УДК 658.012

Б. І. Мороз, доктор технічних наук, професор,
професор кафедри інформаційних систем
та технологій Академії митної служби України
Л. В. Кабак, кандидат технічних наук, доцент
кафедри інформаційних систем та технологій
Академії митної служби України

МЕТОДИ КЛАСИФІКАЦІЇ ІНФОРМАЦІЇ ДЛЯ ОРГАНІЗАЦІЇ РЕПЛІКАЦІЙ У РОЗПОДІЛЕНИХ БАЗАХ ДАНИХ ЗА ОЗНАКАМИ ЦІННОСТІ І СТАРІННЯ

Проведено дослідження та розглянуто основні напрямки розвитку теорії інформації і здійснено аналіз існуючих методів визначення характеристик цінності й старіння інформації. З урахуванням виконаного аналізу запропоновано метод класифікації інформації за допомогою визначення цінності інформації з використанням імітаційних моделей і функцій чутливості. Розроблений метод доцільно використовувати для керування транзакціями під час організації реплікацій у розподіленій базі даних.

Проведено исследование и рассмотрены основные направления развития теории информации и осуществлен анализ существующих методов определения характеристик ценности и старения информации. С учетом выполненного анализа предложен метод классификации информации с помощью определения ценности информации с использованием имитационных моделей и функций чувствительности. Разработанный метод целесообразно использовать для управления транзакциями при организации репликации в распределенной базе данных.

The researches are realized, the basic directions of development of the theory of the information are examined, and the analysis of existing methods of definition of characteristics of value and becoming obsolete information is fulfilled. According to the given analysis there was an offered method of classification of the information by means of definition of value of the information with using of imitating models and functions of the sensitivity; the developed method is necessary for the organization of transaction by means of the replication in the distributed database.

Ключові слова. Розподілені бази даних, реплікація, транзакції, ціна старіння і споживання інформації.

Вступ. Розподілена база даних (Distributed Database – DDB) митної служби України містить фрагменти з декількох баз даних, які розташовуються на різних вузлах мережі комп'ютерів і управляються різними системами керування базами даних. Розподілена база даних, з погляду користувачів і прикладних програм, – звичайна локальна база даних. У цьому змісті слово "розподілена" означає спосіб організації бази даних, але не її зовнішню характеристику (розподіленість бази даних невидима ззовні).

Розподілена база даних передбачає зберігання й виконання функцій керування даними в декількох вузлах і передачу даних між цими вузлами в процесі виконання запитів. Розбиття даних у розподіленій базі даних може досягатися шляхом зберігання різних таблиць на різних комп'ютерах або навіть зберігання різних частин і фрагментів однієї таблиці на різних комп'ютерах. Для користувача або прикладної програми не має значення, яким чином дані розподілені між комп'ютерами. Робота з розподіленою базою даних здійснюється так само, як і з централізованою, тобто розміщення бази даних має бути прозорим [1–4]. На рис. 1 наведено спрощену схему розподіленої бази даних з використання центральної бази даних (ЦБД). Саме така схема даних використовується в митній службі України.

