УДК 004.652

О. М. Верес
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

ОНТОЛОГІЯ ОЧИЩЕННЯ ДАНИХ

Описано етапи процесу очищення даних у СППР. Запропоновано та описано концепти онтології очищення даних. Проведено аналіз методів і технологій очищення даних на кожному з етапів процесу з врахуванням його особливостей. Побудована онтологія очищення даних для методологічної систематизації методів у реалізації функціональних елементів моделі СППР.

Ключові слова: дані, метод, онтологія, сховище даних, прийняття рішення, система підтримки прийняття рішень.

This article describes the steps to clear data in the DSS. The ontology concepts of clear data were proposed and described. The analysis of methods and data cleansing technology were carried out at every stage of the process, taking into account its features. Built The ontology of data cleaning techniques for methodological systematization of functional elements in the implementation model of DSS was built.

Key words: data, method, ontology, Data Warehouse, decision making, Decision Support System.

Вступ. Загальна постановка проблеми

Дані, що використовуються для бізнес-аналізу, найчастіше поганої якості. У них міститься багато помилок: дублювання, протиріччя, пропуски, аномалії і безліч інших проблем. Вилучити їх повністю неможливо: дані потрібно очищати. Для поліпшення якості вихідної інформації використовують всі можливі і організаційні, і програмні способи.

Погана якість даних є однією з найбільших проблем під час побудови аналітичних рішень, тому на основі некоректної інформації робляться неправильні висновки. Навіть найдосконаліші методи аналізу не допомагають, необхідно використовувати спеціальні механізми очищення.

Очищення даних – одне з найактуальніших завдань аналізу. На його виконання витрачається багато часу під час створення рішень. Це необхідний етап робіт у будь-якому проекті. Очищення даних є найважливішим етапом аналітичного процесу і від того, наскільки ефективно воно зроблене, багато в чому залежить коректність результатів аналізу і точність побудованих аналітичних моделей.

Збір, обслуговування та аналіз великих обсягів даних – це гігантські завдання, які вимагають подолання серйозних технічних труднощів, величезних витрат та адекватних організаційних рішень. Системи підтримки прийняття рішень (СППР) – основа IT-інфраструктури різних компаній, оскільки дані системи дають можливість перетворювати бізнес-інформацію в зрозумілі та корисні висновки. Одним з основних структурних елементів архітектури сучасної СППР є сховище даних (СД) (*data warehouse*, DW) [1, 2].

Якість даних, які збираються та консолідуються для аналізу з різних джерел, є однією з найбільших проблем бізнес-аналітики.

Недостатня увага до неї здатна звести нанівець всі переваги найсучасніших і наймогутніших методів аналізу, всі зусилля аналітиків і експертів зі створення аналітичних рішень. А самі аналітичні рішення, отримані на основі неякісних даних, можуть опинитися скільки завгодно далекими від дійсності, спотворити реальну картину досліджуваних бізнес-процесів, показати помилкові закономірності, тенденції і зв'язки між об'єктами бізнесу.

Наслідком цього може стати вироблення неправильних управлінських рішень, які завдадуть бізнесу збитку. Саме тому моніторинг якості даних, а також їхнє перетворення з метою вилучення чинників, які є причиною зниження якості даних, проводяться на всіх етапах аналітичного процесу від витягання даних з джерел до їхнього опрацювання в аналітичній системі. Якщо важливість цього напряму усвідомлюється і аналітичним технологіям на підприємстві приділяється належна увага, то контроль і підтримка якості проводяться на етапі збору даних у самих джерелах: OLTP системах, облікових системах і корпоративних базах даних підприємства.

Аналіз останніх досліджень та публікацій

Наявна множина видів помилок у даних, що не залежать від предметної області. Виділяють шість типів таких помилок: суперечливість інформації; аномальні значення; пропуски даних; шум; невідповідність форматів даних; помилки введення даних або друкарські помилки; дублювання.

Поняття «якість даних» дуже широке, не має чітко обкреслених меж і строгого означення, навіть може трактуватися по-різному залежно від того, в якій сфері інформаційних технологій воно застосовується [3].

