

Міністерство освіти і науки, молоді та спорту України
Національний університет «Львівська політехніка»

Шаховська Наталія Богданівна

УДК 51.001.57+004.652.4+004.827

ОРГАНІЗАЦІЯ ПРОСТОРІВ ДАНИХ У СКЛАДНИХ ІНФОРМАЦІЙНИХ СИСТЕМАХ

05.13.06 – інформаційні технології

Автореферат

дисертації на здобуття наукового ступеня
доктора технічних наук

Львів—2012

Дисертація є на правах рукопису.

Робота виконана в Національному університеті «Львівська політехніка» Міністерства освіти і науки, молоді та спорту України.

Науковий консультант доктор технічних наук, професор
Пасічник Володимир Володимирович,
 Національний університет «Львівська політехніка»,
 професор кафедри «Інформаційні системи та мережі».

Офіційні опоненти:

- доктор технічних наук, професор Сікора Любомир Степанович, Національний університет «Львівська політехніка», професор кафедри автоматизованих систем управління
- доктор технічних наук, професор Теленик Сергій Федорович, Національний політехнічний університет України «Київський політехнічний інститут», завідувач кафедри автоматики і управління в технічних системах
- доктор технічних наук, професор Дружинін Євген Анатолійович, Національний аерокосмічний університет ім. М.Є. Жуковського «Харківський авіаційний інститут», завідувач кафедри інформаційних технологій проектування літальних апаратів

Захист відбудеться «_____» _____ 2012 р. о ____ на засіданні спеціалізованої вченої ради Д35.052.14 у Національному університеті «Львівська політехніка» (79013, м. Львів, вул. С.Бандери, 12).

З дисертацією можна ознайомитись у науково-технічній бібліотеці Національного університету «Львівська політехніка» (79013, м. Львів, вул. С.Бандери, 12).

Автореферат розісланий «___» _____ 2012 р.

Учений секретар спеціалізованої вченої ради,
 кандидат технічних наук, доцент

А.Є. Батюк

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Глобалізаційні аспекти розвитку сучасного суспільства викликають потребу у побудові складних систем функціонування окремих предметних областей. Так, на прикладі університету – це формування рейтингів викладачів та кафедр, визначення показників успішності та якості навчання тощо; на прикладі обласної адміністрації – це визначення критичних показників розвитку регіону на основі даних, отриманих від організацій різної форми власності. Проте це складно зробити у зв'язку з невідповідністю між вимогами, що ставляться до інформаційних систем, та необхідністю організації (пошуку об'єктів, їх систематизації, узгодження, інтеграції даних) різнотипних інформаційних об'єктів у складну інформаційну систему, що проявляється через: слабку структурування зв'язків між об'єктами, потребу включення нових об'єктів у систему, недотримання загальних стандартів організації ведення документообігу, неможливість проведення систематизації через велику кількість об'єктів та їх різну природу. Актуальність дисертаційного дослідження визначається такими обставинами.

Опрацювання інформаційних ресурсів, що використовують різні моделі даних, схеми керування тощо вимагає розроблення уніфікованого методу доступу до них для того, щоб надати можливість користувачу вибирати адекватний інструментарій для вивчення та використання різних засобів опрацювання даних. Необхідність у цьому виникає в організацій, робота яких полягає в опрацюванні великої кількості різнотипних, взаємозалежних 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 в користувачів немає єдиної схеми, за якою вони можуть створювати запити відносно цільових задач.

Внаслідок керування різнотипними даними з метою розв'язання аналітичних задач стратегічного рівня виникає задача якості даних – відповідності 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 використання даних з декількох джерел, наперед неузгоджених та з невідомими структурами, призводить до того, що якість даних різко знижується і вже не може задовольняти потреб користувача через неузгодженість форматів, різне подання, необхідне для вирішення проблеми.

Зміна масштабів і рівня задач – від оперативного опрацювання до аналітичного, призвела до необхідності: опрацювання даних за певною ієрархією; забезпечення цілісності даних – в системах зберігаються метадані, а не самі об'єкти; усунення дублювання даних, що надходять з різних джерел, визначення довіри до джерела даних, що є різними для різних областей та різних груп користувачів.

Опрацюванням різнотипних неузгоджених даних дослідники займалися з 70-х років ХХ ст. Розроблені моделі та метамови опрацювання таких даних. Проте існуючі на сьогодні моделі та методи стосуються або лише наперед відомих типів даних (здебільшого, реляційних баз даних – праці Калініченка Л., Коха К.), або вирішують лише часткові задачі опрацювання різнотипних даних – наприклад,

індексування для пришвидшення пошуку (Спакапетра С). 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, видобування нових знань тощо).

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: інтеграція веб-застосунків (Лагозе К., Ван де Зомпель Г.) та інтеграція на основі сховищ даних (Косман Д., Гелеві А.). Проте проведений аналіз літературних джерел показав, що для опрацювання інформації від усіх об'єктів галузі необхідно поєднати обидва типи інтеграції та вдосконалити наявні моделі даних у зв'язку з формуванням нових вимог до джерел даних та їх динамічному додаванню.

Інформаційним об'єктом, який здатен розв'язати наведені вище задачі, є простір даних (ПД) (множина різнотипних даних, об'єднаних середовищем керування моделями). Проте його спроб формалізації не знайдено.

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

Отже, дисертаційна робота присвячена вирішенню актуальної наукової проблеми організації та інтеграції інформаційних ресурсів у простір даних, його формалізації та розроблення операцій над інформаційними ресурсами з ціллю формування підстав прийняття ефективних рішень стратегічних задач в складних системах.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконувалась в рамках пріоритетного наукового напрямку, затвердженого в числі актуальних проблем Міністерством освіти і науки України, «Перспективні інформаційні технології, прилади комплексної автоматизації, системи зв'язку» за темами: ДБ/ Бізнес «Моделі та методи побудови інтелектуальних систем бізнес-аналітики на основі інтегрованих інформаційних ресурсів», № державного реєстру 0110U001102 (автор розробила методологію організації просторів даних, що дозволило опрацьовувати дані, отримані з різнотипних джерел); «Програма інформатизації Національного університету «Львівська політехніка» (автор розробила інформаційно-аналітичну систему «Випускник працевлаштування»).

Мета і задачі дослідження. Метою дисертаційної роботи є розроблення системної методології організації та інтеграції інформаційних ресурсів у просторі даних та підвищення їхньої якості та змістовності для забезпечення прийняття управлінських рішень.

Мета дисертаційної роботи визначає необхідність вирішення таких задач:

- 1) проаналізувати методи, моделі та засоби опрацювання різнотипних даних, інформаційні технології отримання та аналізу даних;
- 2) розробити математичну модель простору даних для уніфікації доступу до інформаційних ресурсів та встановлення відповідності структур даних;

- 3) розробити модель сховища консолідованих даних за допомогою удосконалення моделі відношення з невизначеністю для забезпечення коректного відображення факту інформаційної невизначеності;
- 4) вдосконалити метод зменшення інформаційної невизначеності у зв'язку з інтеграцією даних з різних джерел, який враховує рівень довіри до них;
- 5) розробити методи та засоби формування відповіді на запит користувача до простору даних;
- 6) розробити критерії оцінювання якості інтегрованих даних з метою визначення можливості прийняття на їх основі керівних рішень;
- 7) розробити та апробувати на основі отриманих результатів прикладні інформаційні системи підтримки прийняття рішень на базі простору даних.

Об'єктом дослідження є процес опрацювання даних з різнотипних джерел даних.

Предметом дослідження є методи та засоби інтеграції даних з різнотипних джерел, простори даних, критерії оцінювання якості даних, прикладні інформаційні системи.

Методи дослідження. Для досягнення поставленої мети використано: методи системного аналізу – для формування концептуальної моделі простору даних; теорію реляційних баз даних, теоретичні основи побудови сховищ даних реляційного типу, методи моделювання інформаційних систем, систему операцій реляційної алгебри – для розроблення операцій алгебричної системи «простір даних»; методи штучного інтелекту – для виявлення закономірностей для усунення невизначеностей у каталозі простору даних; методи об'єктно-орієнтованого аналізу та проектування – для визначення семантичних зв'язків між джерелами даних.

Наукова новизна одержаних результатів. У дисертаційній роботі запропоновано нові моделі та методи вирішення наукової проблеми організації та інтеграції інформаційних ресурсів у простір даних, розроблено операції над інформаційними ресурсами для підвищення якості консолідованих даних, отриманих з різнотипних джерел. Отримано такі нові наукові результати:

вперше

- розроблено алгебричну систему «простір даних» та доведено аналогію простору даних з нормованим простором, що на відміну від існуючих методів організації інформаційних ресурсів у вигляді графа дає змогу їх систематизувати, уніфікувати операції доступу та опрацювання, а це свою чергу дало змогу удосконалити методи консолідації даних з різнотипних джерел за допомогою розширення теоретико-множинних операцій;
- для розв'язання задачі порівняння структур і семантики даних розроблено метод побудови інтелектуального агента визначення структури даних інформаційного ресурсу на підставі порівняння структур даних інформаційних ресурсів, наявних у просторі даних, та тих, які входять до нього, що на відміну від існуючого методу комутативного відображення не вимагає комплексного опису предметної області на етапі проектування, а це дало змогу розробити операції опрацювання даних з інформаційних ресурсів простору даних, незалежно від їх структур даних та методів доступу;

- розроблено системну методологію побудови структури простору даних за допомогою встановлення правил побудови логічних моделей елементів простору даних для уніфікації математичних описів інформаційних продуктів, що на відміну від існуючих методів дає змогу будувати простори даних будь-яких предметних областей, і що дало змогу розробити алгебричну систему «простір даних»;
- **удосконалено** методи інтеграції даних за рахунок попереднього визначення структур даних та узгодження семантики, що на відміну від методів інтеграції даних на рівні сховища даних дозволило інтегрувати дані з джерел з наперед невідомою структурою даних, і що, в свою чергу, дало змогу підвищити ефективність подальшого аналізу даних;

одержали подальший розвиток

- система оцінки ефективності програмного продукту на основі ISO 9126 за допомогою визначення метрик та розроблення методу визначення якості даних простору даних, що є нормою для розроблення методів і засобів побудови просторів даних як інформаційних продуктів;
- методи опрацювання невизначеності у сховищах даних реляційного типу завдяки розширенню схеми каталогом даних, що дало змогу використовувати їх до просторів даних та звести процес усунення невизначеності до задачі класифікації та застосовувати відповідні алгоритми, збільшивши адекватність моделі даних.

Практичне значення одержаних результатів. Практична цінність розроблених методів та засобів полягає в тому, що з їх використанням:

- розроблено архітектуру простору даних та схему каталогу простору даних з врахуванням рівня довіри до джерела даних, що дозволило використати введені операції над інформаційними продуктами та здійснювати консолідацію даних з розрізнених джерел;
- розроблено алгоритм роботи інтелектуального агента визначення структури джерела даних для встановлення розбіжностей з джерелами, наявними у просторі даних, що дало змогу уніфікувати запити до ПД;
- вдосконалено алгоритми консолідації інформаційних ресурсів за допомогою попереднього визначення структури джерел даних та їх узгодження, що дало можливість розробити алгоритм оцінювання якості консолідованих даних;
- розроблено алгоритми пошуку даних за запитом користувача з метою уніфікації алгоритмів опрацювання даних з різнотипних джерел даних, що дозволило покращити якість консолідованих даних;
- розроблено : 1) програмно-алгоритмічний комплекс туристичної сфери, в якому практично втілено та апробовано результати дисертаційних досліджень, зокрема для Львівської та Чернівецької обласних державних адміністрацій; 2) ряд систем консолідації даних – про студентів Національного університету «Львівська політехніка», даних мережі супермаркетів «Сільпо», даних відділу гарантійного обслуговування ТОВ «Стек-комп'ютер», клієнтів ПП «Анат»; 3) систему автоматизованого розбору резюме для «Львівської консалтингової групи»; 4) інформаційно-аналітичну систему формування наукових звітів у Львівському інституті банківської справи Університету банківської справи Національного

банку України; 5) систему збору та опрацювання наукових публікацій для Інституту регіональних досліджень Національної академії наук України та Львівської національної наукової бібліотеки імені В.Стефаника.

Одержані в роботі результати використано під час розроблення інформаційно-аналітичної системи збору та опрацювання інформації для Львівської та Чернівецької обласних державних адміністрацій, інформаційно-аналітичної системи «Випускник-працевлаштування» Національного університету «Львівська політехніка», інформаційно-аналітичної системи «Львівської консалтингової групи», інформаційно-аналітичної системи для мережі супермаркетів «Сільпо», інформаційно-аналітичної системи формування наукових звітів у Львівському інституті банківської справи Університету банківської справи Національного банку України, системи агрегування даних для відділу гарантійного обслуговування ТОВ «Стек-комп'ютер», системи збору та опрацювання наукових публікацій для Інституту регіональних досліджень Національної академії наук України, інформаційно-аналітичної системи для приватного підприємства «Анат», системи аналітичного опрацювання наукових даних для Львівської національної наукової бібліотеки імені В.Стефаника, що підтверджено відповідними актами впровадження. Результати впроваджені також в навчальному процесі в курсах «Сховища даних», «Технології сховищ даних», «Проектування проблемно-орієнтованих систем» у методичних вказівках до вивчення окремих розділів курсу, виконання лабораторних робіт, курсових робіт, а також у 3-х посібниках з грифом Міністерства освіти та науки України. Інформаційна система «Випускник-працевлаштування», впроваджена у Національному університеті «Львівська політехніка», номінована на Другій виставці «Інноватика в освіті України», Львів-Київ-2011 – золотою медаллю, на другій міжнародній виставці «Сучасні навчальні заклади – 2011», Львів-Київ-2011 – гран-прі.

