УДК 004.652

А.Ю. Берко, В.А. Висоцька

Національний університет «Львівська політехніка», Львів

СЕМАНТИЧНА ІНТЕГРАЦІЯ НЕПОВНИХ ТА НЕТОЧНИХ ДАНИХ

У роботі розглянуто основні проблеми семантичної інтеграції різнорідних даних за наявності в них неповних чи неточних даних. Запропоновано застосування онтологій для класифікації та інтерпретації різноманітних типів невизначеностей. Застосування онтологій дозволяє компенсувати зміст відсутніх чи неточних значень в інтегрованих наборах даних

Ключові слова: *Інтеграція даних, синтаксис, структура, семантика даних, невизначеність, Null, онтологія, класифікація, інтерпретація*

Вступ

Сьогодні інтеграція є одним з перспективних напрямів формування інформаційних ресурсів загального користування, зокрема, на основі сервісно-орієнтованої архітектури. Сьогодні на ринку інформаційних технологій багато провідних виробників пропонують власні інструментарії вирішення проблем інтеграції на різних рівнях, таких як інтеграція бізнес-процесів *(Business Ptrocess Integration – BPI)*, інтеграція корпоративних застосувань (Entetrprise Application Integration- EAI), інтеграція корпоративних платформ *(Platform Integration - EPI)* інтеграція даних або, як часто її називають, інтеграція корпоративної інформації *(Enterprise Information Integration – EII)*. Інтеграція даних – це завдання об'єднання даних, отриманих з різних джерел з метою подання користувачеві їх об'єднаного подання [2, 5].

Одним з важливих напрямів інтеграції в багатьох сферах застосування інформаційних технологій є інтеграція даних. Проблеми інтеграції даних є актуальними з огляду на активний розвиток корпоративних та суспільних інформаційних ресурсів, зростання їх обсягів та різноманіття способів подання. При цьому часто виникають проблеми, пов'язані з неповнотою, неточністю чи відсутністю даних у вхідних наборах. Проблеми цього роду мають декілька аспектів – способи позначення неповних чи відсутніх даних, структурування даних за умови їх неповноти, інтерпретація невизначеностей чи відсутності даних.

У запропонованій роботі розглянуто один з аспектів інтеграції інформаційних ресурсів – інтеграцію семантики даних за умов їх неточності, невизначеності чи неповноти.

Семантична інтеграція даних

Головними завданнями інтеграції даних є формування повного і несуперечного набору на основі множини різнорідних вхідних даних, отриманих з різноманітних джерел. Основні теоретичні положення та принципи інтеграції сформульовано у [2]. Для досягнення кінцевої мети інтеграції необхідно забезпечити узгодження їх синтаксису, структури і семантики [5]. У ході вирішення такого роду задач виникає низка проблем, від розв'язання яких залежать властивості кінцевого результату. На рівні інтеграції синтаксису даних:

- це неоднозначність чи суперечність алфавітів;
- невідповідність типів даних та форматів;
- невідповідність синтаксичних обмежень.

На рівні інтеграції структур даних:

- невідповідність способів визначення одиниць даних;
- суперечності видів та способів побудови зв'язків;
- різноманітність способів впорядкування даних [5].

Семантична складова процесу інтеграції є однією з найважливіших та найскладніших, оскільки проблеми синтаксису та структури, загалом, вирішують на технічному та технологічному рівні. Формування узгодженої інтерпретації інтегрованих даних є неможливим без участі людини, а також застосування методів та засобів інтелектуального опрацювання даних.

На рівні інтеграції семантики виникають такі проблеми [4]:

- суперечності у визначенні концептів;
- неоднозначність чи різночитання імен;
- застосування несумісних метрик;
- суперечності у визначенні відношень між даними;
- суперечності обмежень та аксіом;
- неоднозначність інтерпретації значень.

Ці проблеми вирішують різними засобами, такими як тезауруси чи словники даних, семантичний аналіз даних, онтології [5], які вважають найперспективнішим з відомих сьогодні засобом семантичної

інтеграції. Основні принципи та методику інтеграції даних на основі онтологій викладено у [4].

Проблематика формування та застосування онтологій як засобу опису семантики даних у процесах інтеграції гетерогенних інформаційних ресурсів має декілька підходів [4], для кожного з яких характерними є власні моделі, способи та засоби вирішення.

