УДК 004.9

Н.Б. Шаховська, О.Ю. Пшеничний, І.М. Чорней

*Національний університет «Львівська політехніка», Львів*

# ПРОБЛЕМИ ЯКОСТІ КОНСОЛІДОВАНИХ ДАНИХ У ПРОСТОРАХ ДАНИХ

*У статті формалізовано характеристики якості консолідованих даних у просторах даних. Уведено поняття корисності даних, отриманих з джерел даних. Розроблено архітектуру системи оцінювання якості різнотипних даних. Уведено метамову для формування запитів користувачів до різнотипних джерел. Розроблено структури даних для опису інформаційних продуктів та схеми метаданих. Досліджено якість даних у просторі даних туризму та експериментально визначено значення коефіцієнтів характеристик якості даних.*

***Ключові слова:*** *простір даних, консолідація, якість даних.*

## Вступ

Інформаційне суспільство – суспільство, у якому створення, поширення, дифузія, використання, інтеграція і маніпулювання інформації – важлива господарська, політична діяльність. Специфікою до цього виду суспільства є те, задача *консолідації даних* (об'єднання даних, розміщених у різних, наперед неузгоджених джерелах) виникає досить часто.

*Консолідовані дані* – це одержані з декількох джерел та системно інтегровані різнотипні інформаційні ресурси, які в сукупності поділені ознаками повноти, цілісності, несуперечності та складають адекватну інформаційну модель проблемної області з метою її аналізу опрацювання та ефективного використання в процесах підтримки прийняття рішень.

Для прийняття адекватних рішень у певній галузі необхідно, щоб дані, які надаються різними інформаційними ресурсами та у подальшому використовуються для прийняття керівних рішень, задовольняли такі вимоги: були повними, несуперечливими та надходили вчасно; були інформативними, оскільки вони застосовуватимуться для прийняття рішень; були однакової структури, щоб мати можливість завантажити їх у єдине сховище даних та проаналізувати; зберігалися в однакових моделях даних та були незалежними від платформи розроблення, щоб мати можливість використання цих даних іншими засобами.

Тому при опрацюванні даних з різних джерел виникає проблема якості цих даних (відповідності даних вимогам користувачів). На рівні задач, для яких використовується точкове джерело, якість даних цього джерела є достатньою, і задовольняє (повністю чи частково) потреби осіб, що приймають рішення на їх основі. Проте, коли йде мова про використання даних з декількох джерел, наперед неузгоджених, якість таких даних різко знижується і вже не може задовольняти потреб користувача через неузгодженість форматів, різне представлення тощо.

**Постановка задачі.** Інформаційні продукти, що описують певну предметну область, та консолідовані дані становитимуть простір даних. Однією зі задач, яка виникатиме у процесі консолідації, є невизначеність даних, що є результатом дублювання, неточності, відсутності, протиріччя даних.

Над опрацюванням різнотипних даних працювали Colin White, A. Sheth, J. Larson, Maurizio Lenzerini, Frederick Lane, Christoph Koch, Xin Dong, Л.А. Калиниченко, С.А. Ступников, А.В. Фомичев, М.Н. Гриньов, С.Д. Кузнецов та ін. Розроблені моделі та метамови опрацювання різнотипних даних [1 – 9]. Проте, вказані моделі та методи опрацьовують або лише наперед відомі типи даних (здебільшого, реляційні бази даних), або вирішують лише часткові проблеми опрацювання різнотипних даних – наприклад, індексування для пришвидшення пошуку. Тому на сьогодні немає жодної методики опрацювання даних, яка б задовольняла всі наведені вимоги до якості даних, а отже, немає можливості аналізувати стан галузі загалом, використовуючи першоджерела інформації, а не визначені наперед статистичні звіти.

Іншою задачею є визначення та узгодження схем даних інформаційних ресурсів. Існуючі на сьогодні методи [1 – 5] опрацьовують або наперед відомі схеми даних, або вимагають, щоб джерела даних (інформаційні продукти) перебували під жорстким контролем, що не дає змоги встановлювати змінні семантичні зв'язки.

## Основний матеріал

Під *якістю даних* розумітимемо процес компонування даних (включаючи імена ресурсів, схеми даних, особливості доступу та опрацювання тощо), очищення та вдосконалення даних, а також об'єднання з усуненням дублювання та невизначеності.

