一种基于有损连接的个性化隐私保护方法

刘英华

(中国青年政治学院 北京 100089)

摘 要 匿名模型是近年来隐私保护研究的热点技术之一,主要研究如何在数据发布中既能避免敏感数据泄露,又能保证数据发布的高效用性。提出了一种$(\alpha_{[s]},k)$-匿名有损分解模型,该模型通过将敏感属性泛化成泛化树,根据数据发布中隐私保护的具体要求,给各结点设置不同的个性化α约束;基于数据库有损分解思想,将数据分解成敏感信息表和非敏感信息表,利用有损连接生成的冗余信息实现隐私保护。实验结果表明,该模型很好的个性化保护了数据隐私。

关键词 数据发布,隐私保护,数据挖掘,有损连接,k-匿名

中图法分类号 TP311 **文献标识码** A

Personalized Privacy Preserving Method Based on Lossy Join

LIU Ying-hua

(China Youth University for Political Science, Beijing 100089, China)

Abstract Recently, anonymity model is one of the hot topic techniques in privacy preserving research. The mainly research is how to avoid leakage of sensitive data in data publishing, but also ensures the efficient use of data. This paper proposed a personalized$(\alpha_{[s]},k)$ - lossy decomposition anonymity model. This method publishes the personalized data though generalization technology and personalizedα restriction for different code of the generalization tree. Based on the idea of lossy decomposition in database, data is projected into the sensitive information table and non sensitive information table, and then the redundant information can be realized privacy protection. Experimental results show that the model can provide better privacy.

Keywords Data dissemination, Privacy preserving, Data mining, Lossy join, k-anonymity

近年来,基于匿名化操作的限制发布技术在隐私保护领域得到了广泛的关注,并成为了该领域的研究热点和难点。

无论是k-匿名技术、ℓ-多样技术、t-closeness技术、(α,k)-匿名技术及其衍生技术等,均提出了等价类[1-11]的概念,并对等价类中的元组进行相应的约束。k-匿名技术[1-5]约束等价类中必须包含k条记录,ℓ-多样技术[6]约束等价类中的元组必须有ℓ个好表现,t-closeness技术[7]约束等价类内的敏感属性分布与在总体分布的差异小于t。(α,k)-匿名技术[3,9]约束等价类中必须包含k条记录,且约束了敏感属性比例。

匿名技术均采用了泛化技术[12],因此数据集的信息损失度高,数据查询精度也将降低,则为后续的数据挖掘工作提供的信息缺失度也会增加。

数据库中的有损分解思想是将一个数据表分解为2个或多个数据表,这种模型没有对原始数据进行匿名操作,即没有对数据进行泛化,隐私信息损失度低,数据查询精度高,能为后续的数据挖掘工作提供更多的信息。而且基于有损分解思想的模型,数据挖掘者无法重构原始数据表,避免了隐私的泄露[11]。

因此,我们提出一种面向有损连接的个性化隐私保护模型$(\alpha_{[s]},k)$-匿名有损分解模型,这种模型没有对原始数据进行匿名操作,即没有对数据进行泛化,隐私信息损失度低,数据查询精度高,能为后续的数据挖掘工作提供更多的信息。由于采用了有损分解思想,数据挖掘者无法重构原始数据表,避免、降低了隐私的泄露。

1 基本概念

1.1 属性分类

数据发布将待处理的原始数据按属性分为4类:1)个体标识属性(Individually Identifier Attribute),即用来标识个体的属性,例如姓名、身份证号、手机号等。2)准标识属性(Quasi Identifier Attribute),即一组基于背景知识可以标识个体的属性,例如{出生日期,地址,性别}属性集可以标识个体。3)敏感属性(Sensitive Attribute),即包含个体隐私信息的属性,例如薪酬、宗教信仰、疾病等。4)非敏感属性(Not Sensitive Attribute),即可以公开的属性。

