УДК 004.9

Н. Б. Шаховська
Національний університет «Львівська політехніка»
вул. Ст. Бандери, 28, 79013 Львів, Україна

Організація просторів даних для забезпечення якості консолідованих даних

Представлено метод співставлення структур даних, формалізовано характеристики якості консолідованих даних у просторах даних. Уведено поняття корисності даних з джерел даних. Розроблено архітектуру системи оцінювання якості різнотипних даних. Уведено метод уніфікації доступу до текстових і структурованих джерел даних. Уведено модель сховища консолідованих даних та операції зменшення невизначеності.

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

Вступ

На сьогоднішній день задача консолідації даних (системної інтеграції різноманітних неузгоджених інформаційних ресурсів) виникає досить часто. Актуальність роботи визначається такими обставинами.

Опрацювання інформаційних ресурсів, що використовують різні моделі даних, схеми керування тощо вимагає розроблення уніфікованого методу доступу до них для того, щоб позбавити кінцевого користувача необхідності вивчати та використовувати різні засоби опрацювання даних. Необхідність у цьому виникає в організацій, робота яких полягає в опрацюванні великої кількості різнотипних, взаємонезалежних джерел даних, для яких не всі семантичні взаємозв'язки відомі і вказані. У деяких випадках семантичні зв'язки невідомі через невизначену кількість початкових джерел або через брак кваліфікованих людей у визначенні таких зв'язків. У інших випадках не всі семантичні зв'язки необхідні для класифікації послуг користувачам. Тому у користувачів немає єдиної схеми, за якою вони можуть створювати запити.

Унаслідок керування різнотипними даними виникає задача їхньої якості — 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 їхній основі. Проте, використання даних із декількох джерел, наперед неузгоджених і з невідомими

структурами, призводить до того, що якість даних різко знижується і вже не може задовольняти потреб користувача через неузгодженість форматів, різне представлення тощо.

Зміна масштабів і рівня задач — від оперативного опрацювання до аналітичного — призвела до необхідності:

— визначення авторства даних і джерела, з якого вони отримані;

— забезпечення цілісності даних — у системах зберігаються метадані, а не самі об'єкти;

— усунення дублювання даних, що надходять із різних джерел, визначення довіри до джерела даних, що є різними для різних областей та різних груп користувачів;

— зменшення невизначеності, яка виникає внаслідок різнотипної реалізації систем, з яких інтегруються дані.

Аналіз літературних джерел і постановка задачі

Опрацюванням таких даних дослідники займалися з 70-х років 20-го ст. Розроблені моделі та метамови опрацювання різнотипних даних. Проте існуючі на сьогодні моделі та методи стосуються або лише наперед відомих типів даних (здебільшого, реляційних баз даних [1–3, 5]), або вирішують лише часткові задачі опрацювання різнотипних даних — наприклад, індексування для пришвидшення пошуку [6, 8, 9]. Тому виникає необхідність управління розрізненою інформацією, а саме її подання у зрозумілому для користувачів вигляді (навіть якщо вони не знають особливостей організації структур цього джерела даних) та опрацювання (пошуку, інтеграції, видобуванні нових знань тощо) [4, 7, 10].

Одним із базових завдань опрацювання різнотипних даних є їхня інтеграція в сховище. Розроблені на сьогодні методи інтеграції даних за своєю функціональністю поділяються на два типи: інтеграцію веб-застосувань [10] та інтеграцію на основі сховищ даних [7]. Проте проведений аналіз літературних джерел показав, що для опрацювання інформації від усіх об'єктів галузі необхідно поєднати обидва типи інтеграцій та вдосконалити наявні моделі даних.

За усієї важливості відомих результатів, теоретичні та експериментальні дослідження повинні розвиватися в наступному напрямку: розроблення ефективних засобів опрацювання даних з різнотипних інформаційних ресурсів у просторі даних і вироблення засад і критеріїв оцінювання якості інтегрованих даних.

