

一种带频繁项过滤机制的隐私保护新方法

吴泓润, 覃俊

(中南民族大学 计算机科学学院, 武汉 430074)

摘要: 针对差分隐私保护方法的隐私保护过度问题, 提出了一种带频繁项过滤机制的隐私保护新方法, 以提高数据发布结果的准确性。该方法首先对数据源进行预处理, 即对非频繁项进行过滤, 然后执行差分隐私保护算法。从理论上证明了带频繁项过滤机制的隐私保护方法达到差分隐私保护级别, 而且实验结果表明, 在相同的隐私保护度下, 提出的方法数据发布准确性比当前差分隐私保护方法更高。

关键词: 差分隐私保护; 频繁项过滤机制; 数据发布准确性

中图分类号: TP315 文献标志码: A 文章编号: 1001-3695(2012)02-0679-04

doi:10.3969/j.issn.1001-3695.2012.02.074

Novel method of frequent item filtering mechanism in privacy-preserving

WU Hong-run, QIN Jun

(College of Computer Science, South-Central University for Nationalities, Wuhan 430074, China)

Abstract: In order to improve the accuracy of publishing dataset, this paper put forward a novel privacy preserving method that was based on frequent item filtering mechanism, which aimed at to solve the existing issue of over-protected privacy in differential-privacy preserving methods. This method solved the issue by preprocessing the data source, namely, filtering non-frequent items firstly, and then performing the differential-privacy algorithm. The protection level of a frequent item filtering mechanism of privacy preserving method was proved to be differential-privacy in this paper. Experiment results show that the accuracy of the publishing data that of proposed method is higher than the current differential-privacy methods under the same degree of privacy protection.

Key words: differential-privacy protection; frequent item filtering mechanism; accuracy of the publishing data

数据挖掘对当今海量信息的分析和知识发现发挥了积极的作用,但也带来了数据隐私泄露方面的诸多问题。所以,当前数据挖掘过程中一个迫切需要解决的问题就是数据隐私保护问题,而该问题目前已经是挖掘界的一个研究热点^[1]。

目前,隐私保护方法主要有 k-匿名隐私保护方法^[2]、L-多样化方法^[2]、 ϵ -差分隐私保护方法^[3,4]。本文侧重针对当前最流行的差分隐私保护方法进行研究,提出了带频繁过滤机制的隐私保护新方法,并试图从理论和实验上验证带频繁过滤机制的隐私保护方法的有效性。

ϵ -差分隐私保护^[5,6]概念由 Dwork 于 2005 年提出。由于 ϵ -差分隐私保护对发布数据集有很高的隐私级别要求,因此在学术界引起了广大学者的研究兴趣,其中文献[7,8]提出了一些初步的差分隐私保护方法。本文通过对差分隐私保护方法的深入研究,发现 ϵ -差分隐私保护方法过分注重数据的高隐私保护度,忽略了数据发布结果的准确性。例如, T 时刻从某医疗网站日志集中随机选出 7 条敏感属性为“关键字”的记录,其中“癌症”关键字记录在 7 条记录中仅有 1 条,提交“减肥”“体重与疾病”关键字记录均为 3 条,若采用传统的差分隐私保护算法^[7,8],为了达到较高的隐私保护度,会为数据源添加较大的噪声,因此根据隐私保护后的数据发布集,数据挖掘者会得出用户提交“癌症”关键字和提交“减肥”关键字的比例均为 1/3,然而这与实际用户提交的“癌症”关键字的比例(1/

7)相差甚远。本文提出一种新的隐私保护方法——带频繁项过滤机制的隐私保护方法。并通过理论证明和实验验证带频繁项过滤机制的隐私保护方法达到了差分隐私保护级别,并且在数据发布结果的准确性要求较高的场景下,采用本文的带频繁项过滤机制的隐私保护方法明显优于当前的 ϵ -差分隐私保护方法。