Інтеграція даних на основі єдиної онтології. У цьому випадку для явної специфікації семантики різних наборів даних формують єдину глобальну онтологію зі спільними узгодженими розподіленими ресурсами [4].

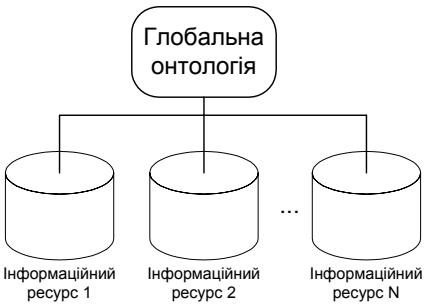

Рис. 1. Схема семантичної інтеграції даних на основі єдиної онтології

Єдину онтологію може бути сформовано двома методами :

- шляхом розподілу – при цьому утворюється глобальний опис концептів, відношень та функцій інтерпретації з розподіленими словниками, який застосовують для специфікації семантики кожного з наборів даних, які підлягають інтеграції;
- шляхом інтеграції – такий спосіб передбачає формування та поповнення глобальної онтології як результатів узгодженого об'єднання словникових ресурсів локальних онтологій, сформованих для наборів даних, які підлягають інтеграції (рис.1).

Особливістю семантичної інтеграції даних на основі єдиної онтології є спільне використання її ресурсів для опису семантики кожного вхідного набору даних.

Інтеграція даних на основі множини онтологій. У такому випадку кожен вхідний набір даних для семантичної інтеграції описують власною онтологією, яка не пов’язана з іншими і оперує власними нерозподіленими словниковими ресурсами [4]. Процес семантичної інтеграції у цьому випадку ґрунтується на узгодженні, взаємодії та обміні ресурсами локальних онтологій (рис. 2).

Проблема інтеграції даних при цьому вимагає застосування таких методів та засобів побудови і опрацювання онтологій, які забезпечують їх спільне застосування у формуванні єдиного семантичного простору інтегрованих даних.

Гібридний підхід до інтеграції даних на основі онтологій. Такий спосіб семантичної інтеграції поєднує особливості двох попередньо описаних методів. За аналогією з єдиною онтологією, в даному випадку створюють спільний, узгоджений розподілений ресурс.

Використання цього ресурсу для специфікації семантики вхідних наборів даних відбувається через їх власні локальні онтології [5] (рис. 3).

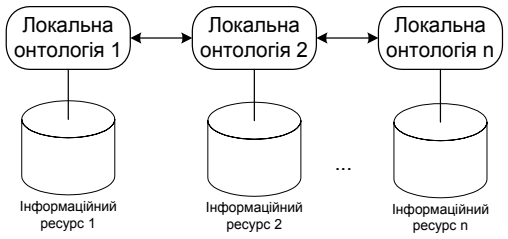

Рис. 2. Схема семантичної інтеграції даних на основі множини онтологій

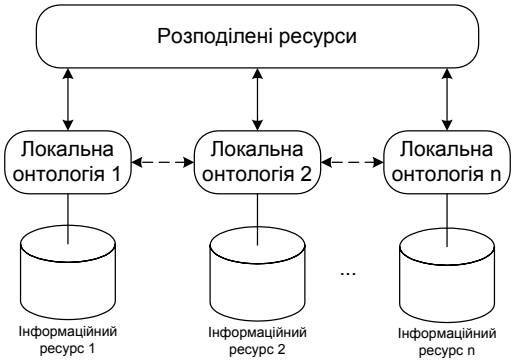

Рис. 3. Схема гібридної семантичної інтеграції даних

Аналогічно до попереднього випадку семантику інформаційного ресурсу, що підлягає інтеграції, описує окрема онтологія. Але для сумісності локальних онтологій створюють глобальні розподілені словникові ресурси, в яких зосереджено базові терміни і поняття, спільні для предметної області інтегрованих даних.

Особливо важливим у процесах інтеграції, які ґрунтуються на онтологіях є коректне опрацювання невідомих, неповних чи відсутніх даних. Неоднозначність їх походження, суті, природи, та порядку опрацювання є потенційним джерелом суперечностей у змісті інтегрованих даних та виконанні запитів до них.