Основний виклад

Для уніфікації роботи з різнорідними джерелами даних введемо ряд означень. Вважатимемо, що усі джерела даних є інформаційним продуктом того чи іншого типу.

Інформаційний ресурс (ІР) — Ir-документи і масиви документів у інформаційних системах, що організовані для багаторазового використання та вирішення проблем користувача.

Структура даних ІР (СДІР) — загальна властивість інформаційного ресурсу, опис складних інформаційних об'єктів засобами простіших типів даних. Характеризується: множиною допустимих значень даної структури; множиною допусти-

мих операцій; характером організованості.

Інформаційний продукт (ІП) — документований інформаційний ресурс, підготовлений відповідно до потреб користувачів і поданий у формі товару. Інформаційними продуктами є програмні продукти, текстові файли, веб-сторінки, електронні таблиці, xml-файли, бази даних, сховища даних та інша інформація $Ip = \langle Ir, Rl \rangle$, де Rl — методи доступу.

Каталог ІП — метадані про ІП: $Pl \cup Rl \to Cg$. Описує місцезнаходження ІП Pl, його СДІР, методи доступу до ІР Rl тощо.

Множину усіх інформаційних продуктів ***Ip*** предметної області становить *простір даних*

$$DS = <\mathbf{DB}, \mathbf{DW}, \mathbf{Wb}, \mathbf{Nd}, \mathbf{Gr}>, \quad (1)$$

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

Стан інформаційного продукту S_{Ip} — зафіксований у певний момент часу його інформаційний ресурс *Ir* та відомості про ІП (каталог даних *Cg*): $S_{Ip} =< Ir, Cg >$.

Стан простору даних S_{DS} — множина станів усіх інформаційних продуктів предметної області та відношень між ними $S_{DS} = S_{Ip1} \cup ... \cup S_{Ip\,n}$.

Говорячи про інформаційний продукт, матимемо на увазі його вміст (інформаційний ресурс *Ir*), а також множину відомостей про нього (каталог даних *Cg*).

Хоча інформаційні продукти, що входять до простору даних (ПД), за своїм характером є різними та керуються різними платформами, проте вони усі *виконують однакову роль*: надають дані для простору даних через фіксацію свого стану та забезпечують виконання притаманних для них операцій, причому ці операції та їхні результати є визначені для усього простору даних.

Для опрацювання інформаційних ресурсів необхідно визначати його вміст і структуру даних [11, 12]. Для роботи з ІП використано операції реляційної алгебри. Визначення СДІР здійснюється за допомогою інтелектуального агента та полягає у доповненні *Cg* новими даними про структури даних ІП:

$$f_{Ip} : Ip \xrightarrow{Agent} DS, \quad (2)$$

де *Cg* — каталог простору даних; *Ip.Cg* — каталог ІП *Ip*.

Агент *Agent* задано кортежем

$$Agent = <Cg, M, Dic, MB, Dif, H>, \quad (3)$$

де *Cg* — інформація про джерела, що вже є у ПД; *M* — середовище керування моделями; *Dic* — база знань агента про власні можливості (терміни-синоніми, що позначають у джерелах одні й ті ж властивості); *MB* — база знань про методи доступу до джерела, $MB: \sigma_{evd=Date()}(Cg) \cap Ip.Rl$ (де $\sigma_{evd=Date()}$ — операція селекції за

сьогоднішньою датою, $Ip.Rl = \pi_{Rl}(Ip)$ — методи доступу Rl до інформаційного продукту Ip); *Dif* — множина розбіжностей, які виявив агент; H — відношення відповідності між новим джерелом Ip_{new}, $Y \subset Ip_{new}(Ir)$ та опрацьованими. Позначимо через $Dic_{X=Y}$ вибірку з $Dic_{X=Y}$ кортежів, для яких значення за атрибутами X,Y співпадають. Відношення відповідності сформовано наступним чином.