1.2 抑制和泛化

匿名技术主要采用泛化和抑制两种操作实现。抑制技术即隐藏属性,这部分不发布的属性,攻击者是无法得到的,即

本文受国家自然科学基金项目(61175048)资助。

刘英华(1975—),女,博士,主要研究领域为数据挖掘、知识发现、隐私保护。

是安全的，无任何隐私泄露。泛化技术即对原始数据的概括，例如用范围代替具体数据值，泛化虽然保护了隐私，但导致了部分数据的缺损。

1.3 有损分解

定义 1(有损分解) 关系模式 $R\langle U,F\rangle$ 的一个模式分解是 $\rho=\{R_1\langle U_1,F_1\rangle,R_2\langle U_2,F_2\rangle,\cdots,R_n\langle U_n,F_n\rangle\}$，若关系模式 $R\langle U,F\rangle$ 的任意一个关系符合 $r=m_\rho(r)$，则称 ρ 是无损分解(Lossless Decomposition)，反之，则称为有损分解(Lossy Decomposition)。其中：

$m_\rho(r)=\bowtie\pi_{R_i}(r),i=1\cdots k$，$\bowtie$ 表示自然连接

$\pi_{R_i}(r)=\{t.U_i\mid t\in r\}$

2 相关工作

k-匿名(k-Anonymity)[1-5]隐私保护模型解决了早期存在的链接攻击问题，但因为没有对敏感数据约束，攻击者可以利用背景知识攻击匿名化后的数据。

ℓ-多样(ℓ-diversity)[6]模型在一定程度上解决了背景知识攻击，但是若原始数据量过大，ℓ 值较小(例如 $\ell=3$)，则产生的各个等价类中包含的数量过大。另外，若两个敏感属性值差异过大，则确定敏感属性值的敏感度是非常困难的。

个性化匿名(Personalized Anonymity)[8]隐私保护模型即针对不同的发布要求可以通过抑制度的不同确定隐私的保护程度，实现了满足个人隐私要求的最小量概括，最大程度地保留了原始数据中信息。

3 $(\alpha_{[s]},k)$-匿名有损分解模型

3.1 k-匿名模型

表 1 是隐藏了身份证、姓名等个体标识属性信息的医疗信息表，{Sex,Age,Zip}是准标识属性，{Disease}是敏感属性。表 2 是选民表。根据表 1 和表 2 很容易知道 Mary 得了肺癌。

表 1 医疗信息表

No.	Sex	Age	Zip	Disease
1	M	21	10095	Flu
2	F	23	10095	Flu
3	F	45	10087	Dyspepsia
4	F	45	10087	Gastritis
5	M	34	10086	Gastritis
6	F	32	10088	Cancer
7	M	45	10078	Gastritis
8	F	43	10078	Dyspepsia

定义 1(等价类) 已知数据集 D 和准标识属性集 QI，QI 属性相同的元组组成的集合成为 D 的一个等价类。表 3 共 4 个等价类，每个等价类中包含 2 个元组，即每个等价类包含两个 QI 完全一致的元组。

表 2 选民表

Name	Sex	Age	Zip
Anne	F	29	10095
Mary	F	32	10088
Sam	M	45	10097

定义 2(k-匿名) 已知数据集 D 和准标识属性集 QI，若 D 中各等价类中的元组个数不小于 k，则数据集 D 满足 k-匿名。表 3 满足 2-匿名。

表 3 2-匿名

GroupID	Sex	Age	Zip	Disease
1	*	[20,25]	10095	Flu
1	*	[20,25]	10095	Flu
2	F	45	10087	Dyspepsia
2	F	45	10087	Gastritis
3	*	[30,35]	1008 *	Gastritis
3	*	[30,35]	1008 *	Cancer
4	*	[40,45]	10078	Gastritis
4	*	[40—45]	10078	Dyspepsia

3.2 ℓ-多样模型

定义 3(ℓ-多样) 已知数据集 D 和准标识属性集 QI，若数据集 D 满足 k-匿名，且数据集 D 中各个等价类中的好表现(well - represented)不同的元组个数不小于 ℓ，则称数据集 D 满足 ℓ-多样。好表现既可以是一个等价类中至少包含不同的 ℓ 个敏感属性值，也可以是熵的 ℓ-多样(Entropy l-Diversity)。表 4 满足 2-匿名，同时也满足 2-多样，表 4 中的每个等价类至少包含 2 个元组。

表 4 2-匿名，2-多样

GroupID	Sex	Age	Zip	Disease
1	*	[20,35]	100 * *	Flu
1	*	[20,35]	100 * *	Flu
1	*	[20,35]	100 * *	Gastritis
1	*	[20,35]	100 * *	Cancer
2	F	45	10087	Dyspepsia
2	F	45	10087	Gastritis
3	*	[40—45]	10078	Gastritis
3	*	[40—45]	10078	Dyspepsia

