УДК 51.001.57+004.652.4+004.827

Н.Б. Шаховська

МЕТОДИ ОПРАЦЮВАННЯ КОНСОЛІДОВАНИХ ДАНИХ ЗА ДОПОМОГОЮ ПРОСТОРІВ ДАНИХ

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

Вступ

Інформаційне суспільство – суспільство, в якому створення, передавання, перетворення, використання, інтеграція і маніпулювання інформацією – важлива господарська, політична і культурна діяльність. Специфікою цього виду суспільства є те, задача консолідації даних (об'єднання даних, розміщених у різних, наперед неузгоджених джерелах) виникає досить часто. Так, для університету прикладом консолідації є формування наукових звітів, визначення показників успішності та якості навчання, формування рейтингу кафедри тощо; для обласної адміністрації – це визначення критичних показників розвитку регіону на основі даних, отриманих з організацій державної та недержавної форми власності.

Постановка проблеми в загальному вигляді

Причини виникнення проблеми подання та опрацювання різнотипових даних:

– глобалізація суспільства – прагнення знайти нові дані шляхом консолідації даних з джерел, призначених для локального застосування;

– прагнення зберігати дані «вічно» – навіть потужні системи керування базами даних (СКБД) також мають обмеження на кількість даних;

– визначення авторства даних;

– опрацювання різнотипових даних – даних, що зберігаються в різних системах з різними методами доступу та особливостям організації даних;

– забезпечення цілісності даних – в системах зберігаються метадані, а не самі об'єкти;

– дублювання даних, що надходять з різних джерел, довіра до джерела даних;

– невизначеність, яка виникає внаслідок різнотипового проектування систем, з яких консолідуються дані;

– визначення операцій, виконання яких привело до зміни даних;

– зміна класу задач дослідників – від статистичних до інтелектуальних (пошук залежностей, «важливих даних»).

Тому при опрацюванні даних з різних джерел та керування ними 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 та з невідомими структурами, якість таких даних різко знижується і вже не може задовольняти потреб користувача через неузгодженість форматів, різне представлення тощо.

Аналіз останніх досліджень

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. [1 – 4]. Розроблені моделі та метамови опрацювання різнотипних даних. 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 виникає проблема керування 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).

Одним із базових завдань опрацювання різнотипних даних є інтеграція. Розроблені на сьогодні методи інтеграції даних за своєю функціональністю поділяються на два типи: інтеграція веб-застосувань та інтеграція на основі сховищ даних. Проте специфіка опрацювання консолідованих даних, а саме [5]:

– наявність великої кількості різнотипних джерел даних, не виключаються протиріччя та суперечливість інформації;
– наявність великої кількості моделей зберігання джерел даних (реляційні бази даних (РБД), сховища даних (СД), напівструктуровані текстові файли, електронні таблиці, статичні та динамічні веб-сайти тощо);
– відсутність або недотримання розробниками стандартів називання елементів систем

вказує на те, що для врахування інформації від усіх об’єктів галузі необхідно поєднати обидва типи інтеграції та вдосконалити наявні моделі зберігання даних.

Проблеми керування інформацією виникають в організацій, робота яких полягає в опрацюванні великої кількості різнотипних, взаємонезалежних джерел даних. Такий тип системи отримав назву простір даних (ПД). 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 і вказані. У користувачів, які працюють з просторами даних, немає єдиної схеми, за якою вони можуть створювати запити. У деяких випадках семантичні зв'язки невідомі через невизначену кількість початкових джерел, які залучені до ПД, або через брак кваліфікованих людей у визначенні таких зв'язків. У інших випадках, не всі семантичні зв'язки необхідні для класифікації послуг користувачам.

Отже, робота присвячена вирішенню актуальної проблеми подання та опрацювання різнотипних джерел та підвищення якості консолідованих даних. Для цього необхідно формалізувати поняття простору даних і визначити операцій над ним.

Постановка задачі

Поняття «консолідація» широко використовується в інших сферах діяльності, зокрема, в керуванні ресурсами. Так, можна зустріти таке визначення консолідації: консолідація (бізнесу) – злиття або поглинання малих компаній у більші. Структуризація процесів керування бізнесом та опрацювання різнотипних даних показана на рис. 1.

Рис. 1. Структуризація процесів керування бізнесом та опрацювання різнотипних даних

Консолідовані дані – це отримані 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.

Наведена аналогія з бізнесом показує, що якість даних у джерелах даних для вирішення задач, для яких це джерело призначене, є достатньою. Але коли йде мова про консолідовані дані, необхідно здійснювати узгодження та перетворення даних, оскільки фізичне об'єднання без попереднього опрацювання різко знижує їх якість.

Інформаційні продукти (ІП) певної предметної області та консолідовані дані становитимуть простір даних. Однією із задач, яка виникатиме у процесі консолідації, є невизначеність даних, що є результатом дублювання, неточності, відсутності, протиріччя даних (рис. 2).