Особистий внесок здобувача. Усі наукові результати, подані у дисертації, одержані здобувачем особисто. У друкованих працях, опублікованих у співавторстві, особистий внесок здобувача такий: [2] – розроблено алгебричну систему «простір даних»; [9] – уведено формальний опис ПД та окреслено його основні задачі; [11] – описано особливості інтеграції даних ВНЗ; [13] – розроблено формальну модель ПД; [14] – розроблено метод оцінювання якості консолідованих даних; [15] – проаналізовано програмні засоби реалізації ПД та окреслено клас задач; [16] – розроблено концептуальні моделі та схеми, діаграми потоків даних сховища даних, а також логічні схеми даних для збереження зв'язків між атрибутами відношень, правил, метаданих; [17] – побудовано алгебричну систему «простір даних»; [18] – проаналізовано методи опрацювання медичних даних; [19] – описано типи невизначеності, що виникають у сховищі консолідованих даних; [20] – розроблено метод реферування з множини документів; [22] – розроблено алгоритм побудови семантичної мережі на основі текстового документа; [21] – визначено методи інтеграції даних; [23] – побудовано схему аналізу туристичних даних, формалізовано елементи простору даних; [24] – розроблено архітектуру ПД, розділену за рівнями та компонентами платформ; [25] – спроектовано простір даних виставкової діяльності; [26] – формалізовано базові операції над інформаційними продуктами; [27] – спроектовано простір даних системи множинного реферування; [28] – форма-

лізовано етапи технології ETL; [29] – спроектовано архітектуру сховища даних ВНЗ; [30] – спроектовано архітектуру системи реферувальника; [31] – описано використання просторів даних для побудови інтелектуальних систем; [32] – реалізовано структури даних для опису онтології для реферування документів; [33] – спроектовано засоби роботи з різнотипними даних у просторі даних Веб-спільноти; [34] – спроектовано простір даних енергетичної сфери; [37] – спроектовано метод опрацювання різнотипних даних; [38] – розроблено операції у просторі даних; [40] – розроблено операції пошуку різнотипних даних; [43] – розроблено засоби інтеграції даних; [44] – уведено операції пошуку у інформаційних продуктах; [46] – розроблено структуру та програмне забезпечення мультимедійного підручника; [47] – спроектовано інформаційну модель верифікації даних; [48] – розроблено метод керування корпоративними даними; [49] – розроблено метод зменшення невизначеності у корпоративних даних; [51] – описано структуру простору даних; [52] – розроблено засіб організації онтологій; [54] – визначено сфери застосування ПД; [55] – визначено структуру простору даних.

Апробація результатів дисертації. Основні результати дисертаційної роботи доповідалися на наступних міжнародних наукових конференціях та семінарах: семінарі Західного наукового центру НАН України і МОН України, відділення інформатики, Львів, 2011; X-XII Міжнарод. наук.-техн. конф. «САПР в машинобудуванні. Проблеми впровадження та навчання CADSM», 2009-2011; Międzynarodowa naukowo-praktyczna konferencja «Nauka: teoria i praktyka – 2009», Przemysł, Nauka i studia, 2009; II-VI Міжнарод. конф. «Комп’ютерні науки та інформаційні технології CSIT», Львів, 2007–2011; X-XIII Міжнарод. наук.-техн. конф. «Системний аналіз та інформаційні технології», Київ, 2008–2011; VI-VII Міжнарод. конф. «Дні науки у Запорізькому класичному університеті», Запоріжжя, 2008, 2009; III Міжнарод. наук.-техн. конф. «Комп’ютерна математика в науці, інженерії та освіті» CMSEE-2009, Полтава, ПолтНТУ ім. Ю.Кондратюка; VI Міжнарод. наук.-техн. конф. «Інтернет – Освіта – Наука – 2008», Вінниця; XII Міжнарод. молодіжному форумі «Радиоэлектроника и молодежь в XXI веке» – Харків, 2008; I-IV Міжнарод. наук. конф. для молодих вчених «Комп’ютерні науки та інженерія CSE», Львів 2007 – 2010; IV Міжнарод. наук.-техн. конф. «Научная мысль информационного века – 2008», Прага, Чехія; IV Міжнарод. наук.-практ. конф. «Достижения высшей школы – 2008», Софія, Болгарія; VIII Міжнарод. наук.-практ. конф. «Математичне та програмне забезпечення інтелектуальних систем», MPZIS–2010, Дніпропетровськ, 2010; на наукових семінарах кафедри інформаційних систем та мереж та щорічних (2007–2011) наукових конференціях викладачів та науковців Національного університету «Львівська політехніка».

Публікації. Результати дисертаційного дослідження опубліковано в 85 наукових публікаціях, серед 55 основних наукових публікацій наведених в авторефераті: 2 монографії, 30 статей у фахових наукових виданнях, 3 статті в інших виданнях та 20 публікацій у матеріалах міжнародних конференцій.

Структура та обсяг роботи. Дисертація складається з вступу, семи розділів, висновків, списку використаної літератури та додатків. Робота містить 298 сторінок основного тексту. Загальний обсяг дисертації – 452 сторінки, 9 таблиць, 117 рисунків, 308 найменувань використаних літературних джерел.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтовано актуальність теми, сформульовано мету та основні задачі досліджень, подано короткий зміст роботи.

У першому розділі проаналізовано принципи організації інформаційних об'єктів у складні системи. Визначено особливості функціонування систем різних масштабів. Проаналізовано методи опрацювання даних з різнотипних джерел (побудова інвертованих індексів, канонічні системи, теорія інтеграції, інтеграція на основі сховищ даних, веб-інтеграція та ін.) і визначено їх обмеження (основною умовою інтеграції є те, що структура джерела даних має бути наперед визначеною). Встановлено, що для задач, які виникають на вищому рівні інтеграції і управлінні ієрархією (велика кількість джерел, консолідація не за графіком) використання сховищ даних призведе до великих затрат часу на написання спеціалізованих програмних засобів та порівняння структур даних в ручному режимі. Проаналізовано методи підвищення ефективності інтегрованих даних та визначено їх частковість; програмні засоби для опрацювання даних у різнотипних джерелах. Визначено, що з технічної точки зору немає перешкод в опрацюванні різнотипної інформації. Основна проблема, що виникає – незнання структур даних та їх семантичного значення, а також можливість виникнення неповноти чи неточності у процесі консолідації інформації з різних джерел, що спотворює умови прийняття рішень.

Інформаційні продукти певної предметної області та консолідовані дані становитимуть простір даних. Однією зі задач, яка виникатиме у процесі консолідації, є невизначеність даних, що є результатом дублювання, неточності, відсутності, протиріччя даних (рис. 1).

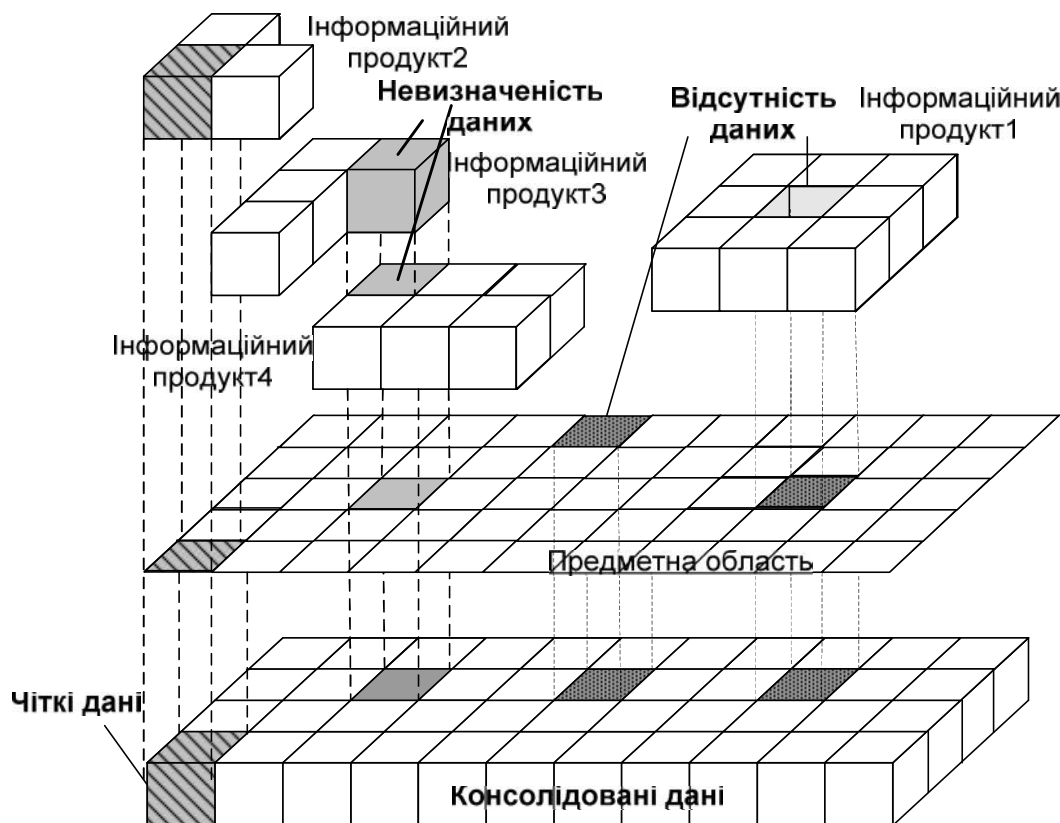


Рис. 1. Схема консолідації даних

Наявні методи опрацювання розрізаних даних потребують вдосконалення, оскільки у просторі даних наперед невідомо усіх учасників та їх структури даних. Також необхідним є розроблення нової інформаційної технології для організації простору даних. На основі аналізу функціонування та взаємодії інформаційних ресурсів та з метою організувати їх у систему уведено два нових принципи: часткової самоорганізації (автоматизованого встановлення зв'язків); самоадаптації (переорганізація у випадку незадовільної якості даних). Встановлено, що для опрацювання та об'єднання різнотипних даних необхідно розробити нову методологію, яка ґрунтується на використанні простору даних як інтеграційного базису формування стратегічного управління.

Результати першого розділу дали змогу сформулювати раніше невирішені задачі та окреслити проблему дослідження.

Основні результати розділу опубліковані у працях [1, 2, 8-10, 29, 31, 37, 41, 45].

У другому розділі введено формальне означення простору даних та описано його елементи. Розроблено алгебричну систему простору даних. Це дозволило уніфікувати операції з різнотипними джерелами даних та коректно виконувати запити користувача. Уведено операції над інформаційними продуктами. Доведено повноту алгебричної системи простору даних. Розроблено метод обчислення довіри до джерела даних. Введено визначення елементів простору даних, які ґрунтуються на ISO/IEC/ IEEE 24765 – Systems and software engineering Vocabulary.

Інформаційний ресурс (IP) I_r – сукупність даних в інформаційних об'єктах. Характеризується структурою даних S_d .

Інформаційний продукт (ІП) – документований інформаційний ресурс, який є результатом функціонування інформаційної технології $I_p = \langle I_r, Cg_{I_p}, Rl \rangle$, де Cg_{I_p} – каталог, Rl – методи доступу. Інформаційними продуктами є програмні засоби, текстові файли, веб-сторінки, електронні таблиці, xml-файли, бази даних, сховища даних, дані інших форматів.

Каталог ІП – метадані про ІП $S_d \cup Pl \cup Rl \rightarrow Cg_{I_p}$. Описує місцезнаходження ІП, його структури даних (СДІР), методи доступу до ІР тощо.

Множина інформаційних продуктів **I_p** предметної області містить найповнішу інформацію про предметну область, а отже якість прийнятих керівних рішень на її основі є найвищою.

З метою подальшої формалізації та розроблення операцій над інформаційними продуктами введено означення простору даних.

Простір даних – це блоковий вектор, що містить множину інформаційних продуктів предметної області, поділену на три блоки: структуровані дані St , напівструктуровані дані $SemS$ та неструктуровані дані Ns . Над цим вектором та його окремими елементами визначено операції та предикати:

$$DS = \langle I_p, Cg, \Omega_p, \Omega_f \rangle, I_p = \langle St, SemS, Ns \rangle, St = \langle DB, DW \rangle, SemS = \langle Wb, Tb \rangle, Ns = \langle Nd \rangle,$$

де **DB , DW , Wb , Tb , Nd** – інформаційні продукти, що подають множини баз даних, сховищ даних, веб-сторінок, електронних таблиць, текстових файлів; Ω_p – множина операцій, Ω_f – множина предикатів, заданих на множині інформаційних продуктів простору даних.

Стан інформаційного продукту S_{Ip} – зафіксований у певний момент часу його інформаційний ресурс Ir та відомості про ІП (каталог даних) Cg : $S_{Ip} = \langle Ir, Cg_{Ip} \rangle$.

Стан простору даних S_{DS} – множина станів усіх інформаційних продуктів предметної області та відношень між ними: $S_{Ip} \cup Cg \rightarrow S_{DS}$.

Для уніфікації інформаційних продуктів та операцій над ними введено алгебричну систему «простір даних»:

$$DS_a = \langle Ip, \Omega_P, \Omega_F \rangle, \quad (1)$$

де $Ip = DS$ – скінченна множина станів інформаційних продуктів, $\Omega_P = \{O_{P0}, O_{Pu}, O_{Pb}\}$ – множина операцій над інформаційними ресурсами, де: O_{P0} – нульарна операція, результатом якої є стан заданого ІП у просторі даних; O_{Pu} – множина унарних операцій над простором даних DS , результатом є зміна стану простору даних; O_{Pb} – множина бінарних операцій над просторами даних, результатом є утворення нового простору даних.

Говорячи про інформаційний продукт, маємо на увазі його вміст (інформаційний ресурс Ir), а також множину відомостей про нього (каталог даних Cg). Уведено операції, які виконуються над інформаційними продуктами залежно від їх структур даних.

Результатом нульарної операції над простором даних DS є заданий інформаційний продукт Ip :

$$O_{P0} : DS \rightarrow Ip. \quad (2)$$

Нульарний оператор є розширенням реляційної операції селекції без задання умови.

Уведено унарні операції алгебричної системи простір даних. Унарними операціями над просторами даних є множина:

$$O_{Pu} = \{Ss, Se, Sm, S_{ac}, f_{Ip}, C, Ag\}, \quad (3)$$

де f_{Ip} – операція визначення СДІР; Ss, Se, Sm – операції пошуку; S_{ac} – операція доступу.

Визначення структури даних ІП здійснюється за допомогою інтелектуального агента і полягає у доповненні Cg новими даними:

$$f_{Ip} : Ip \xrightarrow{Agent} DS, \quad (4)$$

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

Агент $Agent$ задано кортежем:

$$Agent = \langle Cg, M, Dic, MB, Dif, H \rangle, \quad (5)$$

де Cg – інформація про джерела, що вже є у ПД; M – тип джерела; Dic – база знань агента про власні можливості (терміни-синоніми, що позначають у джерелах одні і ті ж властивості); MB – база знань про методи доступу до джерела, $MB: S_{evd=Date()}(Cg) \cap Ip.Rl$; Dif – множина розбіжностей, які виявив агент; H – відношення відповідності між новим джерелом IP_{new} , $Y \subset IP_{new}(IR)$ та опрацьованими. M, H – база станів агента, Cg, Dif, Dic, MB – база знань.