3.3 $(\alpha_{[s]},k)$-匿名有损分解模型

根据泛化技术将敏感属性{Disease}设计成树型结构，疾病是叶子节点，中间节点是其子树各节点疾病的泛化，越接近根节点，泛化程度越高。

针对不同的人群，敏感属性的强弱是不同的，例如家长更关注幼儿园教师是否患某种传染性疾病，例如肺炎，而股民更关注公司高层是否有某种致命疾病，例如癌症。

定义 4 所有敏感属性值 s 组成敏感属性集合 S，按树形结构分类，构造分类树。敏感属性值 s 是分类树中的一个节点，subs(s)是节点 s 的子树，parent(s)是节点 s 的父结点。

表 5 分类树

Disease	Respiratory System	Respiratory infection	Flu
			Pneumonia
			Bronchitis
			…
		…	
	Digestive System	Gastropathy	Gastric
			Gastritis
			Dyspepsia
			…
		…	
	Tumor	Malignant	Cancer
			…
		Benign	…
	…	…	…

定义 5($\alpha_{[s]}$约束) 原始数据集 D 可以划分成 n 个等价类 $E_i(i\in[1,n])$，等价类 E_i 包含 N_i 个元组，N_i 个元组中包含敏感属性 s 的记录数为 n_i，则 $freq(E_i,s)=n_i/N_i$。α 是数

据发布用户设定的关于敏感属性 s 的参数，$\alpha_{[s]} \in (0,1)$，若等价类 E_i 符合 $\alpha_{[s]}$ 约束，则等价类 E_i 的 $freq(E_i,s)<\alpha_{[s]}$。

定义 6（$(\alpha_{[s]},\ell)$-多样 k-匿名模型） 原始数据集 D 匿名后得到数据集 D'，D' 满足 k-匿名，且每个等价类 E_i 至少包含 ℓ 个不同的敏感属性，不同的等价类 E_i 分别满足不同的 $\alpha_{[s]}$ 约束。

表 6 是符合（$\alpha_c=0.3,2$）-多样 2-匿名的匿名数据表，其中每个匿名后的等价类中至少包含 2 个元组，且每个等价类 E_i 至少包含 2 个不同的敏感属性，$freq(E_1,$"Flu"$)=1/4=0.25$，即 $freq(E_1,$"Flu"$)<\alpha$，满足个性化 $\alpha_{[Flu]}$ 约束。

表 6 （$\alpha_{[Flu]}=0.3,2$）-多样 2-匿名

GroupID	Sex	Age	Zip	Disease
1	*	[20,35]	100 * *	Flu
1	*	[20,35]	100 * *	Respiratory infection
1	*	[20,35]	100 * *	Gastritis
1	*	[20,35]	100 * *	Cancer
2	F	45	10087	Dyspepsia
2	F	45	10087	Gastritis
3	*	[40—45]	10078	Gastritis
3	*	[40—45]	10078	Dyspepsia

表 3 是 2-匿名数据表，匿名化后有 4 个元组，为保证等价类中元组不少于 2，6 个元组泛化{Sex}属性，2 个元组泛化{Zip}属性，{Zip}属性只泛化了 1 位。表 4 是 2-匿名 2-多样数据表，为保证约束，匿名化后只有 3 个元组，说明泛化更加严重，6 个元组泛化{Sex}属性，4 个元组泛化{Zip}属性，{Zip}属性泛化了 2 位。表 6 是（$\alpha_{[Flu]}=0.3,2$）-多样 2-匿名数据表，为完成"$\alpha_{[Flu]}=0.3$"约束泛化了一个敏感属性。因此可以得出结论，即随着匿名程度的增加，数据泄露的可能性越小，但数据的缺损度越大，而缺损度的增大意味着挖掘出的信息与基于原始数据表挖掘的效果差异越大，即数据的效用性越差。

定义 7（$(\alpha_{[s]},k)$-匿名有损分解模型） 已知原始数据集 D 和（$\alpha_{[s]},\ell$）-多样 k-匿名模型。将 D 分解成 2 个关系表，即敏感信息表和非敏感信息表，通过等价类编号有损连接发布数据。

