УДК 621.391:519.2:519.7

Ю.І.ГОРБЕНКО, О.С.ТОЦЬКИЙ, В.А.ПОНОМАР

АНАЛІЗ МЕТОДІВ ЗНЕОСОБЛЕННЯ ПЕРСОНАЛЬНИХ ДАНИХ

Вступ

Зараз, в час розвитку електронного документообігу, коли майже вся інформація зберігається та обробляється в електронному вигляді одним з найважливіших питань становиться питання захисту цієї інформації. Особливо це стосується персональних даних.

Надання послуг конфіденційності є однією з головних вимог до органів які зберігають персональні дані. Майже всі вони зберігаються в базах персональних даних і за Законом України «Про захист персональних даних» її власник повинен забезпечити безпеку персональних даних, що в ній знаходяться.

Головною загрозою для бази даних є виток інформації, що в ній зберігається, тому постає задача захисту даних від витоку, тобто забезпечення її конфіденційності. Це можливо за рахунок використання різних програмних, апаратних чи програмно-апаратних засобів.

Але іноді для баз даних різних установ виникають випадки коли потрібно передати персональні дані на обробку, не надаючи доступ до самої інформації. Одним способом вирішення цієї задачі є шифрування інформації перед тим як відіслати її на обробку. Але що робити коли потрібно відіслати базу даних на обробку без доступу до інформації, але зберігши її внутрішні зв'язки та різні статистичні властивості. В даному випадку шифрування не може використовуватися. І саме в таких випадках виконується знеособлення персональних даних.

Мета та задачі знеособлення персональних даних

При роботі з персональними даними трапляються випадки коли потрібно ці дані знеособити. Для того щоб зрозуміти принципи та методику знеособлення, треба проаналізувати мету та задачі знеособлення:

1) Метою знеособлення персональних даних, що містяться у базах даних (БД), є можливість передачі такої БД для виконання тестування із навантаженням за межами України [1].

2) Метод знеособлення персональних даних має бути погоджений із Державною службою України з питань захисту персональних даних [2].

3) Метод знеособлення персональних даних повинен надавати можливість знеособлення персональних даних для БД розміром 150 Тб.

4) Метод знеособлення персональних даних може включати можливість виконання переіндексації даних БД.

5) Метод знеособлення персональних даних повинен залишати структуру БД, розміри даних, внутрішні логічні зв'язки БД.

Методи реалізації знеособлення персональних даних

Загалом знеособлення персональних даних окремих БД виконується за рахунок підстановки замість значення у полі БД частини гами (псевдовипадкової послідовності – ПВП), що генерується за алфавітом притаманним для даного типу даних.

Гама управління в залежності від вимог може генеруватися двома методами. Основний метод відповідає ДСТУ 4145-2002 [3]. Він полягає у тому, що генератор за одне звернення до нього видає випадковий рядок довжини t=0. Як криптографічне перетворення в генераторі застосовується алгоритм криптографічного перетворення згідно з ГОСТ 28147 у режимі простої заміни [4]. Таблиця заміни і особистий ключ цього перетворення повинні відповідати ГОСТ ГОСТ 28147-89. Умови отримання й використання особистого ключа повинні унеможливлювати доступ до нього або його частини, модифікацію, підміну або знищення. Особис-

тий ключ криптографічного перетворення згідно з ГОСТ 28147, що використовується в генераторі випадкових послідовностей, не можна використовувати для іншої мети. Сам алгоритм полягає в наступному:

1) Задається початковий стан генератора – 64 біти.

2) Обчислюється $I = E_k(D)$ (де E_k – шифрування згідно ГОСТ 28147).

3) $x = {}_{Ek}(I \oplus s)$ (де $\oplus$ – сума по модулю 2 бітових рядків)

4) $s = E_k(x \oplus I)$

Необхідний для гами біт – молодший біт бітового рядка *x*. Таким чином генерується необхідна кількість біт гами підстановки.

Другий, перспективний метод, дозволяє генерувати гаму з необхідними властивостями, його відмінність в тому, що алфавіт гами може бути довільний [5]. Він зводиться полягає у тому, що у загальному випадку детермінований генератор ПВП, що функціонує згідно трьох модульного перетворення в розширенні поля Галуа $F(p^n)$ може бути реалізований на основі рекуретного обчислення згідно

$$b_i = \left((\theta_j)^i \left(mod(f(X), p, n), (f_1(X), p_1, n_1), (f_m(X), \vec{m})\right)\right), \tag{1}$$

$$b_i = \left((\theta_j)^{k_0+i} \left(mod(f(X), p, n), (f_1(X), p_1, n_1), (f_m(X), \vec{m})\right)\right). \tag{2}$$

Для знеособлення персональних даних застосовується метод генерування ПВП на основі багатомодульного перетворення елементів тільки в простому полі Галуа GF (p), тобто при n=1.

1) Ввести або генерувати загальносистемні параметри p та θ побудування елементів поля Галуа, де p – просте число, а θ – первісний (породжуючий просте поле Галуа) елемент поля, а також просте число p_1.

2) Ввести параметр генератора (ключ), k=1 ÷ p-1.