Перш за все виконується визначення методу доступу за типом джерела, де джерело за типом може бути структуроване (база даних, сховище даних, xml), напівструктуроване (електронна таблиця, html) або неструктуроване (текст):

$$MB: \begin{cases} struct(Ip.Rl), M = 1, \\ parse(Ip.Rl), M = 2, \\ S(Ip.Rl), M = 3. \end{cases}$$

Для того, щоб забезпечити опрацювання напівструктурованої та текстової інформації, розроблено метод формування семантичної мережі. Особливості структури семантичних мереж: вузли є концептами 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, побудовану на основі аналізу термів напівструктурованого джерела інформації Γ , подано як $\Gamma = \{V, D\}$, 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 (їх важливість). Вершини є елементами сховища консолідованих даних cg' . Функція трансформації напівструктурованого тексту та Веб-сайтів у семантичну мережу: $S(E) \rightarrow N, E \in Wb \vee E \in Nd$.

Між двома будь-якими елементами Y_i, Y_j словника даних Dic , $Y_i \in Dic, Y_j \in Dic$, існує відображення деякого порядку $\forall Y_i: \exists n, \Gamma^n(Y_i) = \{Y_i, i = \overline{1, M}\}$, 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 $\{X_k, 1 \leq k \leq N\} \leftarrow Y_i \leftarrow \{X_l, 1 \leq l \leq N\}$, тут $X_k \leftarrow Y_i = S(X_k, Y_i): Y_i \leftarrow X_k = S(Y_i, X_k)$. Крім цього, в граф так само входять всі вхідні поняття, які використовуються в якості обмежень $\forall Y_i: \Gamma'^k(Y_i)$, де $\{X_j: \exists S(Y_i, X_j) \vee S(X_j, Y_i)\}$. Тоді семантично зв'язані елементи: $\Gamma(Y_i, \Gamma(X_j)) = \{X_k: \exists S(Y_i, X_k) \vee S(X_k, X_j)\}$.

Далі агент встановлює тип відношення відповідності H для визначення характеристик, що мають потрапити у результат запиту.

1) Відношення еквівалентності: $x = y \Rightarrow H(x, y) = 1; P^Y(IP_{new}) = 1$, де $P^Y(IP_{new})$ – довіра до атрибута IP_{new} .

2) Відношення синонімізму:

$$\forall x \in X, \exists y \in Y: X \neq Y, X \subset Dic, Y \subset Dic \Rightarrow H(x, F(x)) = 1, P^Y(IP_{new}) = 1.$$

3) Відношення перетворення:

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

4) Відношення узагальнення: Y – узагальнення X :

$$F: Y \rightarrow X, \forall x \in X, \exists y \in Y: X \subset Dic, Y \subset Dic \Rightarrow H(x, y) < 1,$$

$$P^Y(IP_{new}) = 0,5; P^X(IP_{new}) = 0. Dic \setminus X \cup Y \rightarrow Dic.$$

5) Ізоморфізм доменів: $F : X \rightarrow Y, y \neq F(x), \exists x \in \text{dom}(X), \exists y \in \text{dom}(Y) : X \subset \text{Dic}, Y \notin \text{Dic} \Rightarrow H(x, y) < 1, P^Y(IP_{\text{new}}) = 0,25; P^X(IP_{\text{new}}) = 0,5, \text{Dic} \cup Y \rightarrow \text{Dic}.$

6) Відношення полярності: $F : X \rightarrow Y, G : Y \rightarrow X, X \subset \text{Dic}, Y \notin \text{Dic}, P^Y(IP_{\text{new}}) = 0; X \cup Y \rightarrow \text{Dif}; \forall x \in X, \exists y \in Y : X \subset \text{Dic} \left\{ \begin{array}{l} X \cup Y \rightarrow \text{Dif} \\ P^X(Cg) = 0 \quad H(x, y) = 0 \quad P^X(Cg) = 0, Mb \triangleleft 0 \end{array} \right\} \rightarrow \left\{ \begin{array}{l} X \cup Y \rightarrow \text{Dif} \\ P^Y(IP_{\text{new}}) = 0 \end{array} \right\}.$

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

$$\forall (i, j) : H(x_i, y_j) = 1 \Rightarrow H(a, b) = 1,$$

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

Відношення еквівалентності та синонімізму є рефлексивними, оскільки $\forall x, y \in \text{Dic} : x = x \Rightarrow H(x, x) = 1$. Відношення узагальнення та ізоморфізму є асиметричними, оскільки $\forall x, y \in \text{Dic} : H(x, y) = 1 \Leftrightarrow H(y, x) = 1$, але $\forall x, y \in \text{Dic} : H(x, y) < 1 \neq H(y, x) < 1$. Відношення перетворення є транзитивним, оскільки, $(y = F^{-1}(x)) \wedge (x = F^{-1}(z)) \Rightarrow (y = F^{-1}(z))$.

Встановлення відповідності структур дало змогу розробити решту операцій над інформаційними продуктами.

Консолідація даних – це об'єднання інформаційних ресурсів ІП у сховище консолідованих даних визначеної структури cg' :

$$C : \langle Ip_1.Ir \cup \dots \cup Ip_n.Ir; f_{Ip}(Ip_1.Cg) \cup \dots \cup f_{Ip}(Ip_n.Cg) \rangle \rightarrow cg'. \quad (6)$$

Запит про довільні дані Ss – запит будь-якого елементу даних, незалежно від його формату і моделі даних. Здійснюється на основі множини ключових слів K та каталогу ІП Cg :

$$Ss : S_K(S_{DS}) \cup S_K(Ip_1.Ir) \cup \dots \cup S_K(Ip_n.Ir). \quad (7)$$

Структуровані запити будуються з використанням SQL та подібних мов:

$$Se : S_{Cg, x='struct'}(p_x(S_K(Ip_1.Ir)) \cup \dots \cup p_x(S_K(Ip_n.Ir))). \quad (8)$$

Запити до метаданих:

$$Sm : S_{up}(Cg) \cup S_{up}(P), \quad (9)$$

де up – множина параметрів користувача (вимог до запиту), його профілю, або вимог, які ставляться до рішення, P – рівень довіри до джерела.

Доступ до кожного з ІП залежить від прав користувача. Права доступу кожного із користувачів до заданого Ip_i вказуються у Cg . Профіль користувача – підмножина каталогу даних, яка вказує на ті ІП, до яких користувач має доступ:

$$\text{profile} : S_{ac=Yes}(Cg) \triangleright \triangleleft S_{P < ex}(Dic). \quad (10)$$

ПД вкладаються один в інший (наприклад, ПД району вкладається в ПД області), і вони можуть перекриватися (ПД в сфері туризму перекривається з

просторами даних оздоровчо-лікувальної, історичної сфери та сфери управління природними ресурсами). Тому в ПД містяться правила розмежування доступу.

З метою опрацювання даних з кількох просторів даних (операції над блоковими векторами) уведено бінарними операціями над множинами ІП, які є розширенням теоретико-множинних операцій об'єднання, перетину та різниці.

Уведено бінарну операцію об'єднання просторів даних:

$$DS_3 = DS_1 \cup DS_2: \text{profile}(f_{Ip}(Cg_1) \cup f_{Ip}(Cg_2)), Cg_3 = Cg_1 \cup Cg_2.$$

Операція об'єднання ПД використовується також для додавання нового інформаційного продукту до простору даних:

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

Уведено бінарну операцію перетину просторів даних:

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

Додавання нового інформаційного продукту до простору даних:

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

Уведено бінарну операцію різниці просторів даних:

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

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

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

Далі у розділі розроблено предикати алгебричної системи «простір даних».

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 ін, а також результат порівняння подібності структур даних один з одним.

Для організації роботи з розрізненими джерелами використано словник термінів та понять (ключових слів) *Dic*, який містить синонімічний опис одного і того ж концепту у різних джерелах даних

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

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

$$\Omega_{F0} : s_{Ip}(Dic) \neq \emptyset. \quad (12)$$

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

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

На основі відомої аналогії реляційної моделі з скінченим автоматом доведено, що реляційна модель є підсистемою алгебричної системи «простір даних», а також гомоморфність реляційних операцій операціям простору даних.

Оскільки базисом побудови простору 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. З цією метою розроблено *метод визначення довіри до джерела даних*.

Спершу визначено рівень довіри до характеристики джерела даних:

$$P^{attr}(i, j) = \frac{\sum_{k=1}^n P_k^{attr}(i, j)}{\sum_{t=1}^s a_t \sum_{k=1}^n l_{tk}}, \quad (14)$$

де n – кількість звернень користувача до ресурсу, $P_k^{attr}(i, j)$ – значення відповідності i -му користувачу характеристики $attr$ j -го джерела даних при k -ому зверненні, $attr$ – назва атрибута, S – кількість видів джерел, l_{tk} – кількість звернень до кожного з джерел виду t , a_t – вага джерела виду t , $a_t = |Ip_i \cdot Cg|$.

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

$$P(i, j) = \min(\max(P_1^{attr}(i, j)), \dots, \max(P_k^{attr}(i, j))), \quad (15)$$

Визначено рівень довіри до джерела j :

$$P(j) = \frac{\sum_{i=1}^m P(i, j)}{m}, \quad (16)$$

де m – кількість користувачів, що звертались до ресурсу.

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

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

$$v(r) = \sum_{j=1}^m k_j v_j(r_j), \quad (17)$$

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

Результатом другого розділу є введення алгебричної системи «простір даних», доведення гомоморфності реляційних операцій та операцій простору даних, розроблення методу визначення довіри до джерела даних, що є основою для розробки відповідних алгоритмів консолідації та пошуку даних, а також оцінки їх якості.

Основні результати розділу опубліковані у працях [2, 3, 5, 12, 17, 24, 34, 35, 36, 39, 40, 43, 44, 50, 53].

У третьому розділі введено інформаційну модель сховища консолідованих даних (СКД) як удосконалену модель відношення з невизначеністю, яка використовується для визначення якості даних та їх релевантності запиту для підтримки прийняття рішення. Інформаційна модель складається з сховища консолідованих даних, встановлених залежностей між даними, операцій та методів усунення невизначеності.

Схема сховища консолідованих даних Cg' – множина імен атрибутів $\{C_1, C_2, \dots, C_n\}$; $\{C_unk_1, C_unk_2, C_unk_p\}$ з нечіткими або недермінованими значеннями; множину імен атрибутів $\{Unk_1, Unk_2, \dots, Unk_m\}$, доменами яких є числові дані, що моделюють імовірнісні дані, значення функції приналежності нечітких множин, лінгвістичні оцінки (зберігання значень рівнів довіри $P(j)$, $P^{atr}(j)$); схему словника синонімів Dic та схему каталогу даних Cg :

$$Cg' = \langle \{C_1, C_2, \dots, C_n\}, \{C_unk_1, C_unk_2, C_unk_p\}, \{Unk_1, Unk_2, \dots, Unk_m\}, Dic, Cg \rangle.$$

Схема відношення з невизначеністю Cg' є результатом роботи агента визначення структури джерела даних: $C_i \subset Cg' \leftarrow H > 0$, $C \in Dic$; $C_unk_i = P^{C_i}(p_{ip}(Cg))$.

Кортеж консолідованих даних dc – інформаційний опис об'єкта t джерела даних S , поданий у вигляді множини значень характеристик (атрибутів) про об'єкт, джерело даних та 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:

$$dc = t_{rel} \cup Unk, t_{rel} = \{c_1, \dots, c_n\} \cup \{c_unk_1, \dots, c_unk_m\}.$$

2. Сховище даних :

$$dc = t_{dw} \cup Unk, t_{dw} = \{c_{11}, \dots, c_{1n}\} \cup \dots \cup \{c_{k1}, \dots, c_{kn}\} \cup \{c_{rf_1}, \dots, c_{rf_l}\} \cup \\ \cup \{c_unk_{11}, \dots, c_unk_{1m}\} \cup \dots \cup \{c_unk_{k1}, \dots, c_unk_{ks}\} \cup \{c_unk_{rf_1}, \dots, c_unk_{rf_t}\}'$$

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 атрибута з невизначеністю i -го виміру, $c_unk_{rf_j}$ – значення j -ї характеристики з невизначеністю відношення фактів.

3. Напівструктурований текст:

$$dc = t_{text} \cup Unk, t_{text} = \{c_1, \dots, c_n\} \cup \{c_unk_1, \dots, c_unk_m\}.$$

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

$$dc = \langle C, C_unk, Unk, \{dic\}, \{cg\} \rangle,$$

де C – підмножина значень атрибутів із чіткими значеннями, $C = t_{rel} \cup t_{dw} \cup t_{text}$, C_unk – підмножина значень атрибутів із нечіткими та недетермінованими значеннями, Unk – підмножина значень рівнів довіри до значень атрибутів C_unk $P^{atr}(j)$ та кортежу загалом $P(j)$ і $meta(C_unk, Unk) = 1$, $\{dic\}$ – множина значень словника даних, $\{cg\}$ – множина значень каталога даних. Його формування здійснюється з використанням операції консолідації даних (6).

Тоді сховище консолідованих даних cg' – відношення з схемою Cg' та множиною кортежів консолідованих даних **dc**. Інформаційна модель сховища консолідованих даних містить дані з усіх типів джерел простору даних.

Удосконалено операції зрізу:

$$slice : s^{cons}_{\left(\frac{Unk \Theta unk}{\cup s_c(Dic) \cup s_c(Cg)} \right) \cup (C_unk \Theta c_unk)}(cg') = \left\{ \begin{array}{l} t \in dc | t(Unk) \Theta unk, \\ t(C_unk) \Theta c_unk, meta_{Unk, C_unk} = 1, \\ s_c(Dic) \text{ Is Not NULL,} \\ s_c(Cg) \text{ Is Not NULL, } unk = P(cg') \end{array} \right\},$$

згортання та з'єднання у сховищах даних за допомогою розширення схеми каталогом даних Cg і словником даних Dic , а також врахування наявності рівня довіри до джерела. Доведено збереженість властивостей базових операцій: удосконалений оператор зрізу, як і оператор вибірки, зберігає властивості комутативності та дистрибутивності відносно булевих операцій; удосконалений оператор згортання зберігає властивості оператора проєкції; удосконалений оператор з'єднання комутативний та асоціативний. Розширено типи невизначеності, що виникають на рівні сховища даних. Це дозволило врахувати їх на рівні простору даних.

Удосконалено методи зменшення невизначеності у сховищі консолідованих даних – на основі класифікаційних правил та побудови ієрархії об'єктів.

Перший метод використовується для усунення лінгвістичної невизначеності й відсутності та ґрунтується на методі побудови функціональних залежностей. Доведено, що для класифікаційних правил, за аналогією до функціональних залежностей, збережені аксіоми рефлексивності, поповнення, адитивності.