1 当前主要的隐私保护方法

1.1 k-匿名(k-anonymous)隐私保护方法

如果数据集 T 的每一个记录都与至少 $k-1$ 个关于这个数据集中的准标志符属性记录相同,则称 T 为 k-匿名。虽然 k-匿名可以防止身份泄露,但是仍不能防止属性泄露。

1.2 L-多样化(L-diversity)方法

如果数据集 T 的准标志符组对应的敏感属性中至少有 L 种取值,则称 T 为 L-多样化。L-多样化方法使得攻击者最多以 $1/L$ 的概率识别出攻击目标。然而数据发布者必须知道敏感值的分布情况,但是在很多情况下敏感值分布很难确定。

在面对数据源中添加或者移除了某一位用户此类问题时 k-匿名和 L-多样化隐私保护方法,均会泄露该用户隐私。针对此类问题, ϵ -差分隐私该保护方法可以防止攻击者以较高概率推断出目标对象的敏感属性取值,因此本文重点对 ϵ -差分

收稿日期: 2011-06-23; 修回日期: 2011-07-24

作者简介: 吴泓润(1989-),女,硕士研究生,主要研究方向为数据挖掘(ms.wuhr@gmail.com);覃俊(1968-),女,教授,硕导,主要研究方向为数据挖掘、网络安全。

隐私保护方法进行研究。

1.3 ε-差分隐私(ε-Differential privacy)方法及问题分析

算法 A 使得改变、添加或者删除一个用户记录,对算法输出的数据几乎没有影响的,即算法 A 对用户 U 记录的改变和删除不敏感,则称算法 A 是差分隐私保护方法。但由于当前差分隐私保护方法的过度保护,使得对于数据源中的一些非频繁项^[9],其发布后的数据准确性下降。

针对引言中某医疗网站日志集差分隐私保护例子,有如下定义:Uid 表示用户账户^[10,11]，“关键字”表示用户提交的关键字，“年龄”表示提交关键字的用户的年龄，“所在地”表示提交关键字用户所在地属于哪里。其中关键字为敏感属性,年龄和地区为标志符。若采用传统的差分隐私保护方法发布数据,则在 T 和 T+1 时刻的数据发布结果(表 1、2)如表 3 所示。发布数据中各个关键字出现的概率分别为 p' (减肥) = 1/3, p' (体重与疾病) = 1/3, p' (癌症) = 1/3,而在数据源中各个关键字的概率为 p (减肥) = 3/7, p (体重与疾病) = 3/7, p (疾病) = 1/7。则疾病项数据源和数据发布集相对误差为 $\frac{p'(\text{疾病}) - p(\text{疾病})}{p(\text{疾病})} \times 100\% > 100\%$ 。

表 1 T 时刻数据源

Uid	关键字	年龄	地区
U1	减肥	21	北京
U2	体重与疾病	36	北京
U3	减肥	28	南京
U4	体重与疾病	35	南京
U5	减肥	37	武汉
U6	体重与疾病	37	武汉
U7	癌症	40	福州

表 2 T+1 时刻数据源

Uid	关键字	年龄	地区
U1	减肥	21	北京
U2	体重与疾病	36	北京
U3	减肥	28	南京
U4	体重与疾病	35	南京
U5	减肥	37	武汉
U6	体重与疾病	37	武汉
U7	癌症	40	福州
U8	失眠	53	厦门

表 3 当前差分隐私方法发布数据集

Uid	关键字	年龄	地区
U1	减肥	20-30	华北
U2	减肥	20-30	华北
U3	减肥	20-30	华北
U4	体重与疾病	30-40	华中
U5	体重与疾病	30-40	华中
U6	体重与疾病	30-40	华中
U7	癌症	40-50	华南
U8	癌症	40-50	华南
U9	癌症	40-50	华南