Неповнота та неточність даних

Проблема опрацювання, та інтерпретації невизначеностей виникає як на рівні інтеграції синтаксису та структури даних, так і на рівні інтеграції семантики. Загалом до категорії невизначеностей відносять відсутність значень даних, їх неповноту, неточність чи неможливість отримання [1]. В реляційних базах даних для зображення невизначених понять застосовують умовне позначення – псевдоконстанту Null [3]. Це забезпечує вирішення основних структурних та семантичних проблем, викликаних відсу-

тністю даних. Застосування Null замість значень у рядках таблиць дозволяє забезпечити дотримання цілісності реляційних структур та правил позначення даних. Але при цьому зберігаються проблеми, пов'язані зі специфікою маніпулювання позначенням Null, дії з яким суттєво відрізняються від роботи з явно заданими даними.

У слабкоструктурованих даних, та даних без попередньо описаної структури, таких як тексти, потоки, сенсорні дані, невизначеність подають через відсутність значень відповідних елементів або застосування, так званих, "порожніх" одиниць, наприклад, XML-тег без значень вигляду <*tagexample/*>, порожній символьний рядок – "", порожній елементи списку – ,, тощо. Такі позначення дозволяють описати певні структурні одиниці, значення яких явним чином подати неможливо. Такий підхід узгоджується з концепцією слабкоструктурованих даних, хоча часто і приводить до неоднозначного їх сприйняття.

При наявності неповних чи неточних значень як у базах даних так і в слабкоструктурованих даних можуть виникати певні проблеми на рівні семантики, пов'язані з інтерпретацією невизначеностей. Неоднозначність змісту, який пов'язують з позначенням Null, часом призводить до некоректного виконання запитів при роботі з неповними чи неточними даними [1], а у процесах інтеграції є потенційним джерелом виникнення неоднорідних чи суперечних даних в інтегрованому наборі. Крім того, загалом, немає єдиного трактування поняття Null. Його розглядають або як спеціальне значення даних, константу, що має специфічну інтерпретацію [3], або як умовну позначку, прапорець, що сигналізує про відсутність даних та потребу застосування спеціальних прийомів для їх опрацювання [1]. Окрім того, невизначеності в даних мають різну природу та походження, що впливає на підходи та методи роботи з ними. Зокрема, К. Дейт [1] описує низку найпоширеніших видів невизначеностей у даних:

- значення є неприпустимим;
- значення є невідомим;
- значення не існує;
- значення визначити неможливо;
- значення не є достовірним;
- значення не отримано;
- значенням є порожня множина, тощо.

Цей список може бути продовжено, оскільки в конкретних ситуаціях можуть виникати інші варіанти та причини відсутності даних. Така неоднорідність є особливо суттєвою у процесах семантичної інтеграції даних з невизначеностями, коли необхідно забезпечити єдину з зрозумілу інтерпретацію значень, отриманих з різнорідних джерел інформації.

Застосування онтологій для опису невизначеностей в даних

Проблема неповноти, неточності чи відсутності даних є особливо чутливою у процесах інтеграції, коли необхідно об'єднати та узгодити дані, які мають різну природу різні способи зображення та різну інтерпретацію. При цьому в інтегрованому наборі даних потрібно забезпечити не лише об'єднане подання даних, алей визначити порядок подальшого опрацювання різного роду невизначеностей. Наприклад неприпустимість використання значення не є еквівалентною його тимчасовій відсутності, а невідоме значення не є еквівалентним до значення, котре існує але не надане.

Одним із засобів вирішення проблеми узгодження семантики різнорідних даних та усунення суперечностей в їх змісті і опрацюванні є онтології [4]. Зокрема, їх можна застосувати для вирішення проблеми визначення способів опрацювання та інтерпретації невизначеностей в інтеграційних процесах неповних чи неточних даних. Семантична інтеграція, що ґрунтується на онтологічному підході передбачає побудову для кожного вхідного набору даних побудови власної онтології, яка містить вичерпний опис складу і змісту даних.

Використання онтологій для явного визначення неявних чи прихованих знань про дані та їх зміст дозволяє вирішити значну частину проблем семантичної неоднорідності даних та появу в них неповних, неточних значень чи їх відсутність.