1. Змістовний взаємозв'язок доменів — відношення еквівалентності: $H : X \times Y \to \{0,1\}$, $Dic_{X=Y}$. $H(x,y) = 1$, якщо за атрибутами x і y об'єкти співпадають, $H(x,y) = 0$ — в іншому випадку. Якщо $H(x,y) = 1$ і $Dic_{X \neq Y}$, то доповнюємо *Dic* новими синонімами, $P^Y(IP_{new}) = 1$, де $P^Y(IP_{new})$ — довіра до атрибута IP_{new}.

2. Відношення перетворення: перевіряємо, чи атрибути X, Y належать схемі $Dic : \forall x \in X, \exists y \in Y : X \neq Y, X \subset Dic, Y \subset Dic \Rightarrow H(x, F(x)) = 1$, додаємо зв'язок між існуючими елементами *Dic*, $P^Y(IP_{new}) = 0{,}5$.

3. Відношення узагальнення: Y — узагальнення X (вилучаємо з *Dic* X та додаємо Y):

$$F : X \to Y, y \neq F(x),$$
$$\forall x \in X, \exists y \in Y : X \subset Dic, Y \not\subset Dic \Rightarrow H(x,y) < 1,$$
$$P^Y(IP_{new}) = 0{,}5;\ P^X(IP_{new}) = 0.$$

4. Однозначне співпадіння: X — деталізація Y (Y не додаємо):

$$F : X \to Y,\ F(x) = y,$$
$$\forall x \in dom(X),\ \exists! y \in dom(Y) : X \subset Dic,\ y \in Dic \Rightarrow H(x,y) < 1,$$
$$P^Y(IP_{new}) = 0,\ P^X(IP_{new}) = 0{,}5.$$

5. Ізоморфізм доменів $F : X \to Y$, $F^{-1} : Y \to X$ (додаємо X, Y, $P^Y(IP_{new}) = 0{,}5$; $P^X(Cg) = 0{,}5$).

Об'єкт, що заданий кортежем $a = \{x_1, x_2, ..., x_n\}$ в одній схемі даних, співпадає з об'єктом, що заданий кортежем $b = \{y_1, y_2, ..., y_m\}$ в іншій схемі даних, якщо $H_{ij} : X_i \times Y_j \to \{0,1\}$ виконується рівність $H_{ij}(x_i, y_j) = 1$. Множину пар індексів (i, j), для яких задані функції H_{ij}, позначимо $\Omega = \{(i,j)\}$, i = Num(x), j = Num(y), $x, y \in Dic$. Функція відповідності об'єктів $H : A \times B \to \{0,1\}$ задана таким чином:

$$\forall (i,j) : H(x_i, y_j) = 1 \Rightarrow H(a,b) = 1,$$
$$\forall (i,j) : H(x_i, y_j) \neq 1,\ X \in Dif,\ Y \in Dif \Rightarrow H(a,b) = 0.$$

Для організації простору даних необхідною умовою є уніфікація джерел даних. Проте, як випливає із визначення простору даних (1), джерелами його інформації є також напівструктуровані дані — текст і веб-сайти. Для ефективного пошуку та аналізу напівструктурованої текстової інформації використаємо семан-

тичну мережу. Особливості структури семантичних мереж: вузли семантичних мереж являють собою концепти об'єктів, подій, станів, які, у свою чергу, визначаються із словника *Dic*; вважається, що довільні вузли одного концепту відносяться до різних значень, якщо вони невідзначені; дуги семантичних мереж створюють відношення між вузлами-концептами (помітки над дугами вказуватимуть на тип відношення).

Семантичну мережу, що побудована на основі аналізу термів напівструктурованого джерела інформації Γ, подано як

$$\Gamma = \{V,D\}, \tag{4}$$

де $V = \{v_i\}$ — множина вершин мережі; $V \in Dic$, $D = \{d_j\}$ — множина дуг.

Дуги між елементами визначають взаємозв'язки між вершинами і задають послідовність пошуку концептів (їхню важливість). Вершини є елементами сховища консолідованих даних cg'.