Корисність даних для певного користувача чи групи користувачів залежить також і від ступеня

довіри до джерела даних. Тоді визначення ступеня довіри і-го користувача до j-го джерела даних:

$$\mathrm{Trust}(i, j) = \frac{\sum_{k=1}^{n} \mathrm{Trust}_k(i, j)}{n}, \quad (1)$$

де n – кількість звернень користувача до ресурсу; $\mathrm{Trust}_k(i, j)$ – значення лінгвістичної змінної, що відображає довіру і-го користувача до j-го джерела даних при k-ому зверненні..

Для розрахунку загального ступеня довіри до джерела *j* узагальнимо (1):

$$\mathrm{Trust}_j = \frac{\sum_{i=1}^{m} \left(\mathrm{Trust}(i, j)\right)}{n * m}, \quad (2)$$

де m – кількість користувачів, що звертались до ресурсу.

Ступінь довіри може встановлюватись і до конкретної характеристики джерела даних. Тоді він враховуватиме і ступінь довіри до джерела загалом, і довіру до конкретної характеристики:

$$\mathrm{Trust}^{attr}(i, j) = \mathrm{Trust}(i, j) * \frac{\sum_{k=1}^{n} \mathrm{Trust}_k^{attr}(i, j)}{n}, \quad (3)$$

де attr – назва атрибута, для якого здійснюється визначення ступеня довіри.

Визначимо корисність даних для прийняття рішення.

Нехай є критерій $R_j \in R$ оцінки наслідків рішення $x = (x_1, ..., x_j, ..., x_n)$, розподіл значень якого залежить тільки від компоненти $x_j$ альтернативи $x$.

Введемо такі позначення: $r*j$ – найкраще значення за критерієм $R_j$, j = 1, 2, … , m; rj0 – найгірше значення j-го критерію.

Якщо має місце незалежність критеріїв $R_1$, $R_2$, … , $R_m$ за перевагою, то багатовимірна функція корисності прийнятого рішення v(r) представлена у вигляді:

$$v(r) = \sum_{j=1}^{m} k_j v_j(r_j), \quad (4)$$

де $0<k_j<1$; j = 1, 2, … , m; $\sum_{j=1}^{m} k_j = 1$. Функцію $v_j$, що виражає оцінку значення $r_j$, можна вважати j-ою компонентою функції корисності, а $k_j$ – вагою, що визначає критерій $R_j$. У випадку просторів даних вага джерела даних j визначається як $k_j = \mathrm{Trust}(i, j)$ (див. (1)), де і є заданим і вказує на конкретного користувача.

Для оцінювання якості даних застосуємо загальний методичний підхід до виділення адекватної номенклатури стандартизованих в ISO 9126 базових характеристик і субхарактеристик. Базовими характеристиками якості даних є: функціональна придатність до використання, коректність або достовірність, ресурсна економічність, практичність, супроводжувальність, мобільність.

*Функціональна придатність* визначається, у першу чергу, повнотою накопичених об’єктів об'єктів – відносною кількістю об'єктів або документів, наявних у джерелах даних, до загальної кількості об'єктів, що потрапили у локальне сховище, яке містить консолідовані дані:

$$\mathrm{plenitude} = \frac{\mathrm{Count}(local\_dataware)}{\sum_i \mathrm{Count}(source_i)}. \quad (5)$$

*Коректність або достовірність даних* – це ступінь відповідності даних про об'єкти в базах даних реальним об'єктам в даний момент часу, що визначається змінами самих об'єктів, некоректних записів про їх стан або некоректними розрахунки їх характеристик:

$$\mathrm{identy} = \frac{\mathrm{Count}(\sigma_{\mathrm{Trust}>0.6}(local\_dataware))}{\mathrm{Count}(local\_dataware)}. \quad (6)$$

Використовуваність ресурсів (або *ресурсна економічність*) у стандартах відображається зайнятістю ресурсів центрального процесора, оперативної, зовнішньої та віртуальної пам'яті, каналів введення-виведення, терміналів і каналів зв'язку. Цей показник у роботі не буде аналізуватися, оскільки існують розроблені методи (наприклад, метод критичних робіт) та засоби визначення завантаженості ресурсів.

*Практичність* – визначає функціональну придатність і корисність застосування консолідованих даних для певних користувачів. У просторі даних оцінка практичності здійснюватиметься за допомогою функції корисності прийнятих рішень (4).

*Супроводжуваність* даних відображається зручністю і ефективністю виправлення, удосконалення або адаптації структури та змісту описів даних залежно від змін у зовнішньому середовищі застосування, а також у вимогах і функціональних специфікаціях замовника. Для оцінки супроводжуваності розроблені методи та засоби (наприклад, технологія ETL – витягнення, трансформування, завантаження), тому у роботі ця характеристика даних не розглядається.