Другий метод моделює залежності між концептами предметної області, що існували у вигляді ієрархії інформаційних продуктів. Але, оскільки ця ієрархія не завжди відома, і особливістю простору даних є одноранговість інформаційних продуктів, то врахування згаданої ієрархії здійснюється власне у сховищі консолідованих даних. Моделювання залежностей між інформаційними продуктами організовано як ієрархію кортежів. Для руху по мережі залежних об'єктів, інформація про які наявна у cg' , модифіковано оператор визначення нащадка:

$$Down_c_{X=x_2, s_X(Dic)}(cg') = s_{Y=x_2, s_X(Dic)}(cg').$$

Модифіковано оператор усунення невизначеності у мережевій структурі консолідованих даних:

$$s_{X=x_1, Val=v, s_X(Dic)}(cg') = Heir_c \left(\begin{array}{l} s^{cons}_{X=x_1, Val=v, s_X(Dic)}(cg'), \\ Down_c_{X=x_1, Val=Null, s_X(Dic)}(cg') \\ Cg'.Unk_X = \min(Cg'.Unk_X, P^X(s_X(Dic))) \end{array} \right).$$

Результатом розділу є розроблення інформаційної моделі сховища консолідованих даних та вдосконалення методів зменшення невизначеності даних, що дало змогу зберігати дані, отримані від інформаційних ресурсів, та аналізувати їх з метою подальшої підтримки прийняття рішень.

Основні результати розділу опубліковані у [1, 2, 6, 14, 19].

У четвертому розділі на основі результатів розд. 2 побудовано алгоритми для інтелектуального агента визначення структури джерела, інтеграції даних; розроблено архітектуру каталогу даних *Cg* та словника даних *Dis*. Спроековано методи керування вмістом каталогу простору даних та відповіді на запит користувача. Розроблено мову *метаописів джерел даних* та метод інтерпретації запиту, що дозволило уніфікувати доступ до даних. показано механізм переведення операцій метамови до простору даних у мови джерел даних за типами.

Для опису інформаційних продуктів ПД розроблено структури даних з такими відношеннями: *Tbl_meta* містить інформацію про структури джерел даних, що вже є у ПД, та джерела, що додається до ПД; *Tbl_Path* зберігає шляхи або сторінки усіх джерел даних; *Tbl_Oper* містить перелік операцій, що виконуються над даними, поданими у різних моделях (для кожної операції розроблена процедура, яка запускається за необхідності); *Tbl_Model* містить перелік моделей джерел даних.

Метод інтерпретації запиту полягає в семантичному аналізі запиту, побудові семантичної мережі на основі неструктурованих джерел даних, що входять в запит, інтерпретації та оптимізації запиту, побудові сховища консолідованих даних, та завантаженню даних в нього.

Процедура семантичного аналізу запиту полягає в перевірці коректності задання операндів – їх наявності в одній з множин: каталог даних, словник даних, сформована множина ключових слів заданої предметної області. У випадку невходження у жодну з множин запит вважається невиконаним. Спроековано процедуру опрацювання запитів, яка полягає у виконанні наступних кроків:

- 1) застосування операцій пошуку джерел *Ip*, що містять відповідь на запит користувача;
- 2) запис параметрів запиту у тимчасове сховище параметрів запиту *A*;
- 3) запуск інтелектуального агента *Agent(Ip)*;
- 4) якщо джерело даних структуроване, здійснити відбір даних за запитом $Se(s_A(Ip))$;
- 5) якщо джерело даних напівструктуроване чи неструктуроване, побудувати семантичну мережу $S(Ip)$;
- 6) консолідація даних $C(Ip)$;
- 7) запуск модифікованого оператора зрізу для відсікання непотрібних даних $s_A^{cons}(cg')$;
- 8) відображення вмісту отриманого сховища консолідованих даних.

Для візуалізації запиту користувача з метою його подальшого виконання використовується дерево. Побудова дерева здійснюється відомими методами формату запису виразу. Для візуалізації обрано польський інверсний запис.

Для попередньо отриманих і збережених даних розроблено систему норм і критеріїв перетворення їх до форми, придатної для збереження у сховищі консолідованих даних. Далі розроблено алгоритм підготовки і отримання вхідних даних з декількох джерел з метою їх підготовки до завантаження у сховище консолідованих даних. Для аналізу множини документів будується семантична мережа, аналізується вміст каталогу простору даних та здійснюється їхнє порівняння.

Після того як інтелектуальним агентом дозволено опрацювання документів, необхідно виконати наступне: виключити з розгляду речення, які за своїм змістом вже потрапили у сховище консолідованих даних та виконати підсумкове сортування речень. Для цього розроблено наступний алгоритм.

1. Ініціалізуються дві множини речень $B = \emptyset$ і $A = \{A_i \mid i = 1, 2, \dots, N\}$, де N – кількість речень у всіх текстових документах. Для кожного речення A_i рівень довіри P_i приймається рівним q_i : $P_i = q_i, i = 1, 2, \dots, N$.
2. Речення множини A сортуються відповідно до їх поточного рангу в порядку спадання.
3. Вважаючи, що речення A_i має найвищий ранг, воно переміщується в B . Ранг речень, що залишилися в A розраховується як $P_i = P_i - kq_i$, де $k > 0$ – фактор відсікання подібних речень.
4. Процес повторюється, поки A не стане порожньою множиною.

Далі у розділі описано елементи метамови простору даних. Запит q до простору даних заданий коректно, якщо він складається з елементів, описаних у Cg та Dic :

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

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

Абстрактний синтаксис: алфавіт запиту об'єднує алфавіт усіх джерел даних, до яких направляють запит, а для встановлення характеристик вибираються усі можливі синоніми $\{R\} \cup \{Rel\} \cup \{key\} \cup \{Hd\}$, де описані схеми БД, СД, ключових слів текстових файлів, заголовків веб-документів відповідно. Семантичний синтаксис метамови подано в Dic , схема семантичних функцій – в Ω_p .

Процедура інтерпретації.

1. Визначення типу операції і перевірка, чи позначення такої операції належить до множини операцій Ω_p . Якщо ні – помилка.
2. Визначення призначення параметру за типом операції – елемент каталогу, словника чи сховища консолідованих даних.
3. Визначення предметної області, до якої належить об'єкт $object$.
4. Якщо об'єкт описаний більше, ніж для одної предметної області – помилка.
5. Якщо в каталозі даних не знайдено значення параметру, пошук у словнику синонімів, повернення на крок 3, інакше – повідомлення про відсутність.
6. Пошук $(c_1, \dots, c_n)$ у словнику синонімів. Перерахунок кількості звернень.
7. Виконання операції. Повернення результату.

Розроблено механізм оптимізації запиту. Оптимізація запитів користувача відбувається за таких обмежень:

1. Обмеження за аналогією до реляційної алгебри:

$$Se_{p_1}(Se_{p_2}(object)) \equiv Se_{p_1 \text{ AND } p_2}(object);$$

$$Se_{p_1}(object) \cup Se_{p_2}(object) \equiv Se_{p_1 \text{ OR } p_2}(object);$$

$$Se_{p_1}(object) \cap Se_{p_2}(object) \equiv Se_{p_1 \text{ AND } p_2}(object);$$

$$object - Se_p(object) \equiv Se_{NOT p}(object);$$

$$Sm_{object}(\{Sm_{object}(X, Y, Z), Z\}) \equiv Sm_{object}(Z);$$

$$Sm_{object}(Se_p(X)) \equiv Se_p(Sm_{object}(X));$$

$$DS_1 \cup DS_1 \equiv DS_1; \quad DS_1 \cap DS_1 \equiv DS_1;$$

$$DS_1 \cup (DS_1 \cap DS_2) \equiv DS_1; \quad DS_1 \cap (DS_1 \cup DS_2) \equiv DS_1;$$

$$Ss_{p_1}(Ss_{p_2}(object)) \equiv Ss_{p_1 \text{ AND } p_2}(object).$$

2. Обмеження, не притаманні процесору оптимізації операцій реляційної алгебри:

- використання операції агрегації в межах операції запиту до структурованих даних;
- заміна операції консолідації для текстових даних на операцію трансформації тексту у семантичну мережу;
- заміна операції профілю користувача на запит до метаданих (здійснюється з метою уникнення операції з'єднання, що виконується для операції профілю).

3. Вартісні обмеження:

$$\min \sum_{j=1}^n c_j x_j, \quad \sum_{j=1}^n a_{ij} x_j \leq b_j, \quad i = 1, \dots, m, \quad x_j \in \{0, 1\}, \quad j = 1, \dots, n$$

$$op = \begin{cases} C, x_j = 0, M = 1, \\ S, x_j = 0, M = 2 \vee M = 3, \\ Se, x_j = 1, \end{cases}$$

де c_j – вартість доступу до джерела j ; a_{ij} – вартість завантаження об'єкта i з джерела j ; b_j – допустима межа затрат вартості (часу) для джерела j .

Результатом розділу є розроблення алгоритмів консолідації різнотипних ІІ, структур даних каталогу та словника даних, що стало основою для розроблення прикладних інформаційних систем консолідації даних на основі простору даних. Розроблено метамову простору даних, метод інтерпретації запиту, механізми оптимізації та трансформації операцій метамови в операції, що виконуються на рівні джерела даних, що дало можливість реалізувати операції пошуку та доступу.

Основні результати розділу опубліковані у [1, 2, 4, 7, 16, 21, 22, 24, 27, 28, 30, 32, 38, 48, 51, 52].

У **п'ятому розділі** формалізовано характеристики якості методів консолідації та простору даних загалом. Розроблено метод визначення якості консолідованих даних на основі функції якості та підвищення якості інтегрованих даних. Для оцінювання якості даних застосовано загальний методичний підхід до виділення адекватної номенклатури стандартизованих в ISO 9126 базових характеристик і суб-характеристик та розроблено вирази для оцінки якості даних. Доведено аналогії простору даних і нормованого простору.

Визначено параметри якості даних на основі характеристик стандарту ISO 9126.

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 сховища даних, неефективно застосовувати до простору даних, то визначення функціональної придатності є однією з базових характеристик, що досліджені у роботі:

$$z_1 = \frac{|cg'|}{|Ip_i.Ir|}. \quad (18)$$

Для забезпечення функціональної придатності додаються ІП з відповідністю структур даних 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:

$$z_2 = \frac{|S_P(Ip)|}{|cg'|}. \quad (19)$$

Додаються ІП з відповідністю структур даних еквівалентність.

Практичність – визначає корисність застосування консолідованих даних для певних користувачів. У ПД оцінка практичності здійснюватиметься за допомогою функції корисності прийнятих рішень (17):

$$z_3 = \frac{v(r)}{|cg'|}.$$

Для забезпечення практичності додаються ІП з відповідністю структур даних 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:

$$z_4 = \frac{|S_{meta_upd}(Ip_i.Cg)|}{|cg'|}. \quad (20)$$

Для забезпечення супроводжуваності додаються ІП з відповідністю структур даних синонімізм.

Визначено корисність даних з ІП стосовно прийняття рішення на їх 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 :

$$Q(z_1, \dots, z_4) = \sum_{i=1}^4 \left(\sum_k r_k z_i \prod_{j=1} P_{ij} \right) \rightarrow \max, \quad (21)$$

де j вказує на інформаційний продукт, P_{ij} – рівень довіри до інформаційного продукту j для рішення k , r_k – оцінка рішення k , C_1 – загальна вартість завантаження об'єктів, C_2 – загальна вартість модифікації описів, T – загальний час

завантаження, t_s – середній час завантаження одного об'єкту, c – середня вартість завантаження (модифікації) одного об'єкту:

$$\begin{aligned} 1 \geq z_1 \geq 0.75; z_3 \geq 0; \\ 0.25 \geq z_1 - z_2 \geq 0; 1 \geq z_4 \geq 0.5; \\ z_1 t_s \leq T; z_2 t_s \leq T; z_1 c \leq C_1; z_4 c \leq C_2. \end{aligned}$$

Це задача нелінійної оптимізації з лінійними обмеженнями, яка вирішується певними методами (градієнтним).

Поряд з фактичним оцінювання якості консолідованої інформації (21) проведено оцінювання якості еталонного зразка (22), що дає змогу отримати оптимальне рішення:

$$Q_i^e = \sum_i n_i z_i^e; Q_{const}' = \sum_i k_i Q_i^e, \quad (22)$$

де k_i – ранг важливості, $k_i \in [0; 1]$. Здійснено нормування фактичної оцінки:

$$Q_{const} = Q_{const}' / Q_{const}'^e. \quad (23)$$

(23) використано для прийняття рішень стосовно: зменшення невизначеності; визначення доцільності додавання ІП. Розроблено методику керування елементами простору даних на основі значення функції якості та рівнів довіри:

Set norm: [min, max]

Set ex

Sm($P \leq ex$) $\rightarrow$ Ip

$$f_{man} : \begin{cases} Q > \max(norm) \rightarrow \text{Set norm: [min, max], Set ex} \\ Q > \text{Avg}(norm) \rightarrow \text{profile}_{us} (p_{P_{Ip} \leq ex} (Cg)) \\ \exists(P_{Ip}) = 0, Q \approx \text{Avg}(norm) \rightarrow f_{Ip} (Cg) \\ Q < \text{Avg}(norm) \rightarrow DS - O_{P0} (p_{P_{Ip} \leq ex} (Cg)) \\ Q < \min(norm) \rightarrow f_{Ip_{new}} (Cg); Se(Ip_{new}) \end{cases} \quad (24)$$

Сценарії керування:

- встановити вищі значення граничних показників функції якості,
- перевизначити права користувачам,
- перерахувати рівні довіри, перевизначити відповідності,
- знищити ІП, в якого рівень довіри менший за заданий,
- знайти нові ІП для підвищення якості даних.

Спроековано схему організації просторів даних як складної інформаційної системи та підсистему забезпечення і підтримки якості даних у просторі даних.

Схему опрацювання даних з інформаційних продуктів подано на рис. 2.

