УДК 681.3 http://doi.org/10.15407/pp2023.02.067

Ю.В. Рогушина

НЕЧІТКІ ДАНІ У СЕМАНТИЧНИХ WIKI-РЕСУРСАХ: МОДЕЛІ, ДЖЕРЕЛА ТА МЕТОДИ ОБРОБКИ

Запропонована у роботі онтологічна модель призначена для класифікації різних типів некласичних та семантично некоректних даних, щоб уможливити ефективніше знаходження методів виявлення таких даних та засобів їх обробки. Така обробка, що може розглядатися як одна зі складових Smart data, має зробити дані придатними для автоматичного аналізу та повторного використання в інших інформаційних системах. Онтологічний підхід забезпечує інтеграцію запропонованої моделі з іншими зовнішніми онтологіями, що описують різноманітні методи та програмні засоби аналізу даних. У роботі використано досвід розробки бази знань портальної версії Великої української енциклопедії e-ВУЕ, яка має великий обсяг, складну структуру та містить велику кількість різноманітних гетерогенних інформаційних об'єктів. Участь великої кількості спеціалістів різних наукових напрямків у створенні цього ресурсу викликає розбіжності у розумінні правил подання та структурування даних, і тому виникає необхідність додаткової веріфікації контенту. Для цього потрібні формалізовані та масштабовані рішення для знаходження та опрацювання різноманітних типів нечіткості, неповноти та семантичної некоректності контенту. Запропонований підхід може бути корисним для створення інших великомасштабних ресурсів як на основі семантичних Wiki, так і інших технологічних платформ колаборативної обробки розподілених даних та знань.

Ключові слова: онтологія, семантично некоректні дані, брудні дані, Wiki-ресурс.

Вступ

Щоб дані стали корисними, їх потрібно інтерпретувати та перетворювати на знання. Вже зараз значна частка даних є неструктурованою, містить помилки та суперечності, потребує додаткових уточнень для їх коректного розуміння і таким чином вимагає додаткової обробки перед тим, як ці дані стають придатними для подальшого аналізу та практичного використання. Тому виникає потреба як у розробці методів такої попередньої обробки даних, так і в класифікації тих проблем, які ці методи мають розв'язати.

Попередня обробка дозволяє перетворювати "сирі" дані на "розумні", які більш придатні для автоматизованого здобуття корисних відомостей. Залежно від того, які саме невизначеності потрібно вирішувати, можуть застосовуватися різні процедури попереднього обробки. Тому доцільно класифікувати існуючі види невизначеності даних та розробити відповідну таксономічну модель.

Крім того, на вибір методів обробки впливає як тип самих "сирих" даних, так і предметна область, до якої вони належать. Для деяких типів невизначеності допустимо використання автоматичних методів перетворення, деякі потребують безпосередньої участі людини. Існує велика кількість ситуацій, коли попередня обробка даних може автоматизовано використовувати зовнішні джерела знань. У таких випадках участь людини-експерта може обмежуватися вибором таких джерел або формулюванням умов для їх пошуку. Знання проблеми, для розв'язання якої мають використовуватися дані, а також структури інформаційних об'єктів дозволяють забезпечити перетворення даних на семантичному рівні. Наприклад, оцінюючи семантичну близькість між різними концептами предметної області.

Таке перетворення даних у широкому сенсі відповідає напрямку досліджень, що отримав назву *Smart data.* Але методи та задачі цих перетворень істотно залежать як від предметної області (ПрО), для якої здійснюється аналіз, так і від особливостей власне "сирих" даних. Розпливчастий характер і нечіткість притаманні багатьом типам інформації, що переробляється людиною та інформаційними системами. На відміну від класичних даних (КД), яким властиві повнота, точність, узгодженість

та визначеність, сирі дані не відповідають цим вимогам і можуть бути нечіткими та неповними. Крім того, певна підмножина даних генерується внаслідок м'яких обчислень, що моделюють неоднозначність та непевність міркувань людини на основі методів нечіткої логіки [1]. Такі обчислення дозволяють аналізувати дані, що містять різні види невизначеності, неповноти та помилок.

Об'єднуючи дані з розрізнених джерел, можна отримувати нову інформацію за допомогою їх аналізу. Але якість нової інформації залежить не лише від алгоритмів аналізу, а й від якості даних. Це можуть бути як нечіткі твердження, так і нечіткі продукційні правила. Джерела нечіткості інформації знаходяться всередині самої взаємодії людини з навколишнім світом, тобто обумовлені природою відображення об'єктивної реальності. КД не дозволяють відображати всю існуючу інформацію про реальний світ, або маніпулювати знаннями, які можуть бути неточними, невизначеними, розпливчастими тощо. Результати аналізу таких даних можуть бути ненадійними та некоректними. Тому виникає потреба у ширшій моделі даних, яка б дозволила використовувати брудні дані в інформаційних системах. Дані вважаються *брудними*, якщо користувач або програма, що працюють коректно, не в змозі отримати результат їхньої обробки, або отримує неправильний результат через певні проблеми з даними. Водночас потрібно аналізувати два різні аспекти – з якої причини дані стали брудними та що можна зробити, аби вони стали придатними для аналізу. Наприклад, джерелами брудних даних можуть бути помилки введення, або оновлення даних, помилки передачі даних, або некоректно обрана форма подання даних.

