УДК 004.827

О.Р. Чертов

Національний технічний університет України «Київський політехнічний інститут», Київ

ЗАСТОСУВАННЯ ВЕЙВЛЕТ-ПЕРЕТВОРЕНЬ ДЛЯ ЗАБЕЗПЕЧЕННЯ ГРУПОВОЇ АНОНІМНОСТІ ДАНИХ

Надання публічного доступу навіть до знеособлених даних може призвести до розкриття конфіденційної інформації про груповий розподіл респондентів чи їх атрибутів. В роботі пропонується метод забезпечення групової анонімності даних за рахунок використання вейвлет-перетворень. Можливості методу ілюструються на практичному прикладі.

Ключові слова: *вейвлет-перетворення, групова анонімність, знеособленість інформації.*

Вступ

Питання захисту приватної інформації (privacy) від несанкціонованого доступу завжди були актуальними. І з кожним роком їх актуальність тільки підвищується. Пов'язано це не тільки зі значним зростанням обсягів цифрової інформації [1], але й з розширенням доступу до збираємих даних. Пройшовши нескладну перевірку, наразі можна отримати доступ до первинних даних про лікування пацієнтів госпіталів – так звані клінічні репозиторії (clinical data repositories) [2], до мікрофайлів з широкими вибірками даних переписів чи різних соціологічних досліджень тощо. Найбільш масштабним є проект IPUMS-International [3], в межах якого зібрано та відкрито для доступу дослідникам 279 млн. персональних записів по 130 переписам 44 країн. Звичайно, вся ця інформація надається в знеособленому вигляді і для збереження її конфіденційності застосовуються спеціальні методи анонімізації даних, що регулюється відповідним законодавством.

Отже, загальний ріст обсягів доступної інформації, розширення переліку різноманітних (часто перехресних) джерел, де її можна отримати, розвиток методів інтелектуального аналізу даних (data mining), направлених на пошук неявних взаємозв'язків даних, та необхідність дотримання відповідних регулюючих норм створюють перед власниками даних все нові виклики, які вони повинні розв'язати до оприлюднення інформації.

Роботи з забезпечення конфіденційності даних, що знаходяться у вільному доступі, можна віднести до напрямку, який в західній літературі називається або privacy-preserving data mining, або statistical disclosure control, якщо мова йде про захист даних в мікрофайлах.

Під *мікроданими* розуміють інформацію про окремих респондентів, наприклад, про людей, домогосподарства чи підприємства. *Мікрофайл* – це набір зібраних в один файл мікроданих, котрі містять атрибутивний запис по кожному респонденту.

Метою статистичного контролю за розкриттям інформації (statistical disclosure control) є отримання із вихідного мікрофайлу такого нового (захищеного) мікрофайлу, щоб при цьому [4, с. 339]:

– ризик розголошення конфіденційної інформації був мінімальним чи, принаймні, адекватним важливості інформації, що захищається,

– аналіз даних про респондентів в обох мікрофайлах давав однакові чи подібні результати,

– вартість перетворення мікроданих була прийнятною.

Зазначені вимоги еквівалентні твердженню про те, що втрата інформативності внаслідок застосування операцій із захисту даних, повинна бути допустимою при проведенні аналізу спотворених даних. Іншими словами, викривлення мікроданих повинно бути настільки малим, щоб гарантувати збереження корисності отриманих даних, але одночасно воно повинно забезпечувати конфіденційність приватної інформації про окремих респондентів чи їх групи.

Можливі цілі зазначеного вище аналізу інформації в даній роботі не розглядаються.

Аналіз останніх досліджень та публікацій. Індивідуальна та групова анонімність. Серед усіх методів статистичного контролю за розкриттям інформації можна виділити наступні, які широко застосовуються на практиці:

– *рандомізація* (до даних додається шум, що маскує значення атрибутів записів мікрофайлу) [5];

– *мікроагрегація* – набір вихідних записів розбивається на декілька груп таким чином, щоб записи одного і того ж набору були близькими одне одному і щоб їх було не менше певної заздалегідь визначеної кількості; потім значення кожного атрибуту в записі будь-якого набору замінюється на середнє по відповідному набору [6];

– *обмін даними* (мікрофайл змінюється шляхом обміну значень виділених атрибутів персональних записів) [7];

– *несхвилюючі (non-perturbative) методи* – вони захищають дані, не змінюючи їх, наприклад, за рахунок приховування (вилучення деяких записів із

мікрофайлу) чи розширення діапазону значень (від дат до років тощо) [8].