Функція трансформації напівструктурованого тексту та веб-сайтів у семантичну мережу: $S(E) \to N, E \in \mathbf{Wb} \vee E \in \mathbf{Nd}$. Між двома будь-якими елементами Y_i, Y_j словника даних *Dic*, $Y_i \in Dic$, $Y_j \in Dic$, існує відображення деякого порядку $i = \overline{1,M}, \forall Y_i : \exists n, \Gamma^n(Y_i) = \{Y_i\}$, де $\Gamma(Y_i) = \{Y_j : \exists S(Y_i, Y_j) \vee S(Y_j, Y_i)\}$.

Формуються підграфи для кожного Y_i, такі що в підпункті вузол вихідного параметра один, а інші вузли — це вхідні поняття, що описують обмеження на атрибути $\{X_k, 1 \le k \le N\} \leftarrow Y_i \leftarrow \{X_l, 1 \le l \le N\}$, тут $X_k \leftarrow Y_i \Rightarrow S(X_k, Y_i) : Y_i \leftarrow X_k = S(Y_i, X_k)$. Крім цього, до графу так само входять всі вхідні поняття, які використовуються як обмеження $\forall Y_i : \Gamma'^k(Y_i)$, де $\Gamma'(Y_i) = \{X_j : \exists S(Y_i, X_j) \vee S(X_j, Y_i)\}$, $\Gamma'^2(Y_i) = \Gamma'(Y_i, \Gamma'(X_j)) = \{X_k : \exists S(Y_i, X_k) \vee S(X_k, X_i)\}$.

Після визначення структури даних інформаційного продукту необхідно сформувати сховище даних і здійснити інтеграцію даних. Оскільки у просторі даних усі ІП вважаються рівнозначними (ієрархія або відсутня, або задана неявно) і всі вони можуть бути приймачами даних для тих чи інших задач, то традиційний підхід консолідації даних (ETL) використовувати не можна.

Тому введемо модель сховища консолідованих даних, яка будується на основі співставлення структур даних інтелектуальним агентом.

Схема сховища консолідованих даних Cg' — множина імен атрибутів $\{C_1, C_2, ..., C_n\}$, значення яких є чіткими; $\{C\_unk_1, C\_unk_2, C\_unk_p\}$ з нечіткими або недетермінованими значеннями; множину імен атрибутів $Unk = \{Unk_1, Unk_2, \ldots, Unk_m\}$, доменами яких є числові дані, що моделюють імовірнісні дані, значення функції приналежності нечітких множин тощо (зберігання значень рівнів довіри $P(j)$, $P^{atr}(j)$); схему словника синонімів *Dic* та схему каталогу даних *Cg*:

$$Cg' = <\{C_1, C_2, ..., C_n\}, \{C\_unk_1, C\_unk_2, C\_unk_p\}, \{Unk_1, Unk_2, \ldots, Unk_m\}, Dic, Cg>.$$

Кортеж консолідованих даних cons_data — інформаційний опис об'єкта *t* джерела даних *S*, що поданий у вигляді множини (кортежу) значень характерис-

тик (атрибутів), підмножина значень атрибутів якого містить дані про об'єкт, джерело даних і синонімічні назви об'єкта, причому ці дані можуть бути неповні, нечіткі чи недетерміновані дані.

Приклади кортежу консолідованих даних для різних типів джерел даних.

1. Реляційна база даних — у цьому випадку використовується розширений реляційний кортеж t_{rel} :

$$dc = t_{rel} \cup Unk;\ t_{rel} = \{c_1,...,c_n\} \cup \{c\_unk_1,...,c\_unk_m\},$$

де $\{c_1,...,c_n\}$ — значення чітких атрибутів; $\{c\_unk_1,...,c\_unk_m\}$ — значення атрибутів із невизначеністю.