Термін «якість даних» з'явився задовго до IT-технологій. Під якістю даних розуміли кількість помилок при введенні та форматуванні даних. У контексті сучасних аналітичних технологій якість даних – сукупність їхніх властивостей і характеристик, що визначають ступінь природності для аналізу.

Для підвищення якості даних використовується комплекс методів і алгоритмів, що отримали назву «очищення даних» (*cleaning, refinement*). Щоб правильно підготувати дані до аналізу, необхідно мати стратегію їхнього очищення, яка розробляється на основі знання структури і особливостей джерел, з яких отримують дані, характеру самих даних, методики і мети їхнього аналізу.

Очищення даних (*data scrubbing*) – виправлення успадкованих даних підприємства шляхом виявлення неузгодженостей, дублювання і помилок введення [4–7]. Очищенню можна піддавати дані і з однієї бази даних (БД), і з декількох. Засоби очищення також об'єднують записи.

Очищення даних (*data cleaning, data cleansing* або *scrubbing*) займається виявленням і видаленням помилок і невідповідностей у даних з метою поліпшення якості даних.

Очищення даних – це процес аналізу якості даних у джерелі даних з виконуваним вручну затвердженням або відхиленням рекомендацій, що даються системою, і внесенням змін до даних.

Метод очищення даних має задовольняти низку критеріїв. По-перше, він повинен виявляти і видаляти всі основні помилки і невідповідності, і в окремих джерелах даних, і при інтеграції декількох джерел. Метод повинен підтримуватися певними інструментами, щоб скоротити обсяги ручної перевірки та програмування, і бути гнучким у роботі з додатковими джерелами. Очищення даних не проводиться незалежно від пов'язаних зі схемою перетворення даних, що виконуються на основі складних метаданих. Інфраструктура технологічного процесу має особливо підтримуватися для сховищ даних, забезпечуючи ефективне і надійне виконання всіх етапів перетворення даних для множини джерел і великих наборів даних.

Є безліч засобів, з різною функціональністю, призначених для підтримки подібних завдань, проте часто чималий обсяг роботи з очищення і перетворення доводиться виконувати вручну або програмами низького рівня, які є важкими для написання і використання.

Переважно проблеми виникають: при втраті значень (не введені значення); орфографічних помилках; групових значеннях (декілька значень в одному атрибуті); у разі невідповідності значень полів; при порушенні логічних зв'язків; при дублюванні чи суперечливості записів. Враховуючи, що очищення джерел даних є доволі дорогим процесом, запобігання введенню забруднених даних є важливим кроком у зменшенні проблем. Для цього потрібно відповідно спроектовані схеми бази даних і обмеження цілісності, а також застосування для введення даних [1, 8].

Характер впливу окремих аспектів якості даних на процес аналізу різний. Одні проблеми в даних можуть викликати тільки технічні складності, наприклад, неможливість інтегрування даних або завантаження їх у сховище. Інші не дають можливості виконувати коректний аналіз даних і перешкоджають отриманню достовірних результатів, наприклад, шуми, дублікати і протиріччя.

Технічні проблеми переважно виявляються і за можливості усуваються при консолідації даних. Аналітичні проблеми домінують на етапі безпосередньої опрацювання даних, коли дані вже прив'язані до певної методики і задачі аналізу.

Технічні проблеми впливають і на процес консолідації даних, і на результати аналізу. Аналітичні проблеми впливають тільки на аналіз даних. Якщо дані в тому чи іншому вигляді вдалося завантажити в аналітичний додаток, то, найшвидше, в них залишилися тільки проблеми аналітичного плану.

Основні проблеми в області очищення даних, що підлягають вирішенню під час очищення і перетворення даних, тісно зв'язані і тому повинні вирішуватися в комплексі [9, 10]. Перетворення даних потрібне для підтримки будь-яких змін у структурі, зображенні або змісті даних. Ці перетворення стають необхідні в різних ситуаціях, наприклад, при зміні структури даних, переході на нову інформаційну систему або тоді, коли треба інтегрувати множинні джерела даних. Проблеми рівня схеми відображаються і в елементах даних; їх вирішують за допомогою її поліпшення, трансляції і інтеграції схеми даних. З іншого боку, проблеми рівня елементу даних пов'язані з помилками і невідповідностями у джерелі поточних даних, непомітних на рівні схеми. Вони і є основною метою очищення. Проблеми в окремих джерелах з вірогідністю, що збільшується, характерні і для множини джерел.

