要在 $T[QI]$ 中出现 k 次.

根据K-匿名的定义,如果外表中的一个属性子集出现在与表 T 相关联的准标识符中(即 T 满足K-匿名),则攻击者会通过外表与 T 表的连接推导出任何少于 k 个个体的元组.

推论1 假设有数据表 $T(F_1, F_2, \dots, F_n)$, QI $=\{F_i, \dots, F_j\}$ 是与 T 关联的准标识符, 且 $\{F_i, \dots, F_j\} \subseteq \{F_1, F_2, \dots, F_n\}$. 若 T 满足K-匿名, 则在

$T[F_x]$ 中出现的每一个值序列至少要在 $T[QI]$ 中出现 k 次, 其中 $x = i, \dots, j$.

如果待发布的数据表 T 满足K-匿名, 那么通过准标识符QI连接外表与 T 表, 不能够连接到QI或QI任一子集上, 进而推断出少于 k 个数目的个体. 该特性能够预防表中信息与外部源的连接推理.

3 电子商务隐私信息匿名实现

3.1 电子商务网站所收集消费者个体数据信息构成

电子商务的消费者第一次在某一个商务网站购买商品, 首先要进行网站注册, 在注册过程中要提供如下信息: 用户名、姓名、性别、出生日期、住址、邮编、手机号等. 消费者购买商品时要提供如下信息: 上品名、商品属性值、数量、商品付费额等.

按照K-匿名模型将电子商务交易消费者个体属性信息也分为3类:(1)个体身份标识属性, 姓名和手机号码;(2)准标识属性, 性别、出生日期、住址和邮政编码;(3)敏感属性, 商品名、购买数量和商品单价及付费.

3.2 匿名方法

如果按照K-匿名模型对个体身份标识属性简单删除, 他不符合电子商务数据挖掘与分析的需要, 无法分析个体在购买商品的关联性, 他喜爱买哪一类商品, 在买一个商品的同时还买哪一些商品, 那些经常买, 哪些偶然买等等. 但不改变又是隐私泄密, 所以我们对个体身份标识属性采取综合重编码, 也就是对全部个体身份标识属性值采用一个算法形成一个新的编码值. 对个体身份准标识属性, 逐个进行泛化, 对个体敏感属性保持不变, 以此方法实现身份保持的个体信息匿名.

3.2.1 个体身份标识属性的重编码

对个体身份标识属性采取综合重编码, 把姓名和手机号两个属性值相等的同一个个体给它一个相同的编码, 使得既能标识不同个体, 又无法推出编码与个体的关系. 算法描述如下:

- (1) 输入数据表 T
- (2) 以姓名为第一关键字, 手机号为第二关键字建立索引
- (3) 变量定义:
 - string 存放“姓名+手机号”的字符串;
 - ID 变量对不同个体进行编号, 初始值为1;
- (4) string=姓名+手机号
ID=1
编号='t'+str(ID)

```

skip
while(! EOF)
do
    if string == 姓名+手机号
        编号='t'+str(ID)
    else
        ID=ID+1
        编号='t'+str(ID)
    endif
    string=姓名+手机号
skip
done

```

3.2.2 个体身份准标识属性的泛化

目前泛化方法分为局部重编码和全局重编码两类. 文献[5]中提出了全局重编码算法, 它可以应用到身份保持的泛化方法中, 它每次简单地选择一个准标识属性进行泛化, 直到满足匿名需求. 而局部重编码不要求同一属性的取值泛化到相同层次, Wong 等人在全局编码算法的基础上, 提出了被称为 top-down 的局部重编码算法^[6], 该方法首先将数据表中所有元组泛化成一个等价类, 然后反复地将元组进行特化, 在特化期间, 必须保持 K-匿名. 该过程一直持续到元组不能再被特化为止. 本文使用一种自底向上的全局重编码算法.

由于电子商务进行数据挖掘的目的是为了商品的销售, 通过分析不同地区、不同类型的消费者的购买情况, 进行商品的采购和宣传. 基于此思想, 首先选择准标识属性性别对所有元组分組; 再选择准标识属性邮编进行分組泛化, 泛化到县(市), 例如邮编为 712000, 把它泛化为 7120 * *; 第三选择以出生日期分組, 按年龄段泛化. 例如, 年龄可分小于等于 25、大于 25 且小于等于 35、大于 35 且小于等于 50、大于 50, 4 个年龄段按区间进行泛化. 这样形成一个泛化树, 既达到了匿名的目的又便于电子商务数据分析. 算法描述如下:

输入: 个体隐私表 T;

输出: 发布表 RT;

(1) RT 为对 T 进行个体标识属性重编码的结果表;

(2) 对 RT, 以性别为主关键字索引;

(3) 对 RT, 以邮编为关键字索引; 以邮编属性高 4 位相同进行所有元组泛化, 形成新的属性泛化编码.

(4) 对 PT, 以出生日期为关键字索引; 把生日换

算成年龄期, 并对年龄按区段进行泛化, 形成一个表征区间的泛化值编码.

4 匿名评价指标与分析

实现电子商务消费者信息匿名的目的, 就是使攻击者不能通过获得的外部信息或者通过其他途径获得的外部表, 从发布的数据表中推断出某一个个体, 从而发现他的购买隐私. 但基于 K-匿名模型实现的方法是对个体身份标识属性的重编码和个体身份准标识属性的泛化, 泛化度越高匿名性越强, 而泛化度越高信息损耗越大, 也就是个体身份标识属性的重编码值相同的记录越多, 造成消费者信息的可分析性降低, 所以泛化操作在匿名的同时也产生了信息丢失. 以下根据泛化层分析泛化对信息丢失率和 K 的影响.