（$\alpha_{[s]},k$）-匿名有损分解模型是将一个是数据表有损分解为两个表，一个是包含敏感信息的数据表 SS，另一是包含非敏感信息的数据表 NSS，两个数据表通过增加的等价类属性连接。根据原始数据表 1，首先通过 2-匿名得到表 2，然后通过 2-匿名 2-多样得到表 3，再给定"流感"约束 $\alpha_{[Flu]}=0.3$，得到表 6。

（$\alpha_{[s]},k$）-匿名有损分解模型是在表 6 的基础上得到的，表 6 是一个临时数据表，命名为 D^*，通过临时数据表 D^* 得到了等价类分组的编号。然后 D^* 在 GroupID 和准标识属性集上投影得到敏感属性表 SS，合并敏感属性表 SS 中相同的元组；在 GroupID 和敏感属性上投影得到非敏感属性表 NSS，根据合并后的 SS 修改 GroupID，合并非敏感属性表 NSS 中相同的元组，最后返回合并后的非敏感属性表和合并后的敏感属性表。根据表 6（$\alpha_{[Flu]}=0.3,2$）-多样 2-匿名分解得敏感信息的数据表 7。

表 7 敏感属性表 SS

GroupID	Disease
1	Flu
1	Respiratory infection
1	Gastritis
1	Cancer
2	Dyspepsia
2	Gastritis
3	Gastritis
3	Dyspepsia

不考虑元组顺序，表 7 中的等价类 2 和等价类 3 是完全一致的，因此等价类 2 和等价类 3 是可以合并的，合并后的敏感属性表见表 8。

表 8 合并后的敏感属性表

GroupID	Disease
1	Flu
1	Respiratory infection
1	Gastritis
1	Cancer
2	Gastritis
2	Dyspepsia

根据表 6（$\alpha_{[Flu]}=0.3,2$）-多样 2-匿名数据表和表 1 医疗信息表，分解得到包含非敏感信息的表 9。

表 9 非敏感属性表 NSS

GroupID	Sex	Age	Zip
1	M	21	10095
1	F	23	10095
1	M	34	10086
1	F	32	10088
2	F	45	10087
2	F	45	10087
3	M	45	10078
3	F	43	10078

表 9 非敏感属性表结合表 8 合并后的敏感属性表，将等价类 2 和等价类 3 合并，见表 10。

表 10 新非敏感属性表 NSS

GroupID	Sex	Age	Zip
1	M	21	10095
1	F	23	10095
1	M	34	10086
1	F	32	10088
2	F	45	10087
2	F	45	10087
2	M	45	10078
2	F	43	10078

表 9 非敏感属性表中的等价类 2 包含了 4 个元组，其中存在 2 个元组信息完全一致的情况，因此等价类 2 的这 2 个元组是可以合并为一个元组的，合并后的非敏感属性表见表 11。

表 11 合并后的非敏感属性表 NSS

GroupID	Sex	Age	Zip
1	M	21	10095
1	F	23	10095
1	M	34	10086
1	F	32	10088
2	F	45	10087
2	M	45	10078
2	F	43	10078

对比表 6($\alpha_{[Flu]}$=0.3,2)-多样 2-匿名数据表和表 8 合并后敏感属性 SS 表与表 11 合并后的非敏感属性 NSS,前者仅仅发布 8 个元组,而后者发布 2 个表,共 13 个元组(表 8 包含 6 个元组,表 11 包含 7 个元组),因此($\alpha_{[s]}$,k)-匿名有损分解模型发布的信息量要大于($\alpha_{[s]}$,ℓ)-多样 k-匿名模型。

将表 8 合并后的非敏感属性表与表 11 合并后的敏感属性表进行有损连接,见表 12。

根据表 12 可以看出,将合并后的非敏感属性表 NSS 与合并后的敏感属性表 SS 进行有损分解后连接,无法重构原始数据表,重构数据表大于原始数据表,原始数据隐藏于重构数据表中,数据挖掘者无法定位个体,保护了个体的隐私,避免个体隐私泄露。