У цій статті весь набір неточних, розпливчастих, невизначених, непослідовних, неповних тощо даних, які не можуть бути віднесені до КД, будемо називати *некласичними* даними (НКД). Цей термін близький до брудних даних, але охоплює більший спектр причин, через які аналіз даних не дає того результату, на який він спрямований. Природа та походження таких даних різняться, тому існує потреба в різних технологіях для роботи з НКД.

У широкому сенсі причинами виникнення брудних даних, отриманих із різноманітних джерел даних, є відсутність певної інформації, її неправильність (невідповідність реальному світу або іншим даним) і нестандартні подання самих даних. У деяких випадках вони потребують автоматизованого очищення, в інших до них доцільно застосовувати різні моделі м'яких обчислень. Іноді вони потребують явної перевірки та виправлення, але у всіх цих ситуаціях основою для отримання корисних результатів є визначення типу їх відмінності від класичних даних.

Класифікація НКД допомагає обирати методи роботи з брудними даними та метрики для вимірювання якості даних.

Типи некласичних даних

Існує багато типів брудних даних, які за різними властивостями відрізняються від КД та непридатні для обробки традиційними методами. Такими властивостями є неточність, неповнота, нечіткість та неузгодженість даних, а також неоднозначність їх інтерпретації та вибору моделі подання.

Розглянемо інформаційний об'єкт (ІО) I: де О – об'єкт, що описується даними; А – атрибут, значенням якого є дані; а – значення атрибута А; К – впевненість у виборі атрибуту; k – впевненість у значенні атрибута. Невизначеність даних щодо ІО є характеристикою змісту інформації – а та А , а їх ненадійність – характеристикою істинності інформації к і К, щодо їх відповідності дійсності. Інформація є ненадійною, якщо в інформаційній одиниці I впевненості к і К не можна представити двома значеннями: 1 (істинно) і 0 (хибно). Одна з форм ненадійності – неточність. Вона належить до якості значень фактів. Для обробки таких даних використовують коефіцієнти впевненості, що кількісно оцінюють ступінь впевненості в тому, що атрибут має саме це значення, і це значення належить саме до цього атрибута. Оцінки правдоподібності к і К істотно залежать від суб'єктивно заданих для кожного правила умовних ймовірностей.

Неповні дані (Incomplete data) – це

дані, в яких відсутнє значення певного атрибута. Така неповнота може бути викликана некоректним читанням або відсутністю доступу до інформації [2]. Для обробки неповних даних важливим аспектом аналізу є розуміння того, чи існує взагалі значення такого атрибута для певного інформаційного об'єкта (навіть невідоме на поточний момент), чи воно взагалі не може бути отримане на поточний момент(наприклад, дата смерті для ще живої людини). Одним із поширених способів формалізації й обробки неповних даних, який може бути застосований до даних у відкритому інформаційному середовищі, є запропонований Коддом метод «Null Values» (A-marks) [3], відповідно до якого дані є неповними, якщо значення певної властивості для конкретного об'єкта на поточний момент невідомо, хоча сама ця властивість притаманна об'єкту і може бути довизначена пізніше. Різноманітні логічні системи використовують різні позначення, щоб ідентифікувати тип неповноти даних. Таке невідоме значення позначають спеціальною константою, і будь-яке входження такого значення може бути замінене на конкретне значення з множини припустимих. Для роботи з невідомими значеннями потрібні багатозначні логіки з епістемічними значеннями істинності, такі як тризначна логіки Лукашевича, n-значна логіки Поста. Вони дозволили перейти від двох оцінок істинності – "істинно" або "хибно" – до довільної кількості тверджень (приміром, "істинно", "невідомо", "недоступно" або "хибно") з відповідними таблицями істинності для всіх логічних операцій.

Неузгоджені дані (Inconsistent data) – це дані, інтерпретація яких викликає семантичний конфлікт: їх одночасна істинність не є припустимою. Концепція неузгодженості радше стосується зберігання даних у різних моделях, аніж безпосередньо даних та інтеграції (поєднання) інформації з різних джерел. Наприклад, в одному джерелі даних рік народження особи X – 1985, а в іншому – 1988. Однією з причин неузгодженості може бути використання різних одиниць виміру (наприклад, відстань між A та B наводиться в кілометрах або у милях) або різним порядком введення інформації (наприклад, формат дати "11.05" та "05.11"). В таких випадках перетворення та узгодження даних може бути автоматизоване після аналізу семантики джерела. Значно складніше обробляти дані, в яких використовуються ті самі (або схожі) назви параметрів, але вони мають різний зміст. Наприклад, в двох джерелах вказана кількість публікацій для особи X, але в першому джерелі враховуються всі публікації, а в другому – тільки публікації англійською. Ще одна причина неузгодженості – помилкове введення значення: наприклад, дата народження "33.41.77" не може бути інтерпретована в будь-яких форматах подання дати). Крім того, розбіжності в значеннях даних можуть бути зумовлені часом їх введення. Приміром, у різних джерелах кількість публікацій для особи X може дорівнювати 55 та 78, але в першому випадку відомості введені за 2015 рік, а в другому – за 2020. У таких випадках інтеграція даних повинна базуватися на виборі найновіших даних. Але в цьому випадку теж потрібно враховувати семантику даних – зокрема, значення певних даних можуть тільки зменшуватися, а інших – тільки збільшуватися.