Для моделювання зміни стану простору даних в часі наведено аналогії між простором даних DS та класичним означенням нормованого простору:

- елементом простору є множина станів ІП та стану СКД;
- ортами простору є показники якості $Z_1, \dots, Z_4$; значення орт розраховується на основі параметрів стану простору даних;

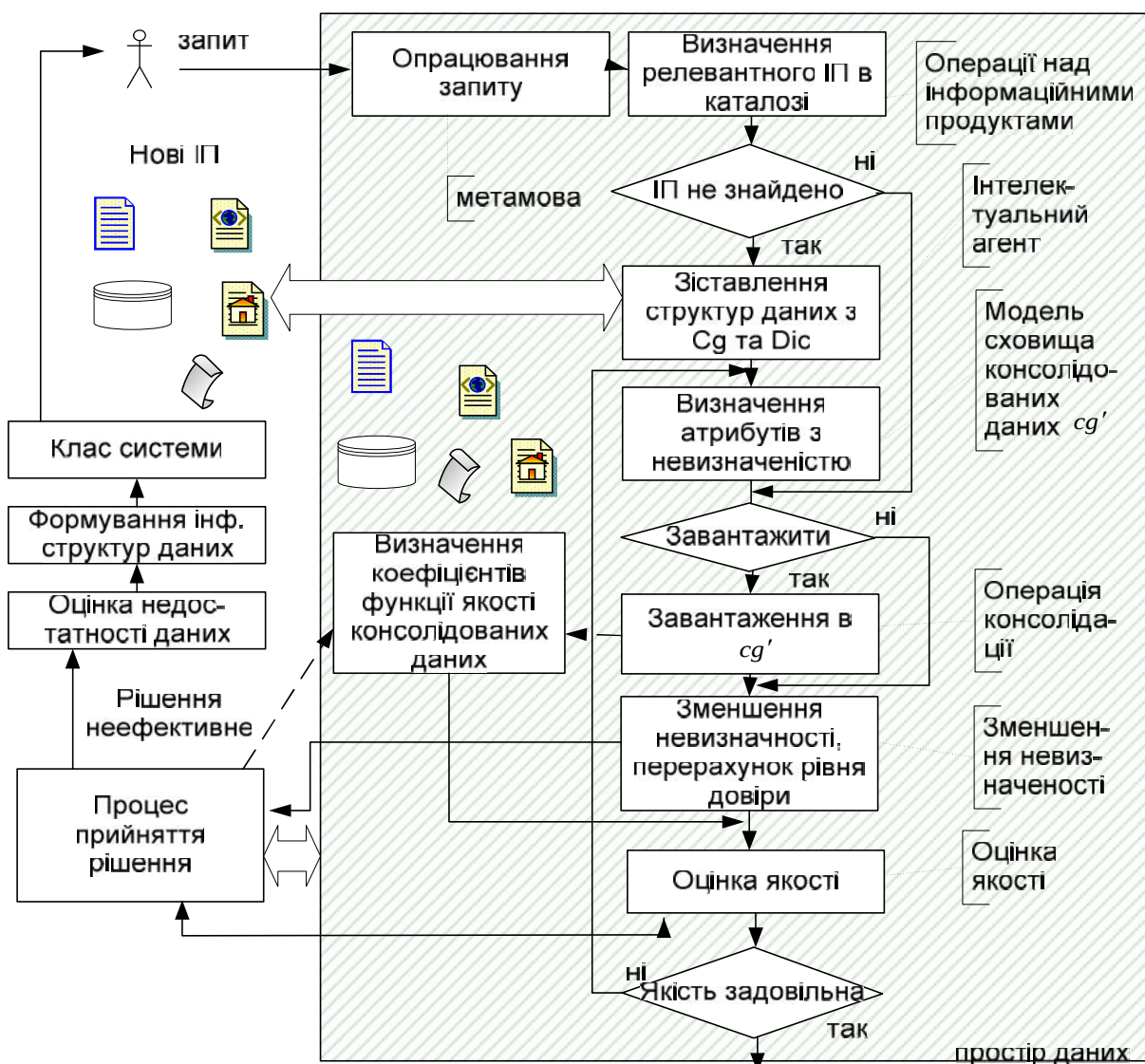


Рис. 2. Організаційна схема функціонування простору даних

- нормою простору є значення функції якості інтегрованих даних; якість даних порожнього простору – нульова;
- унарні операції – операції, що не призводять до зміни значення норми;
- бінарні операції, операції визначення структури даних та побудови семантичної мережі призводять до зміни значення норми (перехід в нову точку).

Наведена аналогія між інформаційним об'єктом «простір даних» та нормованим простором дала змогу досліджувати нові властивості DS :

- залежність показників якості та значення функції якості від кількості джерел даних (визначені як потужність множин DB, DW, Wb, Nd);
- дослідження поведінки користувача від значень коефіцієнтів якості даних;
- прогнозування поведінки користувача у відповідь на включення джерела даних певного типу у простір даних.

Результатом розділу є розроблення методу визначення якості консолідованих даних та сценарію керування простором даних.

Основні результати розділу опубліковані у [1, 6, 14, 42, 55].

У **шостому** розділі спроектовано архітектуру простору даних (рис. 3) та структуру модулів (рис. 4).

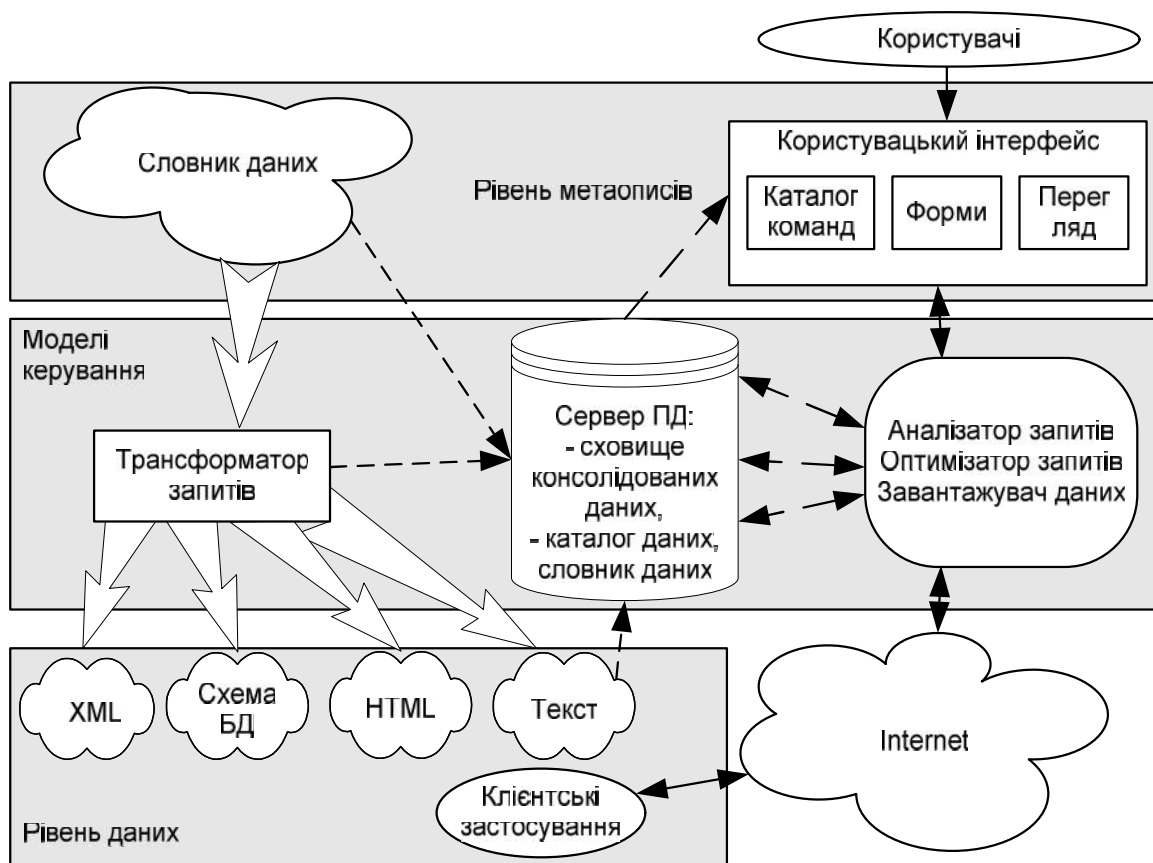


Рис. 3. Архітектура простору даних

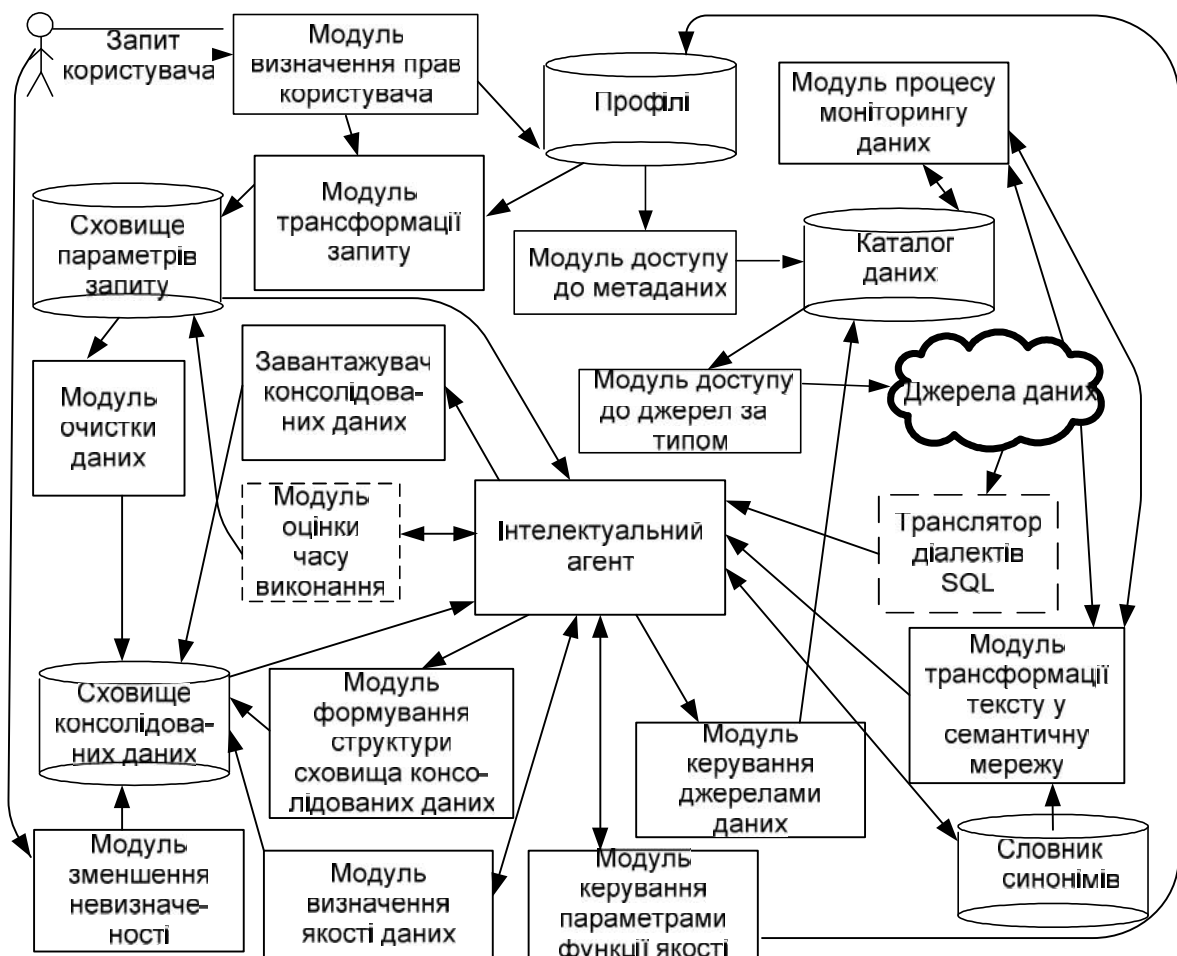


Рис. 4. Структура модулів

Розроблено схему каталогу даних *сг'* та словника синонімів. У схемі передбачено атрибути для зберігання рівня довіри до джерела даних. Схема каталогу даних використовується для реалізації операції запиту до метаданих та визначення профілю користувача. Подано проектне рішення для розроблення інформаційної технології для простору даних. Розроблено структуру класу керування компонентами простору даних та метамодель інтерфейсу формування запиту користувача.

Результатом розділу є розроблення архітектури простору даних та схеми каталогу простору даних з врахуванням рівня довіри до джерела даних. Описано метамову та інтерфейс запиту користувача.

Основні результати розділу опубліковані у [1, 31, 33, 53, 54].

У цьому розділі реалізовано інформаційні технології простору даних та описано прикладні системи, побудовані на їх основі, в яких здійснюються інтеграція та оцінювання якості даних.

Інформатизація ВНЗ викликає появу такої задачі, як інтеграція, оскільки університетом розроблено ряд інформаційних систем, які обмінюються між собою інформацією, а також надають частину інформацію у корпоративне сховище даних ВНЗ з метою подальшого її аналітичного опрацювання. Схему взаємодії основних БД університету зображено на рис. 5.

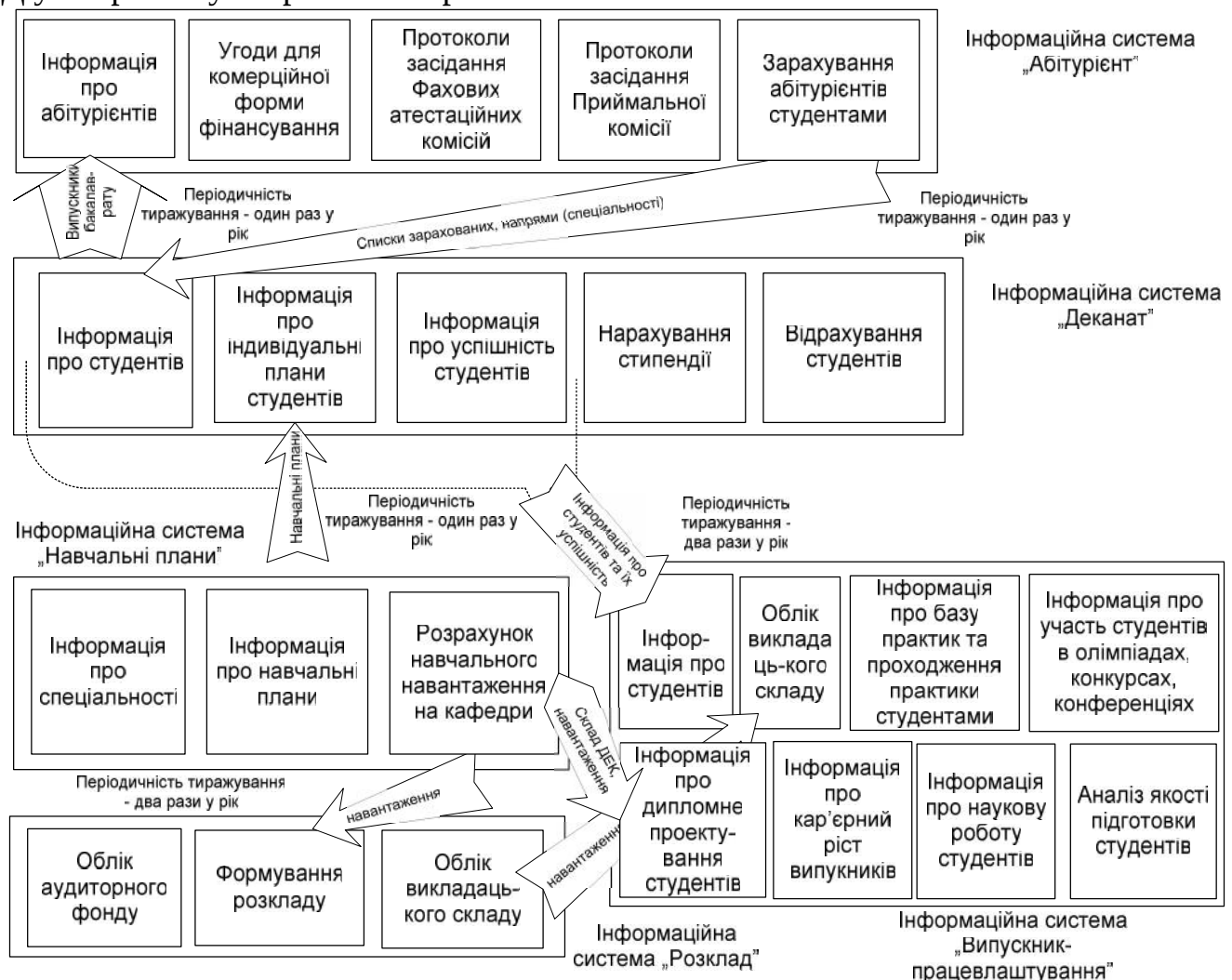


Рис. 5. Схема взаємодії основних БД ВНЗ

Використано три методи обміну даними: синхронізація – порівняння даних; експорт – копіювання даних таблиці за певний період; вибірка – копіювання частин таблиці за певний період за параметрами користувача. Проаналізовано результат інтеграції даних зі систем ВНЗ у двох варіантах: завантаження без попереднього аналізу даних, завантаження після аналізу даних. Проблеми під час завантаження даних традиційним чином виникають тоді, коли з'являються нові спеціальності чи групи, які необхідно додати у відповідні довідники.