В останні 3 – 4 роки для статистичного контролю за розкриттям інформації почали використовувати матричну факторизацію [9] та розкладення по сингулярним числам [10]. Перспективним представляється також застосування вейвлет-перетворення (ВП). Можна використати дискретне ВП для виділення в початкових даних апроксимаційних та деталізуючих коефіцієнтів, а потім замаскувати дані шляхом порогового відсічення високочастотних деталізуючих коефіцієнтів [11].

Але всі перераховані методи направлені, насамперед, на забезпечення анонімності даних про окремих респондентів. В той же час проблема забезпечення анонімності даних про групи респондентів залишається відкритою [4]. Шляхом групової анонімності ми захищаємо важливі розподіли даних чи їх колективні особливості, що не можуть бути виявлені на рівні персональних записів. Забезпечення групової анонімності означає проведення такої перебудови початкових даних, яке буде гарантувати збереження конфіденційності розподілу окремих атрибутних значень по певній групі респондентів. Наприклад, ми можемо захистити територіальний розподіл військових чи деталі поширення наркотиків серед етнічних груп населення.

Тільки на перший погляд такі задачі можуть бути легко розв'язані: наприклад, обміняли округи роботи військовослужбовців з військової бази з округами респондентів зі схожими даними й, тим самим, засекретили місце розташування відповідної військової бази. Проблема полягає в тому, що такий перерозподіл може негативно вплинути на загальну корисність отриманих в результаті захисту даних. Загалом, застосовуючи методи групової анонімності, ми свідомо змінюємо розподіл респондентів відносно значень певного атрибуту. Але при цьому обов'язково потрібно зберегти корисність даних, яка полягатиме, зокрема, в величині пропорційності певних частин (груп) респондентів або відповідних відносних даних. Так, територіальний розподіл військових (чи співробітників спецслужб) є строго конфіденційним. В той же час інформація про їх розподіл за віком чи, скажімо, за кількістю членів родини може становити великий інтерес для соціологів.

Метою даної роботи є розробка методу групової анонімності даних в мікрофайлах. Пропонується застосовувати ВП для пошуку балансу між зміною початкових даних та запобіганням втрати їх корисності. В певному розумінні, це застосовування є протилежним (та взаємодоповнюючим) до того, як ВП використовується для забезпечення індивідуальної анонімності в [11]. Для захисту даних ми будемо перерозподіляти апроксимаційні значення, при цьому зберігаючи середнє початкових даних. Окрім того, ми будемо запобігати втраті корисності даних залишаючи незмінними (чи лише пропорційно змінюючи) їх деталізуючі складові. В цьому випадку пропорційність співвідношень між різними діапазонами атрибутних значень буде збережена. Так, дослідження результатів 44 опитувань суспільної думки громадян Російської Федерації за 1994-2001 рр. показало [12], що саме деталізуючі складові вейвлет-розкладення відображають приховані особливості часових рядів, які можна використовувати для коротко- та середньотермінових прогнозувань соціальних процесів.

Метод групової анонімності

Основні визначення та позначення. Нехай дані мікрофайлу організовані так, як це представлено в табл. 1. При цьому через m позначимо кількість респондентів в мікрофайлі, а через q – кількість атрибутів кожного респондента. Змінна w_j відповідає за j-й атрибут, r_i – за і-й запис, а z_{ij} – за відповідний елемент мікрофайлу.

Таблиця 1

Структура мікрофайлу

	атрибути				
респонденти		w_1	w_2	…	w_q
	r_1	z_{11}	z_{12}	…	z_{1q}
	r_2	z_{21}	z_{22}	…	z_{2q}
	…	…	…	…	…
	r_m	z_{m1}	z_{m2}	…	z_{mq}

В подальшому ми будемо перерозподіляти деякі значення z_{ij} згідно до поставленої мети. Введемо відповідні поняття.

Через S_v позначимо підмножину декартового добутку $w_{v_1} \times w_{v_2} \times ... \times w_{v_u}$ колонок табл. 1, де змінна v_i, $i = \overline{1,u}$ приймає натуральні значення. *Істотностною комбінацією значень* назвемо елемент $s_k^{(v)} \in S_v$, $k = \overline{1,u_v}, u_v \le u$, оскільки така комбінація є суттєвою для розв'язуваної нами задачі. Відповідно, кожен елемент вектору $s_k^{(v)}$ будемо називати *істотностним значенням*, а w_{v_j} – *істотностним атрибутом*.

