УДК 004

**Ярмолюк Руслан Сергійович**

Хмельницький національний університет

# ІНТЕГРАЛЬНА ОЦІНКА ЯКІСНОГО НАПОВНЕННЯ БАЗИ ДАНИХ В СИСТЕМАХ АВТОМАТИЗОВАНОГО УПРАВЛІННЯ ВНЗ

***В роботі наведено підходи до розрахунку оцінки якісного наповнення бази даних в системах автоматизованого управління університетом. За допомогою теорії множин та реляційної алгебри представлено схеми розрахунку інтегральної оцінки. Вказано шляхи застосування інтегральної оцінки наповненості при розрахунку актуальної ціни бази даних.***

***In this paper the approaches to calculating content quality assessment database systems for automated control of the university. By using set theory and relational algebra are presented calculation scheme integrated assessment. The ways of applying integrated assessment of fullness in the calculation of the actual price database.***

**Вступ.** Основним елементом будь-якої автоматизованої системи є персоніфіковані або інші дані організовані в бази даних. Від якості даних напряму залежить якість отриманих аналітичних звітів та прийнятих управлінських рішень. На даний час не існує єдиного підходу, щодо оцінювання наповненості бази даних у системах автоматизованого управління ВНЗ. Єдиним критерієм наповненості, що широко застосовується є потужність бази даних, тобто кількість записів у базі даних.

У свою чергу, при розробці автоматизованих систем управління постає проблема оцінки та визначення, тієї бази даних, на основі якої будувати систему управління. Очевидно, що критерій потужності бази даних є малоінформативним і не відображає інформацію про якісний рівень записів.

Отже, розробка методів та алгоритмів для розрахунку оцінок якісного наповнення бази даних є досить актуальною задачею.

**Постановка проблеми.** За час свого життєвого циклу у базах даних відбуваються певні інформаційно-технологічні процеси, що на пряму впливають на виникнення помилок в записах бази даних. Метод оцінювання полягає в урахуванні всіх типів помилок (тип помилки, можливість автоматичного виправлення, можливість виправлення автоматично, тощо) в значеннях атрибутів, так і ролі даного атрибуту для забезпечення основних функцій (вагові коефіцієнти кожного атрибуту).

У свою чергу поняття «якість даних» різними науковцями трактується по-різному в залежності від сфери використання. У даній роботі поняття «якість даних», відповідно до стандартів ISO для інформаційних систем [1], використовується як деякий критерій відповідності даних або інформації потребам користувача. Тоді якість даних у контексті бази даних – це критерій достовірності даних, тобто чи запис відповідає реальному об'єкту реальності.

Сучасні системи управління баз даних реалізують всі можливості реляційної моделі представлення даних [2]. Далі за допомогою теорії множин та реляційної алгебри опишемо схеми знаходження оцінок якісного наповнення бази даних.

**Оцінка якісного наповнення за типами помилок.** Отже під оцінкою якісного наповнення бази даних будемо розуміти відношення кількості записів (далі кортежів), що відповідають вимогам (не містять помилок певного виду) до загальної кількості кортежів (потужності):

$$\Omega^r = \frac{\left| A^r \right|}{\left| A \right|}$$

де $\Omega^r$ - оцінка якісного наповнення за $r$-м типом помилки ($r \in \mathbb{N}$); $\left| A^r \right|$ - кількість кортежів, що не містять помилок $r$-го типу; $\left| A \right|$ - загальна кількість кортежів. Відповідно: $A^r$ - множина усіх кортежів, що не містять помилки $r$-го типу, а $A$ - загальна множина усіх кортежів.

Вимоги приналежності кортежу до множини $A^r$ за $r$-м типом помилок задаються відповідним граничним значенням сумарної оцінки пошуку помилки у кожному окремому значенні атрибуту кортежу:

$$A^r = \left\{ A_i \in A \left| \sum_j f^r(a_{ij}) \geq g^r \right. \right\},\ i = \overrightarrow{1,\ldots,|A|} \qquad (1)$$

де $A_i$ - $i$-й кортеж; $a_{ij}$ - значення $j$-го атрибуту $i$-го кортежу; $f^r(x)$ - функція перевірки значення $x$ наявність помилки $r$-го типу; $g^r$ - граничне значення сумарної оцінки для помилки типу $r$.