2. Сховище даних — поєднує дані з відношень фактів і вимірів. Множину значень вимірів і характеристик фактів подано як кортеж t_{dw} :

$$dc = t_{dw} \cup Unk; t_{dw} = \{c_{11},...,c_{1n}\} \cup ... \cup \{c_{k1},...,ca_{kn}\} \cup \{c_{rf1},...,c_{rfl}\} \cup$$
$$\cup \{c\_unk_{11},...,c\_unk_{1m}\} \cup ... \cup \{c\_unk_{k1},...,c\_unk_{ks}\} \cup \{c\_unk_{rf1},...,c\_unk_{rft}\},$$

де c_{ij} — значення чіткої *j*-ї характеристики *i*-го виміру; c_{rf_j} — значення *j*-ї характеристики відношення фактів; c_unk_{ij} — значення *j*-го атрибутів із невизначеністю *i*-го виміру; $c\_unk_{rf_j}$ — значення *j*-ї характеристики з невизначеністю відношення фактів.

3. Напівструктурований текст — описується значення вершин семантичної мережі та ступінь належності цих значень до об'єктів, назви яких описані у словнику синонімів t_{text} :

$$dc = t_{text} \cup Unk\ ;\ t_{text} = \{c_1,...,c_n\} \cup \{c\_unk_1,...,c\_unk_m\}.$$

Отже кортеж консолідованих даних *dc* — це множина значень характеристик об'єкта сутності, описана як

$$dc = <\boldsymbol{C}, \boldsymbol{C\_unk}, \boldsymbol{Unk}, \{dic\},\{cg\}>, \quad (5)$$

де $\boldsymbol{C}$ — підмножина значень атрибутів із чіткими значеннями, $C = t_{rel} \cup t_{dw} \cup t_{text}$; $\boldsymbol{C\_unk}$ — підмножина значень атрибутів із нечіткими та недетермінованими значеннями; $\boldsymbol{Unk}$ — підмножина значень рівнів довіри до значень атрибутів $\boldsymbol{C\_unk}$ $P^{atr}(j)$ та кортежу загалом $P(j)$ і $meta(\boldsymbol{C\_unk},\boldsymbol{Unk}) = 1$; $\{dic\}$ — множина значень словника даних; $\{cg\}$ — множина значень каталогу даних.

Тоді *сховище консолідованих даних* cg' — це відношення зі схемою Cg' та множиною кортежів консолідованих даних $\boldsymbol{dc}$.

Модель сховища консолідованих даних містить дані з усіх типів джерел простору даних.

Вважаємо, що між описами об'єктів сховища консолідованих даних можна побудувати мережу (встановити приховані залежності).

Для руху мережею залежних об'єктів, інформація про які наявна у cg', модифіковано оператор визначення предка

$$Up\_c_{X=x_1,\sigma_X(Dic)}(cg') = \sigma_{X=x_1}(\sigma_{X=Y,\sigma_X(Dic)}(cg'))$$

та оператор визначення нащадка

$$Down\_c_{X=x_2,\sigma_X(Dic)}(cg') = \sigma_{Y=x_2,\sigma_X(Dic)}(cg').$$

Ці оператори використано у модифікованому операторі усунення невизначеності у мереженій структурі консолідованих даних:

$$\sigma_{X=x_1,Val=v,\sigma_X(Dic)}(cg') = Heir\_c\begin{pmatrix}\sigma^{cons}_{X=x_1,Val=v,\sigma_X(Dic)}(cg'),\\ Down\_c_{X=x_1,Val=Null,\sigma_X(Dic)}(cg')\\ Cg'.Unk_X = \min\left(Cg'.Unk_X, P^X(\sigma_X(Dic))\right)\end{pmatrix}.$$

Для оцінювання якості даних застосовано загальний методичний підхід до виділення адекватної номенклатури стандартизованих в ISO 9126 базових характеристик і субхарактеристик. Визначено неперервну функцію якості Q.

Функціональна придатність — відносна кількість об'єктів, що потрапили у сховище консолідованих даних, до загальної кількості об'єктів, наявних у ІП. Оскільки методи інтеграції, що застосовуються до СД, неефективно застосовувати до ПД, то визначення функціональної придатності є однією з базових характеристик, що досліджується у роботі:

$$z_1 = \frac{|cg'|}{|Ip_i.Ir|}. \qquad (6)$$

Для забезпечення функціональної придатності додаються ІП з відповідністю СДІР «ізоморфізм доменів» та «узагальнення» (результат інтелектуального агента).