因此,可以从上述简单举例中得出,若对数据源采用当前的差分隐私保护方法,则无论用户 U 是否在数据源中攻击者

得到的数据发布集几乎相同,因此攻击者很难确定目标攻击对象是否在数据发布集中。然而该方法的缺点也不言而喻,由于当前差分隐私方法的过度保护使得数据发布的准确性降低。如何解决数据源中的非频繁项所引起的数据挖掘结果准确性降低这一问题,将关系到企业、网站等数据使用者的切身利益。本文下面给出了一种解决方案,首先对数据源作预处理,将非频繁项除去,然后对预处理之后的数据进行隐私保护。

2 带频繁项过滤预处理机制的差分隐私保护策略

2.1 算法思想

本文提出的带频繁项过滤机制的隐私保护策略总体分为两个阶段:a)设定预处理门限值,删除关键字频数少于门限值的非频繁项;b)对符合 a) 阶段要求的项添加拉普拉斯噪声,并删除小于新的门限值的非频繁项。算法框架如图 1 所示。

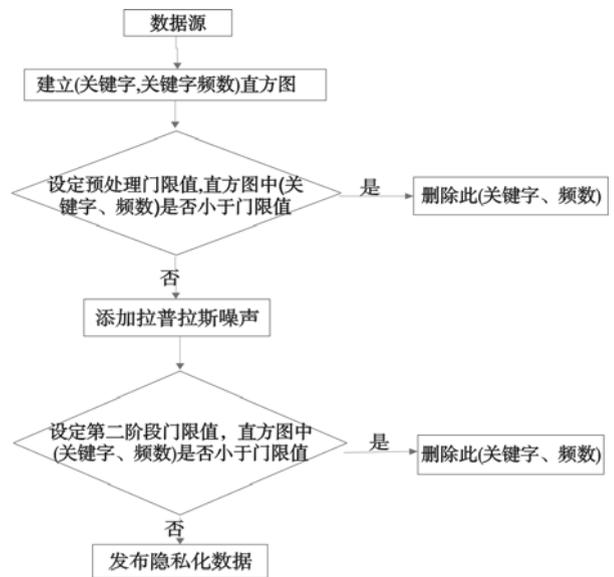


图 1 算法框架

2.2 算法 FFA(frequency-filtering algorithm)的设计

带频繁项过滤机制的隐私保护算法伪代码如下:

- a) 输入: $k, \lambda, \tau, \tau', \Omega$
- b) 随机选择日志记录: for $i = 1$ to u_k
SelectRecordsToHistogram(Ω', u_i)
- c) 过滤非频繁项:
for $i = 1$ to k
 $n_q = \text{Count}(q)$
if $n_q < \tau$ then delete (q, n_q)
else AddToHistogram ($H, (q, n_q)$)
end if
end for
output H
- d) 添加噪声:
for $i = 1$ to k
 $\chi_k = \text{Lap}(\lambda)^4$
 $n_q = n_q + \chi_k$
if $n_q < \tau'$ then delete (q, n_q)
else AddToOutSet ($\Theta', (q, n_q)$)
end if
end for
output Θ'
- e) 发布数据: Publishing Θ'

其中: Ω 为日志集; u 为 Ω 中的每一位用户; u_i 为 u 用户的记录; k 为不同关键字数; u_k 为随机从用户 u 日志集中选择出不

同关键字数目; Ω' 为从 Ω 的每一位用户 u 中, 选出具有 u_k 种不同关键字的日志记录所构成的集合; q 为关键字; n_q 为关键字 q 出现的频数; $H(q, n_q)$ 为 q 在 Ω' 出现的频数; τ 为预处理门限值, 用来过滤非频繁项 (q, n_q); χ_k 为拉普拉斯分布 $\text{Lap}(\lambda)$ ⁴ 产生一个随机数 χ_k ; τ' 为第二个门限值, 用来过滤噪声处理后的非频繁项 (q, n_q); Θ' 为算法结束后输出的结果集。