Методика семантичної інтеграції неповних чи неточних даних на основі онтологій передбачає вирішення таких завдань[5]:

- *визначення порядку застосування онтології* – роль та загальна архітектура онтології, як засобу опису семантики неповних даних значною мірою впливають на способи її формування та подання;
- *вибір способів та засобів зображення онтології* – в залежності від призначення та застосування онтологій, способи їх зображення можуть бути дуже відмінними у різних випадках;
- *застосування рівності понять* – у інтеграційних процесах, в залежності від специфіки даних, які підлягають інтеграції, наявність в них неповних неточних недостовірних значень чи їх відсутність, поняття рівності може мати різну інтерпретації, зокрема, застосовують такі варіанти, як точна рівність, частковий збіг, еквівалентність, подібність, тощо;
- *побудова онтологій* – перед початком процесів семантичної інтеграції неповних чи неточних даних, для кожного локального набору необхідно створити власну онтологію, орієнтовану на взаємодії з онтологіями інших наборів даних.

В загальному випадку, онтологію даних пода-

ють як трійку множин загального виду [4]

$$O = <X, R, F>,$$

де X – множина концептів , які описують певні поняття, їх зміст та властивості; R – множина відношень та зв'язків між концептами; F – множина аксіом (функцій інтерпретації).

За допомогою онтології можна із потрібним рівнем деталізації описати семантику вхідних наборів даних, які підлягають інтеграції.

Для опису невизначеностей, як структурної та семантичної одиниці даних пропонується застосувати спеціальний концепт – клас *"Невизначеність"*, елементами якого є різноманітні варіанти невизначеностей, наприклад

Undefined={U_1≈"неможливо", U_2≈"невідомо", U_3≈"не існує",U_4≈"не визначено", … }.

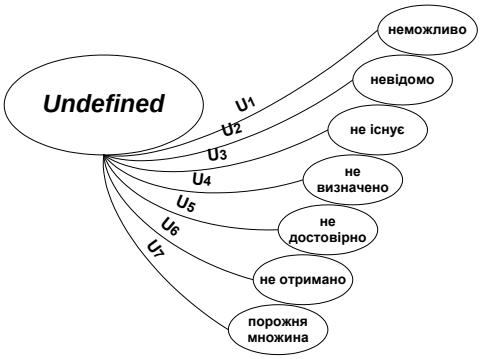

Рис. 4. Структура класу "Навизначеність"

Така класифікація невизначеностей необхідна для уникнення суперечностей, які виникають:

- при побудові глобальної онтології для вхідних наборів даних за умови їх неповноти чи неточності;
- при організації обміну і взаємодії локальних онтологій вхідних наборів даних з невизначеностями;
- при побудові розподілених словників загального користування у поєднанні з локальними онтологіями;
- при визначенні способів подальшого опрацювання невизначеностей;
- при виконанні запитів до інтегрованих даних, в складі яких присутні невизначеності.

Окремо в складі онтології передбачають множину функцій інтерпретації невизначеностей, які встановлюють відповідність між конкретними невизначеностями в даних та елементами, відповідного класу в складі онтології

$$F_i():UN \rightarrow \text{Undefined}, i = 1, 2, 3, \ldots$$

де UN = {null, "", </>, ,, …} – множина можливих позначень невизначеностей; Undefined – відповідний клас онтології, який описує їх природу та характер.

На рис. 4. подано приклад структури класу онтології, який описує семантику та природу невизначеностей в складі даних.

Аналогічний опис може бути подано в складі загального визначення онтології даних у форматі XMl. Такий спосіб застосовують для подання онтологій у формі метаданих за допомогою таких засобах як RDF *(Resource Data Framework)* та OWL *(Web Ontology Language)* [5].

Визначення такого класу для типізованих даних [1] означає створення відповідного типу, який є підтипом усіх можливих типів.

