УДК 51.001.57+004.652.4+004.827

ФОРМАЛЬНЕ ПОДАННЯ ПРОСТОРУ ДАНИХ У ВИГЛЯДІ АЛҐЕБРАЇЧНОЇ СИСТЕМИ

Н.Б. ШАХОВСЬКА

Проаналізовано проблеми опрацювання розрізнених даних. Побудовано формальну модель простору даних та уведено операції над ним. Показано алгебраїчні системи бази даних та сховища даних, які є підкласами алгебраїчної системи «простір даних». Визначено особливості інтеграції даних із різнорідних джерел. Побудовано схему інтеграції даних та засоби обміну даними.

ВСТУП

Інформаційне суспільство — суспільство, в якому створення, поширення, дифузія, використання, інтеграція та маніпулювання інформацією — важлива господарська, політична і культурна діяльність [6].

Специфікою цього виду суспільства є те, що інформаційна технологія є центральною позицією для виробництва, економіки та суспільства загалом. У сучасному суспільстві інформація є найдорожчою цінністю, а індустрія отримання, опрацювання і трансляції інформації — провідною галуззю діяльності, в яку з кожним роком все більше інвестують грошей. Як вважають провідні вчені, інформація є важливим стратегічним ресурсом, відсутність якої призводить до суттєвих втрат в економіці. Інформатизація суспільства виступає одним із вирішальних чинників модернізації економіки на ринкових засадах і запорукою інтеграції України у світове співтовариство.

Для прийняття адекватних рішень у певній галузі необхідно, щоб дані, які надходять із різних джерел і використовуються для прийняття керівних рішень, задовольняли такі вимоги:

- були повними, несуперечливими та вчасно надходили;
- були інформативними, оскільки вони застосовуватимуться для прийняття рішень;
- були однакової структури, для можливості завантажувати їх у єдине сховище даних та аналізувати;
- зберігалися в однакових моделях даних та були незалежними від платформи розроблення, для можливості використання їх іншими засобами.

Однак, на сьогодні немає жодної методики опрацювання даних, яка б задовольняла всі зазначені вимоги до опрацювання даних, тому немає мож-

ливості аналізувати стан галузі загалом, використовуючи першоджерела інформації, а не визначені наперед статистичні звіти. Так, наприклад, для керівництва туристичною галуззю використовуються результати аналізу зведеної форми 1 Тур та надходжень із митниць. Така наявна інформація дозволяє фіксувати факт настання певної причини та її наслідки, але найчастіше не дозволяє визначати причини, оскільки для аналізу використовується обмежена і наперед жорстко визначена частина інформації.

За останні роки спостерігалося зростання потреби в «даних, які застосовуються у всіх сферах», що призвело до виникнення нового типу інформаційної інтелектуальної системи. Найгостріші проблеми керування інформацією виникають в організацій (наприклад, готелів, баз відпочинку, оздоровчих закладів, туристичних агентств), робота яких полягає в опрацюванні великої кількості різнотипних, взаємонезалежних джерел даних. Такий тип системи отримав назву «простір даних». На відміну від систем інтеграції даних, що також пропонують загальноприйнятий доступ до різнорідних джерел даних, простори даних не припускають, що всі семантичні взаємозв'язки між джерелами відомі та вказані. Багато користувачів, які працюють із просторами даних проводять дослідження даних, і немає єдиної схеми, за якою вони можуть створювати запити. Тому важливо, що запити є дозволеними елементами, щоб конкретизувати різні ступені структури, при цьому використання ключового слова робить запит більш структурованим.

Простір даних розглядають як нову абстракцію керування даними [4]. Основоположником ідеї просторів даних був А. Хелеві. Нині розроблюються два проекти, орієнтовані на підтримку просторів індивідуальних даних. Перший з них — проект SEMEX (SEMantic Explorer — система навігації та пошуку по повнотекстових документах), виконується у Вашингтонському університеті під керівництвом А. Хелеві. Другий, називається iMeMex, виконується під керівництвом Йенса-Петера Диттриха в компанії «ETH Zurich». Проте, як видно з аналізу Інтернет-джерел [2, 3], жоден із проектів ще не формалізував поняття простору даних, що, у свою чергу, призводить до розрізненості підходів роботи з ними.

Мета роботи — формалізація та математичний опис простору даних з метою уніфікації описів джерел даних; розроблення алгебраїчної системи класу «простір даних».

Об'єктом дослідження є процес консолідації даних певної галузі за умов наявності різнотипних джерел даних.

Предметом дослідження є методи підвищення якості консолідованих даних, отриманих із різнотипних джерел.

АЛГЕБРАЇЧНА СИСТЕМА КЛАСУ ПРОСТОРУ ДАНИХ

Введемо деякі означення.