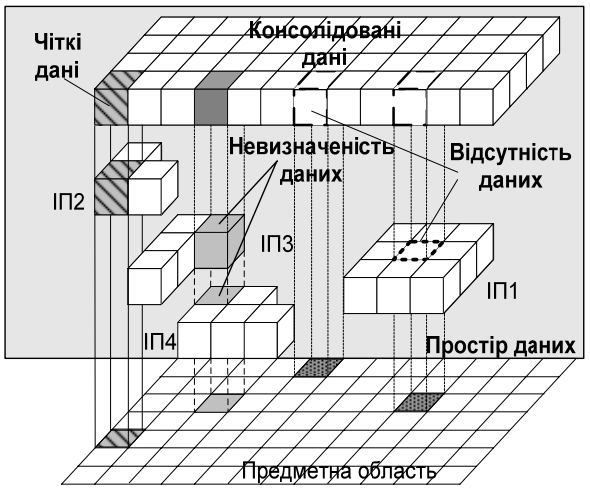

Рис. 2. Схема консолідації даних

Іншою задачею є визначення та узгодження схем даних інформаційних ресурсів. Існуючі методи (теорія інтеграції, канонічні системи, онтологічний пошук) 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. Також одною з перепон використання проаналізованих методів інтеграції є те, що розробники наявних ІП не завжди дотримувалися стандартів під час розроблення схем даних. Аналіз можливостей застосування існуючих стандартів показав, що розроблення словника даних дозволить уникнути цю проблему та частково уніфікувати схеми джерел даних.

Наявні методи опрацювання розрізнених даних потребують вдосконалення, оскільки у просторі даних наперед невідомо всіх учасників та їх структури даних.

Основний матеріал

Введемо ряд означень.

Інформаційний ресурс (ІР) – масиви документів у інформаційних системах: бібліотеках, архівах, фондах, банках даних, інших видах інформаційних систем, організовані для багаторазово використання та вирішення проблем користувача.

Структура даних ІР (СДІР) – загальна властивість інформаційного ресурсу, з яким взаємодіє та або інша програма, опис складних інформаційних об'єктів засобами простіших типів даних. Характеризується: множиною допустимих значень; множиною допустимих 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.

Каталог ІП – метадані про ІП. Описує місцезнаходження ІП, його СДІР, методи доступу до ІР тощо.

Множина інформаційних продуктів **Ip** предметної області містить найповнішу інформацію про предметну область, а отже якість прийнятих керівних рішень на її основі є найвищою. Множину всіх інформаційних продуктів предметної області назвемо *простором даних*.

$$DS = \langle \mathbf{DB}, \mathbf{DW}, \mathbf{Wb}, \mathbf{Nd}, \mathbf{Gr} \rangle,$$

де **DB**, **DW**, **Wb**, **Nd**, **Gr** – інформаційні

продукти, що подають множини баз даних, сховищ даних, веб-сторінок, текстових файлів, електронних таблиць, графічних даних відповідно.

Стан інформаційного продукту – зафіксований у певний момент часу його інформаційний ресурс *Ir* та відомості про ІП (каталог даних) *Cg*. Стан інформаційного продукту будемо позначати:

$$S_{Ip}: S_{Ip} =< Ir, Cg >.$$

Стан простору даних – множина станів усіх інформаційних продуктів предметної області та відношень між ними. Стан ПД позначатимемо S_{DS}.

Множину інформаційних продуктів простору даних, операцій над ІР в них та предикатів на множині ***Ip*** назвемо *алгебраїчною системою сигнатури простір даних* [6].

$$DS_a = <\mathbf{Ip}, \Omega_P, \Omega_F>, \qquad (1),$$

де ***Ip*** = *DS* – скінченна множина станів інформаційних продуктів певної предметної галузі (баз даних DB, сховищ даних DW, статичних Web-сторінок Wb, текстових даних Nd, графічних та мультимедійних даних Gr), $\Omega_P = \{O_{P0}, \boldsymbol{O_{Pu}}, \boldsymbol{O_{Pb}}\}$ – множина операцій над інформаційними ресурсами ІП, де O_{P0} – нульарна операція, результатом якої є стан заданого ІП у просторі даних; $\boldsymbol{O_{Pu}}$ – множина унарних операцій над простором даних *DS*. Результатом цих операцій є зміна стану простору даних; $\boldsymbol{O_{Pb}}$ – множина бінарних операцій над просторами даних. Результатом цих операцій є утворення нового простору даних; Ω_F – множина предикатів, заданих на множині інформаційних продуктів простору даних. Серед предикатів також є нульмісний предикат Ω_{F0}, результатом якого є TRUE, якщо для заданого інформаційного продукту *Ip* відомо його структури даних IP, та FALSE у іншому випадку.

Алгебраїчна система (1) скінченна, оскільки множина інформаційних продуктів *DS* та їх станів є скінченною.

Говорячи про інформаційний продукт, матимемо на увазі його вміст (інформаційний ресурс), а також множину відомостей про нього (розміщення, схема доступу, швидкість оновлення інформації тощо). Також описуватимемо операції, які виконуються над ІР залежно від його СДІР.

Основною операцією, що виконується над вмістом текстових файлів, ефектронних 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

Отже, хоча інформаційні продукти, 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 усього простору даних.