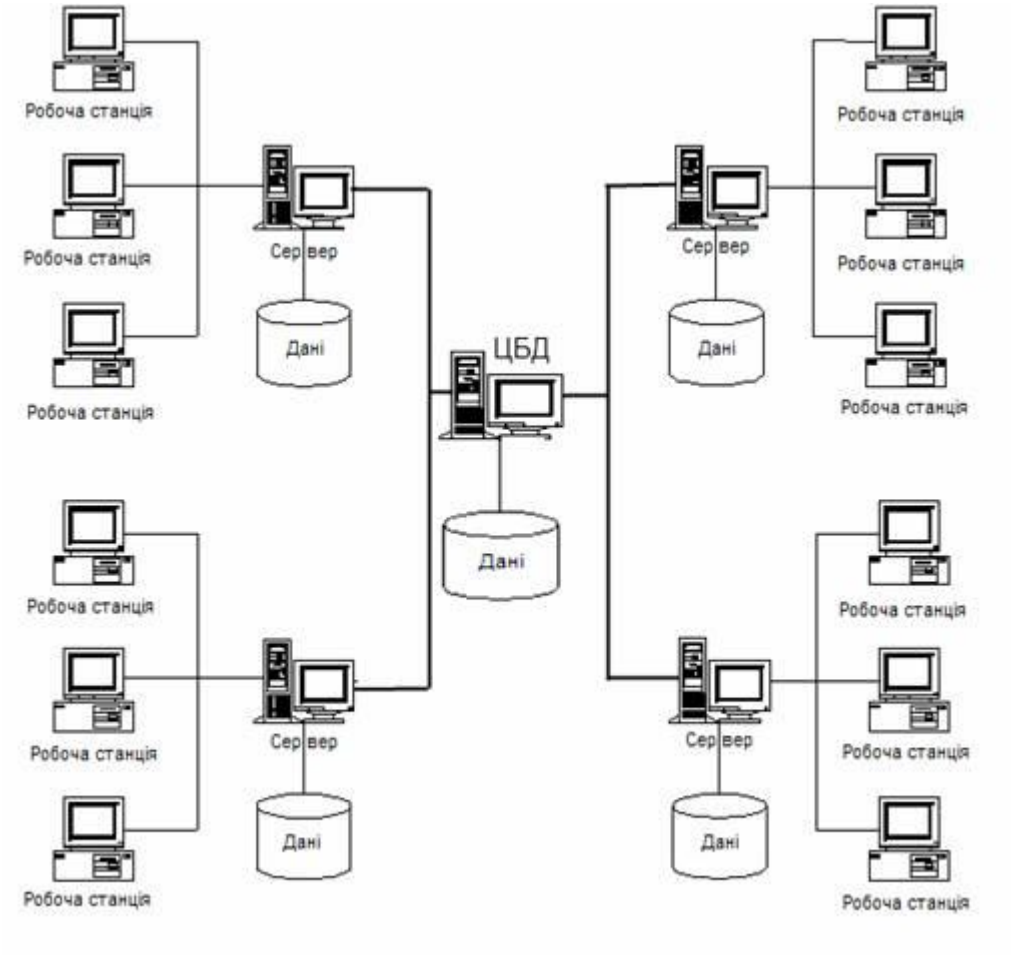

Рис. 1. Спрощена схема розподіленої бази даних

Під час розподіленої обробки робота з базою (подання даних, їхня обробка та інше) ведеться на комп'ютері клієнта, а підтримка бази в актуальному стані – на сервері. При цьому такі бази даних звичайно розташовуються на декількох серверах – різних вузлах комп'ютерної мережі, а деякі дані можуть дублюватися.

Розміщення частин загальної бази даних буває надлишковим або безнадлишковим. За надлишкового розміщення визначають ступінь дублювання частин (фрагментів) єдиної бази даних. Щоб підтримувати цілісність бази даних слід постійно коректувати всі її копії. Переваги дублювання зменшуються, коли збільшується вартість зберігання її частин, що пов'язано з необхідністю забезпечувати стійкість системи.

Створення розподілених баз даних викликано спробою одночасного виконання двох завдань: інтеграції та децентралізації.

Інтеграція має на увазі централізоване керування й ведення баз даних.

Децентралізація забезпечує зберігання даних там, де вони з'явилися й обробляються. При цьому знижується вартість системи й збільшується ступінь її надійності, а також підвищується швидкість обробки даних.

У праці [1] проведено аналіз існуючих методів реплікації інформації в розподілених базах даних та виявлено їх переваги та недоліки. Також набув подальшого розвитку метод визначення цінності і старіння інформації завдяки уведенню критерію ефективності роботи системи. Вперше запропоновано для реплікації баз даних використовувати метод, який
ураховуватиме цінність і старіння інформації під час реплікації даних з метою підвищення ефективності роботи інформаційної системи. Запропоновано впровадження розробленого методу реплікації як елементу єдиної автоматизованої інформаційної системи митної служби України.

У дослідженні [1] розглянуто існуючі системи реплікації, виявлено їхні переваги та недоліки. Було запропоновано для організації реплікації інформації використовувати систему, яка враховує критерії цінності і старіння інформації, наведену на рис. 2. Але не було наведено методів класифікації інформації за ознаками цінності старіння і споживання інфор-
мації. Тому пропонуються деякі підходи, які можуть використовуватись під час організації процесу реплікації в розподілених базах даних, а саме організації транзакцій.