Інформаційний ресурс (ІР) — дані, які можна багаторазово використовувати для вирішення проблем користувача. Прикладами інформаційних ресурсів є текстові файли, веб-сторінки, електронні таблиці, xml-файли, бази даних, сховища даних.

Структура даних ІР (СДІР) — загальна властивість інформаційного об'єкта, з яким взаємодіє та або інша програма. Ця загальна властивість характеризується:

- множиною допустимих значень цієї структури;
- множиною допустимих операцій;
- характером організованості.

Каталог ІР — метадані про ІР. Описує місцезнаходження ІР, його СДІР, методи доступу тощо.

Множина інформаційних ресурсів *Ir* предметної області містить найповнішу інформацію про предметну область. Такий вид інформації називатимемо консолідованою. Якість прийнятих рішень на основі консолідованої інформації є вищою, ніж на основі даних з точкових джерел, оскільки є можливість пошуку прихованих залежностей даних. Множина всіх інформаційних ресурсів предметної області — простір даних.

$$DS = \langle DB, DW, Wb, Nd, Gr \rangle, \qquad (1)$$

де *DB*, *DW*, *ODW*, *Wb*, *Nd*, *Gr* — інформаційні ресурси, що подають множини баз даних, сховищ даних, веб-сторінок, текстових файлів, електронних таблиць та графічних даних відповідно.

Стан інформаційного ресурсу — зафіксований у певний момент часу вміст інформаційного ресурсу (даних) та відомостей про нього. Стан інформаційного ресурсу позначатимемо S_{Ir}.

Стан простору даних — стани всіх інформаційних ресурсів предметної області (множина даних) та відношень між ними. Стан ПД позначатимемо S_{DS}.

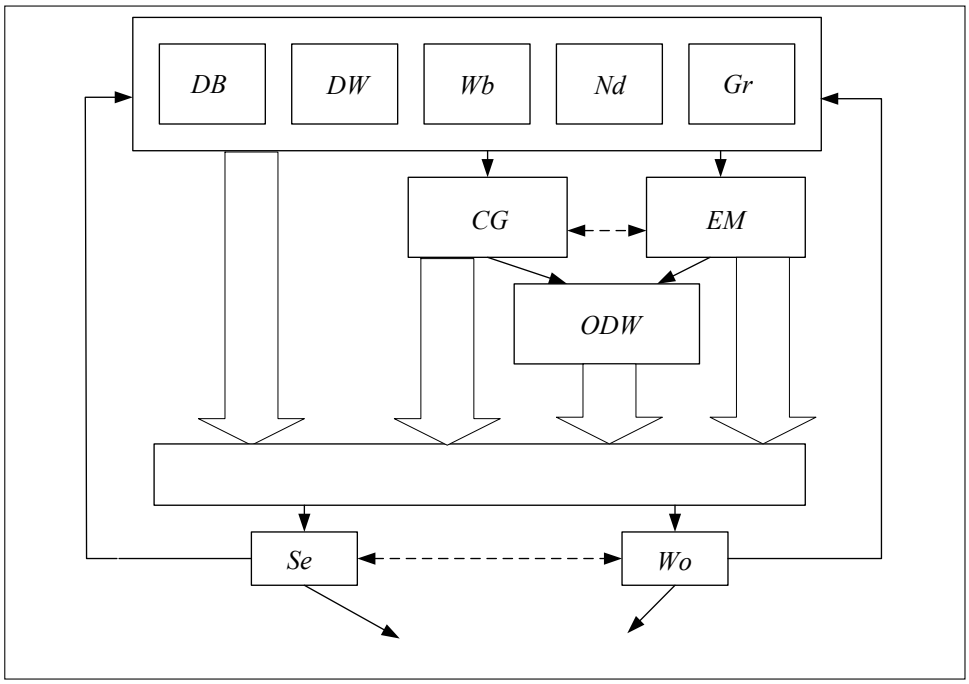

Рис. 1. Схема зв'язку елементів простору даних

Множину інформаційних ресурсів простору даних, операцій над ними та предикатів на множині *Ir* називатимемо алгебраїчною системою класу «простір даних».

$$DS = \langle Ir, \Omega_P, \Omega_F \rangle, \tag{2}$$

де $Ir = DS$ — множини інформаційних ресурсів певної предметної галузі (баз даних DB, сховищ даних DW, статичних Web-сторінок Wb, текстових даних Nd графічних та мультимедійних даних Gr), $\Omega_P = \{O_{P_0}, \boldsymbol{O}_{\boldsymbol{P_u}}, \boldsymbol{O}_{\boldsymbol{P_b}}\}$ — множина операцій над інформаційними ресурсами, де: O_{P0} — нульарна операція, результатом якої є стан заданого інформаційного ресурсу; $\boldsymbol{O}_{\boldsymbol{P_u}}$ — множина унарних операцій над простором даних DS. Результатом цих операцій є зміна стану простору даних; $\boldsymbol{O}_{\boldsymbol{P_b}}$ — множина бінарних операцій над просторами даних. Результатом цих операцій є утворення нового простору даних. Ω_F — множина предикатів, заданих на множині інформаційних ресурсів простору даних. Серед предикатів також є нульарний предикат Ω_{F_0}, результатом якого є TRUE, якщо для заданого інформаційного ресурсу Ir відомого його структури даних, та FALSE в іншому випадку.