Результатом нульарної операції над простором даних *DS* є стан заданого інформаційного продукту *Ip*:

$$S_{Ip} = O_{P0}(DS, Ip). \qquad (2)$$

Нульарний оператор є розширенням реляційної операції селекції без задання умови.

Уведено унарні операції алгебраїчної системи сигнатури простір даних. Унарними операціями над просторами даних є шістка:

$$\boldsymbol{O_{Pu}} = \{Se_{simple}, Se_{structured}, Se_{meta}, \sigma_{access}, Agent, Consolid, Ag\}, \qquad (3)$$

де *Agent* – операція визначення СДІР; $Se_{simple}, Se_{structured}, Se_{meta}$ – операції пошуку; σ_{access} – операція доступу.

Визначення СДІР здійснюється за допомогою інтелектуального агента (ІА) і полягає у доповненні *Cg* новими даними про СДІР ІП

$$f_{Ip}(DS) \xrightarrow{Agent} Cg \cup Ip.Cg, \qquad (4)$$

де *Cg* – каталог простору даних, *Ip*.*Cg* – каталог ІП *Ip*.

Агент *Agent* задано кортежем:

$$Agent = \left\langle \begin{matrix} Cg, EM, Dic, \\ Experience\_Base, \\ Solver, Effector \end{matrix} \right\rangle, \quad (5)$$

де *Cg* – 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 агент 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 за декількома джерелами, приведення результатів запитів за джерелами до єдиної структури, відмова у запиті).

В основі роботи агента лежить інформація про джерела, які вже є у просторі. Його задачею є порівняння структур даних джерела даних, що входитиме у простір, із структурами даних джерел у ПД, та визначення різниці. Це дозволило автоматизувати формування запитів, що виконуються у просторі даних. Чим більше джерел здатний «розрізнити» агент, тим точніше буде інформація в *DS* і тим ефективніше можна буде проводити процедури консолідації, пошуку та опрацювання даних у ПД.

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

Консолідація даних – це об'єднання інформаційних ресурсів ІП у сховище консолідованих даних визначеної структури DW.*rel* з метою подальшого опрацювання для прийняття керівних рішень:

$$DW.rel = \langle Ip_1.Ir \cup \ldots \cup Ip_n.Ir;$$
$$Ip_1.Cg \cup \ldots \cup Ip_n.Cg \rangle \xrightarrow{consolid} S_{DS}. \quad (6)$$

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 повинна бути можливість запиту будь-якого елемента даних, незалежно від його формату і моделі даних. Здійснюється на основі множини ключових слів *keyword* та каталогу ІП *Cg*.

$$Se_{simple} : \sigma_{keyword}(Cg). \quad (7)$$

Стуктуровані запити будуються з використанням SQL та подібних мов. За допомогою каталогу визначається джерело, в якому здійснюватиметься пошук, що містить структуровану інформацію. Запит виконується безпосередньо до джерела даних.

$$Se_{structured} : \sigma_{Cg.x='structured'}(\pi_x(\sigma_{keyword}(Ip_1)) \cup \ldots$$
$$\ldots \cup \pi_x(\sigma_{keyword}(Ip_n))). \quad (8)$$

Запити до метаданих мають забезпечуватися можливостями:

– отримання даних про джерело відповіді та місцезнаходження джерела;

– визначення елементів даних у просторі даних, що можуть залежати від заданого елемента даних і підтримка гіпотетичних запитів;

– визначення рівня невірогідності відповіді.

$$\mathrm{Se}_{\mathrm{meta}} : \sigma_{user\_param}(Cg), \quad (9)$$

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

Доступ до кожного з ІП залежить від прав користувача. Права доступу кожного із користувачів до заданого Ip_i вказуються у *Cg*. Під профілем користувача будемо розуміти підмножину каталогу даних, яка вказує на ті ІП, до яких користувач має доступ.

$$\mathrm{profile} : \sigma_{\mathrm{access}=\mathrm{Yes}}(\mathbf{Cg}). \quad (10)$$

ПД можуть вкладатися одне в інший (наприклад, ПД району вкладається

в ПД області), і вони можуть перекриватися (наприклад, ПД в сфері туризму перекривається з ПД оздоровчо-лікувальної, історичної сфери та сфери керування природними ресурсами). Тому в ПД містяться правила розмежування доступу.

Бінарними операціями над множинами ІП є розширені теоретико-множинні операції об'єднання, перетину та різниці:

$$\boldsymbol{O}_{Pb} = \{\cup, \cap, -\}.$$

Уведено бінарну операцію об'єднання просторів даних:

$$DS_3 = DS_1 \cup DS_2:$$
$$\text{profile}(Agent(\mathbf{Cg}_1) \cup Agent(\mathbf{Cg}_2)),$$
$$Cg_3 = Cg_1 \cup Cg_2.$$

Операція об'єднання ПД використовується також для додавання нового інформаційного продукту до простору даних: оскільки використовується множинне представлення інформаційних продуктів ПД, то множина ІП ПД даних може складатися і з одного інформаційного продукту:

$$DS_2 = DS_1 \cup \{Ip\}:$$
$$\text{profile}(Agent(Cg_1) \cup Agent(Ip.Cg)),$$
$$Cg_3 = Cg_1 \cup Ip.Cg.$$