Коректність або достовірність даних — відносна кількість описів об'єктів з ІП, які не містять дефектів і помилок, до загальної кількості об'єктів у просторі даних:

$$z_2 = \frac{|\sigma_P(Ip)|}{|cg'|}. \qquad (7)$$

Для забезпечення коректності додаються ІП з відповідністю СДІР «еквівалентність».

Використовуванність 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 проаналізовано, оскільки існують розроблені

методи (наприклад, метод критичних робіт) та засоби визначення завантаженості ресурсів.

Практичність — визначає корисність застосування консолідованих даних для певних користувачів. У цю групу показників входять субхарактеристики, які відображають зрозумілість, зручність освоєння, системну ефективність і простоту використання даних. Деякі субхарактеристики оцінюють економічними показниками — витратами праці і часу спеціалістів на реалізацію певних функцій взаємодії з даними. У ПД оцінка практичності здійснюватиметься за допомогою функції корисності прийнятих рішень:

$$z_3 = \frac{\sum_{j=1}^{m} k_j v_j(r_j)}{|cg'|},$$

де $0 < k_j < 1$, $j = 1,2,\ldots,m$; $\sum_{j=1}^{m} k_j = 1$. Функцію v_j, що виражає оцінку значення r_j, вважаємо j-ю компонентою функції корисності, а k_j — вагою, що визначає критерій R_j.

Окрім того, цей показник враховує залежність прийнятого рішення від рівня довіри. Для забезпечення практичності додаються ІП з відповідністю СДІР «еквівалентність» та «перетворення».

Супроводжуваність даних — зручність і ефективність виправлення, удосконалення або адаптації структури та змісту описів даних залежно від змін у зовнішньому середовищі застосування. У просторах даних характеристика супроводжуваність пов'язана зі зміною даних про ІП у каталозі:

$$z_4 = \frac{\left|\sigma_{meta\_upd}\left(Ip_i.Cg\right)\right|}{|cg'|}. \tag{8}$$

Для забезпечення супроводжуваності додаються ІП з відповідністю СДІР «співпадіння».

Мобільність характеризується тривалістю і трудомісткістю їхньої інсталяції, адаптації та заміщення при перенесенні на інші апаратні та операційні платформи. Для оцінки супроводжуваності розроблено методи та засоби, тому в роботі ця характеристика даних не розглядається.

Визначимо корисність даних з ІП стосовно прийняття рішення на їхній основі. Оцінку корисності даних здійснено наступним чином. Є множина керованих змінних $Z = (z_1, z_2, z_3, z_4)$. Визначено неперервну функцію якості Q. *Твердження*: цільова функція якості при обмеженнях має глобальний максимум:

$$Q(z_1,\ldots,z_4) = \sum_{i=1}^{4}\left(\sum_{k} r_k z_i \prod_{j=1} P_{ij}\right) \to \max, \tag{9}$$

$$1 \geq z_1 \geq 0,75,\ \ z_3 \geq 0,\ \ 0,25 \geq z_1 - z_2 \geq 0,$$
$$1 \geq z_4 \geq 0,5,\ \ z_1 t_s \leq T,\ \ z_1 v \leq V_1,\ \ z_4 c \leq V_2,$$

де j вказує на інформаційний продукт; P_{ij} — рівень довіри до інформаційного продукту j для рішення k; r_k — оцінка рішення k; V_1 — загальна вартість завантаження об'єктів; V_2 — загальна вартість модифікації описів; T — загальний час завантаження; t_s — середній час завантаження одного об'єкта; v — середня вартість завантаження (модифікації) одного об'єкта.

Це задача нелінійної оптимізації з лінійними обмеженнями, яка вирішується певними методами (наприклад, методом релаксації).