2.3 差分隐私保护级别的理论证明

定理 1 如果算法 F 满足 $\frac{\text{pr}(F(S) = \Theta)}{\text{pr}(F(S') = \Theta)} \leq e^\varepsilon$ 且 $\frac{\text{pr}(F(S') = \Theta)}{\text{pr}(F(S) = \Theta)} \leq e^\varepsilon$, 则称算法 F 达到 ε -差分隐私保护级别^[5,6]。其中, 算法 F 为本文算法 FFA 的简称, S 和 S' 分别表示两个相邻的数据源, Θ 表示数据源 S 和 S' 对应的输出数据集。

证明 S 和 S' 在伪码步骤 b) 分别建立的直方图为 H 和 H' 。 $|\Delta|$ 为 H 和 H' 对应的关键字频数差。 k_i 为 Δ 中的第 i 个关键字, d_i, d'_i, d_o 分别为关键字 k_i 在直方图 H, H', Ω' 的频数。由差分隐私算法敏感度要求可知 $|d_i - d'_i| = 1, |\Delta| \leq 2k$ 。 $E_i, E'_i, E_i^*, E_i'^*$ 分别表示关键字 k_i 在直方图 H, H', Θ 中出现的频数 d_i, d'_i, d_o 事件。所以可得如下关系:

$$\frac{\text{pr}(F(S) = \Theta)}{\text{pr}(F(S') = \Theta)} = \prod_{i \in [1, |\Delta|]} \frac{\text{pr}(E_i^* | E_i)}{\text{pr}(E_i'^* | E_i')}$$

那么接下来将需要证明 $\frac{\text{pr}(E_i^* | E_i)}{\text{pr}(E_i'^* | E_i')} \leq e^{1/\lambda}$ 。

为了证明上述结论需分析 d_i, d_o, τ 之间的关系, 因此有以下三种情况: a) $d_i \geq \tau, d_o \geq \tau$; b) $d_i < \tau, d_o < \tau$; c) $d_i = \tau, d_o = \tau - 1$ 。

针对情况 a), 若 $d_o > 0$, 因为对于任意 i 有 $|d_i - d'_i| = 1$, 所以有

$$\frac{\text{pr}(E_i^* | E_i)}{\text{pr}(E_i'^* | E_i')} = \frac{\frac{1}{2\lambda} e^{-|d_o - d_i|/\lambda}}{\frac{1}{2\lambda} e^{-|d_o - d'_i|/\lambda}} = e^{(|d_o - d'_i| - |d_o - d_i|)/\lambda} \leq e^{|d_i - d'_i|/\lambda} = e^{1/\lambda}$$

同理, 如果 $d_o = 0$, 则 $\frac{\text{pr}(E_i^* | E_i)}{\text{pr}(E_i'^* | E_i')} \leq e^{1/\lambda}$ 。

针对情况 b), 如果 $d_i < \tau$, 算法在步骤 e) 将小于门限值 τ 的项删除, 因此 $d_o = 0$, 并且 $\text{pr}(E_i^* | E_i) = 1, \text{pr}(E_i'^* | E_i') =$

$$\begin{cases} 1, d_i \leq \tau \\ 1 - \frac{1}{2} e^{-|d_i - d'_i|/\lambda}, d_i > \tau \end{cases} \frac{\text{pr}(E_i^* | E_i)}{\text{pr}(E_i'^* | E_i')} \leq \frac{1}{1 - \frac{1}{2} e^{-|d_i - d'_i|/\lambda}} \leq$$

$$\frac{1}{1 - \frac{1}{2} e^{-|d_i - d'_i|/\lambda}} \leq \frac{1}{1 - \frac{1}{2} e^{-\ln(2 - 2e^{-1/\lambda})}} = e^{1/\lambda} \quad (\tau \text{ 和 } \tau' \text{ 满足: } \tau' \geq \tau -$$

$\lambda \ln(2 - 2e^{-1/\lambda})$ 和 $\tau = \lceil 2k/\varepsilon \rceil$ 时, 情况 b) 成立)。