Очищують дані і перед їхнім завантаженням у сховище, і в аналітичному додатку безпосередньо перед аналізом. При цьому основне очищення проводиться в аналітичному додатку, оскільки деякі проблеми (наприклад, дублікати і протиріччя) неможливо виявити до завершення консолідації даних. Крім того, вимоги до якості даних можуть бути різними для різних методів і алгоритмів аналізу. Тому більшість аналітичних додатків містить розвинений комплекс засобів очищення даних.

Сьогодні наявна величезна кількість методів очищення даних від помилок і неточностей. Виявити найефективніший складно, оскільки кожен метод абсолютно по-різному підходить до цієї проблеми.

Невирішені раніше частини загальної проблеми. Сьогодні у великих корпораціях отримують та опрацьовують величезну кількість даних, особливо персональних, що зібрані з усіх філій компанії. У кожній філії своя структура бази даних і після інтеграції в СД, як єдиному джерелі даних, виникає проблема видобування достовірних даних з причини розрізнених даних у різному поданні, які необхідно надалі використовувати для аналізу. Такі дані будуть низької якості, оскільки в них допускалися помилки, і їхня опрацювання втрачає сенс. Тому для отримання реальних висновків з наявних даних застосовують різні методи їхньої корекції, вилучення дублікатів та очищення. Отже, завдання очищення даних у корпоративних інформаційних системах підтримки прийняття рішень є актуальним.

Формулювання мети

Основним завданням статті є розроблення онтології очищення даних для спрощення побудови моделі СППР та її функціональних компонент. Необхідно подати означення етапів процесу очищення даних, а також описати основні методи, алгоритми і підходи до реалізації функції кожного з них.

Побудова онтології очищення даних

Онтологія – це точна специфікація деякої предметної області [11–15]. Вона забезпечує словник для подання та обміну знаннями про цю предметну область і множину зв'язків, встановлених між термінами в даному словнику. У простому випадку побудова онтології зводиться до:

- виділення концептів – базових понять даної предметної області;
- побудови зв'язків між концептами – визначенню співвідношень і взаємодій базових понять.

Оскільки знання мають особистісний характер, то одну і ту саму предметну область можна описати різними онтологіями. Особливо це стосується предметних областей, які погано формалізуються, або за наявності великої кількості спірних запитань.

Даючи можливість будувати моделі предметних областей, онтології автоматизують обробку семантики даних з метою її ефективного використання (подання, перетворення, пошуку). За

формою онтологія є зручною, як доволі складно організована знакова структура елементарних знань про предметну область, з іншого, як вихідний матеріал для отримання нових емпіричних знань у процесах діяльності.

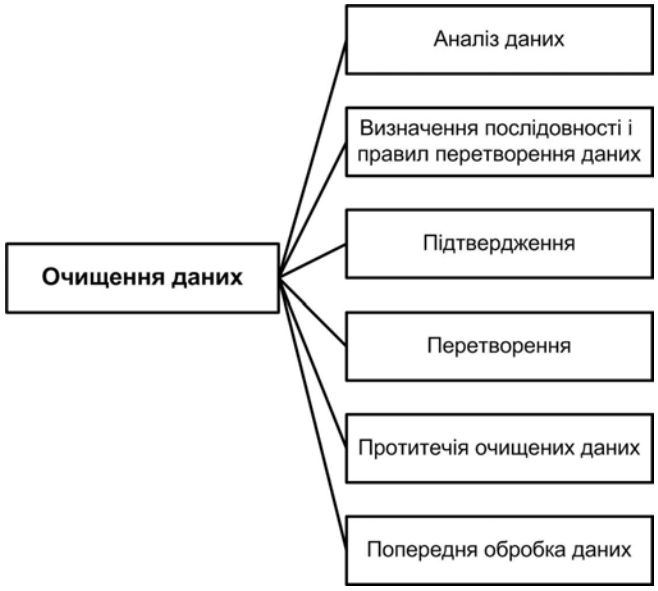
Потреба в розробленні онтологій виникає для:

- спільного використання людьми або програмними агентами загального розуміння структури даних;
- повторного використання знань про предметну область;
- перетворення припущень в явні зв'язки або залежності;
- виділення знань про предметну область від оперативних знань;
- аналізу знань про предметну область.