Оцінимо якість консолідованої інформації. Для цього реалізовано порівняння з еталонними даними:

$$Q_i^e = \sum_i n_i z_i^e,\quad Q'^e_{const} = \sum_i k_i Q_i^e, \tag{10}$$

$$Q_{const} = Q'_{const} / Q'^e_{const}. \tag{11}$$

Поряд із фактичним оцінюванням якості консолідованої інформації (9) необхідно провести оцінювання якості еталонного зразка (10), що відображає найкраще прийняте рішення. Потім виконується нормування фактичної оцінки за формулою (11).

Спроектовано підсистему забезпечення і підтримки якості даних у просторі даних, яка призначена для реалізації алгоритмів і процедур, що 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. Процес оцінювання якості консолідованих даних складається з трьох стадій: встановлення вимог до якості консолідованих даних, підготовка до оцінювання та процедура оцінювання.

Укрупнений алгоритм визначення відповідності рішення еталонному подано так.

1. Отримання параметрів вибірки еталонних і консолідованих даних.
2. Визначення критеріїв оптимальності.
3. Визначення найкращого значення за критерієм.
4. Визначення найгіршого значення за критерієм.
5. Пошук прямо пов'язаних даних (через відношення *Dic*).
6. Групування вибраних даних.
7. Обрання тих консолідованих даних, у яких агреговані кількісні характеристики, рівні середньому значенню критеріїв 2 і 3.
8. Визначення джерела даних, з якого отримано інформацію, що задовольняє 7.

Кількість параметрів співставлення n є різною для кожного типу рішення, що приймається. Обрання того чи іншого параметра фізично означатиме, що при співставленні знайдених даних та еталону за обраним атрибутом буде виконуватись операція агрегації для визначення релевантності. Чим більше параметрів буде включено до агрегування, тим точнішим буде отриманий результат співставлення.

Обрання усіх параметрів означає максимальний рівень довіри до отриманих результатів співставлення.

Можуть бути отримані такі результати співставлення v:

— еталон не має аналога, $v = 0$;

— знайдені дані не відповідають жодному еталону, $v = 1$;

— часткове співпадіння (при агрегації даних еталону та знайдених даних отримано кількісні характеристики, які не рівні між собою) , $v = 2$;

— повне співпадіння, $v = 3$.

Схема опрацювання інформаційних продуктів у просторі даних подана на рисунку.