表 12 合并后的非敏感属性表与合并后的敏感属性表有损连接

GroupID	Sex	Age	Zip	Disease
1	M	21	10095	Flu
1	M	21	10095	Respiratory infection
1	M	21	10095	Gastritis
1	M	21	10095	Cancer
1	F	23	10095	Flu
1	F	23	10095	Respiratory infection
1	F	23	10095	Gastritis
1	F	23	10095	Cancer
1	M	34	10086	Flu
1	M	34	10086	Respiratory infection
1	M	34	10086	Gastritis
1	M	34	10086	Cancer
1	F	32	10088	Flu
1	F	32	10088	Respiratory infection
1	F	32	10088	Gastritis
1	F	32	10088	Cancer
2	F	45	10087	Gastritis
2	F	45	10087	Dyspepsia
2	M	45	10078	Gastritis
2	M	45	10078	Dyspepsia
2	F	43	10078	Gastritis
2	F	43	10078	Dyspepsia

而($\alpha_{[s]}$,ℓ)-多样 k-匿名模型重建的数据量将是相当大的,在表 6 中对{Sex}属性、{Age}属性、{Zip}属性和{Disease}属性均进行了泛化操作,数据缺损严重,如果对其泛化,则:

{Sex}属性:可以细化为 M、F;

{Age}属性:范围[20,35]可以细化为 20,21,22,23,24,25;

{Zip}属性:100 * * 可以细化为 10000~10099,共 100 个细化值;

{Disease}属性:根据表 12,属性值 Respiratory infection 可以细化为其子节点的值,即 Flu、Pneumonia、Bronchitis 等等。

综上可知($\alpha_{[s]}$,ℓ)-多样 k-匿名模型重构原始表是可行的,但数据量巨大,因此($\alpha_{[s]}$,k)-匿名有损分解模型的重构概率要低于($\alpha_{[s]}$,ℓ)-多样 k-匿名模型的重构概率。

($\alpha_{[s]}$,k)-匿名有损分解模型的安全性与($\alpha_{[s]}$,ℓ)-多样 k-匿名模型类似。首先,可以通过合并后的非敏感属性表 NSS 与合并后的敏感属性表 SS 推演出($\alpha_{[s]}$,ℓ)-多样 k-匿名模型数据表,即通过表 10 和表 11 共同推演出表 6。表 11 的等价类 1 包含 4 个元组,对其准标识属性{Sex,Age,Zip}泛化后可以得到元组{ * ,[20,35],100 * * },与表 6 中等价类 1 的元组准标识属性{Sex,Age,Zip}值一致。

其次,表 11 与表 8 连接后即得到 4 个元组{ * ,[20,35],100 * * ,Flu},{ * ,[20,35],100 * * ,Respiratory infection },{ * ,[20,35],100 * * ,Gastritis },{ * ,[20,35],100 * * ,Cancer },这 4 个元组与表 6 的第 1 个等价类完全一致。

综上,($\alpha_{[s]}$,k)-匿名有损分解模型的安全性与($\alpha_{[s]}$,ℓ)-多样 k-匿名模型类似,($\alpha_{[s]}$,k)-匿名有损分解模型的重构概率要低于($\alpha_{[s]}$,ℓ)-多样 k-匿名模型,最重要的是($\alpha_{[s]}$,k)-匿名有损分解模型的数据缺失度要远远小于($\alpha_{[s]}$,ℓ)-多样 k-匿名模型。生成($\alpha_{[s]}$,k)-匿名有损分解模型的算法如下。

输入:原始数据集 D,QI,$\alpha_{[s]}$,k,ℓ。

输出:符合($\alpha_{[s]}$,k)-匿名有损分解模型的匿名数据表(敏感属性表 SS,非敏感属性表 NSS)。

过程:

1)产生符合 k-匿名,ℓ-多样的数据集 D′;

2)生成临时数据集 D*

```
while(i<n)
  while(freq(Ei,s)> α[s])
    任取等价类 Ei 中的一个元组,用 parent(s)替代 s
  end while
  i=i+1
end while
```

3)得到 GroupID;

4)D* 在 GroupID 和敏感属性上投影得到敏感属性表 SS;

5)合并敏感属性表 SS 中相同的元组;

6)D* 在 GroupID 和准标识属性集上投影得到非敏感属性表 NSS;

7)根据合并后的敏感属性表 SS 修改 GroupID,合并非敏感属性表 NSS 中相同的元组;