Бінарна операція перетину просторів даних:

$$DS_3 = DS_1 \cap DS_2:$$
$$\text{profile}(Agent(Cg_1) \cap Agent(Cg_2)),$$
$$Cg_3 = Cg_1 \cap Cg_2.$$

Бінарна операція різниці ПД:

$$DS_3 = DS_1 - DS_2: \text{profile}(Agent(Cg_1) -$$
$$- Agent(Cg_2)),\ Cg_3 = Cg_1 - Cg_2.$$

Операція різниці використовується також для вилучення інформаційного продукту з простору даних:

$$DS_2 = DS_1 - \{Ip\}: \text{profile}(Agent(Cg_1) -$$
$$- Agent(Ip.Cg)),\ Cg_3 = Cg_1 - Ip.Cg.$$

Розширені операції об'єднання, перетину та різниці означають теоретико-множинне об'єднання, перетин чи різницю каталогів даних просторів даних. При цьому доступ користувачів до ІП з просторів даних DS_1 та DS_2 визначається профілем, сформованим на основі нового каталогу Cg_3.

Предикати на інформаційних продуктах – реєстр ІП, що містить базову інформацію про кожного з них: джерело, ім'я, місцезнаходження в джерелі, розмір, дату створення, власника та інше, а також результат порівняння подібності структур даних один з одним.

Для організації роботи з 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 (*ODW* – сховище консолідованих даних).

$$\text{Metadata}(DS) \cup Dic \Rightarrow ODW. \quad (11)$$

Зміна стану ПД полягає не тільки у зміні наповнення інформаційних ресурсів ІП, 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 зберігаємо інформацію у реєстрі продуктів, змінивши його стан.

Розроблено предикати алгебраїчної системи сигнатури простір даних.

Нульмісний предикат Ω_{F0}: повертає TRUE, якщо для заданого інформаційного продукту *Ip* відомо його структури даних IP, та FALSE в іншому випадку:

$$\Omega_{F0}(Ip, Dic) : \sigma_{Ip}(Dic) \neq \varnothing. \quad (12)$$

Предикат порівняння структур даних інформаційних ресурсів ІП використовується для визначення відмінностей та подібностей у структурах даних інформаційних ресурсів, що входять до складу простору даних:

$$\Omega_{eq}(Ip_1, Ip_2) \rightarrow Dic.$$

Для аналізу інформації, що зберігається у різних джерелах, користувачі ПД, виходячи з їхнього профілю формуватимуть алгебраїчні вирази. Вони задаватимуть необхідні їм

операції з множини Ω_P над елементами *DS*. Оскільки профіль визначає перелік джерел, до яких користувач має доступ, та операції над ними, то це дозволить уникнути проблеми ведення додаткової раціоналізації виразів за умов певної розмитості у визначенні операцій.

Алгебраїчні вирази – це запити, які формує користувач для отримання необхідних йому даних. Оскільки основою побудови ПД є підтримка подальшого процесу прийняття рішень на основі консолідованих даних, то необхідно проаналізувати вплив цих даних на якість прийнятого рішення. Критерій кращого чи гіршого рішення залежить від предметної області та конкретної задачі. Прикладами критеріїв є: співпадіння прогнозованого плану з реальним, мінімізація кількості вхідних даних, згортка параметрів тощо.

Корисність даних для певного користувача чи групи користувачів залежить також і від ступеня довіри до джерела даних. Тоді визначення ступеня довіри *i*-го користувача до *j*-го джерела даних:

$$Trust(i,j) = \frac{\sum_{k=1}^{n} Trust_k(i,j)}{n}, \qquad (13)$$

де *n* – кількість звернень користувача до ресурсу, $Trust_k(i,j)$ – значення лінгвістичної змінної, що відображає довіру довіри *i*-го користувача до *j*-го джерела даних при *k*-у зверненні.

Для розрахунку загального ступеня довіри до джерела *j* узагальнено формулу (13):

$$Trust_j = \frac{\sum_{i=1}^{m} \left(Trust(i,j)\right)}{n * m}, \qquad (14)$$

де *m* – кількість користувачів, що звертались до ресурсу.

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

$$Trust^{attr}(i,j) = Trust(i,j)\frac{\sum_{k=1}^{n} Trust_k^{attr}(i,j)}{n}, \qquad (15)$$

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

$$v(r) = \sum_{j=1}^{m} k_j v_j(r_j), \qquad (16)$$

де $v_j(r_{j0}) = 0$; $v_j(r_{j*}) = 1$; $0 < k_j < 1$; $j=1,2,\ldots,m$; $\sum_{j=1}^{m} k_j = 1$. Функцію v_j, що виражає оцінку значення r_j, можна вважати *j*-ю компонентою функції корисності, а k_j – вагою, що визначає критерій R_j. У випадку просторів даних вага джерела даних *j* визначається як $k_j = Trust(i,j)$, де *i* є заданим і вказує на конкретного користувача.

Для оцінювання якості даних у ПД застосовано загальний методичний підхід до виділення адекватної номенклатури стандартизованих в ISO 9126 базових характеристик і субхарактеристик.