Алгебраїчна система (2) скінченна, оскільки множина інформаційних ресурсів DS є скінченною [1].

Інформаційні ресурси простору даних

Говорячи про інформаційний ресурс, матимемо на увазі його вміст (дані). Також описуватимемо операції, які виконуються над даними залежно від його СДІР.

Як вже було перераховано вище, інформаційними ресурсами простору даних є текстові файли, електронні таблиці, веб-сторінки, графічні файли (карти, об'єкти на яких задані точково або векторно), бази даних та сховища даних.

Основною операцією, що виконується над вмістом текстових файлів, електронних таблиць та веб-сторінок є операція пошуку. Структури даних цих інформаційних ресурсів є простими, і як відомо, називаються типами даних, тому детально описуватись не будуть.

Реляційна база даних — це алгебраїчна система, в якої носієм є множина реляційних відношень r, множиною операцій — реляційна алгебра $\Re$, множиною предикатів — словник даних (схема даних бази даних) R [2].

$$DB = \langle r, \Re, R \rangle, \tag{3}$$

$$\Re = \{\pi, \sigma, \triangleright\triangleleft, \cup, \cap, -\}.$$

Сховищем даних (СД) назвемо шістку

$$DW = \langle DB, rf, RF, rm, RM, func \rangle,$$

де DB — множина вхідних баз даних (реляційних, багатовимірних, об'єктно-орієнтованих, ненормалізованих тощо, або множина відношень, їх схем та обмежень цілісності, які містять інформацію з вхідних баз даних), rf — множина відношень фактів, RM — схема rf, rm — множина відношень метаданих, RM — схема rm, $func$ — множина процедур прийняття рішень.

Метадані — дані, що містять опис структури сховища даних, джерел та приймачів даних тощо (дані про дані).

Тоді нові дані (або рішення) — це результат застосування функцій сховища даних над відношенням фактів:

$$Design = func(rf, user\_param),$$

де $user\_param$ — множина параметрів користувача або вимог, які ставляться до рішення.

Відношення між вимірами — відношення, яке є зв'язком між певними вимірами та відношенням фактів:

$$V_1 \times V_2 \times ... \times V_n \times rf \to rel.$$

У відношенні фактів виміри подаються за допомогою зовнішніх ключів, а самі значення — за допомогою атрибутів агрегації. У свою чергу, *rel* можуть бути параметрами для інших відношень між вимірами і тим самим створювати ієрархію вимірів.

Отже, хоча інформаційні ресурси, що входять в ПД, за своїм характером є різними та керуються різними платформами, проте вони всі виконують однакову роль: надають дані для простору даних через фіксацію свого стану та забезпечують виконання притаманних для них операцій, причому ці операції та їх результати є визначені для всього простору даних.

ОПЕРАЦІЇ АЛГЕБРАЇЧНОЇ СИСТЕМИ КЛАСУ «ПРОСТІР ДАНИХ»

Нульарна операція. Результатом нульарної операції над простором даних *DS* є стан заданого інформаційного ресурсу:

$$S_{Ir} = O_{P0}(DS).$$

Наприклад, нульарний оператор поверне стан заданої бази даних i: $S_{DBi} = O_0(DS)$.

Унарні операції. Унарними операціями над просторами даних є шістка:

$$\boldsymbol{O_{Pu}} = \{Agent(\aleph), Se_{\text{simple}}, Se_{\text{structured}}, Se_{\text{meta}}, \sigma_{\text{access}}, Agent\},$$

де $\aleph = \{\Re, \xrightarrow{\text{consolid}}, \sigma_{\text{fed}}, Ag, func\}$ — операції над інформаційними ресурсами з відомими СДІР, $Agent(\aleph)$ — операції над інформаційними ресурсами з попереднім визначенням СДІР.