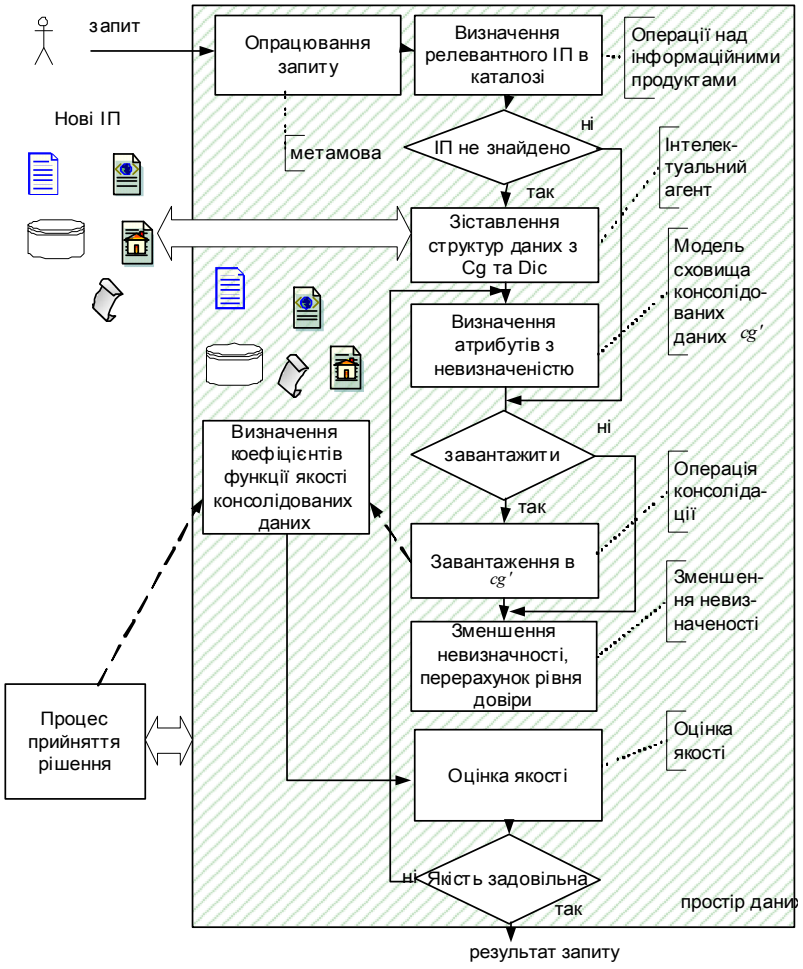

Схема опрацювання інформаційних продуктів

Висновки

У роботі розроблено методологію опрацювання різнотипних джерел даних з метою підвищення якості консолідованих даних шляхом використання розроблених теоретичних засад і програмних засобів організації просторів даних як множини інформаційних продуктів та операцій над ними. У результаті виконання цієї роботи одержані наступні результати.

1. Описано підхід до побудови інтелектуального агента визначення структури джерела даних шляхом порівняння структур джерел даних, наявних у ПД, із структурами джерел даних, які входитимуть у ПД, що дозволило сформувати єдиний тип запитів до джерел даних з урахуванням ступеня довіри до джерела та отримати коректні відповіді на сформовані запити.

2. Розроблено модель сховища консолідованих даних, що дозволило корегувати подання даних для особи, що приймає рішення, та контролювати процес завантаження даних.

3. Розроблено метод визначення якості консолідованих даних на основі формалізації стандарту ISO 9126, що дозволило визначати придатність цих даних для подальшого прийняття рішень.

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 компьютерных и информационных технологий. — 2007. — № 1. — С. 26–33.

2. *Основные* концепции и подходы при создании контекстно-поисковых систем на основе реляционных баз данных [Електронний ресурс]. — Режим доступу: http://www.citforum.ru/database/articles/search_sys.shtml.

3. *Особенности* построения хранилищ данных [Електронний ресурс]. — Режим доступу: http://citforum.uar.net/seminars/cis99/sch.shtml/

4. *Michael J. Franklin.* A First Tutorial on Dataspaces / Michael J. Franklin, Alon Y. Halevy, David Maier // PVLDB. — 2008. — **1**(2). — P. 1516–1517.

5. *Калиниченко Л.А.* СИНТЕЗ — язык определения, проектирования и программирования интероперабельных сред неоднородных информационных ресурсов / Л.А. Калиниченко // ИПИ РАН. — 1993. — 115 с.

6. *Ступников С.А.* Формальные методы и модели в композиционных инфраструктурах распределенных информационных систем / С.А. Ступников // Системы и средства інформатики. Специальный выпуск. — ИПИ РАН, 2005. — 304 с.

7. M. Franklin, A. Halevy and D. Maier: From Databases to Dataspaces: A New Abstraction for Information Management. ACM SIGMOD Record 34, N 4 (December 2005), P. 27–33.

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

9. *Alon Y. Halevy.* Answering Queries Using Views: A Survey / Alon Y. Halevy // The VLDB Journal. — 2001. — P. 270–294.

10. *Оперативная* интеграция данных на основе XML: системная архитектура BizQuery / К.В. Антипин, А.В. Фомичев, М.Н. Гринев [и др.] // Труды Института системного программирования РАН. — 2004. — Т. 5. —С. 157–174.

11. *Шаховська Н.Б.* Математичне та програмне забезпечення сховищ та просторів даних: монографія / Н.Б. Шаховська. — Львів, 2010. — 237 с

12. *Шаховська Н.Б.* Формалізація простору даних за допомогою алгебраїчної системи / Н.Б. Шаховська // Радіоелектроніка, інформатика, управління. — 2010. — № 1. — С. 102–109.

Надійшла до редакції 02.06.2011