3) Обчислити початкове значення генератора a_0:

$$a_0 = \theta^k (mod\ p) = R_p(\theta^k). \tag{3}$$

4) Обчислити z елементів a_i ПВП:

$$a_i = a_{i-1}\theta (mod\ p) = R_p(a_0\theta^i)_{i=1,2,3,\ldots,z} \tag{4}$$

5) Обчислити елемент b_i послідовності {B}, що генерується:

$$b_i = a_i (mod\ p_1) = R_{p_1}(a_i) = R_{p_1}(R_p(a_i\theta^i)). \tag{5}$$

Таким чином одержуємо z блоків ПВП.

Для заміни необхідно використання гами підстановки – ПВП із необхідними властивостями необоротності, нерозрізнюваності, непередбачуваності та заданим (гарантованим) періодом повторення. При цьому статистичні властивості символів гами підстановки повинні співпадати зі статистичними властивостями символів вихідної мови, а метод генерування гами повинен опиратись на діючі в Україні стандарти криптографічного захисту інформації.

В знеособлює мій базі підстановки повинні виконуватися для таких алфавітів:

1) цифри (0-9), основа алфавіту m=10;

2) український алфавіт, заголовні букви (А-Я), основа алфавіту m=33;

3) український алфавіт, заголовні та прописні букви (А-Я, а-я), основа алфавіту m=66;

4) український алфавіт, заголовні та прописні букви, цифри (А-Я, а-я, 0-9), основа алфавіту m=76;

5) український алфавіт, заголовні букви, англійський алфавіт, заголовні букви, цифри (А-Я, A-Z, 0-9), основа алфавіту m=69;

6) англійський алфавіт, заголовні та прописні букви, символи (A-Z, a-z, \), основа алфавіту m=53;

7) англійський алфавіт, заголовні та прописні букви, цифри (A-Z, a-z, 0-9), символи (,_.,), основа алфавіту m=64.

Метод генерування ПВП c_i з алфавітом m зводиться до виконання таких кроків:

1. Якщо m =10, то генеруємо байти, тобто числа в діапазоні 0 -255 діляться на 10 інтервалів і число із відповідного інтервалу заміняється на відповідну цифру.

2. Якщо m =33 (український алфавіт), то згідно табл. 1 частот появи 33 символів української мови формуємо випадкові послідовності символів згідно таблиці і заміни випадкового байта на символ алфавіту з урахуванням частоти його появи.

Таблиця 1

Прог./0,134	о/0,082	н/0,070	а/0,070	и/0,056	т/0,051	в/0,046
е/0,043	р/0,038	і/0,037	с/0,036	к/0,036	м/0,033	д/0,028
л/0,028	у/0,027	п/0,026	я/0,021	з/0,019	ь/0,015	г/0,013
ч/0,011	б/0,010	х/0,010	ц/0,009	ю/0,009	ж/0,008	й/0,007
ї/0,006	є/0,006	ф/0,005	ш/0,005	щ/0,003	ґ/0,000	

3. Якщо m =66 (український алфавіт, заголовні та прописні букви), то згідно табл. 1 часто появи символів української мови формуємо випадкові послідовності символів згідно сформованої таблиці і заміни випадкового байта на символ алфавіту з урахуванням частоти його появи. Частота появи заголовної букви береться в залежності від даних поля БД.

4. Якщо m =76 (український алфавіт, заголовні та прописні букви, цифри), то згідно табл. 1 частот появи символів української мови формуємо випадкові послідовності для символів української мови згідно таблиці. Частота появи цифр чи заголовних букв береться в залежності від даних поля БД, а вірогідність появи відповідної цифри вважати рівно ймовірною відносно інших.

5. Якщо m =69 (український алфавіт, заголовні букви, англійський алфавіт, заголовні букви, цифри 0-9). Основа алфавіту m =69, то частоти появи 33 заголовних букв української мови визначаємо згідно табл.1, частоти появи 26 заголовних букв англійської мови згідно табл.. 2, появу цифр вважати рівно ймовірною. Частота появи букви відповідної мови чи цифри залежить від даних поля БД.

Таблиця 2

Прог./0,137	e/0.127	t/0.097	i/0.075	a/0.073	o/0.068	n/0.067
s/0.067	r/0.064	h/0.049	c/0.045	l/0.040	d/0.031	p/0.030
y/0.027	u/0.024	m/0.024	f/0.021	b/0.017	g/0.016	w/0.013
v/0.008	k/0.008	x/0.005	q/0.002	z/0.001	j/0.001	

6. Якщо m =53 (англійський алфавіт, заголовні та прописні букви, та символ '/'), то згідно з табл. 2 частот появи символів англійської мови формуємо випадкові послідовності для символів української мови згідно з таблицею. Частота появи заголовних букв береться в залежності від даних поля БД, а символ '/' в процесі знеособлення будемо залишати незмінним.