Визначення СДІР даних здійснюється за допомогою інтелектуального агента

$$EM\,(CG) \xrightarrow{Agent} ODW. \qquad (4)$$

Агент *Op* подається сімкою об'єктів:

$$Agent = \langle CG, EM, Dic, Experience\_Base, Solver, Effector \rangle, \qquad (5)$$

де *CG* — ідентифікатор внутрішнього стану агента (інформація про джерела, що вже є у ПД); *EM* — компонента агента, що відповідає за сприйняття

середовища (сенсор), тобто середовище керування моделями; *Dic* — база знань, що містить знання агента про власні можливості (терміни-синоніми, що позначають у джерелах одні й ті ж властивості);

Experience_Base — база накопиченого досвіду агента, що містить «історію» впливів на агента з боку середовища й відповідної їм реакції агента ($Experience\_Base = \sigma_{evdate=Date()}(Dic)$; *Solver* — компонента, що відповідає за навчання (подає список розбіжностей, які виявив агент); *Effector* — компонента, яка відповідає за дії агента (формування запиту по декількох джерелах, приведення результатів запитів по джерелах до єдиної структури, відмова у запиті).

В основі роботи агента лежить інформація про джерела, які вже є у просторі. Його завданням є порівняння структур даних джерела даних, що входитимуть у простір, із структурами даних джерел, що вже є у просторі, а також визначення різниці. Це дозволить автоматизувати формування запитів, що виконуватимуться у просторі даних.

Чим більше джерел здатний «розрізнити» агент, тим точніше буде інформація в *ODW* і тим ефективніше можна буде проводити процедури інтеграції, пошуку та опрацювання даних у просторі даних *DS*.

Розглянемо завдання порівняння інформації з двох схем даних для однакових фізичних сутностей. При цьому допускається, що схеми мають різні системи кодування, тобто той самий об'єкт може мати в цих схемах різні ідентифікатори. Допускається, що назви таблиць, атрибутів і розподіл атрибутів по таблицях можуть розрізнятися. Але передбачається, що між схемами існують взаємозв'язки, які можуть бути задані експертами. Наше завдання — класифікувати типи можливих взаємозв'язків і знайти необхідні умови для рішення різних завдань інтеграції даних на основі цих взаємозв'язків.

Нехай деяка сутність описується в першій схемі даних відношенням *A*, що містить кортежі $\{x_1, x_2, ..., x_n\}$, а в другій схемі даних відношенням *B*, що містить кортежі $\{y_1, y_2, ..., y_m\}$. Відношення *A* і *B* можуть бути як окремими таблицями в реляційній схемі даних, так і переглядами. Запишемо формально умову, що *A* і *B* містять однакові фізичні сутності. Будемо вважати, що в цьому випадку існують взаємозв'язки між окремими атрибутами x_i та y_i.

Розглянемо різні типи таких взаємозв'язків між двома скалярними атрибутами x та y, визначеними на скінченних доменах X та Y відповідно.

- Змістовний взаємозв'язок доменів. Найзагальнішим типом взаємозв'язку можна вважати випадок, коли ми хоча б можемо визначити, чи співпадають об'єкти по атрибутах *x* та *y*, або не співпадають і чи співпадають назви-синоніми у словнику термінів *Dic*. Тобто, задана функція змістовної еквівалентності: $P: X \times Y \to \{0,1\}$, $Dic_{X=Y}$. $P(x, y) = 1$, якщо по атрибутах *x* та *y* об'єкти співпадають, $P(x, y) = 0$ в іншому випадку. Якщо $P(x, y) = 1$ і $Dic_{X \neq Y}$, то доповнюємо Dic новими синонімами.

- Існує відображення, що конвертує X та Y за умови, якщо для будь-якого значення $x \in X$ існує значення $y \in Y$ таке, що по атрибутах x та y об'єкти будуть співпадати. Тобто, існує відображення $F: X \to Y$ таке, що для всіх $x \in X$ виконується рівність

$$P(x, F(x)) = 1, Dic_{X \neq Y}. \quad (6)$$

- Існує узагальнююче відображення з X в Y (Y — узагальнення X) за умови, якщо для будь-якого значення $x \in X$ існує рівно одне значення $y \in Y$ таке, що по атрибутах x та y об'єкти будуть співпадати. Тобто, існує відображення $F : X \to Y$ таке, що для всіх $x \in X$ виконується умова (2.5) і нерівність

$$P(x, y) < 1, Dic_X, Dic_Y \text{ для всіх } y \neq F(x). \quad (7)$$

- Існує узагальнююче відображення X на Y (X — деталізація Y) за умови, якщо для будь-якого значення $x \in X$ існує лише одне значення $y \in Y$, і для будь-якого y існує хоча б одне значення x таке, що по атрибутах x і y об'єкти будуть співпадати. Тобто, існує відображення $F : X \to Y$ таке, що для всіх $y \in Y$ існує $x \in X$, такий що $F(x) = y$; і для всіх $x \in X$ виконуються умови (8) і (9).

- Ізоморфізм доменів існує за умови, якщо є відображення $F : X \to Y$, що задовольняє умовам (8) і (9), і зворотне до нього $F^{-1} : Y \to X$, також задовольняючим умовам (6) і (7).