Пропонована онтологія побудована відповідно до підходу METHONTOLOGY [9], який відображає процес ітеративного проектування. У життєвому циклі створення онтології виділяють такі процедури: управління проектом (планування, контроль і гарантії якості), власне розроблення (специфікація, концептуалізація, формалізація і реалізація) та підтримка (супровід). Побудова онтологій – послідовність підпроцесів створення проміжних подань.

Причому підпроцеси виконують не послідовно, а визначаються за повнотою та точності накопичених знань. Спочатку будують глосарій термінів (*Glossary Of Terms*), потім дерева класифікації концептів (*Concept Classification Trees*) і діаграми бінарних відношень (*Binary Relation Diagrams*), тільки потім – решта проміжних подань.

За методологією METHONTOLOGY глосарій термінів містить всі терміни (концепти і їхні екземпляри, атрибути, дії), важливі для очищення даних [6], і їхні природно-мовні описи.



Ієрархія процесу очищення даних

Загалом очищення даних у СППР містить декілька етапів (див. рисунок).

Аналіз даних – виявлення видів помилок і невідповідностей, що підлягають видаленню. Поряд з ручною перевіркою даних або їхніх шаблонів, треба використовувати аналітичні програми для отримання метаданих про властивості даних і виявлення проблем якості даних.

Визначення послідовності і правил перетворення даних. Залежно від кількості джерел даних, ступеня їхньої неоднорідності та забрудненості даних, вони можуть вимагати достатньо широкого перетворення та очищення. Іноді для відображення джерел для загальної моделі даних використовується трансляція схеми; для сховищ даних, зазвичай, використовується реляційне зображення. Перші кроки з очищення даних можуть скоригувати проблеми окремих джерел даних і підготувати дані для інтеграції. Подальші кроки спрямовані на інтеграцію схеми/даних та усунення проблем множинності елементів, наприклад, дублікатів. Для сховищ даних у процесі ETL (*Extract,*

Transform, Load – «видобування, перетворення, завантаження» [4]) визначаються методи контролю і потік даних, що підлягає перетворенню та очищенню. Перетворення даних, що пов'язане зі схемою, визначається за допомогою мови декларативного запиту (маппірування, Mapping Composition), забезпечуючи, у такий спосіб, автоматичну генерацію коду перетворення. У процесі перетворення має бути можливість запуску написаного користувачем коду очищення і спеціальних засобів. Етапи перетворення можуть вимагати зворотного зв'язку з користувачем по тих елементах даних, для яких відсутня вбудована логіка очищення.

Підтвердження – правильність і ефективність процесу і визначення перетворення. Це здійснюється шляхом тестування та оцінювання. Під час аналізу, проектування та підтвердження може знадобитися безліч ітерацій, наприклад, з огляду на те, що деякі помилки стають помітні тільки після певних перетворень.

Перетворення – виконання перетворень або в процесі ETL для завантаження і оновлення сховища даних, або при відповіді на запити з множини джерел. Процес перетворення вимагає великих обсягів метаданих – наприклад, схем, характеристик даних рівня схеми, означень технологічного процесу тощо. Для узгодженості, гнучкості та спрощення використання в інших випадках, ці метадані повинні зберігати в репозиторії на основі СУБД. Для підтримки якості даних ґрунтовна інформація про процес перетворення має записуватися і в репозиторій, і в трансформовані елементи даних, особливо інформація про повноту та актуальність початкових даних і походження інформації про першоджерело трансформованих об'єктів та проведені з ними зміни.

Протитечія очищених даних – заміна забруднених даних у першоджерелах на очищені. Після того, як помилки (окремого джерела) видалені, очищені дані мають замістити забруднені дані в початкових джерелах, щоб покращені дані потрапили і в успадковані застосування і надалі при витяганні не вимагали додаткового очищення. Для сховищ даних очищені дані містяться в області зберігання даних.