Групова анонімність може бути забезпечена шляхом перерозподілу записів з виділеними істотностними комбінаціями значень. Наприклад, при перерозподілі записів, що відповідають жінкам середнього віку в якості істотностних атрибутів потрібно взяти "вік" та "стать".

Також через S_p позначимо підмножину елементів z_{ip} мікрофайлу, що відповідають p-тому атрибуту, причому $p \ne v_i \ \forall i = \overline{1,u}$. Елементи вектору $s_k^{(p)} \in S_p$, $k = \overline{1,u_p}$, $u_p \le u$ будемо називати *пара-*

метризуючими значеннями, а p-й атрибут – *параметризуючим атрибутом*, бо всі вони використовуються для розділення записів мікрофайлу на категорії.

Наприклад, взявши "адміністративно-територіальний об'єкт" як параметризуючий атрибут, ми отримаємо групи респондентів, сформовані за місцем проживання останніх. Таким чином, забезпечення групової анонімності фактично означає перерозподіл записів з істотностними комбінаціями значень відносно різних значень параметризуючого атрибуту.

Наступним кроком після визначення атрибутів є підрахунок кількості записів мікрофайлу з певною парою істотностної комбінації значень та параметризуючого значення. В багатьох випадках абсолютні значення не надають суттєвої інформації про розподіл даних і не є репрезентативними. Тому їх безпосередня модифікація може забезпечити збереження конфіденційності даних, але не гарантує запобігання втрати їх корисності.

В зв'язку з цим нам потрібно розділити абсолютну кількість записів з задовольняючими значеннями атрибутів на загальну кількість записів у цій же групі респондентів. Наприклад, якщо ми хочемо захистити розподіл жінок середнього віку, то ми можемо розділити їх кількість на загальну кількість жінок, осіб одного віку чи, навіть, людей взагалі. Отримані відносні кількості будемо збирати в масив $c = (c_1, c_2, \ldots, c_n)$, який назвемо *концентраційним сигналом*.

Для вирішення задачі забезпечення групової анонімності нам потрібно побудувати новий концентраційний сигнал $\tilde{c} = (\tilde{c}_1, \tilde{c}_2, ..., \tilde{c}_n)$. А для одночасного збереження корисності даних необхідно забезпечити незмінність середнього значення концентраційних сигналів, тобто $\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \tilde{c}_i$, та незмінність (або пропорційну зміну) деталізуючих складових концентраційного сигналу.

Модифікація апроксимаційної складової сигналу та фіксація деталізуючої. Наведемо необхідні для подальшого викладення відомості з теорії ВП. Більш детальну інформацію про ВП та їх класичні сфери застосування можна почерпнути, наприклад, з [13].

Позначимо через $s = (s_1, s_2, \ldots, s_n)$ значення дискретного сигналу, а через $h = (h_1, h_2, \ldots, h_x)$ та $g = (g_1, g_2, \ldots, g_x)$ – високочастотний та низькочастотний вейвлетні фільтри відповідно.

Якщо згортку позначити через *, а диадичне проріджування (downsampling) через $\downarrow_{2n}$, то вейвлет-розкладення першого рівня запишеться як:

$$a_1 = s *\downarrow_{2n} g;\ d_1 = s *\downarrow_{2n} h, \qquad (1)$$

де a_1 – це масив апроксимаційних коефіцієнтів першого рівня, а d_1 – відповідно масив деталізуючих коефіцієнтів цього ж рівня.

Можна застосувати формулу (1) до масиву a_1. В результаті, отримаємо апроксимуючі та деталізуючі коефіцієнти другого рівня. Загалом, застосовуючи формулу (1) до апроксимуючих коефіцієнтів рівня k–1, будемо мати відповідні коефіцієнти k-го рівня:

$$a_k = a_{k-1} *\downarrow_{2n} g = ((s *\downarrow_{2n} g)_{\text{k раз}} \ldots *\downarrow_{2n} g);$$

$$d_k = a_{k-1} *\downarrow_{2n} h = ((s *\downarrow_{2n} h)_{\text{k раз}} \ldots *\downarrow_{2n} h).$$

При цьому початковий сигнал s завжди можна представити як наступну суму:

$$s = A_k + \sum_{u=1}^{k} D_u, \qquad (2)$$

де A_k – це апроксимаційна складова сигналу на k-му рівні вейвлет-розкладення, а D_u – деталізуюча складова u-го рівня відповідно.