Будемо вважати, що об'єкт, заданий кортежем $a = \{x_1, x_2, ..., x_n\}$ в одній схемі даних, співпадає з об'єктом, заданим кортежем $b = \{y_1, y_2, ..., y_m\}$ в іншій схемі даних, якщо вони співпадають за всіма взаємозалежними атрибутами, тобто для всіх функцій взаємозв'язку відношень $P_{ij} : X_i \times Y_j \to \{0,1\}$ має місце рівність $P_{ij}(x_i, y_j) = 1$. Множину пар індексів (i, j), для яких задані функції P_{ij}, позначимо $\Omega = \{(i, j)\}$, $i = Num(x)$, $j = Num(y)$, $x, y \in Dic$. Тоді можна задати функцію відповідності об'єктів $P : A \times B \to \{0,1\}$ таким чином:

$$P(a,b) = 1, \text{ якщо } P_{ij}(x_i, y_j) = 1 \text{ для всіх } (i, j) \in \Omega; \quad (8)$$

$$P(a,b) = 0, \text{ якщо існує } (i, j) \in \Omega \text{ такі, що } P_{ij}(x_i, y_j) \neq 1. \quad (9)$$

Перейдемо до класифікації взаємозв'язків між схемами даних.

1. Відповідність об'єктів. Якщо Ω не порожня, і задана функція $P : A \times B \to \{0,1\}$, будемо говорити, що встановлено відповідність об'єктів. Нехай X_1 і Y_1 є первинними ключами відношень A і B. Тоді, якщо вибрати всі пари $\{x_1, y_1\}$, для яких $P(a,b) = P(\{x_1, x_2, ..., x_n\}, \{y_1, y_2, ..., y_m\}) = 1$, одержимо таблицю відповідності Dic із заголовком $\{\langle x_1 : X_1 \rangle, \langle y_1 : Y_1 \rangle\}$. Маючи таку таблицю, можна робити запити, що отримують дані з обох схем таким чином:

Select $x_1, x_2, ..., x_n, y_1, y_2, ..., y_m$

From A, B, Dic

Where $Dic.X_1 = A.X_1$ and $Dic.Y_1 = B.Y_1$

2. За кортежем a із відношення A можна швидко знайти у відношенні B кортеж b такий, що $P(a,b) = 1$, не створюючи й не використовуючи таблицю відповідності.

3. За кортежем з A можна однозначно визначити кортеж у B.

4. Відношення A і B синхронізовані. Якщо за кортежем з A можна однозначно визначити кортеж у B і за кортежем із B можна однозначно визначити кортеж в A, будемо говорити, що відношення A і B синхронізовані. Зміст цієї умови полягає в тому, що якщо перенести деякий кортеж a із A в B, а потім назад, то гарантовано не буде створено нового запису, що дублює a.

Отже, результатом роботи агента є встановлення взаємозв'язку між схемами даних.

Продемонструємо результат роботи агента. Користувач відсилає запит такої структури:

Вибрати тури, де рейс = «Пам'ятник Шевченка»

У словнику даних властивість «рейс» описана як race_id.

Т а б л и ц я . Словник *Dic*

Код	Властивість	Назва
1	рейс	race_id
2	тур	tour_id

Нехай є такі дві бази даних туристичних організацій (рис. 2, 3) та вебсайт турагентсва (рис. 4). Завданням агента є визначення туристичної фірми, що надає рейси, в які входить відвідування пам'ятника Шевченку.