Попереднє опрацювання даних – комплекс методів і алгоритмів, які застосовуються в аналітичному додатку з метою підготовки даних до виконання конкретного завдання і приведення їх у відповідність до вимог, що обумовлені специфікою завдання і способами його виконання.

Очищення даних в аналітичному додатку – лише один з аспектів попередньої опрацювання даних, хоча ці два процеси часто ототожнюються. Однак очищення не є синонімом попереднього опрацювання даних. Більше того, якщо в даних, завантажених в аналітичний додаток, відсутні проблеми, які потребують очищення, або їхній вплив на якість рішення оцінюється як мінімальний, то очищення даних у процесі їхнього попереднього опрацювання може взагалі не проводитися. Попереднє опрацювання здійснюється в будь-якому випадку. Необхідність застосування цього етапу полягає в тому, що проблеми виникають і можуть бути виявлені тільки на певному етапі збору та консолідації даних.

У OLTP системах, в процесі ETL і в СД проводиться очищення даних від помилок, які зазвичай мають технічний характер. Порушення структури, цілісності, повноти і некоректні формати даних треба виправити до завантаження в СД, оскільки дані з такими проблемами можуть взагалі не завантажитися в сховище. Фіктивні значення також простіше розпізнати і виправити в OLTP системі: як тільки дані вийдуть за її межі, то виявити, що значення помилкове, буде набагато важче, особливо якщо воно має правдоподібний вигляд і коректний формат. Що стосується пропусків, то з'ясувати їхню причину і відновити пропущені значення також легше в тій системі, де вони були створені.

Проблеми, пов'язані з якістю даних, зручніше розпізнавати і усувати в першоджерелі. Тут простіше дізнатися причини і закономірності появи проблем, оскільки для консультацій є доступним персонал, що працював з даними, а також первинні бухгалтерські та інші документи, з яких можна узяти правильні та відсутні значення.

Методи аналізу даних. Основні методи аналізу даних – *математична статистика*, *еволюційне моделювання* і *машинне навчання*.

Прості методи (регулярні вирази, строгі формальні правила тощо) дуже примітивні і можуть виконати завдання очищення даних тільки частково, тому застосовують **математичну статистику**.

Розраховуються необхідні показники за всіма наявними даними, тобто охоплюють весь діапазон значень і прийнятих ознак. На основі отриманих результатів одні методи можуть виділити підозрілу інформацію, яка дуже відрізняється від інших, а інші – обчислити величини, які ймовірно найбільше схожі на справжні. У такий спосіб, аналізуючи відомості за допомогою статистичних характеристик, оцінюють загальну картину даних і вже на її ґрунті визначають можливі помилки з подальшим їхнім виправленням на підібрані схожі значення. Для цього використовують методи: оцінки параметрів розподілу; перевірки статистичних гіпотез; дисперсійний аналіз; кореляційний аналіз; регресивний аналіз; аналіз часових рядів; багатовимірний аналіз. Сьогодні як засоби **еволюційного моделювання** використовують генетичні алгоритми і штучні нейронні мережі.

Метою методів **машинного навчання** є отримання простих класифікаційних. Дерева рішень (*decision trees*) є одним з найпотужніших засобів розв'язання задачі віднесення будь-якого об'єкта (рядки набору даних) до одного з заздалегідь відомих класів. Дерево рішень – це класифікатор, отриманий з навчальної множини, що містить об'єкти та їхні характеристики, на основі навчання. Дерево складається з вузлів і листків, що вказують на клас. За допомогою *логістичної регресії* можна оцінювати ймовірність того, що подія настане для конкретного випробуваного. Оцінити якість логістичної регресії як класифікатора можна на основі таблиці спряженості.

Для отримання реальних (модернізованих) метаданих з характеристиками даних або незвичайними моделями значень важливо аналізувати реальні приклади. Такі метадані полегшують пошук проблем якості даних. Вони сприяють ідентифікації відповідностей атрибутів між початковими схемами (зіставлення схем) залежно від того, які автоматичні перетворення даних проводяться. Наявні два пов'язані між собою методи аналізу: *профілізація* даних і *data mining*. Профайлінг даних орієнтований на зразковий аналіз окремих атрибутів. При цьому відбувається отримання такої інформації, як тип, довжина, спектр значень, дискретні значення даних і їх частота, зміна, унікальність, наявність невизначених значень, типових моделей рядків (наприклад, для номерів телефонів) тощо, що дає змогу забезпечити точне представлення різних аспектів якості атрибуту. За допомогою метаданих виявляють такі проблеми в якості даних: неприпустимі значення, орфографічні помилки, втрачені значення, різне представлення значень, дублікати.