Функціональна придатність визначається повнотою накопичених об’єктів – відносною кількістю об'єктів або документів, наявних у джерелах даних, до загальної кількості об'єктів, що потрапили у сховище консолідованих даних. Оскільки методи інтеграції, що застосовуються до СД, не можуть застосовуватись до ПД, то визначення функціональної придатності є однією з базових характеристик, що досліджується у роботі:

$$plenitude = \frac{Count(ODW)}{\sum_i Count(source_i)}. \qquad (17)$$

Коректність даних – це ступінь відповідності даних про об'єкти в базах даних реальним об'єктам у заданий момент

часу, що визначається змінами самих об'єктів, некоректних записів про їх стан або некоректними розрахунки їх характеристик. Вибір та встановлення вимог до коректності даних оцінюють за ступенем покриття накопиченими, актуальними і достовірними даними стану і зміни зовнішніх об'єктів, які вони відображають. Оскільки у роботі джерелами даних є не об'єкти предметної області, а ІП, то під коректністю даних будемо розуміти кількісну характеристику, що відображає відносну кількість описів об'єктів з джерел даних, які не містять дефектів і помилок, до загальної кількості об'єктів у ПД:

$$identy = \frac{Count(\sigma_{Trust>0.6}(ODW))}{Count(ODW)}. \qquad (18)$$

Ресурсна економічність у стандарті відображено зайнятістю ресурсів центрального процесора, оперативної, зовнішньої та віртуальної пам'яті тощо. Цей показник у роботі не проаналізовано, оскільки існують розроблені методи (наприклад, метод критичних робіт) та засоби визначення завантаженості ресурсів.

Практичність – важко формалізоване поняття, яке визначає функціональну придатність і 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 можна оцінювати 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 (16).

Супроводжуваність даних відображається зручністю і ефективністю адаптації структури та змісту описів даних залежно від змін у зовнішньому середовищі застосування, а також у вимогах і функціональних специфікаціях замовника. Узагальнено якість супроводжуваності консолідованих даних можна оцінювати потребою ресурсів для її забезпечення і для реалізації. Для оцінки супроводжуваності розроблені методи та засоби (наприклад, технологія ETL – витягнення, трансформування, завантаження), тому в 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 ПД характеристика мобільності пов'язана зі зміною даних про джерела даних у каталозі:

$$actuality = \frac{Count(\sigma_{metadata\_update<30}(ODW))}{Count(ODW)}. \qquad (19)$$

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

$$q = s_1 quality + s_2 v(r) \to Max, \qquad (20)$$

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

$quality = n1 \cdot plenitude + n2 \cdot identy + n3 \cdot actuality$

s_1 – коефіцієнт важливості повноти накопичення даних,

$v(r)$ – значення багатовимірної функції корисності, s_2 – коефіцієнт важливості якості прийнятих рішень,

$$s_1 + s_2 = 1.$$

Схемою сховища консолідованих даних Cg' назвемо скінченну множину імен атрибутів $\{A_1, A_2, ..., A_n\}$, значення яких є чіткими; $\{A\_unk_1, A\_unk_2, A\_unk_p\}$ з нечіткими або недермінованими значеннями; множину імен атрибутів $\{Unk_1, Unk_2, \ldots, Unk_m\}$, доменами яких є числові дані, що моделюють імовірнісні дані, значення функції приналежності нечітких множин; схему словника синонімів *Dic* та схему каталогу даних *Cg*:

$$Cg' = <\{A_1, A_2, ..., A_n\}, \{A\_unk_1, A\_unk_2, A\_unk_p\}, \{Unk_1, Unk_2, \ldots, Unk_m\}, Dic, Cg>.$$

Кортежем консолідованих даних

consolid_data назвемо інформаційний опис об'єкта *t* джерела даних *S*, поданий у вигляді кортежу значень характеристик, підмножина значень якого містить дані про об'єкт, джерело даних та синонімічні назви об'єкта, причому ці дані можуть бути неповні, нечіткі чи недетерміновані дані.

Наведемо приклади кортежу консолідованих даних для різних типів джерел даних.

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

$$consolid\_data = t_{rel} \cup Unk,$$
$$t_{rel} = \{a_1,...,a_n\} \cup \{a\_unk_1,...,a\_unk_m\},$$

де $\{a_1,...,a_n\}$ – значення чітких атрибутів, $\{a\_unk_1,...,a\_unk_m\}$ – значення атрибутів з невизначеністю.

2. Сховище даних – множина значень вимірів та характеристик фактів подано як кортеж t_{dw}:

$$consolid\_data = t_{dw} \cup Unk,$$
$$t_{dw} = \{a_{11},...,a_{1n}\} \cup ... \cup \{a_{k1},...,a_{kn}\} \cup$$
$$\cup \{a_{rf_1},...,a_{rf_l}\} \cup \{a\_unk_{11},...,a\_unk_{1m}\} \cup ...$$
$$... \cup \{a\_unk_{k1},...,a\_unk_{ks}\} \cup$$
$$\cup \{a\_unk_{rf_1},...,a\_unk_{rf_t}\},$$

де a_{ij} – значення чіткої *j*-ї характеристики *i*-го виміру, a_{rf_j} – значення *j*-ї характеристики відношення фактів, a_unk_{ij} – значення *j*-го атрибутів з невизначеністю *i*-го виміру, $a\_unk_{rf_j}$ – значення *j*-ї характеристики з невизначеністю відношення фактів.