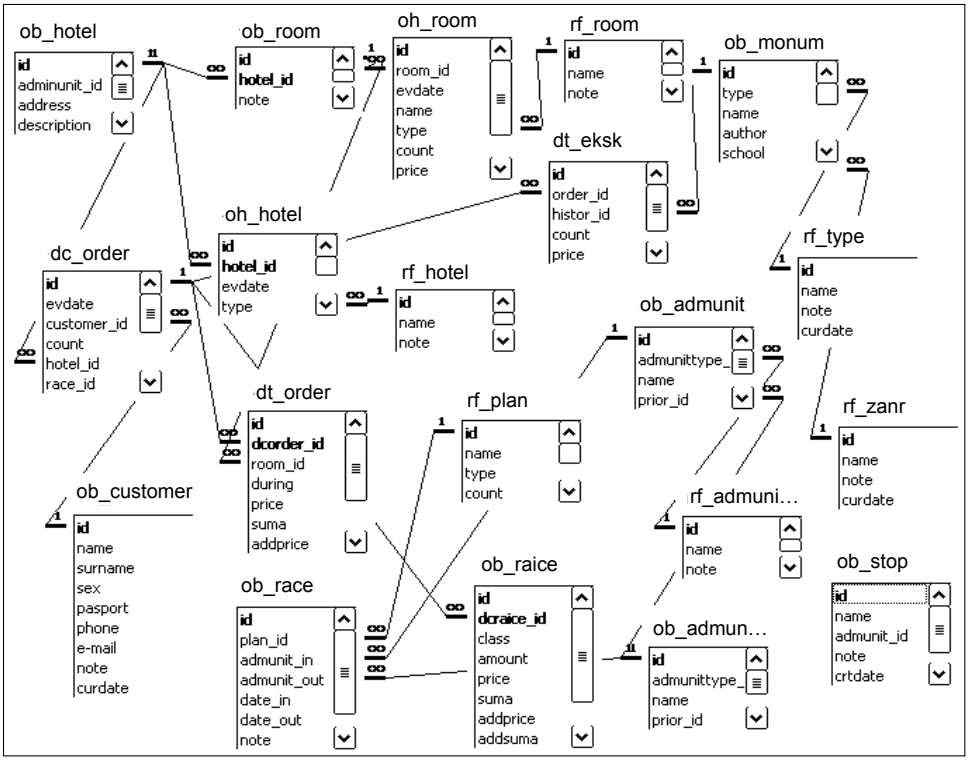

Рис. 2. Схема бази даних туристичного агентства 1

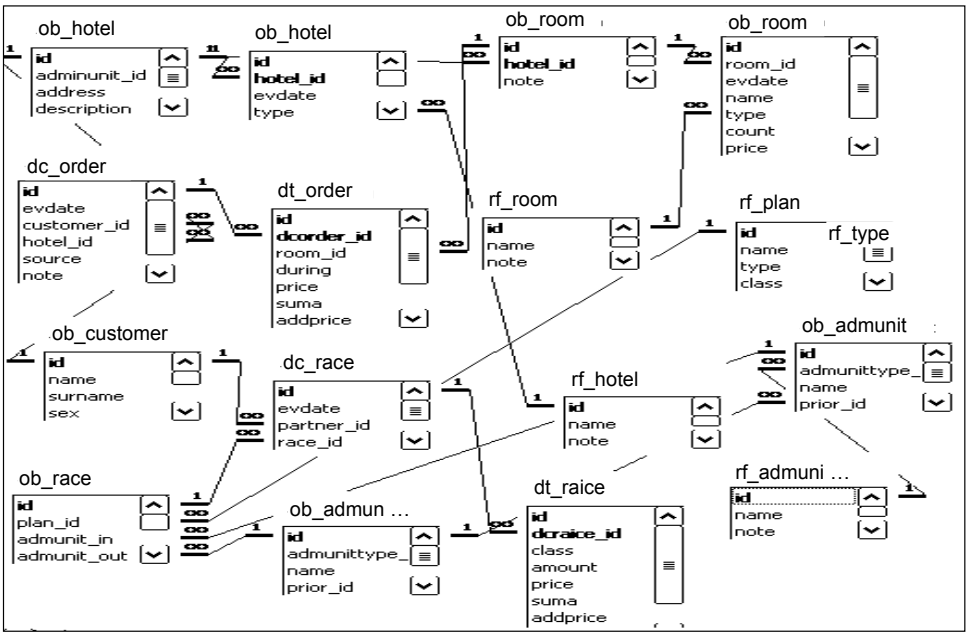

Рис. 3. Схема бази даних туристичного агентства 2

```
<?xml version="1.0" encoding="UTF-8"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:od="urn:schemas-microsoft-com:officedata">
<xsd:element name="dataroot">
<xsd:complexType>
……….
</xsd:complexType>
</xsd:element>
<xsd:element name="dc_order">
<xsd:annotation>
<xsd:appinfo>
<od:index index-name="PrimaryKey" index-key="id " primary="yes"
unique="yes" clustered="no"/>
<od:index index-name="customer_id" index-key="customer_id " primary="no"
unique="no" clustered="no"/>
<od:index index-name="race_id" index-key="race_id " primary="no" unique="no"
clustered="no"/>
<od:index index-name="ob_hoteldc_order" index-key="hotel_id " primary="no"
unique="no" clustered="no"/>
……..
</xsd:complexType>
</xsd:element>
</xsd:schema>
```

Рис. 4. XML-файли туристичного агентства 3

Перш за все, агент визначає, чи є вказаний атрибут (рейс) у перерахованих у каталозі джерел простору даних та в якому відношенні. Визначено, що такий атрибут існує в першій із наведених баз даних у відношенні

dc_race. Далі здійснюється порівняння схеми цієї бази даних із наступною реляційною базою даних та xml-файлом.

Результат порівняння вказаних джерел подано на рис. 5. Тут показано таблиці, які відсутні у джерелах даних, а також відмінності у таблицях із однаковими назвами.