Data mining допомагає знайти специфічні моделі даних у великих наборах даних, наприклад, залежність між декількома атрибутами. Для цього використовують *описові моделі data mining*, зокрема групування, узагальнення, пошук асоціацій і послідовностей. При цьому можуть бути отримані обмеження цілісності в атрибутах, наприклад, функціональні залежності, або характерні для конкретних застосувань бізнес-правила, які можна використовувати для заповнення втрачених і виправлення неприпустимих значень, а також для виявлення дублікатів записів у джерелах даних.

Для очищення даних також можна використовувати *бази знань*, які побудовано на ґрунті набору високоякісних даних. Очищення даних виконується в чотири етапи: зіставлення, коли ідентифікується джерело даних, що підлягають очищенню, і зіставляється з необхідними доменами в базі знань; автоматизованого очищення, коли застосовують базу знань до даних, що підлягають очищенню, а також пропонуються і вносяться зміни у вихідні дані; інтерактивного очищення, коли диспетчери даних можуть аналізувати зміни даних, а також приймати/відхиляти ці зміни даних; нарешті, експорту, на якому очищені дані експортуються.

Методи визначення перетворень даних. Процес перетворення даних, зазвичай, складається з безлічі кроків, кожен з яких може виконувати і перетворення рівня схеми, і перетворення рівня елементу даних (маппірування). Щоб забезпечити системі перетворення і очищення даних можливість генерації програмного коду для перетворення і тим самим зменшити обсяг програмування, треба описати необхідні перетворення відповідною мовою, наприклад, підтримуваній графічним інтерфейсом користувача. Різні засоби ETL містять таку можливість, підтримуючи власні мови правил. Загальний і гнучкіший підхід полягає в застосуванні стандартної мови запитів SQL для виконання перетворень даних і використання можливостей розширень мов застосувань, особливо функцій, які визначаються користувачем (UDFs). UDFs можуть бути реалізовані в SQL або мовою програмування загального призначення, що містить вбудовані оператори SQL. Вони дають змогу реалізовувати широкий спектр перетворень даних і підтримують

просте використання різних перетворень і опрацювання запитів, а їхнє виконання СУБД може понизити вартість доступу даним.

Складніші реструктуризації схеми (наприклад, згортання і розгортання атрибутів) не підтримуються взагалі. Для повної підтримки пов'язаних зі схемою перетворень, необхідні розширення мови – такі, як директива SCHEMASQL. Очищення даних на рівні елементу даних може також виграти від використання спеціалізованих розширень мови – таких, як оператор Match, що підтримує "приблизні об'єднання". Системна підтримка для таких могутніх операторів може істотно спростити програмування перетворень даних.

Методи вирішення конфліктів. Необхідно визначити і дотримуватися послідовності кроків перетворення для опрацювання різних проблем з якістю даних рівня схеми і елементу даних, відображених в зовнішніх джерелах даних. Деякі типи перетворень треба виконувати на окремих джерелах даних, враховуючи проблеми окремого джерела і готуючи його до інтеграції з іншими джерелами. Крім можливої трансляції схеми, такі підготовчі етапи, зазвичай, містять [5].