Складові формули (2) можна підрахувати за допомогою коефіцієнтів ВП:

$$A_k = ((a_k *\uparrow_{2n} g)_{\text{k раз}} \ldots *\uparrow_{2n} g); \qquad (3)$$

$$D_k = (((d_k *\uparrow_{2n} h) *\uparrow_{2n} g)_{\text{(k-l) раз}} \ldots *\uparrow_{2n} g). \qquad (4)$$

В формулах (3) і (4) коефіцієнти a_k і d_k спочатку диадично інтерполюються (upsampling), тобто вдвічі розширюються шляхом додавання нулів на парних місцях, що позначено символом $\uparrow_{2n}$, а потім до них застосовується згортка з відповідним вейвлетним фільтром.

Згідно формули (3) модифікувати апроксимаційну складову сигналу можна лише за рахунок зміни апроксимаційних коефіцієнтів. Але без додаткової інформації неможливо спрогнозувати – яким саме чином вплине ця зміна на апроксимаційну складову.

В якості джерела необхідної додаткової інформації будемо використовувати матриці вейвлетного відновлення (wavelet reconstruction matrix) M_{rec}, які, як відомо [14], дозволяють переписати рівність (3) наступним чином:

$$A_k = M_{rec} \cdot a_k. \qquad (5)$$

Матриця вейвлетного відновлення будується шляхом послідовного добутку відповідних матриць інтерполяції та згортки. Її структура дозволяє досить просто підібрати нові апроксимаційні коефіцієнти $\tilde{a}_k$. Після цього за формулою (5) знаходимо нову апроксимаційну складову $\tilde{A}_k$. Все що залишається – це додати до неї деталізуючі складові початкового концентраційного сигналу c, отримавши новий сигнал $\tilde{c}$.

Інколи, сигнал $\tilde{c}$ може мати від'ємні значення.

Але оскільки концентрації не можуть бути від'ємними, то потрібно модифікувати сигнал $\tilde{c}$, додавши до всіх його значень якесь прийнятне додатне число. Проте в результаті цієї операції обов'язково зміниться середнє початкового сигналу. Для його відновлення потрібно сам сигнал, що згідно формули (2), є сумою апроксимаційної та деталізуючих складових, помножити на відповідний коефіцієнт. Отже, деталізуючі складові сигналу зміняться пропорційно. Це гарантує нам остаточне рішення поставленої задачі забезпечення групової анонімності.

Приклад застосування методу групової анонімності

Для демонстрації запропонованого методу та особливостей його практичного застосування візьмемо з [3] дані мікрофайлу перепису населення Великої Британії за 2001 р. Цей мікрофайл надає різноманітну інформацію про більш ніж 1,8 млн. респондентів. Поставимо за мету захист територіального розподілу науковців, виходячи з тих міркувань, що їх значна концентрація в місцях, де відсутні відомі науково-дослідницькі чи педагогічні центри, буде свідчити про наявність закритих наукових лабораторій. Шляхом формування в концентраційному сигналі фіктивних (неіснуючих) максимумів, ми забезпечимо приховування справжнього місцерозташування таких об'єктів.

Для цього візьмемо атрибут "REGNUK", що відповідає регіонам Великої Британії, в якості параметризуючого атрибуту. Цей атрибут належить до категоріального типу. Кожне значення атрибуту відповідає одному із 16 існуючих регіонів. Але, оскільки, в зазначеному мікрофайлі були відсутні дані по регіонах "North", "East Anglia" і "Rest of South East", то фактично маємо лише 13 параметризуючих значень.

В якості істотностного атрибута візьмемо атрибут "OCC", котрий означає "occupation", тобто "вид діяльності". Цей атрибут також має категоріальний тип, причому науковцям в ньому відповідають лише два значення: "211" для професійних вчених та "311" для "наукових та інженерно-технічних фахівців". Ці значення візьмемо в якості істотностних.