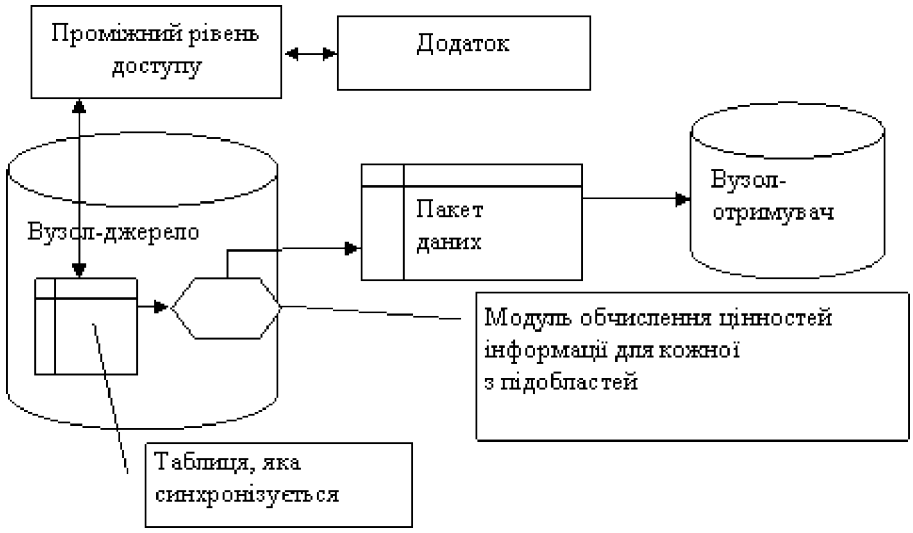

Рис. 2. Модель роботи системи з урахуванням критеріїв цінності і старіння інформації

Постановка завдання. Мета дослідження полягає в розробці методу класифікації інформації за ознаками цінності, старіння і споживання з метою організації реплікацій інформації в розподілених базах даних.

Результати дослідження. Організація обробки інформаційних потоків передбачає, що процес обробки інформації має бути доповнений двома етапами:

1) накопичування і класифікація інформації за допомогою критеріїв цінності, старіння і споживання;

2) організація процесу обробки інформації за допомогою цих же критеріїв.

Попередня оцінка характеристик розглянутих етапів обробки інформації показує, що за складністю реалізації і використання ресурсів вони майже порівнянні з етапами безпосередньої обробки інформації.

Класифікацію транзакцій за критеріями цінності, старіння і споживання пропонується проводити за такими ознаками:

α – ознака належності транзакції до деякого порогового часу старіння $T_{порог.з}$.

$$\alpha = \{\alpha_1, \alpha_2, ..., \alpha_i\}, \tag{1}$$

тут i – визначає значення $T_{порог.з_i}$;

$t_г$ – час генерації транзакції джерелом;

$$\beta = \{\beta_{вф}, \beta_{max}, \beta_{п}, \beta_{вл.}\} \tag{2}$$

група ознак, що характеризують цінність транзакції,

$\beta_{вф.}$ – ознака належності функції апроксимації зміни в часі цінності транзакції до певного виду (табл. 1),

де i – вид функцій: $\beta_{вф.} = \{\beta_{вф1}, \beta_{вф2}, ..., \beta_{вфi.}\}$;

β_{max} – ознака належності транзакції до деякого максимального значення функції апроксимації $Ц_{max}(t)$ (рис. 1),

$$\beta_{max} = \{\beta_{max1}, \beta_{max2}, ..., \beta_{maxi}\}, \tag{3}$$

тут i – визначає значення $Ц_{max}(t)$;

$\beta_{П}$ – ознака належності транзакції до деякого порогового значення часу втрати цінності $T_{порог.ц}$ (рис. 3),

$$\beta_{П} = \{\beta_{П1}, \beta_{П2}, ..., \beta_{Пi}\}, \tag{4}$$

тут i – визначає значення $T_{порог.ц}$;

$t_{max.ц}$ – час максимального значення цінності інформації;

$\beta_{вл.}$ – ознака належності транзакції до часу ефективного використання (рис. 3).

$$\beta_{e.в.} = \{\beta_{e.в.1}, \beta_{e.в.2}, ..., \beta_{e.в.i}\}, \quad (5)$$

γ – ознака належності транзакції до споживача (вузла РБД) з максимальною ефективністю використання,

$$\gamma = \{\gamma_1, \gamma_2, ..., \gamma_i\}, \quad (6)$$

тут i – споживач транзакції.

Транзакції за ознаками α, β_{max}, $\beta_{П}$, $\beta_{e.в.}$ зручно класифікувати на кількісній шкалі (рис. 4.) Функція $\bar{f}(\beta_{max})$ являє собою щільність розподілу ознаки β_{max}. Це дає можливість провести якісний аналіз стану черг транзакцій у часі методами математичної статистики в автоматизованому режимі. Для виконання автоматизованої класифікації із застосуванням функції $\bar{f}(\beta_{max})$ можна скористатися двома основними методами.