定义 3 (信息丢失率) 如果给定表 $T(F_1, F_2, \dots, F_a)$, 准标识符 $|QT| = a$. RT 为 T 的匿名表, G_{Fi} (其中 $i = 0, 1, 2, \dots, h$) 表示属性 F 的泛化层, h 为泛化层的最大高度. 则当 RT 满足 K-匿名时信息的丢失率为:

$$loss(RT) = 1 - \frac{\sum_i^a \sum_1^h \frac{G_{Fi}}{G_{F(i-1)}}}{h * a} \quad (1)$$

如果 $T = RT$, 则 $G_{Fi} = G_{F(i-1)}$, $loss(RT) = 0$, 没有信息丢失; 如果泛化层达到最高级 h, 定义为没有任何有效数据成员, 即 $G_{Fh} = 0$, 则 $loss(RT) = 1$.

从美国欧文大学机器学习 adult 数据库中随机抽取 native-country 属性值为 United-States 的记录, 考虑单属性泛化的丢失率, 将属性 fnlwgt 泛化 5 层, 得到泛化层与丢失率之间的关系. 如图 1 所示.

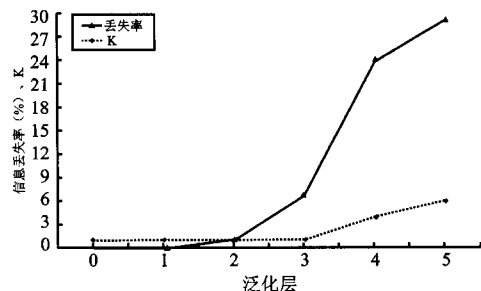


图 1 泛化层对信息丢失率、K 的影响

根据数据分析过程获得的经验, 随着记录数的增加, 经过泛化后的信息丢失率相对减小, 而 k 随之快速增长, 即匿名度提高. 数据持有者通过丢失率和 k 两条曲线选择合适的 k, 对数据进行泛化后发布, 使得结果在保证数据可分析性的同时达到匿名的目的.

5 结束语

当今,网上购物已成为流行的商业运营模式.商业机构掌握了客户的大量购物信息,包括用户的身份标识、性别、地址、邮政编码等信息.当进行信息分析时,为了保护消费者的隐私,防止外部攻击,商业机构在发布信息时,既要保护消费者的隐私,又要对数据进行分析.

本文在K-匿名定义的基础上,给出了数据发布前的处理方法,即使用重编码方法对标识客户身份的信息进行隐藏;采用信息泛化方法保护标识符,以Adult数据库为数据源分析了泛化层、信息丢失率和 k 之间的关系.商家依据了具体的数据库,按照重编码、泛化、分析等步骤选取合适的泛化结果,可以在保证隐藏消费者信息的同时,提供数据的可分析性.

参考文献:

- [1] Samarati P. Protecting respondents identities in microdata release[J]. IEEE Trans on Knowledge and Data Engineering, 2001, 13(6): 1010-1027.
- [2] Sweeney L. Achieving K-anonymity privacy protection using generalization and suppression[J]. Int'l Journal on

Uncertainty, Fuzziness and Knowledge Based Systems, 2002, 10(5): 571-588.

- [3] Sweeney L. K-Anonymity: A model for protecting privacy[J]. Int'l Journal on Uncertainty, Fuzziness and Knowledge Based Systems, 2002, 10(5): 557-570.
- [4] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression[J]. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002: 571-588.
- [5] LeFewe KDeWm DJ, Ramakrishnan R. Incognito: Efficient full-domain K-anonymity[J]. Ozcan F, ed. Proc of the Int'l Conf. on Management of Data. Maryland: ACM Press, 2005: 49-60.
- [6] Wong RC, Li J, Fu AW, Wang K. (α, κ) -Anonymity: An enhanced K-anonymity model for privacy-preserving data publishing[C]//Eliassi-Rad T, Ungar LH, Craven M, Gunopulos D, eds. Proc of the 12th Int'l Conf on Knowledge Discovery and Data Mining. New York: ACM Press, 2006: 754-759.

作者简介:

解争龙 男, (1961—), 硕士, 教授. 研究方向为计算机网络与信息安全.

(上接第135页)

波处理, 获取GPS信号传递过程中的退化程度, 并对其进行补偿. 根据马尔科夫相关理论, 建立GPS通信模型, 从而实现小区域内的GPS通信. 实验结果表明, 这种算法能够有效提高乘性噪声干扰下的小区域GPS通信的准确性, 取得了很好的效果.

参考文献:

- [1] 何小卫, 王爱华, 马跃. 基于GPRS的GPS车载终端通信技术研究[J]. 计算机应用, 2008(11): 2952-2954.
- [2] 冯正, 韩焱, 王黎明. 多线程串口通信技术在GPS导航中的应用[J]. 现代电子技术, 2009(5): 28-30.

- [3] 龚福祥, 王庆, 张小国. 通信网络中的GPS/TDOA混合定位算法[J]. 中国惯性技术学报, 2011(2): 209-213.
- [4] 曹洁, 郭春禹. GSM模块对终端GPS数据的通信实现[J]. 电子测量与仪器学报, 2010(11): 1068-1073.
- [5] 姚西峰, 钱峰, 田蔚风. 基于粒子滤波的GPS/DR组合导航算法研究[J]. 计算机仿真, 2008(7): 262-264.
- [6] 朱星, 许强, 周建斌. 基于GPS/GPRS滑坡位移监测系统的研制及应用[J]. 计算机测量与控制, 2011(10): 2373-2376.

作者简介:

董俊 女, (1975—), 硕士, 讲师. 研究方向为计算机应用.