7. Якщо m =64 (англійський алфавіт, заголовні та прописні букви, цифри, та символи '_' та '.'), то згідно з табл. 2 частот появи символів англійської мови формуємо випадкові послідовності для символів української мови згідно таблиці. Частота появи цифр чи заголовних букв береться в залежності від даних поля БД, вірогідність появи відповідної цифри вважати

рівно ймовірною відносно інших, а символи '_' та '.' в процесі знеособлення будемо залишати незмінним.

Таким чином, у залежності від значення *m*, генерування послідовності заміни здійснюється у такій послідовності:

- генерується двійкова ПВП необхідної довжини;
- двійкова послідовність перетворюється в послідовність необхідного числа бітових блоків, довжина яких задається в залежності від вимог до знеособлення та властивостей способу генерування ПВП;
- у відповідності з частотами появи букв та цифр бітові блоки замінюються на необхідні букви, цифри, символи відповідних алфавітів.

Оскільки зв'язки БД виконанні через ідентифікатори, які не несуть в собі якусь інформацію про данні, то вони знеособленню не підлягають і через це зв'язки БД не руйнуються.

Висновки

На сьогодні реалізація знеособлення персональних даних є одним за важливіших питань в даній сфері. Найважливішими вимогами до знеособлення персональних даних є вимога зберігання зв'язків БД, що виконується майже завжди, оскільки ідентифікатори не несуть якоїсь інформації про данні і тому не виникає потреба в їх знеособлені. Друга вимога полягає в тому, щоб знеособлена БД мала ті ж статистичні властивості, що й її оригінал і саме цю вимогу виконати найскладніше. Для зрозуміння цієї проблеми наведемо приклад роботи методу з найкращими статистичними даним – він виробляє алфавіт не для поля, а для символу при цьому використовується лише три алфавіти: англійський, український, цифри. Заміна символу в ньому виконується на символ того ж алфавіту, крім того заголовна буква замінюється на заголовну, прописна – на прописну. Різні інші символи залишаються незмінні. Результат роботи цього методу наведено в табл. 3.

Таблиця 3

№	Тип даних	Приклад даних	Приклад результату
1	Паспортні дані		
а	ПІБ	Іванов Владислав Андрійович	Еунг оуЯмтсохваняАарезидитп
б	Серія та № паспорта	МН987430	РО259692
в	Адреса реєстрації	61172,м.Харків, вул. Зубарєва 61, кв.15	56112,ио.Рбиачо,пктт. Е в нгосл74,ову.35
г	Дата народження	21.02.1992	83.03.8334
2	Номер телефону (можливе збереження префіксу у початковому вигляді)	0983728814	8510353088
3	Імена облікових записів користувачів системи	biks_iv_2.narod.ua	n__fa84a0.x_u_r.om
4	E-mail адреса абонентів та користувачів системи	Vladlv@gmail.com	afcr_l@bninp.ocd

Як результат при знеособленні дійсної БД статистичні властивості результату максимально наближені до оригіналу. Великим мінусом цього методу є його швидкодія, оскільки перевіряється кожний символ, то при знеособленні БД об'ємом в декілька Тб її обробка може зайняти декілька місяців, коли метод який був наведений в попередніх пунктах буде працювати вдвічі швидше оскільки алфавіт вибирається не на символ, а на поле БД (а то й колонку даних).

Є ще способи підвищення швидкодії знеособлення. По-перше, при розбивці блоку гами на інтервали (щоб виконати підстановку символу згідно частоти його появи) зробити масив (таблицю) підстановки і далі не перевіряти значення блоку гами, в який інтервал він потрапляє, а використовувати його як індекс таблиці підстановки. Це збільшить швидкодію, але потребує значно більший об'єм пам'яті. що використовується при знеособленні. По-друге,

перед знеособленням провести відповідну підготовку БД. Вивести дані з спільним алфавітом в окремий розділ і забрати їх назад після знеособлення – це збільшить швидкодію бо потрібно буде вибирати алфавіт лише на початку розділу, а не при кожному переході до наступного поля. Недоліками цього способу є те, що його потрібно застосовувати в дійсно великих БД (1 Тб і більше) інакше він навпаки зменшить швидкодію, крім цього реалізація цього способу на великих даних досить складна.

Наведені методи – це методи знеособлення персональних даних шляхом шифрування його змісту. Вони залишають всі властивості БД незмінними, але вилучають інформацію, що надає змогу ідентифікувати об'єкти персональних даних, що відповідає вимогам закону України про захист персональних даних.

Список літератури: 1. *Закон України* про захист персональних даних №2297 від 01.06.2010. 2. *Закон України* "Про захист інформації в інформаційно-телекомунікаційних системах" (із змінами). 3. *ДСТУ 4145-2002* «Інформаційні технології. Криптографічний захист інформації. Цифровий підпис, що ґрунтується на еліптичних кривих. Формування та перевіряння». 4. *ДСТУ ГОСТ 28147:2009* «Системы обработки информации. Защита криптографиеская. Алгоритм криптографического преобразования». 5. *Горбенко, І.Д., Горбенко, Ю.І.* Прикладна криптологія. Теорія. Практика. Застосування : монографія. – Харків : Форт, 2012.

Харківський національний
університет радіоелектроніки

Надійшла до редколегії 12.09.2012