Таблиця 1

Ознака належності функції апроксимації зміни в часі цінності транзакції до певного виду

Ознака належності функції апроксимації до певного виду	Вид функції апроксимації
$\beta_{e\phi 1}$	$Ц(t)=a^t$
$\beta_{e\phi 2}$	$Ц(t)=\lg t$
…	
$\beta_{e\phi i}$	$Ц(t)=\lg t + a^t$

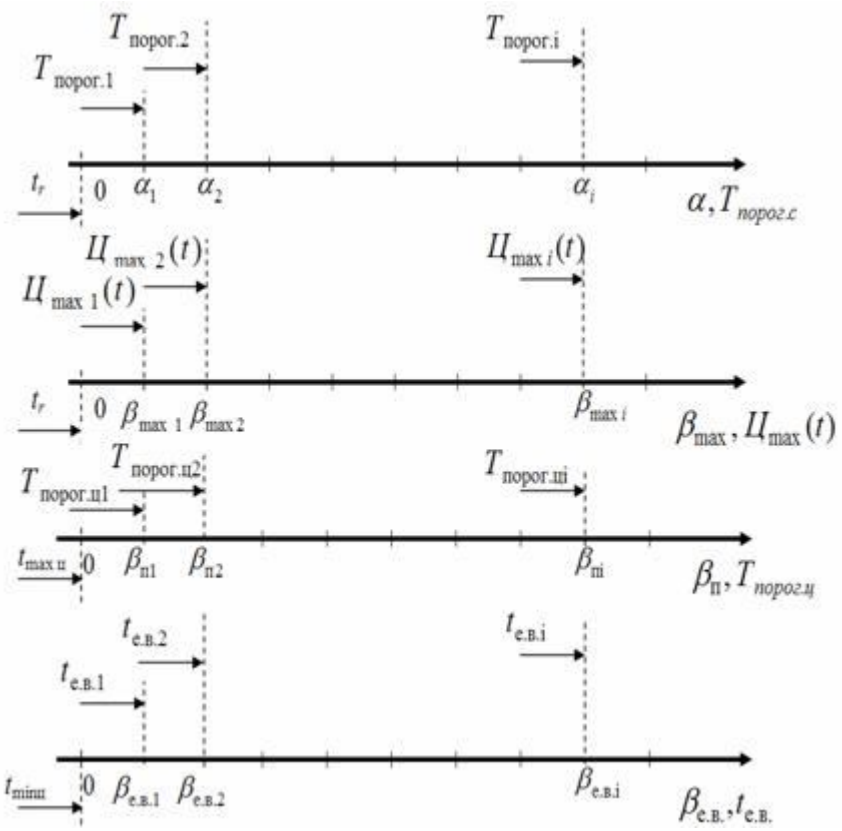

Рис. 3. Ознаки класифікації інформації

Перший метод полягає в знаходженні всіх локальних мінімумів функції $\bar{f}(\beta_{max})$. Усі транзакції, що потрапили в інтервал між i-м і (i + 1)-м локальними мінімумами, належать до класу $\beta_{max\,i}$.

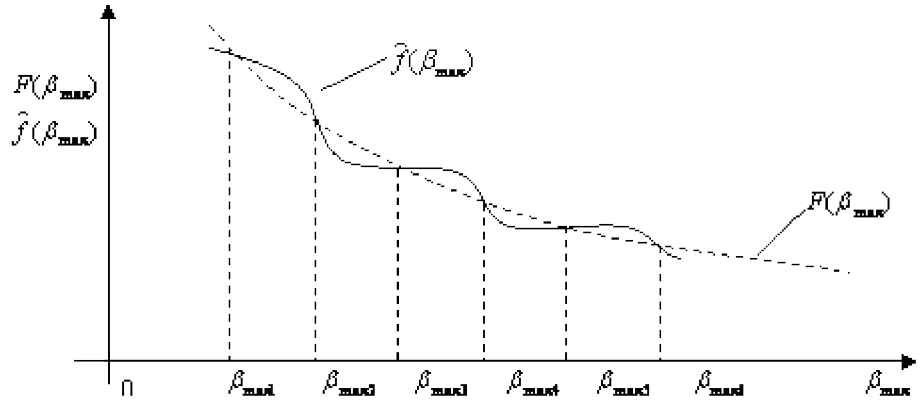

Рис. 4. Класифікація інформації за допомогою кількісної шкали

Другий метод грунтується на таких діях:
1. Визначається деяка функція

$$F(\beta_{max}) = \frac{1}{2a} \int_{\beta_{max}-\Delta}^{\beta_{max}+\Delta} \varphi(\beta_{max}) \hat{f}(\beta_{max}) d\beta_{max} \qquad (7)$$

де $\varphi(\beta_{max})$ – коефіцієнт згладжувального полінома;

2. Знаходять корені функцій

До транзакцій *i*-го класу належать транзакції, що містяться в інтервалі [β_{maxi}, β_{maxi+1}].