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

$$consolid\_data = t_{text} \cup Unk,$$
$$t_{text} = \{a_1,...,a_n\} \cup \{a\_unk_1,...,a\_unk_m\}.$$

Кортеж консолідованих даних *consolid_data* – це множина значень характеристик об'єкта сутності, описана як

$$consolid\_data = <\boldsymbol{C}, \boldsymbol{C\_unk}, \boldsymbol{Unk}, \{dic\}, \{cg\}>,$$

де $\boldsymbol{C}$ – підмножина значень атрибутів із чіткими значеннями, $\boldsymbol{C} = t_{rel} \cup t_{dw} \cup t_{text}$, $\boldsymbol{C\_unk}$ – підмножина значень атрибутів з нечіткими значеннями, $\boldsymbol{Unk}$ – підмножина значень атрибутів із ступенями істинності значень атрибутів $\boldsymbol{C\_unk}$ і $meta(\boldsymbol{C\_unk}, \boldsymbol{Unk})=1$, $\{dic\}$ – множина значень словника даних, $\{cg\}$ – множина значень каталога даних.

Сховищем консолідованих даних cg' назвемо відношення з схемою Cg' та множиною кортежів консолідованих даних ***consolid_data***. Модель сховища консолі дованих даних містить дані з усіх типів джерел ПД. Для опису інформаційних продуктів ПД розроблено структури даних (рис. 3). Відношення tbl_meta містить інформацію про структури джерел даних, що вже є у ПД (In_DS встановлено в TRUE) та джерела, що додається до ПД. Tbl_Path зберігає шляхи усіх джерел даних. Tbl_Oper містить перелік операцій, що виконуються над даними, поданими у різних моделях.

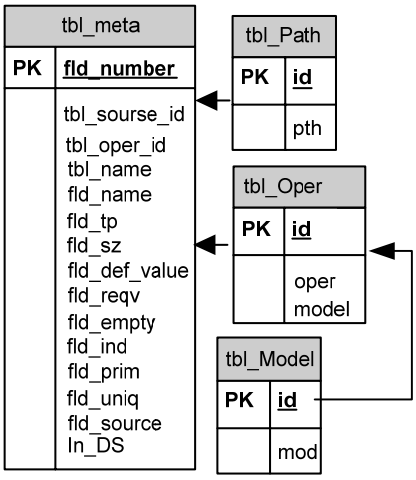

Рис. 3. Схема відношень бази знань для аналізу структури джерел даних

Для кожної операції розроблена процедура, яка запускається за необхідності. При надходженні нового джерела можна взнати, як доступатися до його даних. Tbl_Model містить перелік моделей даних, з якими співпрацюємо у ПД. Алгоритм роботи ІА визначення структури джерела, формального поданого в (5), показаний на рис. 4.

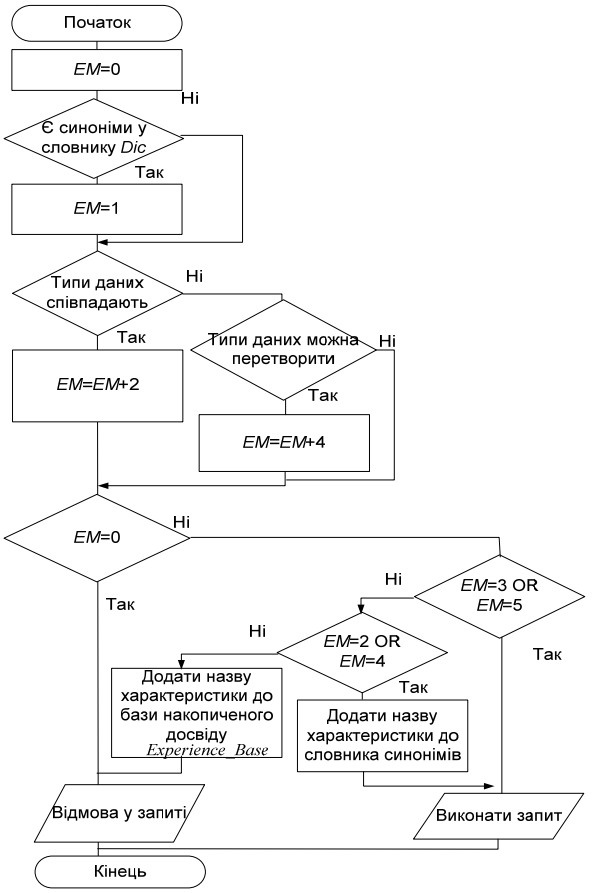

Рис. 4. Алгоритм роботи агента

Для консолідації даних в сховищі консолідованих даних *cg′* вико ристовується каталог, схема якого показана на рис. 5.

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.

Рис. 5. Схема каталогу простору даних

Для встановлення ступеню довіри до джерела використовуватиметься лінгвісттична змінна Trust (рис. 6).