*Мобільність* характеризується тривалістю і трудомісткістю їх інсталяції, адаптації та заміщення при перенесенні на інші апаратні та операційні платформи. У просторах даних характеристика мобільності пов’язана зі зміною даних про джерела даних у каталозі:

$$\mathrm{actuality} = \frac{\mathrm{Count}\left(\sigma_{upd<30}\left(local\_dataware\right)\right)}{\mathrm{Count}\left(local\_dataware\right)}. \quad (7)$$

Отже, під *якістю консолідованих даних у просторі даних* будемо розуміти інтегральну характеристику, яка відображає повноту накопичення даних, коректність та мобільність та корисність прийнятих рішень:

$$s_1 \text{quality} + s_2 v(r) \to \text{Max}, \qquad (8)$$

де quality – інтегральний безрозмірний показник характеристик якості даних, $0 \le \text{quality} \le 1$,

$\text{quality} = n1 \cdot \text{plenitude} + n2 \cdot \text{identy} + n3 \cdot \text{actuality}$;

$s_1$ – коефіцієнт важливості повноти накопичення даних; $v(r)$ – значення багатовимірної функції корисності; $s_2$ – коефіцієнт важливості якості прийнятих рішень, $s_1 + s_2 = 1$.

Для консолідації даних в сховищі консолідованих даних використовується каталог, схема якого подана на рис. 1.

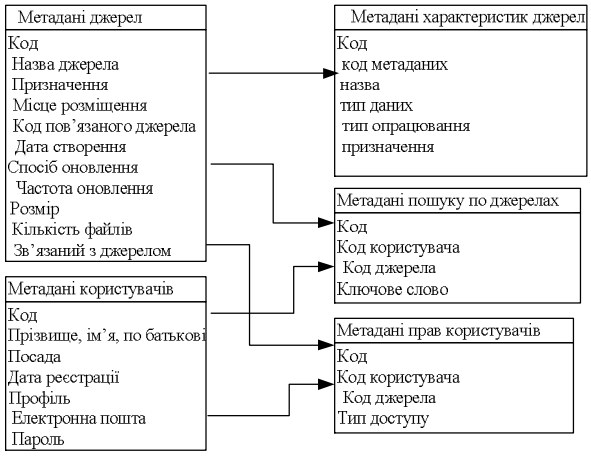


Рис. 1. Схема каталогу простору даних

Також необхідно передбачити той факт, що у різний термін часу джерела можуть мати різний ступінь довіри. Мається на увазі, що не завжди інформація, отримана з джерела даних, буде достовірною. Це особливо притаманно Веб-ресурсам.

Для встановлення ступеню довіри до джерела використовуватиметься лінгвістична змінна *Trust* (рис. 2 – див. (1)).

| **інтервал** | **Назва** |
|---|---|
| [0; 0.20) | не вірю |
| [0.20; 0.40) | неправдоподібно |
| [0.40; 0.60) | треба перевірити |
| [0.60; 0.80) | цілком можливо |
| [0.80; 1) | вірю |

Рис. 2. Лінгвістична змінна *Trust*.

Кожен користувач може вказувати власний ступінь довіри до джерела. Також ступінь довіри розраховується на основі статистичного аналізу задоволеності користувачів результатами запиту, який виконувався у заданому джерелі. Для вказання задоволеності також використовується лінгвістична змінна *Trust*.

Також необхідно розробити структуру даних для словника даних. Вона подана на рис. 3.

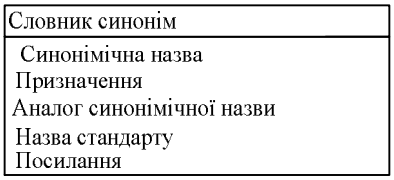


Рис. 3**.** Структура даних словника синонімів

Далі введено елементи метамови простору даних. Вважатимемо, що запит $q$ до простору даних заданий коректно, якщо він складається з елементів, описаних у *Cg* та *Dic*.

$$q_{\text{object}(c_1 \ldots c_n)} : \text{par} = \begin{Bmatrix} \text{object} \in \text{Cg}, \text{Trust}_{\text{object}} > 0, \\ (c_1, \ldots, c_n) \in \text{Dic} \end{Bmatrix} : \text{par},$$

де object – об'єкт, про який йде мова у запиті; $(c_1, \ldots, c_n)$ – назви характеристик об'єкта; par – список параметрів запиту. Залежно від типу джерела параметри можуть відігравати роль: параметрів пошуку – у текстових даних; умови вибору – для структурованих даних.