Результат порівняння традиційного завантаження та з попереднім аналізом подано в табл. 1. Попередній аналіз даних дає змогу завантажувати весь обсяг інформації без втрат (перший та останній стовпці).

Таблиця 1.

Аналіз функціональної придатності – порівняння результату традиційного завантаження та з попереднім аналізом

К-сть студентів	К-сть нових груп	К-сть завантажених студентів традиційно	Кількість завантажених студентів після попереднього опрацювання даних
17324	0	17324	17324
17324	2	17272	17324
17324	4	17211	17324
17324	7	17001	17324

Для систем автоматизації навчального процесу ВНЗ характеристика якості «практичність» $z_3 = 1$, що і продемонстровано в табл. 1.

Також проаналізовано показник коректності та здійснено порівняння отриманих результатів з засобами BizTalk (немодифікована консолідація), MS Integration Services (рис. 6).

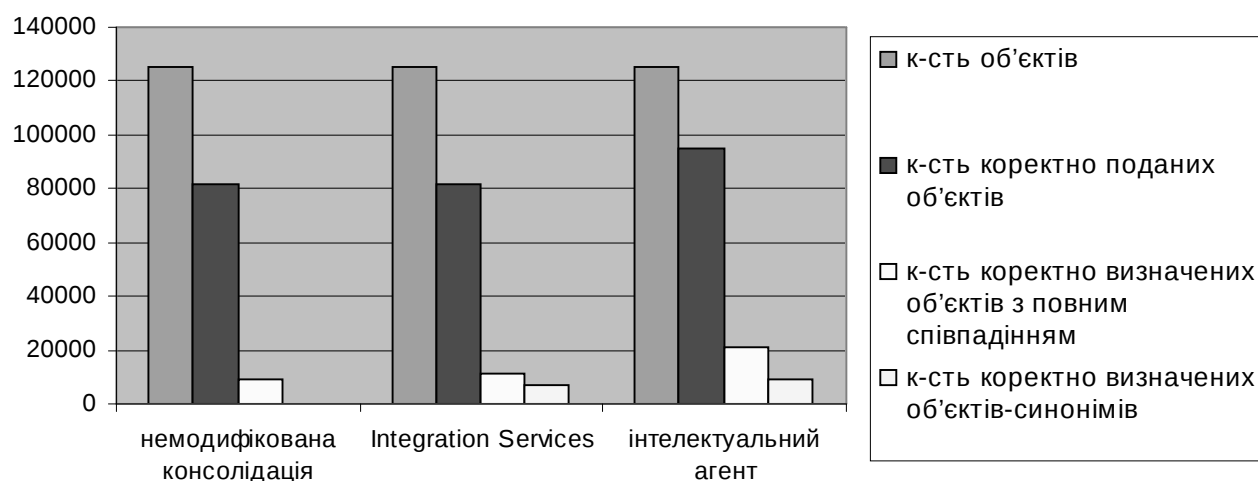


Рис. 6. Аналіз коректності

Окрім систем, джерелами даних є також локальні текстові файли користувачів з відомостями про студентів заочної форми навчання. Для текстових файлів будується семантична мережа. Правила її побудови задані у відношенні (рис. 7). Заповнення цього відношення – напівавтоматичне (спочатку адміністратор, а у разі успішного завантаження даних – інтелектуальним агентом).

id	name	type	division	includes	not_includes	synonic_id	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	SPECIAL		spets_id
6	група	STRING	ENTER	SYMBOL	SPECIAL	5	spets_id
7	прізвище	STRING	TR	SYMBOL	NUMBER	1	

Рис. 7. Приклад заповненого відношення словника синонімів

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

Таблиця 2.

Аналіз коректності даних з локальних джерел

К-сть джерел	К-сть похибок	Відсоток
5	3	60 %
12	6	50 %
23	12	52 %
31	13	42 %
45	15	37 %
67	23	34 %
75	24	32 %

У просторі даних туристичної сфери на якість даних впливають і повнота накопичених об'єктів, і корисність прийнятих рішень. Повнота накопичених описів об'єктів — відносна кількість об'єктів або документів, наявних у джерелах даних, до загальної кількості об'єктів, що потрапили у локальне сховище (проілюстровано на рис. 8). На поданій діаграмі проаналізовано роботу алгоритму модифікованої консолідації. Роботу алгоритму порівняно з роботою засобу інтеграції ETL, застосованому в Oracle Data Integrator.

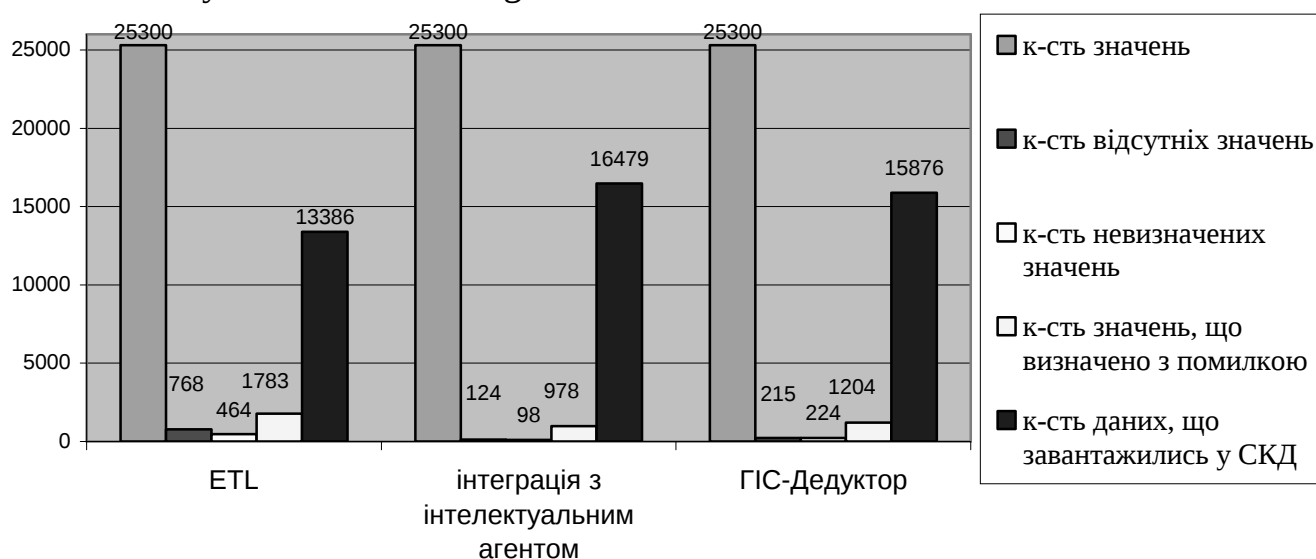


Рис. 8. Аналіз повноти накопичених описів об'єктів

Дані у сховище консолідованих даних потрапляють з баз даних туристичних агенцій, схеми яких наперед невідомі. Кількість записів вхідних баз даних – 15000.

Без попереднього визначення структур джерел частина даних не була завантажена у сховище консолідованих даних (кількість відсутніх даних на діаграмі), частина не могла бути завантажена через неспівпадіння структури локального сховища та структури джерела (невизначені дані на діаграмі). Кількість записів, завантажених за допомогою консолідації з інтелектуальним агентом, є найбільшою.

Проаналізовано, як зміниться корисність прийнятих рішень від кількості джерел, дані з яких використовувались для його підтримки. Результати тестування показали, що співпадіння показників із плановими є вищим після застосування алгоритмів усунення невизначеностей, що доводить доцільність роботи.

На основі тестових даних розраховано коефіцієнти важливості кожного з показників для розроблення комплексного показника якості простору даних. Розрахунок здійснено наступним чином. Коефіцієнт актуальності встановлено в 0, коефіцієнти інших показників визначено як усереднене значення за параметрами плану. Як показали тестові дані, найбільша точність плану (за формулою (22)) досягається не тоді, коли у сховище консолідованих даних потрапляє більшість описань об'єктів, а тоді, коли вони надходять точними (без невизначеності).

	План по точних даних	План з показником «функ.придатність»>0.9	План з показником «коректність»>0.9
Працівники	92	67	86
Тур. послуга	93	91	92
Відпочиваючі	87	74	81

Визначення показника якості: $n_i = \sum_{j=1}^k coef_j^i / \sum_{j=1}^k plan_j^i$, де n_i – показник, що розраховується, $plan_j^i$ – значення плану за точними даними, $coef_j^i$ – значення плану за інтегрованими даними. Для поданого прикладу $n_1 = 0.84$, $n_2 = 0.95$.

На прикладі простору даних наукових досліджень показано результат налаштування параметрів функції якості особою, що приймає рішення (ОПР). Визначено значення функції якості та показано рішення, що прийняла ОПР стосовно відхилення чи підтвердження джерела даних (див. табл. 3).

Таблиця 3.

Результат налаштування параметрів функції якості особою, що приймає рішення

Джерело	Значення функції якості	Відхилено	Джерело	Значення функції якості	Відхилено
1	0,82	1	3	0,7	0
2	0,62	0	3	0,72	0
2	0,67	0	4	0,5	1
2	0,7	0	6	0,8	1

Розв'язано задачу інтеграції даних у режимі реального часу з робочих місць касирів супермаркетів, які розділені географічно. Однією з проблем, що виникала з використанням традиційних способів завантаження, була надмірне навантаження на сервер. Прийняття тактичних рішень ускладнювалися у зв'язку з недостовірною інформацією. Стратегічні рішення щодо керування мережею супермаркетів приймаються на основі консолідованих даних, і потреби у надоперативності даних немає. Для визначення якості інтегрованих даних проаналізовано, наскільки

співпадають прийняті рішення щодо стратегічного розвитку на основі еталонних даних, консолідованих даних та даних, отриманих традиційним шляхом. Було визначено такі показники відсотків інвестицій:

- | | |
|------------------------------|----------------------------------|
| 1) у персонал, | 5) у розвиток інфраструктури, |
| 2) у розвиток транспорту, | 6) у рекламу, |
| 3) в закупівлю продукції, | 7) в управлінську сферу, |
| 4) у програмне забезпечення, | 8) в енергозберігаючі установки. |

Результат порівняння реальних показників та показників по консолідованих даних подано на рис. 9. По осі *У* відкладено частку інвестицій у показники.

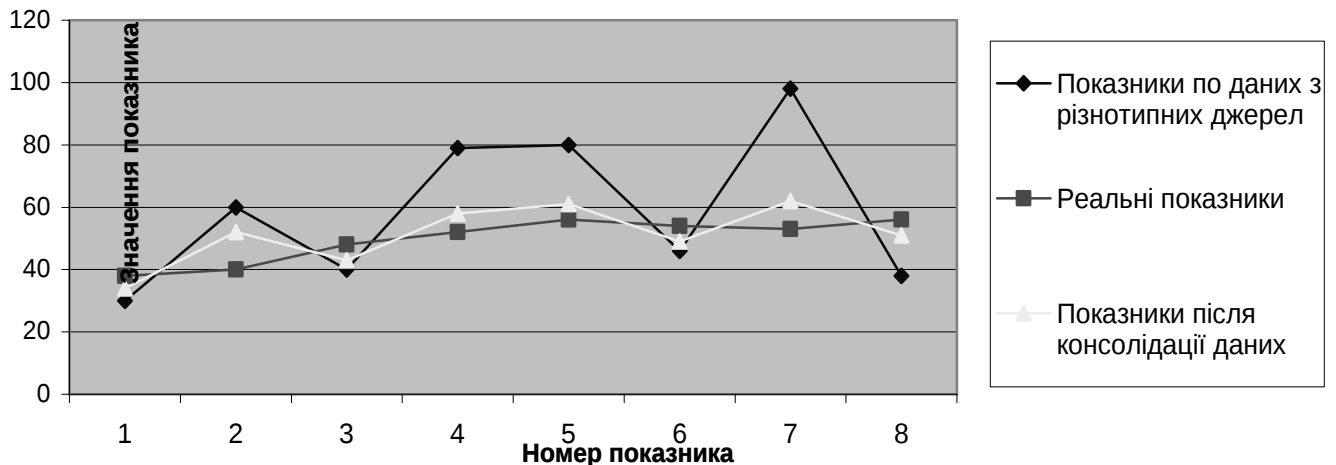


Рис. 9. Дослідження якості прийняття рішень по інтегрованих даних

Проаналізовано результат роботи агента для всіх розглянутих вище предметних областей. Увесь період роботи агента розділено на фази: навчання; експлуатації.

Фаза навчання супроводжується заповненням словника та каталогу даних. До початку фази навчання вони містять початкові відомості про об'єкти простору даних:

- фізичне розміщення об'єктів – для подальшого визначення структури даних;
- стандарти називання характеристик предметної області – для формування бази правил;
- встановлення початкових рівнів довіри до даних – для налаштування формування відношень відповідності;
- визначення основних операцій над об'єктами – для визначення сценаріїв розвитку системи;
- визначення еталонних даних, по яких прийматимуться рішення для налаштування коефіцієнтів функції якості.