针对情况 c), 若 $d_i = \tau, d_o = \tau - 1$ 。算法的步骤 e) 将小于门限值 τ 的项删除, 故 $d_o = 0, \text{pr}(E_i^* | E_i) = 1$ 。因此

$$\frac{\text{pr}(E_i^* | E_i)}{\text{pr}(E_i'^* | E_i')} = \text{pr}(E_i^* | E_i) \leq e^{1/\lambda}。$$

综上所述可得出 $\frac{\text{pr}(E_i^* | E_i)}{\text{pr}(E_i'^* | E_i')} \leq e^{1/\lambda}$ 。

又因为 $|\Delta| \leq 2k$, 所以可以得出: $\frac{\text{pr}(F(S) = \Theta)}{\text{pr}(F(S') = \Theta)} =$

$$\prod_{i \in [1, |\Delta|]} \frac{\text{pr}(E_i^* | E_i)}{\text{pr}(E_i'^* | E_i')} \leq e^{|\Delta|/\lambda} \leq e^\varepsilon。$$

同理可证 $\frac{\text{pr}(F(S') = \Theta)}{\text{pr}(F(S) = \Theta)} \leq e^\varepsilon$ 。

证毕。

3 实验结果与分析

3.1 实验参数设置

算法 F 中有参数 $k, \lambda, \tau, \tau', \varepsilon$ 。其中, 参数 τ, τ' 对数据发布准确性、隐私保护度、数据发布率均有至关重要的意义。因为若 τ 设置过大, 算法 F 会删除很多非频繁项, 会导致数据发布率极低; 若 τ 设置过低, 则会导致数据准确度降低; 同理, 若 τ' 设置过大, 算法 F 会删除很多非频繁项, 会导致数据发布率过低; 若 τ' 设置过低, 则会导致数据隐私保护度过低。因此, 设置合适的 τ, τ' 参数对数据发布准确度极为重要。

由 2.3 节的证明可知:

$$\tau = \lceil 2k/\varepsilon \rceil \text{ 时, } \tau' \text{ 取最小} \quad (1)$$

$$\tau' \geq \tau - \lambda \ln(2 - 2e^{-1/\lambda}) \quad (2)$$

其中: ε, λ, k 参数由数据发布者自行选定, 本部分选取的实验参数为 $\varepsilon = \ln 10, \lambda = 10, k = 3$ 。

根据式 (1) (2), 计算得: 当 $\tau = 4$ 时, τ' 取得最小值, 且 $\tau' = 77.31$ 。

因此, 在下面的数据发布率和数据发布结果准确性实验中 本文使用参数为 $\varepsilon = \ln 10, k = 3, \tau = 4, \tau' = 77.31, \tau = 4, \tau' = 77.31$ 。

3.2 实验结果比较

为了比较当前差分隐私保护方法和本文提出的带频繁项过滤机制的隐私保护方法的数据发布率、隐私保护度、数据发布准确性间的区别, 本文给出以下计算公式:

数据发布率计算式为

$$P(r) = \frac{K_{ir}}{K_{io}}, (1 \leq i \leq k) \quad (3)$$

其中: K_{ir} 为从数据源中选出的记录集中关键字 i 的频数, K_{io} 为数据发布集中关键字 i 的频数。

攻击者攻击成功的概率计算式为

$$P(s) = \frac{t}{K} \quad (4)$$

其中: t 为攻击者识别出目标攻击对象, $K = \Theta'_1 \cap \Theta'_2 \cap \Theta'_3 \cap \dots \cap \Theta'_N$, 即 K 为 N 次发布的数据集得交集。