Далі, підраховуємо абсолютні кількості істотностних значень по регіонам (див. четвертий рядок Таблиці 2) та будуємо концентраційний сигнал (п'ятий рядок тієї ж таблиці), розділивши підраховані кількості на загальне число працюючих у відповідному регіоні. Дані будемо наводити з точністю до трьох знаків після коми, бо цього достатньо, щоб продемонструвати отримані результати.

Максимум отриманого концентраційного сигналу досягається в передостанньому регіоні. Замаскуємо його, створивши ще декілька фіктивних максимумів концентраційного сигналу.

Оскільки сигнал має непарну кількість елементів, то симетрично його розширимо на одне значення ліворуч, щоб можна було скористатися ВП.

В якості низькочастотного фільтру візьмемо відомий фільтр Добеші другого порядку [13]:

$$g \equiv \left(\frac{1+\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{3-\sqrt{3}}{4\sqrt{2}}, \frac{1-\sqrt{3}}{4\sqrt{2}} \right).$$

Застосувавши його до початкового сигналу у відповідності з формулою (1), отримаємо концентраційний сигнал апроксимуючих коефіцієнтів першого рівня вейвлет-розкладення:

$$a_1 = (a_1(1);\ a_1(2);\ a_1(3);\ a_1(4);\ a_1(5);\ a_1(6);\ a_1(7)) =$$
$$= (0,019;\ 0,019;\ 0,018;\ 0,019;\ 0,018;\ 0,014;\ 0,022).$$

При цьому відповідна матриця вейвлетного відновлення буде мати вигляд, наведений на рис. 1.

$$M_{rec} = \begin{pmatrix}
0,837 & 0 & 0 & 0 & 0 & 0 & -0,123 \\
0,224 & 0,483 & 0 & 0 & 0 & 0 & 0 \\
-0,123 & 0,837 & 0 & 0 & 0 & 0 & 0 \\
0 & 0,224 & 0,483 & 0 & 0 & 0 & 0 \\
0 & -0,123 & 0,837 & 0 & 0 & 0 & 0 \\
0 & 0 & 0,224 & 0,483 & 0 & 0 & 0 \\
0 & 0 & -0,123 & 0,837 & 0 & 0 & 0 \\
0 & 0 & 0 & 0,224 & 0,483 & 0 & 0 \\
0 & 0 & 0 & -0,123 & 0,837 & 0 & 0 \\
0 & 0 & 0 & 0 & 0,224 & 0,483 & 0 \\
0 & 0 & 0 & 0 & -0,123 & 0,837 & 0 \\
0 & 0 & 0 & 0 & 0 & 0,224 & 0,483 \\
0 & 0 & 0 & 0 & 0 & -0,123 & 0,837 \\
0,483 & 0 & 0 & 0 & 0 & 0 & 0,224
\end{pmatrix}$$

Рис. 1. Матриця вейвлетного відновлення для прикладу, що розглядається

Таблиця 2

Розподіл даних по регіонам

Номер колонки	1	2	3	4	5	6	7
Код регіону	11	13	14	21	22	31	33
Працюючі	48591	129808	96152	83085	101891	108120	161395
Вчені (спочатку)	695	1672	1176	1163	1171	1524	2294
Сигнал с (спочатку)	0,014	0,013	0,012	0,014	0,015	0,014	0,014
Вчені (остаточно)	699	1867	1170	876	1358	1616	2495
Сигнал с (остаточно)	0,014	0,014	0,012	0,011	0,013	0,015	0,016

Номер колонки	8	9	10	11	12	13	Середні
Код регіону	40	51	52	60	70	80	-
Працюючі	97312	54861	86726	99890	55286	33409	-
Вчені (спочатку)	1246	422	871	1589	927	369	1163
Сигнал с (спочатку)	0,013	0,008	0,010	0,016	0,017	0,011	0,013
Вчені (остаточно)	1582	514	395	1182	877	480	1162,4
Сигнал с (остаточно)	0,016	0,010	0,005	0,012	0,016	0,014	0,013

Тепер згідно формули (5) можна підрахувати апроксимаційну складову сигналу с:

A_1 = (0,013; 0,013; 0,013; 0,013; 0,013; 0,013; 0,013; 0,013; 0,013; 0,011; 0,009; 0,014; 0,017; 0,014),

а згідно формули (4) – відповідну деталізуючу складову:

D_1 = (0,001; 0,001; –0,000; –0,001; 0,001; –0,002; 0,001; 0,001; 0,000; –0,003; 0,001; 0,002; –0,001; –0,003).