Фаза експлуатації характеризується переважно стабільним характером роботи агента, за виключенням збурень. Збуреннями (атаками) є додавання до простору даних інформаційних продуктів з такою структурою даних, що приводить до появи невизначеності у словнику даних (багатозначність, недетермінованість). Це, у свою чергу, викликає зниження якості консолідованих даних, а, отже, зменшення ефективності управлінських рішень (рис. 10).

По осі *У* на рис. 10 відкладено відношення значення функції якості даних до значення функції якості еталонних даних, по осі *X* – час експлуатації агента.

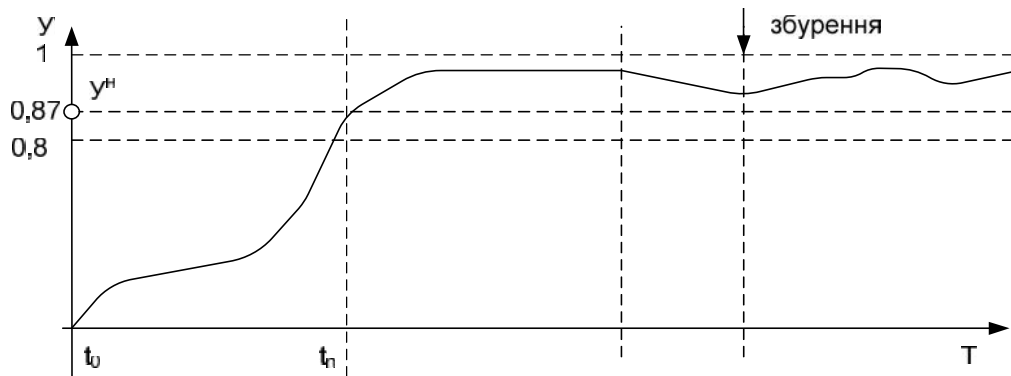


Рис. 10. Дослідження роботи агента співставлення структур даних в часі

Незважаючи на наявність збурень, досягається певний рівень стабільності роботи агента y^n . Експериментально визначено, що це значення не опускається нижче, ніж 0,87 (шум менший, ніж 13%), що, у свою чергу, є краще, ніж заявлений у Клермонтському з'їзді рівень допустимості шуму 25%.

У розділі продемонстровано результати розроблених методів опрацювання різнотипних даних та визначено якість інтегрованих даних.

Основні результати розділу опубліковані у [1, 2, 11, 13 15, 18, 20, 23, 25, 26, 47,49,52].

У **додатку** наведені акти впровадження результатів дисертаційної роботи.

ВИСНОВКИ

У дисертаційній роботі вирішено важливу наукову проблему організації та інтеграції різнотипних джерел даних у простір даних з метою підвищення якості консолідованих даних за допомогою використання розроблених теоретичних засад, системної методології та програмних засобів опрацювання множини інформаційних продуктів. У результаті виконання цієї роботи одержані наступні результати.

1. Здійснено аналіз проблеми подання та опрацювання різнотипних джерел даних. Обґрунтовано актуальність розв'язання цієї проблеми на основі введення нової абстракції керування даними – простору даних, що дало змогу виділити невирішені раніше проблеми з опрацювання та консолідації даних з наперед невідомих джерел.
2. Розроблено алгебричну систему «простір даних», яка складається з множини інформаційних продуктів, предикатів та операцій на них. Це дало змогу розробити операції консолідації та пошуку даних з різнотипних джерел, структура даних яких наперед невідома.
3. Розроблено інтелектуальний агент визначення структури джерела даних за допомогою порівняння структур джерел даних, наявних у ПД, із структурами джерел даних, які входять у ПД, що уможливило формування єдиного типу запитів до джерел даних з врахуванням ступеня довіри до джерела та отримання коректних відповідей на сформовані запити.
4. Розроблено структури даних каталогу даних та синонімічного словника та методи розрахунку ступеню довіри користувача до джерел даних, що дало змогу збільшити релевантність відповіді та розробити метод визначення якості консолідованих даних.
5. Розроблено метод інтеграції даних із розрізнених джерел на базі модифікованого

оператора консолідації, що дало змогу не тільки збільшувати якість інформації у ПД, але й інтерпретувати результати моделювання. Проаналізовано коректність розроблених засобів та порівняно їх результати з існуючими. Розроблений інтелектуальний агент та операція інтеграції дає кращий результат від існуючих засобів на 1,4%.

6. Розроблено метод пошуку даних за запитом користувача та валідації отриманих результатів з метою уніфікації роботи з різнотипними джерелами даних, що дало змогу покращити якість консолідованих даних.
7. Розроблено метод трансформації запиту користувача у запит до джерела даних довільного типу та оптимізації запиту, що уможливило забезпечення коректної відповіді на запит.
8. Розроблено метод визначення якості консолідованих даних на основі формалізації стандарту ISO 9126, що уможливило визначати придатність цих даних для подальшого прийняття рішень.
9. Запропоновані методи та алгоритми використані для розроблення:
 - а) програмно-алгоритмічної системи ПД, яка впроваджена у Центрах управління туризмом Львівської та Чернівецької областей,
 - б) у процесі інформатизації Національного університету «Львівська політехніка»,
 - в) у навчальному процесі Національного університету «Львівська політехніка» для викладання курсів «Проектування інформаційних систем», «Технології сховищ даних»,
 - г) для аналізу резюме у Львівській консалтинговій групі,
 - д) для консолідації даних мережі супермаркетів «Сільпо»,
 - е) для формування наукового звіту в Львівському інституті банківської справи Університету банківської справи Національного банку України,
 - ж) для консолідації даних у відділі гарантійного обслуговування ТОВ «Стек-комп'ютер»;
 - з) аналітичної системи збору та опрацювання наукових публікацій для Інституту регіональних досліджень Національної академії наук України та Львівської національної наукової бібліотеки імені В.Стефаника;
 - и) для консолідації інформації про клієнтів для приватного підприємства «Анат».

СПИСОК ОСНОВНИХ ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Шаховська Н. Б. Програмне та алгоритмічне забезпечення сховищ та просторів даних : монографія / Н. Б. Шаховська; Міністерство освіти і науки України, Національний університет «Львівська політехніка». – Львів : Видавництво Львівської політехніки, 2010. – 194 с.
2. Шаховська Н. Б. Сховища та простори даних : монографія / Н. Б. Шаховська, В. В. Пасічник; Міністерство освіти і науки України, Національний університет «Львівська політехніка». – Львів : Видавництво Національного університету «Львівська політехніка», 2009. – 240 с.
3. Шаховська Н. Б. Формальне подання простору даних у вигляді алгебраїчної системи / Н. Б. Шаховська // Системні дослідження та інформаційні технології = System research & information technologies: міжнародний науково-технічний

- журнал / Національна академія наук України, Інститут прикладного системного аналізу. – Київ, 2011. – № 2. – С. 128 – 140.
4. Шаховська Н. Б. Простір даних області наукових досліджень / Н.Б. Шаховська // Моделювання та інформаційні технології : збірник наукових праць / Національна академія наук України, Інститут проблем моделювання в енергетиці ім. Г.Є.Пухова. – 2008. – Вип. 45. – С. 132 – 140.
 5. Шаховська Н. Б. Формалізація простору даних за допомогою алгебраїчної системи / Н. Б. Шаховська // Радіoeлектроніка, інформатика, управління / Запорізький національний технічний університет. – Запоріжжя : Видавництво ЗНТУ, 2010. – №1. – С. 102 – 109.
 6. Шаховська Н. Б. Організація просторів даних для забезпечення якості консолідованих даних / Н. Б. Шаховська // Реєстрація, зберігання і обробка даних : науково-технічний журнал / Інститут проблем реєстрації інформації НАН України. – 2011. – № 3. – С.86 – 97.
 7. Шаховська Н. Б. Методи опрацювання консолідованих даних за допомогою просторів даних / Н. Б. Шаховська // Проблеми програмування: науково-технічний журнал / Національна академія наук України, Інститут програмних систем НАН України. – 2011, № 4. – С. 72–84.
 8. Шаховська Н. Особливості моделювання просторів даних / Н. Шаховська // Вісник Національного університету «Львівська політехніка». – Л. : Вид-во Нац. ун-ту "Львів. політехніка", 2008. – № 608 : Комп'ютерна інженерія та інформаційні технології. – С. 145 – 154.
 9. Шаховська Н. Б. Інформаційні компоненти інтелектуальних систем / В. В. Пасічник, Н. Б. Шаховська // Вісник Національного університету «Львівська політехніка». – Л. : Вид-во Нац. ун-ту «Львів. політехніка», 2008. – № 514 : Інформаційні системи та мережі. – С.191–203.
 10. Шаховська Н. Б. Простори даних: гносеологія, концепції та тенденції розвитку / Н. Б. Шаховська // Складні системи і процеси / Запорізький інститут державного та муніципального управління. – 2008. – №1. – С. 73 – 86.
 11. Шаховська Н. Б. Особливості інтеграції даних інформаційних систем Національного університету «Львівська політехніка / Н. Б. Шаховська, Д. О. Тарасов // Складні системи і процеси / Запорізький інститут державного та муніципального управління. – 2009. – №2. – С. 98 – 109.
 12. Шаховська Н. Б. Алгебраїчна система класу „простір даних» / Н. Б. Шаховська // Науковий вісник Чернівецького університету. Серія Фізика. Електроніка : Збірник наукових праць. – Чернівці, 2009. – № 479 : Тематичний випуск «Комп'ютерні системи та компоненти». – С. 48 – 51.
 13. Шаховська Н. Б. Простори даних туристичної сфери / Н. Б. Шаховська, Д. І. Угрин // Відбір і обробка інформації : міжвідомчий збірник наукових праць / Національна академія наук України, Фізико-механічний інститут ім. Г.В. Карпенка. – Львів : Фізико-механічний інститут ім. Г.В. Карпенка НАН України, 2008. – Вип. 13. – С. 126 – 134.
 14. Шаховська Н. Б. Проблеми якості консолідованих даних у просторах даних / Н. Б. Шаховська, О. Пшеничний, І. Чорней // Системи обробки інформації : збірник наукових праць. – Харків, 2011. – Вип. 3 (93). – С. 80 – 84.

15. Чорней І. Математичне забезпечення моделювання пандемії грипу / І. Чорней, Н. Шаховська // Вісник Національного університету «Львівська політехніка». – Л.: Вид-во Нац. ун-ту "Львів. політехніка", 2011. – № 710: Комп'ютерна інженерія та інформаційні технології. – С.122 – 128.
16. Шаховська Н. Б. Вирішення проблем комплексного аналізу в просторах даних туризму / Н. Б. Шаховська, Д. І. Угрин // Автоматизированные системы управления и приборы автоматики: Всеукраинский межведомственный научно-технический сборник / Министерство образования и науки Украины; Харьковский национальный университет радиоэлектроники. – 2007. – №140. – С. 110–117.
17. Шаховська Н. Б. Формалізація операцій над джерелами даних у просторі даних / М. О. Медиковський, Н. Б. Шаховська // Оптико-електронні інформаційно-енергетичні технології: Міжнародний науково-технічний журнал. – Вінниця, 2009. – №1(17). – С. 49–58.
18. Шаховська Н. Б. Аналіз методів підтримки прийняття рішень у лікувальних системах / Н. Б. Шаховська, Н. І. Мельникова // Математичні машини і системи / Інститут проблем математичних машин і систем. – Київ, 2011. – № 2. – С. 62–72.
19. Шаховська Н. Б. Аналіз задачі опрацювання відсутності та неповноти інформації у сховищі даних / Н. Б. Шаховська, Д. І. Угрин // Вісник Національного університету «Львівська політехніка». – Л.: Вид-во Нац. ун-ту «Львів. Політехніка», 2008. – № 519 : Інформаційні системи та мережі. – С. 215 – 225.
20. Шаховська Н. Б. Основні підходи до розроблення програмного комплексу автоматичного реферування текстових документів / В. Я. Крайовський, В. В. Литвин, Н. Б. Шаховська // Інститут проблем моделювання в енергетиці: Збірник наукових праць / Національна академія наук України, Відділення фізико-технічних проблем енергетики, Інститут проблем моделювання в енергетиці. Збірник наукових праць. – Київ, 2009. – Вип. 51. – С. 178 – 186.
21. Шаховська Н. Б. Опрацювання даних для інформаційних технологій туристичного бізнесу / Н. Б. Шаховська, Д. І. Угрин // Відбір і обробка інформації: міжвідомчий збірник наукових праць / Національна академія наук України, Фізико-механічний інститут ім. Г.В. Карпенка. – Львів: [Фізико-механічний інститут ім. Г.В. Карпенка НАН України], 2008. – Вип. 16. – С. 106 – 114.
22. Шаховська Н. Б. Простір даних веб-спільнот / Н. Б. Шаховська, Ю. О. Серов // Автоматизированные системы управления и приборы автоматики: Всеукраинский межведомственный научно-технический сборник / Министерство образования и науки Украины; Харьковский национальный университет радиоэлектроники. – 2008. – № 143. – С. 74 – 80.
23. Шаховська Н. Б. Організація побудови просторів даних туристичної сфери / Н. Б. Шаховська, Д. І. Угрин // Штучний інтелект / Національна академія наук України; Інститут проблем штучного інтелекту. – Донецьк, 2009. – № 2. – С.82–90.
24. Шаховська Н. Б. Інтеграція розрізнених даних про клієнта з використанням інтелектуального агента визначення структури джерела / Н. Б. Шаховська, Д. І. Угрин // Моделювання та інформаційні технології: збірник наукових праць / Національна академія наук України, Інститут проблем моделювання в енергетиці ім. Г.Є.Пухова. – 2009. – № 50. – С.165 – 173.