数据发布结果准确性计算式为

$$P(t) = \frac{p_r(i)}{p_o(i)} \times 100\% \quad (1 \leq i \leq k) \quad (5)$$

其中: $p_r(i)$ 为数据源中关键字 i 的概率, $p_o(i)$ 为数据发布集中关键字 i 的概率。

实验中随机选择某医院网站日志集中的 3000 条日志记录, 根据以上参数和式 (3) ~ (5) 比较带频繁项过滤机制的隐私保护方法和当前主流的差分隐私保护方法^[12]。本文根据以上计算, 选定实验参数为 $\varepsilon = \ln 10, k = 3, \tau = 4, \tau' = 77.31$, 分别使用带频繁项过滤机制的隐私保护方法和差分隐私保护方法对数据源隐私化, 得出如图 2 ~ 4 所示实验结果。图 2 表明, 本文提出的带频繁项过滤机制的隐私保护方法和差分隐私保护方法在相同参数下隐私保护度一致; 图 3 表明前者得出的数据发布率略低于后者; 图 4 表明前者得出的数据准确性远高于后者。综上实验分析可得, 本文提出的带频繁项过滤机制的隐私保护方法达到了差分隐私保护级别, 虽然本方法得出的数据发布率略低于后者, 但是本方法得出的数据发布结果准确性远高

于当前主流的差分隐私保护方法。因此,在相同隐私保护级别下,如果数据发布者更关注数据发布结果准确率,本文提出的方法就优于传统的差分隐私保护方法。

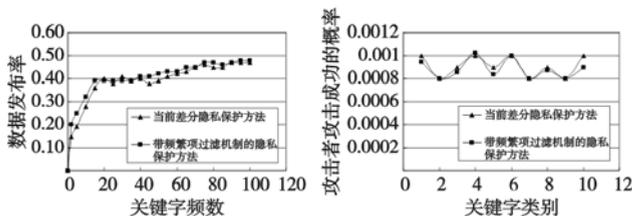


图2 数据发布率比较

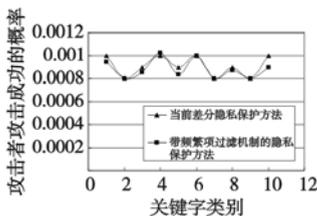


图3 隐私保护级别比较

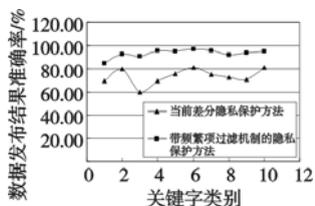


图4 数据发布结果准确率比较

4 结束语

本文就现有差分隐私保护算法由于对数据隐私保护过度而导致数据发布准确性降低这一问题进行了深入研究,分析了差分隐私保护方法隐私保护过度根本原因,提出了一种带频繁项过滤机制的隐私保护新方法,并通过理论和实验验证证明了带频繁项过滤机制的隐私保护方法达到了差分隐私保护级别,可以有效克服差分隐私方法对数据保护过度这一问题。虽然带频繁项过滤机制的隐私保护方法的数据发布率略低于传统的差分隐私保护方法,但前者的数据发布准确性比后者有很大程度上的提高。因此,在数据发布者更关注数据发布结果准确率情况下,本文的方法优于传统的差分隐私保护方法。

在未来的工作中,将在降低对数据发布率影响的情况下,

进一步提高数据发布准确性,并将进一步探索如何提高数据发布率。

参考文献:

- [1] VERYKIOS V S, BERTINO E, FOVINO I N. State of the art in privacy preserving data mining[J]. ACM SIGMOD Record, 2004, 33(1):50-57.
- [2] 阙莹莹,曹天杰.一种增强的隐私保护 K-匿名方法——(α, L) 多样化 K-匿名[J]. 计算机工程与应用, 2010, 46(21):148-151.
- [3] 周水庚,李丰,陶宇飞,等.面向数据库应用的隐私保护研究综述[J]. 计算机学报, 2009, 32(5):848-858.
- [4] 任静涵,张保稳,陈晓桦. 隐私保护数据挖掘研究进展[J]. 信息安全与通信保密, 2008, 27(8):2823-2827.
- [5] DWORK C. Differential privacy: a survey of results[C]//Proc of the 5th Annual Conference on Theory and Applications of Models of Computation. 2008:1-19.
- [6] DWORK C, SMITH A. Differential privacy for statistics: what we know and what we want to learn[J]. Journal of Privacy and Confidentiality, 2009, 2(1):135-154.
- [7] LIU Tan-tan, WANG Fan, ZHU Jie-dan, et al. Differential analysis on deep Web data sources[C]//Proc of the 2010 IEEE International Conference on Data Mining Workshops. 2010:33-40.
- [8] ZHANG Ning, LI Ming. Distributed data mining with differential privacy[J]. ACM SIGMOD Record, 2010, 15(1):493-502.
- [9] 刘彩虹,刘强,李爱平. 基于向量内积的非频繁项挖掘算法研究[J]. 计算机工程与科学, 2011, 33(2):92-96.
- [10] 鲍钰,黄国兴. 基于 Web 日志的隐私保护关联规则挖掘方法[J]. 计算机科学, 2009, 38(8):220-223.
- [11] 雷红艳,邹汉斌. 限制隐私泄露的隐私保护聚类算法[J]. 计算机工程与设计, 2010, 31(7):1444-1446.
- [12] KOROLVA A, KENTHAPADI K, MISHRA N, et al. A releasing search queries and clicks privately[C]//Proc of International World Wide Web Conference Committee. 2009:171-180.

(上接第 678 页)协议具有成本低、效率高、安全性和隐私性好等优点。对应用此安全协议的可行性进行了分析研究,其能够解决隐私、重放、前向安全性、同步性、不可分辨性与位置跟踪、拒绝服务式攻击等安全问题。但是协议中使用的流密码加密,本文并没有给出具体算法,这需要考虑具体的硬件设计,现有的算法有 A5、 η 等^[13]。设计一个合适的密码算法是今后的研究方向。

参考文献:

- [1] LAURIE A. Practical attacks against RFID[J]. Network Security, 2007(9):4-7.
- [2] YUKIYASU T, TERUO S, TOMOYASU S, et al. Cryptanalysis of DES implemented on computers with cache[C]//Proc of the 5th International Workshop on Cryptographic Hardware and Embedded Systems. [S. l.]: Springer-Verlag, 2003:62-76.
- [3] SARMA S, WEIS S, ENGELS D. RFID systems, security and privacy implications, MIT-AUTOID-WH-014[R]. [S. l.]: Auto-ID Center, MIT, 2002.
- [4] JUELS A, RIVEST R L, SZYDLO M. The blocker tag: selective blocking of RFID tags for consumer privacy[C]//Proc of ACM Conference on Computer and Communications Security. 2003:103-111.
- [5] 周永彬,冯登国. RFID 安全协议的设计与分析[J]. 计算机学报,

- 2006, 29(4):581-589.
- [6] GARFINKEL S L, JUELS A, PAPPU R. RFID privacy: an overview of problems and proposed solutions[J]. IEEE Security & Privacy Magazine, 2005, 3(3):34-44.
- [7] JUELS A. RFID security and privacy: a research survey[J]. IEEE Journal on Selected Areas in Communications, 2004, 24(2):381-394.
- [8] KIRN H S, OH J H, CHOI J Y. Analysis of the RFID security protocol for secure smart home network[C]//Proc of International Conference on Hybrid Information Technology. 2006:356-363.
- [9] DIMITRIOU T. A light weight RFID protocol to protect against traceability and cloning attacks[C]//Proc of the 1st International Conference on Security and Privacy for Emerging Areas in Communication Networks. 2005:137-145.
- [10] LE T V, BURNMESTER M, De MEDEIROS B. Universally composable and forward secure RFID authentication and authenticated key exchange[C]//Proc of the 2nd ACM Symposium on Information, Computer and Communications Security. 2007:242-252.
- [11] 王信,薛小平,张思东. RFID 系统数据安全研究[J]. 信息技术与信息化, 2006(1):51-53.
- [12] 杨波. 现代密码学[M]. 2版. 北京:清华大学出版社, 2007.
- [13] MENEZES A J, Van OORSCHOT P C, VANSTONE S A. Handbook of applied cryptography[M]. Boca Raton: CRC Press, 1997.