Враховуючи значення елементів матриці M_{rec} на різних її рядках, завжди можна підібрати такі коефіцієнти $\tilde{a}_1$, що добуток згідно формули (5) матиме максимальні значення в нових (інших) регіонах. Наприклад, якщо ми візьмемо наступні значення апроксимаційних коефіцієнтів:

$$\tilde{a}_1 = (0,019; 0,019; -2; 0; 1; -5; 0,022),$$

то отримаємо максимальні значення в 9-му, 13-му та 3-му (його значення на десятитисячні більше, ніж в 1-му та 2-му) елементах масиву апроксимаційних складових:

$\tilde{A}_1$ = (0,013; 0,013; 0,013; -0,962; -1,675; -0,448; 0,259; 0,483; 0,837; -2,191; -4,312; -1,110; 0,666; 0,014).

Додавши початкові деталізуючі складові, сформуємо новий концентраційний сигнал:

$\tilde{c} = \tilde{A}_1 + D_1$ = (0,014; 0,014; 0,013; -0,967; -1,674; -0,450; 0,260; 0,484; 0,837; -2,194; -4,311, -1,108; 0,666; 0,011).

Очевидно, що деякі значення цього сигналу є від'ємними.

Тому додамо до кожного його значення число, більше за мінімум сигналу, наприклад, число 6,311 (вибір цього числа фактично довільний):

$\hat{\tilde{c}}$ = (6,325; 6,325; 6,324; 5,344; 4,637; 5,861; 6,571; 6,795; 7,148; 4,117; 2,000; 5,203; 6,977; 6,322).

Залишилося лише забезпечити збереження середнього значення сигналу. Для цього помножимо сигнал $\tilde{c}$ на коефіцієнт

$$\sum_{i=2}^{14} c_i \Big/ \sum_{i=2}^{14} \hat{\tilde{c}}_i = 0,002 .$$

Тут, ми брали до розгляду тільки останні 13 елементів, оскільки перший елемент має штучний характер і був доданий лише для того, щоб можна було скористатися відомою процедурою вейвлет-розкладення. Результуючий концентраційний сигнал наведено в останньому рядку табл. 2.

Для переходу до абсолютного числа вчених потрібно отримані концентрації помножити на загальну кількість працюючих у кожному регіоні. Кількість вчених, звичайно, може бути лише натуральним числом, тому результати множення повинні бути округлені до цілого (див. шостий рядок табл. 2).

Хоча розраховані кількості вчених повністю відрізняються від їх початкових кількостей, проте для них обох були збережені як середні значення, так і деталізуючі складові вейвлет-розкладення.

Важливо зазначити, що округлення, яке здійснюється при переході до цілих кількостей, може привести до зміни деталізуючих складових. Але в абсолютній більшості випадків ці зміни є несуттєвими і не впливають на корисність отриманих замаскованих даних.

Після підрахунку нової кількості вчених в кожному регіоні потрібно сформувати новий мікрофайл. Це завжди можна зробити, змінюючи істотностні значення записів мікрофайлу.

Висновки

В роботі був запропонований новий метод анонімізації даних в мікрофайлах, який є фактично взаємодоповнюючим до метода, викладеного в [11]. Але на відміну від зазначеного та інших існуючих методів, нашим методом вирішується задача групової анонімності даних.

Для цього ми перерозподіляємо апроксимаційні значення вейвлет-розкладення початкового сигналу (с незмінним його середнім), в той же час або залишаючи деталізуючі складові всіх рівнів незмінними, або змінюючи їх, але пропорційно. Образно кажучи, ми змінюємо рельєф певної території, зберігаючи незмінним локальний розподіл виділених даних.

Отже, запропонований метод, як зберігає конфіденційність розподілу групи респондентів, так і гарантує збереження корисної інформації в замаскованих даних.

Новий метод породжує нові питання, розв'язання котрих дозволить більш ефективно застосовувати його на практиці. Перспективними для наступного дослідження ми вважаємо:

– розробку методів захисту розподілу не тільки абсолютних чи відносних значень, але й їх різниць;

– введення та формалізацію поняття міри забезпечення групової анонімності;

– вивчення залежності цієї міри від вибору вейвлетного базису.