8)返回合并后的非敏感属性表和合并后的敏感属性表。

4 实验结果与分析

本文基于 Weka 平台封装了($\alpha_{[s]}$,k)-匿名有损分解模型算法,对 UCI 机器学习数据集中的 adult 数据集进行了($\alpha_{[s]}$,k)-匿名有损分解模型算法和 k-匿名模型算法的对比实验。

Adult 数据集来源于 http://kdd.ics.uci.edu,包含 32000 余个美国人口普查信息,首先删除存在缺失数据的记录,然后随机选择 10000 条记录,本实验选择属性集{age,work class,education,gender,country}作为准标识属性,为 adult 数据集中添加一个敏感属性字段 illness。字段 illness 的取值是{flu,pneumonia,bronchitis,gastriculcer,dyspepsia,gastritis}。

表 13 实验数据基本信息

	No.	Attribute	Distinct Values	Generalizations	Height
QI	1	age	79	<20,<40,<60,≥60	4
	2	work class	7	分类树	3
	3	education	16	分类树	4
	4	gender	2	*	1
	5	country	40	分类树	3

图 1 显示了 $\alpha_{[s]}$、QI 取值相同的条件下,k 值的取值对相对正确率的影响。随着 k 值的增大,两种模型的相对正确率均有下降,这是因为随着准标识属性集合中元组数量的增加,

导致泛化度的增加。$(\alpha_{[s]},k)$-匿名有损分解模型的相对聚类查询正确率高于k-匿名模型。

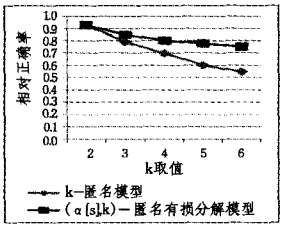

图 1　k 值对两种匿名模型相对正确率的影响图

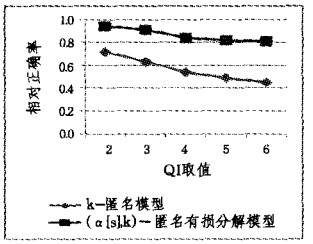

图 2　QI 值对两种匿名模型相对正确率的影响

图 2 显示了在 $\alpha_{[s]}$、k 取值相同的条件下，QI 个数对相对正确率的影响。随着准标识属性 QI 个数的增多，泛化度增加，$(\alpha_{[s]},k)$-匿名有损分解模型和 k-匿名模型的相对查询正确率均有下降，因为$(\alpha_{[s]},k)$-匿名有损分解模型通过敏感属性表 SS 和非敏感属性表 NSS 的有损分解连接的方式发布数据，两个表中尽可能多地保留了原始数据的取值，泛化度低，因此$(\alpha_{[s]},k)$-匿名有损分解模型的相对查询正确率高于 k-匿名模型。

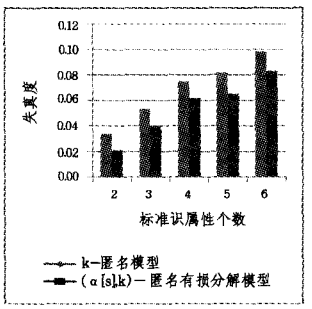

图 3　两种匿名模型的失真度比较

$(\alpha_{[s]},k)$-匿名有损分解模型算法中采用的是 Top Down 算法，此算法中若父节点符合 k-匿名，则在子分支中无需细化，所以图 4 显示 k 值和 QI 不变时，随 $\alpha_{[s]}$ 值的逐步增大，运行时间逐步下降。当 $\alpha_{[s]}$ 达到一定程度时，$\alpha_{[s]}$ 的约束没有意义，运行时间主要受 k 值约束。同理，随着 k 值的逐步增大，生成的等价类减少，需要泛化的数据就越少，因此运行时间逐步降低。

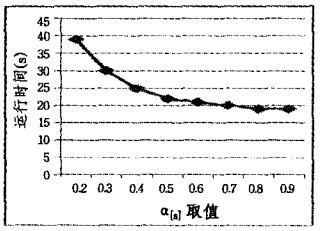

图 4　$\alpha_{[s]}$取值与运行时间的关系

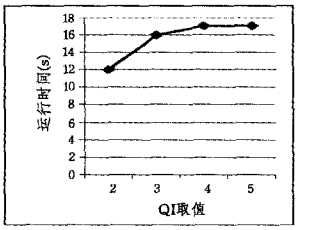

图 5　QI 取值与运行时间的关系

图 5 显示，当 $\alpha_{[s]}$ 和 k 值不变时，随着 QI 的增加，运行时间变化不大，因为无论 QI 如何变化，对应的都是记录集中的一条记录，仅仅是增加了一条记录种多个 QI 泛化的时间，而这个时间是相当少的。