В складі такого типу визначають множину відповідних констант, кожна з яких слугує позначенням належності даних до тої чи іншої категорії невизначеностей з можливістю їх подальшої інтерпретації.

```
<?xml version="1.0"?>
<Ontology ontologyname= "O1">
            …
<Class name="Undefined">
 <Subclass Subclassname="U1" Subclasstype="неможливо">
Undefined type1 </Subclass>
 <Subclass Subclassname="U2" Subclasstype="невідомо">
Undefined type2 </Subclass>
 <Subclass Subclassname="U3" Subclastsype="не існує">
Undefined type3 </Subclass>
<Subclass Subclassname="U4" Subclasstype="не визначено">
Undefined type4 </Subclass>
 <Subclass Subclassname="U5" Subclasstype="не  достовір-
но">
Undefined type5 </Subclass>
  <Subclass Subclassname="U5" Subclasstype="не отримано">
Undefined type6 </Subclass>
  <Subclass Subclassname="U6" Subclasstype="порожня
множина">
Undefined type7 </Subclass>
</Class>
…
</Ontology>
```

Рис. 5. Приклад опису класу "Невизначеність " в складі онтології

Залежно від стратегії семантичної інтеграції даних [4] – на основі спільної онтології, на основі локальних онтологій чи на основі гібридного підходу клас *"Невизначеність"* може бути застосовано різним чином. У випадку інтеграції на основі спільної єдиної онтології клас є спільним для опису невизначеностей в усіх вхідних наборах даних, які підлягають інтеграції.

Такий підхід забезпечує однорідність інтегрованих даних та однозначність інтерпретації невизначеностей для усіх вхідних наборів. Однак побудова спільної глобальної онтології не завжди є можливою [4].

У випадку застосування локальних онтологій для семантичної інтеграції даних, у кожній з них описують клас *"Невизначеність"*. При цьому можливими є варіанти:

- клас типу *"Невизначеність"* є однаковим в усіх локальних онтологіях;
- клас типу *"Невизначеність"* є специфічним в кожній локальній онтології;

- класи типу *"Невизначеність"* є спільними в частині локальних онтологій.

В першому випадку спільний клас *"Невизначеність"* забезпечує єдину інтерпретацію та єдиний порядок опрацювання невизначеностей як у всіх вхідних наборах даних так і в інтегрованому наборі. Такий спосіб значною мірою збігається з випадком семантичної інтеграції на основі єдиної онтології.

У другому випадку для спільного опрацювання даних описаних за допомогою специфічних локальних онтологій виникає потреба формування механізму їх взаємодії, зокрема, в частині інтерпретації та опрацювання невизначеностей. Це може бути реалізовано шляхом визначення відповідності між підкласами класів типу *"Невизначеність"* та їх атрибутами за принципом

$$O_i.UT_i.UST^{k}_{i} \rightarrow O_j.UT_j.UST^{k}_{j},$$

де $O_i.UT_i.UST^{k}_{i}$ – посилання на підклас UST^{k}_{i} класу *"Невизначеність"* UT_i онтології O_i; для $i,j = 1, 2, \ldots, N$, $i \neq j$, де N – кількість локальних онтологій вхідних наборів даних.

Така відповідність дозволяє узгодити інтерпретацію невизначеностей у різних вхідних наборах даних та визначити єдину стратегію їх опрацювання в інтегрованому наборі.

У випадку збігу та відмінностей у визначенні класу *"Невизначеність"* в різних локальних онтологіях, для узгодження роботи з ними і інтегрованих даних застосовують комбінацію двох попередніх підходів.

Гібридний підхід до семантичної інтеграції неповних чи неточних даних, який поєднує принципи інтеграції на основі глобальної та локальних онтологій дозволяє застосувати такі варіанти опрацювання невизначеностей як специфічного класу:

- формування класу *"Невизначеність"* як спільного розподіленого ресурсу,
- формування класу *"Невизначеність"* як власного специфічного ресурсу локальних онтологій,
- формування класу *"Невизначеність"* як у складі локальних так і розподілених ресурсів.

Перший випадок використання такого опису невизначеностей повністю збігається за принципами зі стратегією семантичної інтеграції на основі єдиної глобальної онтології.

Другий випадок реалізують за принципом опрацювання невизначеностей при застосування стратегії локальних онтологій із визначенням специфічних класів типу *"Невизначеність"* для кожної з них.

Третій випадок інтеграції даних з невизначеностями передбачає можливості як локального так і глобального їх опису. При цьому способи опрацювання класу *"Невизначеність"* потребують узгодження як на рівні локальних онтологія як це описано вище, так і на рівні взаємодії локальних і розподілених ресурсів.