Алфавіт запиту об'єднує алфавіт усіх джерел даних, до яких направляють запит, а для встановлення характеристик вибираються усі можливі синоніми:

$$\text{Dic} = \{R\} \cup \{\text{Rel}\} \cup \{\text{key}\} \cup \{H\},$$

де описані схеми баз даних, сховищ даних, ключових слів текстових файлів, заголовків вебдокументів відповідно.

Розроблено систему визначення якості консолідованих даних

Підсистема забезпечення і підтримки якості даних у просторі даних призначена для реалізації алгоритмів і процедур, які забезпечують оцінку якості даних, збір та обробку інформації для підтримки якості даних. Відповідно з цим визначенням до складу підсистеми входить аналітичний центр, засоби збору та передачі даних (рис. 4). Джерелами даних є інформаційні продукти (ІП).

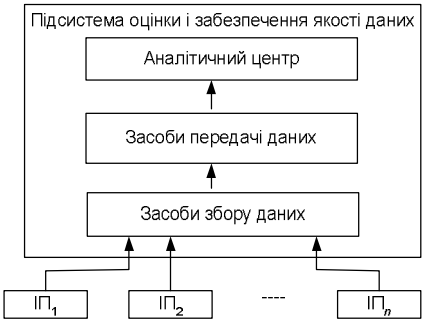


Рис. 4. Архітектура підсистеми оцінки і забезпечення якості консолідованих даних

Процес оцінювання якості консолідованих даних складається з трьох стадій: встановлення вимог до якості консолідованих даних, підготовка до оцінювання та процедура оцінювання (рис. 5).

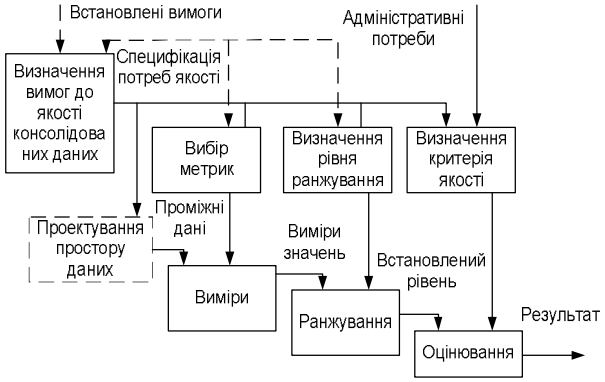


Рис. 5. Схема процесу оцінювання якості консолідованих даних

Поряд з фактичною оцінкою якості консолідованої інформації $K_i$ необхідно провести оцінку якості еталонного зразка $K_i^e$, що відображає найкраще прийняте рішення. Потім проводиться нормування фактичної оцінки за формулою (9):

$$K_i = \sum_i \alpha_i M_i ;\ K'_{const} = \sum_e q_i K_i ;$$

$$K_i^e = \sum_i \alpha_i M_i^e ;\ K'^{e}_{const} = \sum_e q_i K_i^e ,$$

де $\alpha_i$ – ранг важливості, $m_i$ – відносне значення і-ї характеристики, $k_i$ – значення і-о показника якості, що визначається експертно, причому $k_i \in [0; 1]$,

$$K_{const} = K'_{const} / K_{const}^{e} . \qquad (9)$$

## Дослідження якості даних у просторі даних туризму

Для простору даних туристичної сфери на якість даних впливають повнота накопичених об'єктів (рис. 6) і корисність прийнятих рішень.

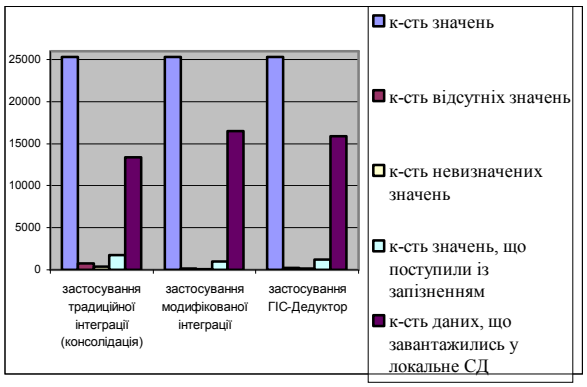


Рис. 6. Аналіз повноти накопичених описів об'єктів

На поданій діаграмі проаналізовано роботу алгоритму модифікованої консолідації та порівняно її з роботою алгоритму немодифікованої інтеграції, застосованому в Oracle Data Integrator. Дані у локальне сховище потрапляють з баз даних туристичних агенцій, структури даних яких наперед невідомі. Кількість записів вхідних баз даних, що мають потрапити у локальне сховище даних - 15000. Без попереднього визначення структур джерел частина даних не була завантажена, а частина не могла бути завантажена через неспівпадіння структури локального сховища та структури джерела.