Таблиці, яких нема в жодній із баз

dc_race	нема у прешій базі
dt_eksk	нема у другій базі
ob_monum	нема у другій базі
ob_stop	нема у другій базі
oh_race	нема у другій базі
rf_type	нема у другій базі
rf_zanr	нема у другій базі

Поля, яких нема в жодній із баз

dc_order	count	відсутнє у другій базі
dc_order	race_id	відсутнє у другій базі
dc_order	realiz	відсутнє у другій базі
dc_order	recr_name	відсутнє у другій базі
dc_order	type_recr	відсутнє у другій базі
rf_plan	count	відсутнє у другій базі

Таблиця	Поле	База	Властивості	
dt_raice			Тип поля	Джерело стрічок
	class	1	nvarchar (50)	«віп»; «стандарт»
		2	nvarchar (30)	«віп»; «стандарт»; «євро»

Рис. 5. Результат роботи оператора

Встановлено відношення з однаковою схемою та з однаковим характером наповнення: dc_race у першій базі даних та dc_order у другій. Встановлено атрибути-синоніми: race_id та dcrace_id. Хоча у xml-файлі описано відношення з атрибутом race_id, то встановлено, що вміст цього відношення не відповідає вмісту аналогічного у першій базі даних.

Інтеграція даних — це об'єднання даних, які знаходяться у різних системах (Базах даних). Існують такі методи інтеграції [4, 5]:

- консолідація даних — це збір даних із територіально віддалених або різноплатформенних джерел DB_i даних в єдине сховище даних DW з метою їх подальшого опрацювання та аналізу:

$$DW.rel = DB_1.r \cup \ldots \cup DB_n.r \xrightarrow{\text{consolid}} S_{DS}; \qquad (10)$$

- операція федералізації даних полягає у витяганні даних з первинних систем на підставі зовнішніх вимог. Усі необхідні перетворення даних здійснюються при їх витяганні з первинних файлів.

$$S_{DS} : S_{DS} : \sigma_{fed}\ rm = DB_1.r(DB_1.r) \cup \ldots \cup \sigma_{fed}\ rm = DB_n.r(DB_n.r). \quad (11)$$

Агрегація даних — це обчислення узагальнених значень на основі даних відношень вимірів для підтримки стратегічного або тактичного керування з детальних даних:

$$rel = Ag(DB_1.r, \ldots, DB_n..r).$$

Запит про довільні дані Se_{simple} — у користувачів має бути можливість запиту будь-якого елементу даних, незалежно від його формату та моделі даних. Здійснюється на основі ключових слів key_word та каталогу IP Cg:

$$Se_{\text{simple}} : \sigma_{key\_word(Cg)}. \tag{12}$$

Приклад запиту: вибрати інформацію про журнали, у навзі яких є слово «Системні». Інформація зберігається у напівструктрованому вигляді.

Структуровані запити будуються з використанням SQL та подібних мов. За допомогою каталогу визначається, чи містить джерело, у якому здійснюватиметься пошук, структуровану інформацію. Якщо це так, то виконується запит безпосередньо до джерела даних. В іншому випадку запит продовжується виконуватись по каталогу даних у вигляді пошуку ключових слів:

$$Se_{\text{structured}} : \sigma_{key\_word(Cg)}, \sigma(Source). \tag{13}$$

Приклад запиту: Select * from tour where race_id = «Пам'ятник Шевченку». Перш за все, агент визначає джерела, де зберігається інформація про рейс, співставляє їх, а потім вибираються ті, де за характеристикою рейсу є Пам'ятник Шевченку.

Запити до метаданих мають забезпечувати можливості:

- отримання даних про джерело відповіді та місцезнаходження джерела;
- визначення елементів даних у просторі даних, що можуть залежати від заданого елементу даних, і підтримка гіпотетичних запитів;
- визначення рівня невірогідності відповіді.

$$\text{Se}_{\text{meta}} : \sigma_{user\_param}(Cg), \tag{14}$$

де *user_param* — множина параметрів користувача (вимог до запиту), його профілю або вимог, які ставляться до рішення.

Приклад запиту: знайти розміщення всіх джерел, які мають більше, ніж три спільних відношення.

Бінарні операції. Простори даних можуть вкладатися одне в одне (наприклад, простір даних району вкладається в простір даних області), і вони можуть перекриватися (наприклад, простір даних у сфері туризму перекривається з просторами даних оздоровчо-лікувальної, історичної сфери та сфери управління природними ресурсами).

Бінарними операціями над множинами IP є операція об'єднання ПД та операція перетину ПД: $\boldsymbol{O_{Pb}} = \{\cup, \cap\}$.