У такий спосіб для кожного невизначеного елемента у вхідних наборах даних може бути задано часткову інтерпретацію, яка не приводить до усунення невизначеності, але дає можливість охарактеризувати її природу, походження, зміст, вплив на інші значення, а також визначає способи і засоби опрацювання відповідних даних в інтегрованому наборі.

Згідно з таким підходом, в інтегрованому наборі даних від позначень типу Null, порожніх чи відсутніх елементів пропонується перейти до їх умовного позначення відповідно з належністю до відповідного підкласу класу *"Невизначеність"*. Тобто кожне значення буде подано у вигляді відповідної константи (псевдоконстанти) як величини, що належить до спеціального типу, який є підтипом усіх можливих типів даних.

Отже часткова інтерпретація невизначеностей за допомогою визначення спеціальних класів в онтологіях даних переводить проблему їх опрацювання на принципово інший рівень – від опрацювання невідомих, неповних, недостовірних понять, до опрацювання явно визначених концептів.

Висновки

Запропонована класифікація та інтерпретація невизначеностей дозволяє не лише їх охарактеризувати і підвищити рівень детермінованості даних. Такий спосіб дозволяє уникнути можливих суперечностей при інтеграції інформаційних ресурсів, що можуть виникати за рахунок різної природи невизначеностей, а також визначити шляхи і процедури опрацювання інтегрованих даних, що містять в своєму складі невизначеності.

Описаний в цій роботі онтологічний підхід до інтеграції неповних та неточних є універсальним для різних інтеграційних методів, моделей та технологій, а також проблемно- та об'єкто-незалежним. Це, в свою чергу, забезпечує семантичну інтероперабельність даних за умови наявності в них невизначеностей, які мають різну природу, зміст та порядок опрацювання.

Список літератури

1. Date C.J. Database in Depth: Relational Theory for Practitioners / Cristopher Date. – CA.: O'Reilly, 2005. – 240 p.

2. Lenzerini M. Data Integration: A Theoretical Perspective / Marco Lenzerini // Proc. of the ACM Symp. on Principles of Database. Systems, 2002. – P. 233-246.

3. Rubinson C. Nulls, Three-Valued Logic, and Ambiguity in SQL : Critiquing Date's Critique. / Claude Rubinson // SIGMOD Record Vol. 36, No. 4, December 2007, 2007. – P. 137-143.

4. Wache H. Ontology-Based Integration of Information – A Survey of Existing Approaches / H. Wache, T. Vogele,

U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hubner // Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing, Seattle. – USA, August 4-5, 2001. – P. 108-118.

5. Берко А.Ю. Методи інтеграції синтаксису різнорідних даних у системах електронного контент-бізнесу / Андрій Берко // Інформаційні системи та мережі: Вісник Націон. ун-ту «Львівська політехніка». – Львів, 2008. – № 621. – С. 19-28.

Надійшла до редколегії 10.03.2009

Рецензент: д-р техн. наук, проф. Я.М. Матвійчук, Інститут підприємництва та перспективних технологій, Львів.

СЕМАНТИЧЕСКАЯ ИНТЕГРАЦИЯ НЕПОЛНЫХ И НЕТОЧНЫХ ДАННЫХ

А.Ю. Берко, В.А. Высоцкая

В работе рассмотрены основные проблемы семантической интеграции разнородных данных при наличии в них неполных или неточных данных. Предложено применения онтологий для классификации и интерпретации разнообразных типов неопределенностей. Применение онтологий разрешает компенсировать содержание отсутствующих или неточных значений в интегрированных наборах данных

Ключевые слова: *интеграция данных, синтаксис, структура, семантика данных, неопределенность, Null, онтология, классификация, интерпретация.*

SEMANTIC INTEGRATION OF INCOMPLETE AND IMPERFECT DATA

A.Y. Berko, V.A. Vysotska

In work are considered main problems to semantic integration heterogeneous given at presence in them incomplete or inexact data. Offered using ontology for categorization and interpreting the varied types of the uncertainties. Using ontology allows to compensate the contents being absent or inexact importances in integrated set.

Keyword*: integration data, syntax, structure, semantics data, uncertainty, Null, ontology, categorization, interpretation.*