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. 6. Лінгвістична змінна *Trust*

Тоді схема каталогу ПД, поданої на рис. 5, доповнюється відношенням про довіру.

Метадані довіри до джерел даних
Код
Код користувача
Код джерела
Дата
Ступінь довіри

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, яка показана на рис. 7.

Словник синонім
Синонімічна назва
Призначення
Аналог синонімічної назви
Назва стандарту
Посилання

Рис. 7. Структура даних словника синонімів

Введемо елементи метамови ПД. Вважатимемо, що запит q до ПД заданий коректно, якщо він складається з елементів, описаних у *Cg* та *Dic*.

$$q_{object(c_1...c_n)} : par = \left\{ \begin{array}{l} object \in Cg, \\ Trust_{object} > 0, \\ (c_1,...,c_n) \in Dic \end{array} \right\} : par,$$

де *object* – об'єкт, про який йде мова у запиті, $(c_1,...,c_n)$ – назви характеристик об'єкта, *par* – список параметрів запиту. Залежно від типу джерела параметри можуть відігравати роль: параметрів пошуку – в 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:

$$Dic = \{R\} \cup \{\mathrm{Re}l\} \cup \{key\} \cup \{H\},$$

де описані схеми баз даних, сховищ даних, ключових слів текстових файлів, заголовків веб-документів відповідно.

Практична реалізація

Інформатизація ВНЗ викликає задачу консолідації, оскільки університетом розроблено ряд інформаційних систем, які мають обмінюватися між собою інформацією, а також надавати частину інформацію у корпоративне сховище даних ВНЗ з метою подальшого її аналітичного опрацювання:

– пошуку залежностей між отриманими оцінками студентів по предметах та за результатами вступу;

– пошуку дисциплін, у яких показники «Успішність», «Якість» або дуже високі, або дуже низькі;

– пошуку залежностей між результатами наукової діяльності студентів та їх практичними здобутками у вигляді проходження практик, участі в олімпіадах, конкурсах робіт тощо.

Схему взаємодії основних БД університету показано на рис. 8.

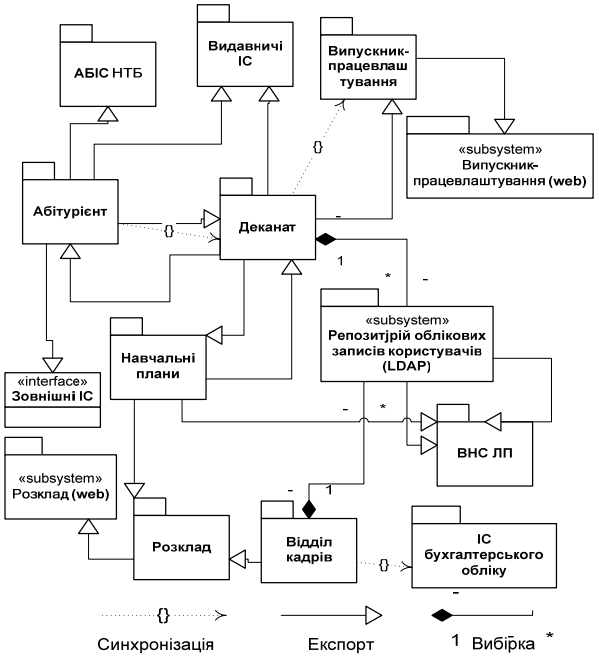

Рис. 8. Схема взаємодії основних БД «Львівської політехніки»

Продемонструємо результат завантаження даних з інших систем. Тут розглянемо два варіанти: традиційне завантаження (без попереднього аналізу даних), завантаження після аналізу даних. Проблеми при завантаженні даних традиційним чином виникають тоді, коли з'являються нові спеціальності чи групи, які необхідно додати у відповідні довідники. Окрім того, додатково необхідно визначати, яка група закріплена за якою кафедрою. Також проблемою є наявність суперечностей: так, є записи про студентів одної групи, що навчаються на різних спеціальностях. Відсоток неспівпадінь – 12 %. Результат порівняння традиційного завантаження та завантаження з попереднім аналізом наведено в табл. 1.

Таблиця 1. Аналіз функціональності – порівняння результату традиційного завантаження та з попереднім аналізом

Кількість студентів	Кількість нових груп	Кількість об'єктів, завантажених традиційно	Кількість завантажень з агентом
17324	0	17324	17324
17324	2	17272	17324
17324	4	17211	17324
17324	7	17001	17324

Попередній аналіз даних дозволяє завантажувати весь обсяг інформації без втрат (рис. 9). Для систем автоматизації навчального процесу ВНЗ $s_1 = 1$ (див. (20)).