25. Шаховська Н. Б. Простори даних у виставковій діяльності / Н. Б. Шаховська, О. Лозицький // Вісник Національного університету «Львівська політехніка». – Л. : Вид-во Нац. ун-ту «Львів. політехніка», 2008. – № 631 : Інформаційні системи та мережі. – С.219 – 226.
26. Шаховська Н. Б. Аналіз інформаційних систем оброблення даних туристичної сфери / Н. Б. Шаховська, Д. І. Угрин // Науковий вісник Національного лісотехнічного університету України : Збірник науково-технічних праць. – 2008. – № 18. – С. 258–263.
27. Шаховська Н. Б. Про задачу автоматизованого анотування події на основі простору даних / В. В. Литвин, Н. Б. Шаховська // Науковий вісник Чернівецького університету. Серія Фізика. Електроніка : Збірник наукових праць. – Чернівці, 2008. – № 426 : Тематичний випуск «Комп'ютерні системи та компоненти». – С. 58–63.
28. Шаховська Н. Б. Технологія ETL в інтеграції даних туристичного бізнесу / Н. Б. Шаховська, Д. І. Угрин // Автоматизированные системы управления и приборы автоматики: Всеукраинский межведомственный научно-технический сборник / Министерство образования и науки Украины; Харьковский национальный университет радиоэлектроники. – 2008. – Вып.145. – С. 68 – 73.
29. Кісь Я. П. Інтелектуальні геоінформаційні системи. Міжнародний досвід та шляхи розвитку в Україні / Я. П. Кісь, Н. Б. Шаховська, О. Б. Вальчук // Вісник Національного університету «Львівська політехніка». – Л.: Вид-во Нац. ун-ту «Львів. політехніка», 2008. – №653: Інформаційні системи та мережі. – С. 139 –145.
30. Шаховська Н. Інформаційна система реферування множини документів, поданих у різних форматах, базована на онтології / Н. Шаховська, В. Литвин, В. Крайовський // Вісник Національного університету «Львівська політехніка». – Л. : Вид-во Нац. ун-ту "Львів. політехніка", 2010. – № 672 : Комп'ютерні науки та інформаційні технології. – С. 63 – 72.
31. Шаховська Н. Б. Сховища та простори даних – інформаційний фундамент систем прийняття рішень / Н. Б. Шаховська, В. Я. Крайовський // Науковий вісник Чернівецького університету. Серія Фізика. Електроніка : Збірник наукових праць. – Чернівці, 2011. – Т2, вип. 1 : Тематичний випуск «Комп'ютерні системи та компоненти». – С. 50–64.
32. Литвин В. В. Реферування текстових документів на основі зважування міри TF-IDF онтологією предметної галузі / В. В. Литвин, Н. Б. Шаховська, В. Я. Крайовський // Вісник Національного університету «Львівська політехніка». – Л. : Вид-во Нац. ун-ту «Львів. політехніка», 2010. – № 689 : Інформаційні системи та мережі. – С. 154 – 164.
33. Schahovs'ka N. Web-community dataspace / N. Schahovs'ka, Yu. Syerov // International e-journal 'Information Technology for Economics and Management'. [Electronic resource] – Mode of access: WWW/URL: – <http://www.item.woiz.polsl.pl/issue4.1/journal4.1.htm>. – Last access: 15.04.2009. Title from screen.

34. Shakhovska N. Modeling of energy system dataspace / N. Shakhovska, M. Medykovsky // Computational Problems of Electrical Engineering. – Lviv, 2011. – Vol. 1, N. 1. – P. 77 – 83.
35. Шаховська Н. Б. Методи опрацювання даних про клієнта з використанням інтелектуального агента визначення структури джерела даних / Н. Б. Шаховська // Актуальные проблемы экономики. – 2011. – № 7 (120). – С. 338–346.
36. Shakhovska N. Algebraic system of dataspace / N. Shakhovska // Proc. of International Conference on Intellectual Systems for Decision Making and Problems of Computational Intelligence «ISDMCI'2011»: Conference proceedings in two volumes, 16-20 May 2011, Yevpatoria. – Vol. 1. – Kherson, 2011. – P. 14–18.
37. Dragan Ya. P. System analysis and problem of dataspace concept in information technologies / Ya. P. Dragan, M. O. Medykjovskyi, N. B. Shakhovska // Proc. of the VI International Scientific and Technical Conference 'CSIT-2011', 15-19 November 2011, Lviv. – Lviv, 2011. – P. 178–180.
38. Shakhovska N. Dataspace class algebraic system for modeling integrated processes / Shakhovska N., Lipinski J., Medykovsky M., Lytvyn V. // Proc. of the XIV International Conference 'System Modelling and Control', June 27-29, 2011, Łódź, Poland. – Łódź, 2011. – P. 23.
39. Shakhovska N. Consolidated processing for differential information products / Shakhovska N. // Proc of the 7 International Conf. on Perspective Technologies and Methods in MEMS Design 'MEMSTECH-2011', Polyana, 11-14 May 2011. – Lviv, 2011. – P. 176.
40. Lytvyn V. Searching the relevant precedents based on adaptive ontology / V. Lytvyn, N. Shakhovska, D. Dosyn, V. Pasichnyk // Proc. of the 12th International Workshop «Computational Problems of Electrical Engineering», Kostryna, Ukraine, 5-7 September, 2011. – Lviv, 2011. – P. 43 – 44.
41. Shakhovska N. Datawarehouse and dataspace – information base of decision support system / N. Shakhovska // Proc. of the 7 International Conf. on The Experience of Designing and Application of CAD Systems in Microelectronics 'CADSM-2011', 23-25 Feb. 2011, Polyana-Svalyava. – Lviv, 2011. – P. 46 – 48.
42. Шаховська Н. Б. Параметри якості консолідованих даних у просторах даних / Н. Б. Шаховська // Матеріали 1 міжнародної науково-практичної конференції «Актуальні проблеми інформаційних технологій, економіки та права», 23-24 лютого 2011 р., Чернівці. – Чернівці, 2011. – С. 98 – 102.
43. Shakhovska N. Web-community dataspaces / N. Shakhovska, J. Lipiński, M. Medykowski, Yu. Syerov // Proc. of the XVI Intern. Conf. on Information Systems and Networks «Sieci i Systemy Informatyczne», Poland, 2008. – Łódź, 2008. – P. 131–132.
44. Shakhovska N. Web-community model, based on dataspaces / N. Shakhovska, J. Lipiński, M. Medykowski, Yu. Syerov // Proc. of International Conference on Intellectual Systems for Decision Making and Problems of Computational Intelligence «ISDMCI'2009»: Conference proceedings in two volumes, 18-22 May 2009, Yevpatoria. – Vol. 2. – Kherson, 2009. – P. 9–13.
45. Шаховська Н. Б. Простори даних: поняття та призначення / Н. Б. Шаховська // Матеріали II міжнародної наукової конференції «Комп'ютерні науки та інфор-

- маційні технології»CSIT-2007, 27-29 вересня 2007 р., Львів. – Львів, 2007. – С. 269–277.
46. Bobalo Yu. Structure and programme implementation of Electronic textbook «Principles of electronic circuits theory» / Yu. Bobalo, P. Stakhiv, B. Mandziy, N. Shakhovska, R. Holoschuk // Proc. of the 12th International Workshop «Computational Problems of Electrical Engineering», Kostryna, Ukraine, 5-7 Sept., 2011. – Lviv, 2011. – P.12-13.
 47. Пшеничний О. Ю. Попередній аналіз даних по захворюваності на панкреатит у м. Львів у 2000-2010 рр. / О. Ю. Пшеничний, Н. Б. Шаховська, А. С. Мельник // Системний аналіз та інформаційні технології : Матеріали міжнародної науково-технічної конференції (SAIT 2011), Київ, 23-28 травня 2011 р. / ННК ІПСА НТУУ КПІ. – К. : НТУУ КПІ, 2011. – С. 143 – 144.
 48. Шаховська Н. Б. Інтеграція даних з розрізнених джерел в туризмі / Н. Б. Шаховська, Д. І. Угрин // Materiały V międzynarodowej naukowo-praktycznej konferencji «Nauka: teoria i praktyka – 2009» / Volume 6, Matematyka. Nowoczesne informacyjne technologie. Fizyka. Chemia i chemiczne technologie, Przemysł, 2009. – Przemysł, 2009. – P. 13 – 17.
 49. Шаховська Н. Б. Моделі об'єктів в просторах даних туризму / Н. Б. Шаховська, Д. І. Угрин // Научная мысль информационного века – 2008 : Матеріали IV міжнародної науково-практичної конференції. – Прага, 2008. – № 13. – С.28–31.
 50. Шаховська Н. Б. Операції над просторами даних / III міжнародна науково-технічна конференція „Комп'ютерна математика в науці, інженерії та освіті» CMSEE-2009. – Полтава : ПолтНТУ імені Юрія Кондратюка, 2009. – С. 32 – 34.
 51. Шаховська Н. Б. Альтернативні рішення ситуацій опрацювання даних з розрізнених джерел в просторах даних туризму / Н. Б. Шаховська, Д. І. Угрин // Дни науки – 2008 : Матеріали IV міжнародної науково-практичної конференції. – Софія, 2008. – № 13. – С. 28 – 31.
 52. Шаховська Н. Б. Алгоритм реферування документів, поданих у різних форматах, з використанням онтології / Н. Б. Шаховська, В. В. Литвин, В. Я. Крайовський // Системний аналіз та інформаційні технології : матеріали 12-ї Міжнародної науково-технічної конференції (SAIT 2010), Київ, 25–29 травня 2010 р. / ННК ІПСА НТУУ КПІ. – К.: ННК ІПСА НТУУ КПІ, 2010. – С. 505 – 506.
 53. Шаховська Н. Б. Математичне та програмне забезпечення просторів даних / Н. Б. Шаховська // V юбилейная научно-практическая конференция с международным участием «Математическое и имитационное моделирование систем (МОДС 2010), 21-25 июня 2010 г., Киев. – Киев, 2010. – С.58 – 62.
 54. Shakhovska N. The main areas of dataspace realization / N. Shakhovska, M. Medykovsky, L. Lipinska // Proc. of the 12th International Workshop «Computational Problems of Electrical Engineering», Kostryna, Ukraine, 5-7 Sept., 2011. – Lviv, 2011. – P.28–29.
 55. Шаховська Н. Б. Аналіз складових проблеми опрацювання різнотипних даних / Н. Б. Шаховська, М. О. Медиковський // Международ. научно-технич. конференция «Информационные технологии и информационная безопасность в науке, технике и образовании "ИНФОТЕХ-2011"», 05-10 сентября 2011, Севастополь. – Севастополь, 2011. – С. 64–65.

АНОТАЦІЇ

Шаховська Н.Б. Організація просторів даних у складних інформаційних системах. – На правах рукопису.

Дисертаційна робота на здобуття наукового ступеня доктора технічних наук за спеціальністю 05.13.06 – «інформаційні технології». – Національний університет «Львівська політехніка» МОНмолодьспорту України, Львів, 2012.

У дисертаційній роботі вирішено важливу наукову проблему організації різнотипних джерел даних у простір даних з метою підвищення якості інтегрованих даних за допомогою використання розроблених теоретичних засад, системної методології та програмних засобів опрацювання множини інформаційних продуктів.

Здійснено аналіз проблеми подання та опрацювання різнотипних джерел даних. Обґрунтовано актуальність розв'язання цієї проблеми на основі введення нової абстракції керування даними – простору даних, що дозволило виділити невирішені раніше проблеми з опрацювання та консолідації даних з наперед невідомих джерел. Розроблено алгебричну систему сигнатури простір даних, яка складається з множини інформаційних продуктів, предикатів та операцій на них. Це дозволило розробити операції консолідації та пошуку даних з різнотипних джерел, структура даних яких наперед невідома. Розроблені методи та алгоритми використані для створення ряду інформаційно-аналітичних систем.

Ключові слова: простір даних, алгебрична система, інтеграція, інтелектуальний агент, функція якості консолідованих даних, дані з невизначеністю.

Shakhovska N.B. Dataspace organization in complex information systems. – Manuscript.

Dissertation for obtaining doctor degree in technical science on speciality 05.13.06 information technologies. – Lviv Polytechnic National University Ministry of education and science, youth and sport of Ukraine, Lviv, 2012.

This thesis solved a major scientific problem of diverse data sources in dataspace to improve the quality of integrated data by using the developed theoretical principles, systematic methodology and software tools for processing a set of information products. The analysis of the problem of diverse data sources representation and processing is given.

In *first* chapter there is actuality solve this problem by introducing a new abstraction data management – dataspace, which allowed to select unsolved problems with processing and consolidation of data from unknown sources in advance.

In the chapter *two* an algebraic dataspace system is described. It consists of a set of information products, predicates and operations on them. It is possible to develop consolidation operation, operation of search data from diverse sources, denote of data structure which one is unknown. Firstly intelligent agent determining the structure of the data source is developed. It works by comparing the structures of data sources available in the dataspace with data sources that will include in the dataspace. This agent allowed forming a single type of queries to data sources, taking into account the degree of confidence in the source and getting the correct answers to formed queries.

The substantiated data structures and data directory synonymous dictionary is given. The solve methods the trust of user data sources are given, which increased the relevance of responses and develop a method of determining the quality of consolidated data.

The *third* chapter there is introduced consolidated datawarehouse model as an

improved model of the uncertainty relation. It is simulated the physical object – a datawarehouse, which one indicating it attributes with clear and undefined values – to reduce uncertainty and taking into account the existence of public confidence in the data source to increase the effectiveness of management decisions. There is analyzed the causes of uncertainties in storage and data spaces. Among them are highlighted: the uncertainty in the scheme of the mediator; uncertain schema mapping, data uncertainty consolidated data repository.

In the *next* chapter the algorithms integrate data from disparate sources on the basis of the modified operator consolidation are described. That algorithm is enabling not only increase the quality of information in the data space, but also to interpret the simulation results. The methods and algorithms used to create a number of information-analytical systems. Next, there is build the method of transformation commands to dataspace in commands to data sources. The method of optimizing a query is build.

In the *sixth* chapter the dataspace architecture is given. The modules schema is built. There are described metadata schema, main classes and functions. There is defined the principles of functional making of dataspace. The query language of dataspace is described.

In the *seventh* chapter there are implemented application systems in which is used the data integration built on the dataspaces. There are tested the results of methods, that are built in dissertation.

Keywords: dataspace, algebraic system, integration, intelligence agent, quality function of consolidated data, data with uncertainty.

Шаховская Н.Б. Организация пространств данных в сложных информационных системах. – На правах рукописи.

Диссертация на соискание ученой степени доктора технических наук по специальности 05.13.06 – «информационные технологии». – Национальный университет «Львовская политехника» МОНмолодежьспорта Украины, Львов, 2012.

В диссертационной работе решена важная научно-прикладная проблема организации разнотипных источников данных в пространство данных с целью повышения качества интегрированных данных путем использования разработанных теоретических основ, системной методологии и программных средств обработки множества информационных продуктов.

Осуществлен анализ проблемы представления и обработки разнотипных источников данных. Обоснована актуальность решения этой проблемы на основе введения новой абстракции управления данными – пространства данных, что позволило выделить нерешенные ранее проблемы по обработке и консолидации данных из заранее неизвестных источников. Разработана алгебраическая система пространства данных, которая состоит из множества информационных продуктов, предикатов и операций на них. Это позволило разработать операции консолидации и поиска данных из разнотипных источников, структура данных которых заранее неизвестна. Разработанные методы и алгоритмы использованы при создании ряда информационно-аналитических систем.

Ключевые слова: пространство данных, алгебраическая система, интеграция, интеллектуальный агент, функция качества консолидированных данных, данные с неопределенностью.