Функція перевірки $f^r(x) \in \mathbb{N}$ і приймає значення на проміжку $[0;1]$:

$$f^r(x) = \begin{cases} 0, & \text{якщо значення } x \text{ містить помилку типу } r; \\ 1, & \text{якщо значення } x \text{ не містить помилку типу } r. \end{cases}$$

Граничне значення сумарної оцінки $g^r$ задає кількість гранично допустимих помилок типу $r$ у значеннях атрибутів кортежу. Очевидно, $g^r \in \mathbb{N} \quad \leq n$, де $n$ - кількість атрибутів у кортежі. При $g^r = 0 \Rightarrow A^r \equiv A$, якщо $g^r = n$, то $A^r \equiv A$, лише у випадку відсутності типу помилок $r$ у загальній множині записів бази даних.

**Інтегральна оцінка якісного наповнення.** Різні типи помилок у записах по різному впливають на можливість виконання автоматизованою системою управління своїх функцій. Відсутність певних значень атрибутів записів, синтаксичні та стилістичні помилки, наприклад, не впливають на інформаційно-аналітичні можливості системи автоматизованого управління.

Для розрахунку сумарної оцінки якісного наповнення потрібно враховувати значення впливу певного виду помилок на можливість виконання основних функцій, а також можливість автоматичного виправлення певного виду помилок без людського втручання. Тому, для кожного типу помилок задається відповідний коефіцієнт впливу $c^r \in \mathbb{R} \quad \leq 1 \quad r = \overrightarrow{1,\ldots,m}$. Відсортовані за коефіцієнтом впливу типи помилок задають схему ієрархії помилок.

Якщо задати схему ієрархії помилок, тобто $A \equiv A^0 \supseteq A^1 \supseteq ... \supseteq A^r \supseteq ... \supseteq A^m$, де $m$ - кількість типів помилок, що мають вплив на функціональність

$$A^{r+1} = \left\{ A_i \in A^r \left| \sum_j f^{r+1}(a_{ij}) \geq g^{r+1} \right. \right\}, \; i = \overrightarrow{1,...,|A^r|}; \; r = \overrightarrow{1,...,m}, \tag{2}$$

тоді інтегральна оцінка має вигляд:

$$\Omega_+^{all} = \frac{|A^m|}{|A|}. \tag{3}$$

У випадку відсутності ієрархії за типом помилок:

$$\Omega_-^{all} = \frac{\left| \bigcap \right.}{|A|}, \tag{4}$$

де $A^r$ визначається за формулою (1).

Запропоновані оцінки якісного наповнення не враховую вплив (наявність або відсутність помилок) окремо кожного атрибута на функціональність електронного каталогу бібліотеки. Для врахування даного фактору, а також ймовірнісної складової функції потрібно застосовувати теорію нечітких множин [3,4].

**Застосування оцінки якісного наповнення при економічних розрахунках.**

Університетам потрібні дані про затрати на створення ними окремих продуктів і послуг. З іншого боку для розробки кошторису створення систем автоматизованого управління та при прийняті рішення про роботу зі сторонніми організаціями (аутсорсінг) необхідно вказувати обґрунтовану цінність даних, що пропонуються у базі даних.

Актуальну ціну бази даних $S_{current}$ пропонуємо обраховувати за формулою:

$$S_{current} = \delta \cdot |A| - \sum_r \alpha_r \left( |A| - |A^r| \right), \tag{5}$$

де $\delta$ - усереднений показник витрат на створення одного запису; $\alpha_r$ - усереднений показник витрат на усунення помилки типу $r$ з запису.

У випадку відсутності інформації про усереднені показники, або при заявленій зі сторони ціни $S_{declared}$ наповненості бази даних покупець може коригувати її на відповідну інтегральну оцінку якісного наповнення, тобто:

$$S_{current} = S_{declared} \cdot \Omega^{all}_{+(-)}. \quad (6)$$

**Висновки.** Запропонований підхід для розрахунку, як узагальненої (інтегральної) оцінки всієї бази даних, так і оцінки якості окремих записів (кортежів) бази даних за кожним типом помилок. Дана оцінка може бути застосовна для об'єктивного оцінювання стану бази даних при фінансових операціях та у процесах створення систем автоматизованого управління ВНЗ.

**Література**

1. ISO 8000-110:2009, Data quality — Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification.

2. Мейер Д. Теория реляционных баз данных / Д. Мейер. – М. : Мир, 1987. – 608 с.

3. Кофман А. Введение в теорию нечетких множеств / А. Кофман - М.: Радио и связь, 1982. - 432 с.

4. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений / Л.Заде. - М.: Мир, 1976. - 166с.