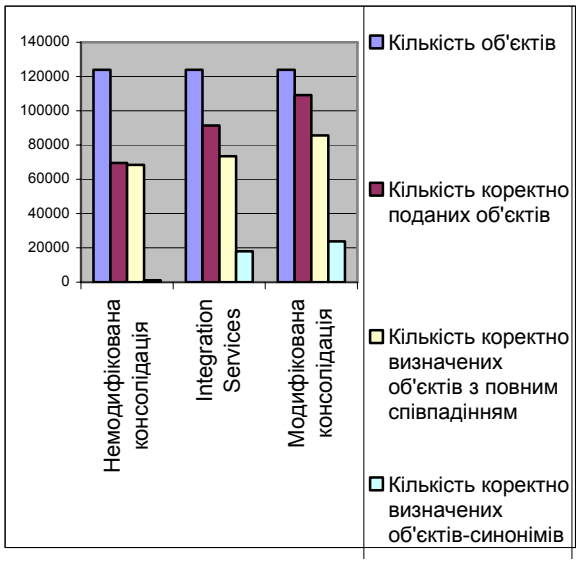

Рис. 9. Аналіз коректності - системи

Окрім систем, джерелами даних є також локальні файли користувачів. У них міститься інформація про студентів заочної форми навчання. Перш за все продемонструємо відбитки вхідних даних для нормалізованої бази даних (рис. 10).

Завантаження інформації здійснюється у сильно формалізовані відношення. Далі використовуємо інтелектуальний агент визначення структур даних, що будує семантичну мережу для текстових вхідних джерел. Правила побудови мережі показано на рис. 11.

Список заочників

8.080403 Програмне забезпечення автоматизованих систем
МАГІСТРИ (ЗАОЧНА ФОРМА)

1.	Царик Володимир Богданович	78,01	ДБ	
2.	Кришталь Галина Ярославівна	77,26	ДБ	
3.	Скляр Павло Анатолійович	60,18	К	
4.	Кирюха Євген Юрійович	75,41	К	
5.	Чоп Андрій Євгенович	59,23	К	

Екстернат Магістр

1. ***Шнайдер Роман Валерійович***

Рис. 10. Приклад джерел даних для системи ВНЗ (текстові дані)

describe

id	name	type	divi-dion	includes	not includes	syno-nic	must be
1	surname	string	enter	symbol	number		
2	prizvyw4e	string	tab	symbol	number	1	
3	firstname	string	prob	symbol	number	1	
4		string	td	symbol	number	1	
5	group	string	enter	symbol	spesial		spets_id
6	група	string	enter	symbol	spesial	5	spets_id
7	прізвище	string	tr	symbol	number	1	

Рис. 11. Правила побудови мережі

Заповнення цього відношення – напівавтоматичне. Перш за все, визначено студентів, які вже є в базі даних, але позначені як відраховані. Їх додатково вносити не потрібно, а лише змінити їх історію. Далі визначаються атрибути, у які агент спробує записати вхідні дані.

Проаналізуємо якість консолідованих даних. Відсоток помилок, що робить агент, зменшується з ростом кількості джерел (табл. 2).

Таблиця 2. Аналіз коректності даних – текст

Кількість джерел	Кількість похибок	Відсоток
5	3	60 %
12	6	50 %
23	12	52 %
27	12	44 %
45	15	33 %
67	23	34 %
75	24	32 %

Висновки

У роботі вирішено науково-прикладну проблему опрацювання різнотипних джерел даних з метою підвищення якості консолідованих даних шляхом вико ристання розроблених теоретичних засад та програмних засобів організації просторів даних як множини інформаційних продуктів та операцій над ними.

У результаті виконання цієї роботи отримані наступні результати.

1. Розроблено алгебраїчну систему сигнатури ПД, яка складається з множини ІП, предикатів та операцій на них. Це дозволило розробити операції консолідації

та пошуку даних з різнотипних джерел, структура даних яких наперед невідома.

2. Розроблено інтелектуальний агент 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 та отримати коректні відповіді на сформовані запити.

3. Розроблено структури даних каталогу даних і синонімічного словника та методи розрахунку ступеню довіри користувача до джерел даних, що дозволило збільшити релевантність відповіді та розробити метод визначення якості консолідованих даних.

1. *Qi Su, Jennifer Widom,* "Indexing Relational Database Content Offline for Efficient Keyword-Based Search," ideas // 9th International Database Engineering & Application Symposium (IDEAS'05). – 2005. – P. 297 – 306.
2. *Аграновский А.В., Арутюнян Р.Э.* Индексация массивов документов. – [Електронний ресурс]. - [Режим доступу]: http://www.scandocs.ru/page.jsp?pk=node_1185787748359.
3. *Denoyer L, Gallinari P.* The Wikipedia XML Corpus. SIGIR Forum, 2006.
4. *DeRose P., Shen W., Chen F., Lee Y., Burdick D., Doan A., Ramakrishnan R.* DBLife: Acommunity information management platform for the database research community. In CIDR, 2007.
5. *Dong X., Halevy A.* A Platform for Personal Information Management and Integration. In CIDR, 2005.
6. *Мальцев А.И.* Алгебраические системы. – М., 1970. – 392 с.

Отримано 21.03.2011

Про автора:

Шаховська Наталія Богданівна,
кандидат технічних наук, доцент,
доцент кафедри інформаційних систем та мереж.

Місце роботи автора:
Національний університет
«Львівська політехніка»,
м. Львів, вул. С. Бандери, 28.
Тел.: (032) 2582404,
natalya233@gmail.com