$(\alpha_{[s]},k)$-匿名有损分解模型算法中随着 $\alpha_{[s]}$ 值的逐步增大，信息损失度逐步减少。因为随着 $\alpha_{[s]}$ 值的逐步增大，约束越少，需要泛化的数据就越少，因此信息损失度逐步减少。随着 k 值的逐步增大，信息损失度逐步增大。随着 k 值、QI 的逐步增大，要求每个等价类中的元组数增加，需要泛化的数据就越多，因此信息损失度逐步增大。

结束语　随着数据发布中个性化要求的提高，本文分析了现有匿名模型算法，发现各种 k-匿名模型算法的问题和缺陷。针对 k-匿名模型中敏感属性的个性化设置，利用有损分解思想，提出$(\alpha_{[s]},k)$-匿名有损分解模型，该模型针对不同等价类的个性化 $\alpha_{[s]}$ 的约束，将数据分解成敏感信息表和非敏感信息表，利用有损连接生成的冗余信息实现隐私保护。实验对比显示，其具有更好的隐私保护能力和灵活性。

参 考 文 献

[1] Sweeney L. Achieving K-Anonymity Privacy Protection Using Generalization and Suppression [J]. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 571-588

[2] Sweeney L. K-anonymity: a model for protecting privacy [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570

[3] Wong R, LI J, et al. (α,k)-anonymity: an enhanced k-anonymity model for privacy-preserving data publishing[C]// Proc of the 12th ACM SIGMOD Int'l Conf. New York, 2006: 754-759

[4] Truta, Vinay B. Privacy protection: p-sensitive k-anonymity property[C]// Proc of 22nd IEEE Int'l Conf. on Data Engineering Workshops. Washington DC: IEEE computer Society, 2006: 94-103

[5] Nergiz M E, Cliftion C. MultiRelational k-Anonymity [C]// Proc of the IEEE 23rd Int'l Conf. 2007: 1417-1421

[6] Machanavajjhala A, Kifer D, Gehrke J, et al. ℓ-diversity: Privacy beyond k-anonymity[C]// Proc of the 22nd Int'l Conf. on Data Engineering. New York: ACM, 2006: 24-35

[7] Li Ning-hui, Li Tian-cheng. T-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity [C]// Proc of 23rd Int'l Conf. on Data Engineering. 2007: 106-115

[8] XiaoXiao-kui, Tao Yu-fei. Personalized privacy preservation [C]// Proc of ACM SIGMOD Conf. on Management of Data. Chicago USA, 2006: 229-240

(下转第 383 页)

计算环境、资源和网络访问的安全和控制。

信道加密：采用密码算法实现移动终端到安全隔离区端到端的通信加密，保证内网信息在传输过程中的机密性和完整性。加密信道建立在通信运营商提供的 APN 专线之上。

认证接入：实现移动终端和安全隔离区接入设备之间的双向身份认证，保证持有合法身份证书的移动终端才能接入安全隔离区。

访问控制：保证内网信息资源只能被授权的终端访问，并对异常的访问进行阻断。

网闸隔离：实现外网和内网之间的网络隔离，对出入内网的数据进行协议剥离和内容过滤。

网络隔离采用专用通信硬件、专有安全协议、加密验证机制及应用层数据提取和鉴别认证技术进行不同安全级别网络之间的数据交换[5]，彻底阻断了网络间的直接 TCP/IP 连接，同时对网间通信的双方、内容、过程施以严格的身份认证、内容过滤、安全审计等多种安全防护机制，从而保证了网间数据交换的安全、可控，杜绝了由于操作系统和网络协议自身漏洞带来的安全风险。

4.2 3G 移动办公安全解决方案

根据 3G 移动办公安全体系架构，基于网络隔离技术组建一个端到端、安全可靠的移动办公解决方案[6]，将 3G 移动办公网络分成 3 个不同的区域：外网（包括移动终端、移动通信网）、安全隔离区和内网；网络部署见图 3。