- **Витягання значень з атрибутів вільного формату (розщеплювання атрибутів).** Атрибути вільного формату часто містять безліч окремих значень, що підлягають витяганню для підвищення точності подання і підтримки подальших етапів очищення, таких, як зіставлення елементів даних і вилучення дублікатів. Парсинг – це граматичний або лексичний аналіз тексту. При виконанні парсингу ведеться поділ полів на атомарні значення.
- **Перевірка допустимості і виправлення.** На цьому етапі кожен елемент даних джерела даних досліджується на наявність помилок, а виявлені помилки за можливості автоматично виправляються. Перевірка орфографії на основі перегляду словника потрібна для ідентифікації і виправлення помилок у написанні слів. Атрибутивні залежності (дата народження – вік, загальна вартість – ціна за шт., місто – регіональний телефонний код тощо) можуть використовуватися для виявлення проблем і заміни втрачених або виправлення неправильних значень.
- **Стандартизація.** Для співвідношення та інтеграції елементів даних, значення атрибутів треба перетворити в узгоджений і уніфікований формат. Текстові дані можуть бути стиснені та уніфіковані за допомогою виявлення кореня, видалення префіксів, суфіксів і ввідних слів. Абревіатури і зашифровані схеми підлягають узгодженому розшифруванню за допомогою спеціального словника синонімів або застосування зазначених правил конверсії.

Методи попереднього опрацювання даних. Типовий набір інструментів **попереднього опрацювання даних** та підготовки даних до аналізу, що поставляється з більшістю аналітичних платформ, містить такі засоби [3].

- Очищення від шумів і згладжування рядів даних.
- Відновлення пропущених значень.
- Редагування аномальних значень.
- Опрацювання дублікатів і протиріч.
- Зниження розмірності вхідних даних.
- Усунення незначущих факторів.

Згладжування даних можливо виконати за допомогою вейвлет-перетворення і видалення шуму. У процесі парціальної опрацювання відновлюються пропущені дані, редагуються аномальні значення, проводиться спектральне опрацювання. Використовуються алгоритми, в яких кожне поле набору, що аналізується, обробляється незалежно від інших полів, тобто дані обробляються по частинах. З цієї причини таке попереднє опрацювання отримало назву парціального.

Метою факторного аналізу є зменшення розмірності простору факторів. Зменшити розмірності треба у випадках, коли вхідні фактори корельовано один з одним, тобто вони взаємозалежні. Факторний аналіз використовується у випадках, коли у дуже великому вихідному наборі даних є багато полів, деякі з яких є взаємозалежні.

Кореляційний аналіз застосовується для оцінки залежності вихідних полів даних від вхідних факторів та усунення незначущих факторів. Якщо кореляція (ступінь взаємозалежності) між вхідним і вихідним факторами менша від порогу значимості, то відповідний фактор відкидається як незначущий.

У процесі аналізу іноді виникає проблема виявлення дублікатів і протиріч у даних. Дедублікація ґрунтується на пошуку однакових і схожих об'єктів за певними стратегіями з метою усунення повторів.

За допомогою операції фільтрації можна залишити в таблиці бази даних тільки ті записи, які задовольняють задані умови, а решту видалити.

Головна відмінність очищення даних при попередній обробці від очищення в OLTP системах і в процесі ETL полягає в тому, що вона проводиться з урахуванням планованої методики аналізу, його цілей, необхідної достовірності результатів і може коригуватися залежно від отриманих попередніх результатів.

Розроблення технологій очищення даних в OLTP системах і в процесі ETL є результатом спільної діяльності програміста і аналітика. При безпосередній підготовці до аналізу, очищення даних є завданням користувача аналітичної програми, але не має вимагати втручання технічних працівників. Цілі і методи попереднього опрацювання даних повністю визначаються користувачем і обмежуються лише комплексом засобів, які надані йому системою.

Глосарій термінів онтології очищення даних містить означені вище терміни, які можна семантично розбити на три групи: структура завдання (етапи очищення, зв'язки), дані, що наповнюють завдання (методи, що застосовують на кожному з етапів), і результати обчислень (очищені дані).

Висновки та перспективи подальших наукових розвідок

Якість персональних даних є проблемою, що значно знижує результативність аналізу. Приймати обґрунтовані рішення можна, тільки ґрунтуючись на повних і достовірних відомостях.

Застосування спеціалізованих інструментів і методів дає змогу перетворити зібрані в облікових системах дані в цінну інформацію, що використовується в процесі прийняття рішень.

Однією з переваг використання онтологій як інструменту пізнання є системний підхід до вивчення предметної області. При цьому досягаються:

- систематичність – онтологія надає цілісний погляд на предметну область;
- однотипність (стандартність) – матеріал, поданий в єдиній формі, набагато краще сприймається та відтворюється;
- науковість – побудова онтології дає змогу відновити необхідні, але відсутні логічні зв'язки у всій їхній повноті.