Проаналізуємо, як зміниться корисність прийнятих рішень від кількості джерел, дані з яких використовувались для його підтримки. Результати тестування показали, що співпадіння показників із плановими є вищим після застосування алгоритмів усунення невизначеностей, що доводить доцільність роботи. На основі тестових даних розрахуємо коефіцієнти важливості кожного з показників для розроблення комплексного показника якості простору даних. Розрахунок здійснюватимемо наступним чином. Коефіцієнт актуальності встановимо в 0, коефіцієнти інших показників визначимо як усереднене значення по параметрах плану.

| | План по точних даних | План з показником "повнота" >0.9 | План з показником "коректність" >0.9 |
|---|---|---|---|
| Працівники | 92 | 67 | 86 |
| Тур. послуга | 93 | 91 | 92 |
| Відпочиваючі | 87 | 74 | 81 |

$$n_i = \frac{\sum_{j=1}^{k} coef_j^i}{\sum_{j=1} plan_j^i} ,$$

де $n_i$ – показник, що розраховується; $plan_j^i$ – значення плану за точними даними; $coef_j^i$ – значення плану за інтегрованими даними. Для нашого прикладу $n_1 = 0.84$, $n_2 = 0.95$.

## Висновки

У роботі розроблено метод визначення якості консолідованих даних, що дозволило визначати придатність цих даних для подальшого прийняття рішень. Науковою новизною роботи є формалізація характеристик якості програмних продуктів, описаних в ISO 9126. Практичною цінністю є визначення важливості показників якості.

## Список літератури

*1. Рогушина Ю.В. Формирование тезауруса предметной области как средства моделирования информационных потребностей пользователя при поиске в Интернете / Ю.В. Рогушина, А.Я. Гладун // Вестник компь-*

*ютерных и информационных технологий. – М., 2007. – № 1. – С. 26-33.*

*2. Kalinichenko L.A. Compositional Specification Calculus for Information Systems Development Proceedings of the East-West Conference on Advances in Databases and Information Systems (ADBIS'99) / L.A. Kalinichenko. – Maribor, Slovenia, September 1999, Springer Verlag, LNCS*

*3. Кулик Б.А. Представление логических систем в вероятностном пространстве на основе алгебры кортежей. 1. Основы алгебры кортежей / Б.А. Кулик // Автоматика и телемеханика. – 1997. – № 1. – С. 126-136.*

*4. Xin Dong. Indexing Dataspaces // Xin Dong, Alon Halevy. – SIGMOD'07, June 11–14, 2007, Beijing, China.*

*5. Lenzerini M. Data Integration: A Theoretical Perspective [Електронний ресурс] / Maurizio Lenzerini // PODS 2002. – 2002. – P. 233-246. – Режим доступу до ресурсу: http://www.dis.uniroma1.it/~lenzerin/homepagine/talks/TutorialPODS02.pdf.*

*Надійшла до редколегії 21.04.2011*

**Рецензент:** д-р техн. наук, проф. В.В.Пасічник, Національний університет «Львівська політехніка», Львів.

## ПРОБЛЕМЫ КАЧЕСТВА КОНСОЛИДИРОВАННЫХ ДАННЫХ В ПРОСТРАНСТВЕ ДАННЫХ

Н.Б. Шаховская, А.Ю. Пшеничный, И.М. Чёрней

*В статье формализованы характеристики качества консолидированных данных в пространствах данных. Введено понятие полезности данных, полученных из источников данных. Разработана архитектура системы оценки качества разнотипных данных. Введен метаязык для формирования запросов пользователей к разнотипным источникам. Разработаны структуры данных для описания информационных продуктов и схемы метаданных. Исследовано качество данных в пространстве данных туризма и экспериментально определены значения коэффициентов характеристик качества данных.*

***Ключевые слова:*** *пространство данных, консолидация, качество данных.*

## CONSOLIDATED DATA QUALITY PROBLEMS IN DATASPACE

N.B. Shahovskaya, A.Yu. Pshenichniy, I.M. Chyorney

*The paper formalizes quality characteristics of consolidated data in dataspaces. There is described the notion of usefulness of data from data sources. There is described an architecture for a system of quality evaluation of different types of data. There is entered metalanguage for the formation of user requests for different types of sources. A data structure for describing the information products and metadata schemas is entered. There is research data quality in dataspace tourism and experimentally determined values of the characteristics of data quality*

***Keywords:*** *dataspace, consolidation, data quality.*