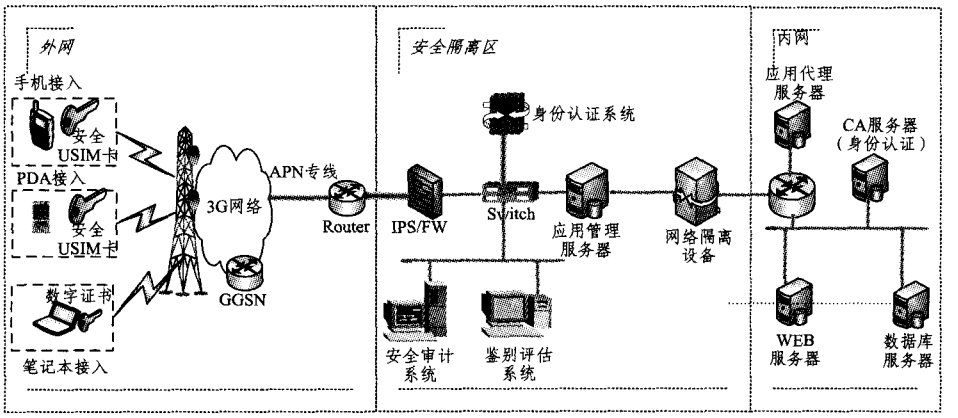

图 3 移动办公网络组网方案

全隔离区配置了入侵监测（IPS）、安全审计、身份认证、鉴别评估等边界安全防护措施，在移动终端接入内网之前进行安全防护；网络隔离设备从物理链路上断开内网与外网之间不可信任的直接网络连接。安全隔离区充分解决了移动办公的安全问题，在为合法访问提供方便的同时，还能防止内网信息资源被非法窃取。

结束语 随着 3G 移动办公在各行业的广泛应用，3G 移动办公的安全问题迫切需要系统的解决方案。本文引入成熟先进的网络隔离技术，建立了一个全方位、多层次的安全服务体系，提出了 3G 移动办公安全解决方案，解决了公网传输、内网保护等方面的安全需求。实践证明，这个安全解决方案不但能够满足内网、外网安全及物理隔离的要求，还能满足内外网信息实时传输的要求，为建设面向服务、安全高效的 3G 移动办公业务提供了有力的保障。

但是，由于在网络部署中增加了隔离网闸，导致组网节点增加，增大了网络时延，同时隔离网闸对所有的交换数据进行全方位、细颗粒的内容过滤，对系统资源有一定的影响，因此网络隔离技术在安全解决 3G 无线接入的同时，对网络的传输效率有一定的影响，这还有待进一步改进和完善。

参 考 文 献

[1] 刘道群，孙庆和. 信息敏感行业 3G 移动办公安全解决方案[J]. 电信科学，2011(S1)

[2] 王珺，李立新，李福林. 物理隔离和网闸的技术原理浅析[J]. 微计算机信息，2007，8(3)

[3] 邓霄博，杜勇，朱伟光. 基于 3G 网络的企业数据通信安全方案[J]. 电信科学，2010(8)

[4] 张红江. 网闸技术在社会保障信息系统种的应用[J]. 电子工程师，2007(11)

[5] 陈强，付强，张勇. 浅谈网络隔离技术[J]. 北方交通，2010(4)：195-197

[6] 张羽，冯朝辉. 安全网闸在公安信息化工作中的应用探讨[J]. 网络安全技术与应用，2007(05)

[7] "网络隔离"安全技术发展方向概述[EB/OL]. http://www.huacolor.com/article/737.html

（上接第 353 页）

[9] Wong R，Li J，Fu A，et al. (alpha，k)-anonymity：An enhanced k-anonymity model for privacy-preserving data publishing[C]// Proc of KDD 2006. New York：ACM，2006：754-759

[10] Xiao Xiao-kui，Tao Yu-fei. m-Invariance：Towards Privacy Preserving Re-publication of Dynamic Datasets[C]// Proc of the ACM SIGMOD Int'l Conf. on Management of Data. 2007：689-700

[11] 刘玉葆，黄志兰，傅慰慈，等. 基于有损分解的数据隐私保护方法[J]. 计算机研究与发展，2009，46(7)：1217-1224

[12] 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)：847-860