Для досягнення поставленої мети здійснено аналіз етапів очищення даних у корпоративних СППР. Подано означення базових понять, а також описано основні методи реалізації завдань очищення даних на кожному з етапів. Також розглянуто особливості очищення даних на рівні аналітичних додатків як структурних елементів корпоративної СППР. Можливість безпосереднього управління підготовкою даних важлива для аналітика ще й тому, що в деяких випадках характер його втручання може йти врозріз з формальними процедурами.

Описані засоби є достатньо сучасними, але вони не вирішують всіх проблем і все ще вимагають додаткового опрацювання вручну або додаткового програмування. Очищення даних потрібне не тільки для сховищ даних, але і для опрацювання запитів за неоднорідними джерелами даних, наприклад, в інформаційних WEB-системах. Це середовище має істотніше обмеження для очищення даних, які потрібно враховувати при виборі відповідних методів. Також актуальним для вивчення сьогодні є моделювання очищення даних в Big Data. Подальші дослідження також будуть присвячені дослідженню методів очищення частково структурованих даних, оскільки структурні обмеження постійно знижуються, а обсяги XML-даних стрімко зростають.

1. Верес О.М. Компоненти концептуальної моделі системи підтримки прийняття рішень / О. М. Верес // Комп'ютерні науки та інформаційні технології. Вісник Нац. ун-ту "Львівська політехніка". – 2010. – № 686. – С. 103–112. 2. Спирли Э. Корпоративные хранилища данных. Планирование, разработка, развитие / Эрик Спирли. – М. : Вильямс, 2001. – 400 с. 3. Паклин Н. Б. Бизнес аналитика: от данных к знаниям : учебн. пособие / Н. Б. Паклин, В. И. Орешков. – 2-е изд., испр. – СПб.: Питер, 2013. – 704 с. 4. Арустамов А. Предобработка и очистка данных перед

загрузкой в хранилище [Електронний ресурс] / Алексей Арустамов. – Режим доступу : http://www.basegroup.ru/library/dw_olap/dataclearing/ 5. Фоурино Р. Электронное качество данных: скрытая перспектива очистки данных [Електронний ресурс] / Роналд Фоурино. — Режим доступа: http://www.olap.ru/basic/el_data_quality.asp/ 6. Рам Э. Очистка данных: проблемы и актуальные подходы [Електронний ресурс] / Эрхард Рам, Хонг Хай До. — Режим доступа: http://www.olap.ru/basic/data_clean.asp/ 7. Некипелов Н. Онтология анализа данных [Електронний ресурс] / Николай Некипелов, Акобир Шахиди. – Режим доступу : http://www.basegroup.ru/ library/methodology/ontology/ 8. Ralph Kimball, The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses. John Willey & Sons, New York, 1996. 9. Гаврилова Т. А. Базы знаний интеллектуальных систем / Т. А. Гаврилова, В. Ф. Хорошевский. – СПб. : Издательство «Питер», 2000. – 384 с. 10. White C. Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise // DMReview. 2005. – № 11. – P. 25–53 11. Гаврилова Т. А. Онтология для изучения инженерии знаний // Труды Международной научно-практической конференции KDS-2001. – 2001. 12. Гаврилова Т. А. Онтологический подход к управлению знаниями при разработке корпоративных информационных систем // «Новости искусственного интеллекта». – 2003. – № 2. – С. 24–30. 13. Буров Є. В. Формальна модель подання знань у системі онтологічного моделювання задач / Є. В. Буров // Вісн. Нац. ун-ту «Львівська політехніка». Серія : Інформаційні системи та мережі. – № 770. – 2013. – С. 21–30. 14. Литвин В.В. Метод використання онтологій у петлі OODA / В.В. Литвин // Вісн. Нац. ун-ту «Львівська політехніка». Серія : Інформаційні системи та мережі. – № 783. – 2014. – С. 137–145. 15. Литвин В. В. Проблема автоматизованої розбудови базової онтології / В. В. Литвин, Т. М.Черна // Вісн. Нац. ун-ту «Львівська політехніка». Серія : Інформаційні системи та мережі. – № 805. – 2014. – С. 306–315.