Під час класифікації транзакцій процес надходження потоків інформації в систему, який починається в момент часу *t,* будемо вважати випадковим, що являє собою потік однорідних та неоднорідних транзакцій, котрі надходять через випадкові проміжки часу. Ці транзакції направляються в буфер і стають у загальну чергу. Далі вони обробляються за допомогою алгоритмів класифікації з метою розподілу вмісту буфера по чергах, причому кожній черзі відповідає конкретне значення ознаки.

У результаті проведених досліджень алгоритм класифікації має такий вигляд.

1. Із загальної черги формуються черги за ознаками належності транзакції до споживача з максимальною ефективністю використання. Транзакції, що належать споживачам 1, 2, …, *i,* які характеризуються ознаками $\gamma_1, \gamma_2, ..., \gamma_i$ відповідно, розподіляються у свої черги.

2. З кожної черги з ознаками $\gamma_1, \gamma_2, ..., \gamma_i$ вибираються транзакції за ознакою належності їх деякому максимальному значенню функції апроксимації зміни цінності в часі β_{max}.

3. З кожної черги з ознаками $\beta_{max1}, \beta_{max2}, ..., \beta_{maxl}$ вибираються транзакції за ознакою належності їхнім різним значенням часу використання $\beta_{v.v1}, \beta_{v.v2}, ..., \beta_{v.v.i}$.

4. З кожної черги з різними ознаками належності транзакцій вчасного ефективного використання формуються черги з транзакцій з різними значеннями часу втрати цінності $\beta_{Ш.i}$.

5. Наступною ознакою, за якою транзакції формуються в черги, є ознака належності транзакції якому-небудь граничному часу старіння $\alpha_1, \alpha_2, ..., \alpha_i$.

Загальну методику класифікації інформації із критеріїв цінності, старіння й споживання можна подати у вигляді графа системи ієрархічної структури, вершинами якого є множина транзакцій, що надійшли, а ребрами – ознаки, за якими відбувається класифікація. Підпорядкованість ознак може бути різною залежно від цілей класифікації. Мета даної класифікації – розподіл вмісту буфера за чергами для подальшого визначення послідовності виконання завдань. Якщо класифікація за всіма ознаками, крім $\beta_{max}+\Delta$, відбувається в порядку зростання їхніх значень, а за $\beta_{max}-\Delta$ в порядку убування, то, грунтуючись на цьому, одержимо оптимальний порядок обробки транзакцій у кожній черзі. Далі можна використовувати дисципліни обслуговування потоків інформації, запропоновані вище.

Висновки. У результаті проведеного дослідження розглянуто основні напрямки розвитку теорії інформації та проаналізовано існуючі методи визначення характеристик цінності
й старіння інформації. З урахуванням виконаного аналізу набув подальшого розвитку
алгоритм класифікації інформації за допомогою визначення цінності інформації з використанням імітаційних моделей і функцій чутливості.

Література

1. Мороз Б. І. Методи та моделі реплікації інформації в розподілених базах даних з урахуванням якісно-кількісних характеристик інформації [Текст] / Б. І. Мороз, Л. В. Кабак, О. П. Буланий // Вісник АМСУ. Серія: “Технічні науки ”. – 2009. – № 2. – С. 13–24.

2. Date C. J. An Introduction to Database Systems [Text] / C. J. Date. – Reading, Mass. : Addison-Wesley, 1984. – V. 1. – 4th ed. – 639 c.

3. Performance Enhancements to a Relational Database System [Text] / M. Stonebraker, L. Woodfil, J. Ranstrom etc. // ACM Trans. Database Syst. – 1983. – № 2. – С. 189–222.

4. Tridgell A. The Rsync Algorithm. Technical Report TR-CS-96-05, Department of Computer Science [Text] / A. Tridgell, P. Mackerras / The Australian National University. – Canberra, 1996.