Список літератури

1. *Gantz J.F. As the Economy Contracts, the Digital Universe Expands [Електронний ресурс] / J.F. Gantz, D. Reinse. – An IDC Multimedia White Paper, 2009. – Режим доступу до док. : .http://www.emc.com/collateral/ demos/microsites/idc-digital-universe/iview.htm*
2. *Data mining and clinical data repositories: Insights from a 667,000 patient data set / I. Mullins, M. Siadaty, J. Lyman et. al. // Computers in Biology and Medicine. – 2006. – Vol. 36(12). – P. 1351-1377.*
3. *Minnesota Population Center. Integrated Public Use Microdata Series International [Електронний ресурс]. – Режим доступу до док.: https://international.ipums.org/ international/.*
4. *Chertov O. Statistical Disclosure Control Methods for Microdata / O. Chertov, A. Pilipyuk // International Symposium on Computing, Communication and Control. – Singapore: IACSIT, 2009. – P. 338-342.*
5. *Agrawal R. Privacy-Preserving Data Mining / R. Agrawal, R. Srikant // ACM SIGMOD International Conference on Management of Data. – Dallas: ACM Press, 2000. – P. 439-450.*
6. *Domingo-Ferrer J. Practical Data-oriented Microaggregation for Statistical Disclosure Control / J. Domingo-Ferrer, J.M. Mateo-Sanz // IEEE Transactions on Knowledge and Data Engineering. – 2002. – 14 (1). – P. 189-201.*
7. *Fienberg S. Data Swapping: Variations on a Theme by Dalenius and Reiss / S. Fienberg, J. McIntyre // Technical Report, National Institute of Statistical Sciences. – 2003.*
8. *Domingo-Ferrer J. A Survey of Inference Control Methods for Privacy-Preserving Data Mining / J. Domingo-Ferrer // Aggarwal C.C., Yu P.S. (eds.): Privacy-Preserving Data Mining: Models and Algorithms. – New York: Springer, 2008. – P. 53-80.*
9. *Wang J. NNMF-based factorization techniques for high-accuracy privacy protection on non-negative-valued datasets / J. Wang, W. Zhong, J. Zhang // The 6th IEEE Conference on Data Mining, International Workshop on Privacy Aspects of Data Mining. – IEEE Computer Society Press. – 2006. – P. 513-517.*
10. *Singular value decomposition based data distortion strategy for privacy protection / S. Xu, J. Zhang, D. Han, J. Wang // Knowledge and Information Systems. – 2006. – Vol. 10 (3). – P. 383-397.*
11. *Liu L. Wavelet-based data perturbation for simultaneous privacy-preserving and statistics-preserving / L. Liu, J. Wang, J. Zhang // 2008 IEEE International Conference on Data Mining Workshops. – IEEE Computer Society Press. – 2008. – P. 27-35.*
12. *Давыдов А.А. Вейвлет-анализ социальных процессов / А.А. Давыдов // Социологические исследования. – 2003. – № 11. – С. 89-101.*
13. *Малла С. Вэйвлеты в обработке сигналов / С. Малла. – М.: Мир, 2005. – 671 с.*
14. *Strang G. Wavelet and Filter Banks / G. Strang, T. Nguyen. – Wellesley: Wellesley-Cambridge Press. – 1997. – 445 p.*

Надійшла до редколегії 22.03.2010

Рецензент: д-р техн. наук, проф. О.А. Молчанов, Національний технічний університет України «Київський політехнічний інститут», Київ.

ПРИМЕНЕНИЕ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЙ ДЛЯ ОБЕСПЕЧЕНИЯ ГРУППОВОЙ АНОНИМНОСТИ ДАННЫХ

О.Р. Чертов

Предоставление публичного доступа даже к обезличенным данным может привести к раскрытию конфиденциальной информации о групповом делении респондентов или их атрибутов. В работе предлагается метод обеспечения групповой анонимности данных за счет использования вейвлет-превращений. Возможности метода иллюстрируются на практическом примере.

Ключевые слова: *вейвлет-превращение, групповая анонимность, обезличенность информации.*

USING WAVELET TRANSFORMS FOR GROUP ANONYMITY PROVIDING OF DATA

Chertov O.R.

The grant of public access even to the depersonalized information can result in opening of confidential information about the group division of respondents or their attributes. The method of providing of group anonymity of the wavelet-transformations given due to the use is offered in work. Possibilities of method are illustrated on a practical example.

Keywords: *wavelet-transformation, group anonymity, depersonalized of information.*