Уведемо бінарну операцію об'єднання просторів даних:

$$DS_1 \cup DS_2 =< DB_1 \cup DB_2, DW_1 \cup DW_2, Wb_1 \cup Wb_2, Nd_1 \cup Nd_2,$$

$$Cr_1 \cup Cr_2,\ ODW_1 \cup ODW_2 >$$

$$Cg = profile(Agent(Cg_1) \cup Agent(Cg_2)),$$

$$Int = Int_1 = Int_2,$$

$$Se = Se_1 = Se_2,$$

$$EM = EM_1 = EM_2.$$

Уведемо операцію перетину просторів даних:

$$DS_1 \cap DS_2 =< DB_1 \cap DB_2,\ DW_1 \cap DW_2,\ Wb_1 \cap Wb_2,\ Nd_1 \cap Nd_2,$$

$$Cr_1 \cap Cr_2,\ ODW_1 \cap ODW_2, >$$

$$Cg = Cg_1 \cap Cg_2,$$

$$Wo = Wo_1 \cap Wo_2,$$

$$Int = Int_1 \cap Int_2,$$

$$Se = Se_1 \cap Se_2,$$

$$EM = EM_1 = EM_2.$$

Предикати на інформаційних ресурсах

Предикати на інформаційних ресурсах — реєстр ресурсів, що містить найбільш базову інформацію про кожного з них: джерело, ім'я, місцезнаходження в джерелі, розмір, дату створення і власника та ін., а також результат порівняння подібності структур даних один із одним.

Для організації робот із розрізненими джерелами використовують словник термінів та понять (ключових слів) *Dic*, який містить синонімічний опис одного і того ж концепту в різних джерелах даних. Заповнення словника даних на початку здійснюється за допомогою розробленої онтології предметної області, пізніше — автоматизовано: Metadata(DS) $\cup$ $Dic \Rightarrow ODW$.

Зміна стану простору даних полягає не тільки у зміні наповнення інформаційних ресурсів, але й зміні стану інформації про них. Наприклад, якщо за допомогою агента визначення структури джерела ми визначаємо схему даних певної бази даних, то тим самим ми зберігаємо інформацію у реєстрі ресурсів, змінивши його стан.

Виділимо предикати алгебраїчної системи класу «простір даних».

Нульарний предикат Ω_{F_0} : повертає TRUE, якщо для заданого інформаційного ресурсу *Ir* відомого його структури даних, та FALSE у іншому випадку.

Предикат порівняння структур даних інформаційних ресурсів $\Omega_{eq}(Ir_1, Ir_2) \to Dic$.

Формування алгебраїчних виразів

Алгебраїчні вирази формуватимуть користувачі ПД для аналізу інформації, що зберігається у різних джерелах, виходячи з їхнього профілю. Вони зада-

ватимуть необхідні їм операції з множини $\wp$ над елементами множини *DS*. Оскільки профіль визначає перелік джерел, до яких користувач має доступ, та операції над ними, то це дозволить уникнути проблеми ведення додаткової раціоналізації виразів в умовах певної розмитості у визначенні операцій.

ВИСНОВКИ

Розроблено алгебраїчну систему класу «простір даних», яка складається з множини інформаційних продуктів, предикатів та операцій на них. Це дозволило розробити операції консолідації та пошуку даних із різнотипних джерел, структура даних яких наперед невідома. Розроблено інтелектуальний агент визначення структури джерела даних шляхом порівняння структур джерел даних, наявних у ПД, із структурами джерел даних, які входитимуть в ПД, що дозволило сформувати єдиний тип запитів до джерел даних.

Новизна роботи полягає в поданні простору даних як алгебраїчної системи. Уведено операції над просторами даних.

Практична цінність полягає у визначенні основних задач і компонент простору даних та зв'язки між ними.

Подальші дослідження стосуватимуться формалізації методів пошуку неструктурованих, напівструктурованих та суворо структурованих даних.

ЛІТЕРАТУРА

1. *Мальцев А.И.* Алгебраические системы. — М.: Наука, 1970. — 392 с.
2. *Аграновский А.В., Арутюнян Р.Э.* Индексация массивов документов. — http://www.scandocs.ru/page.jsp?pk=node_1185787748359.
3. *Su Q., Widom J.* Indexing Relational Database Content Offline for Efficient Keyword-Based Search. Proceedings of the Ninth International Database Engineering and Applications Symposium (IDEAS), 25–27 july. — Canada, Montreal. — 2005. — P. 297–306.
4. *Шаховська Н.Б.* Простори даних: поняття та призначення // Матеріали конф. CSIT-2007. — Львів. — 2007. — С. 269–277.
5. *Шаховська Н.Б.* Простір даних області наукових досліджень // Моделювання та інформаційні технології. — 2009. — № 45. — С.132–140.
6. *Чернов А.А.* Становление глобального информационного общества: проблемы и перспективы: монографія. — М.: «Дашков и К», 2003. — 232 с.

Надійшла 25.03.2009