Непевні дані (Uncertain data) – це дані, для яких неможливо визначити точно їх істинність (через недостатню інформованість, або через відсутність точного значення). Скажімо, експерт дає суб'єктивну оцінку твердження, в якій він не впевнений, але намагається оцінити ймовірність того, що така інформація буде істинною або хибною на деякому інтервалі значень (зазвичай – [0, 1] і [0, 100], де перше та останнє значення ідентифікують правдиву та неправдиву інформацію відповідно) [4]. Ймовірність істинності може залежати від кількості неузгоджених записів у базі даних, від рейтингу експертів, від статистичних прогнозів, від індивідуальної точності інструментів вимірювання, від кількості оброблених даних тощо. Крім того, непевні дані можуть бути результатом обробки інших непевних даних.

Неоднозначні дані (Ambiguous data) – дані, які припускають кілька різних варіантів інтерпретації. Неоднозначність

даних може бути викликана: 1. використанням абревіатур та скорочень, які можуть розшифровуватися різними способами; 2. неповним контекстом (наприклад, відсутністю явного визначення одиниць виміру; 3. різним порядком елементів даних.

Нечіткі або розпливчасті дані (Fuzzy or vague data) – дані, для яких неможливо чітко й точно визначити значення, подані за допомогою лінгвістичних змінних, що досить суб'єктивно описують нечіткі множини об'єктів [5]. Для обробки таких даних, можна застосувати спеціальні математичні механізми (наприклад, нечітку логіку).

Невизначені дані (Imprecise data) – це дані, в яких замість одного значення міститься певний набір або інтервал можливих значень. Вони не є неправдивими або помилковими і не порушують цілісність інформаційної системи, якщо їх властивості викликані існуванням значення, яке неможливо виміряти з достатньою точністю.

У класифікації НКД вадливо визначити критерії, за якими здійснюється поділ даних. Тому важливо не тільки визначити основні типи даних та їхні підтипи, а й розробити таксономію таких даних. Зазвичай таксономія брудних даних базується на ієрархічній декомпозиції їх основних проявів – відсутності даних, їх неправильності (у різних значеннях) та непридатності для подальшого аналізу та використання. Така таксономія включає лише атомарні типи брудних даних і не розглядає їхні різноманітні комбінації. Верхні рівні запропонованої у даній роботі таксономії даних не можуть містити інших підкласів, тому що вони враховують усі можливі альтернативи (але не їхні комбінації). Якщо таксономія використовується для більш конкретної сфери, або стосується певної підмножини даних, то деякі її підкласи можуть бути видалені, а інші – розширені додатковими підкласами нижчого рівня.

Такий підхід до класифікації та аналізу НКД пропонується в [6], де розглядаються брудні дані (dirty data), що проявляються у різний спосіб, і внаслідок цього такі НКД поділяються на *неправильні дані, неповні непридатн*і для аналізу. У цій роботі розроблено комплексну таксономію брудних даних, яка дозволяє зрозуміти, звідки вони виникають, як виявляються і як можуть бути очищені для забезпечення якіснішого аналізу даних. Ця класифікація дуже детальна та багаторівнева, але, на жаль, занадто складна для сприйняття та практичного використання. Важливо, що автори дослідження звертають увагу на критерії класифікації на кожному рівні та переконуються у відсутності інших гілок таксономії.

Ця таксономія обмежується деякими підтипами даних – розглядаються тільки числові та текстові дати, не аналізуються мультимедійні дані та метадані. Але у більш узагальнених випадках аналізу НКД потрібно враховувати й такі дані, а також дані у більш специфічних форматах (як-от, потокові дані від різноманітного обладнання) та частково структуровану інформацію (зокрема, метадані без повної стандартизації). Наприклад, деякі аспекти некоректності даних, пов'язаних із обробкою метаданих та інтеграцією компонент програмної інженерії, які у свою чергу пов'язані із застосуванням методів Data Mining до мультимедійних даних у реальному часі, розглянуто в [7]. Важливо підкреслити, що таксономія НКД значною мірою залежить від тієї інформаційної технології (ІТ), що використовується для створення, збереження та обробки даних. Саме можливості ІТ визначають, які дані можна ввести та зберегти так, щоб надалі вони розглядалися як брудні, на яких етапах знаходження та виправлення брудних даних виконується автоматично та які зовнішні засоби аналізу даних можуть бути інтегровані до базової технології.

Постановка задачі

Перетворення брудних даних на Smart data, придатних для аналізу та здобуття знань, є важливим етапом розробки інформаційних ресурсів великого обсягу та складної структури. Це дозволяє ефективно застосовувати інформацію з таких ресурсів та підвищує якість її семантичної обробки. Класифікація НКЗ забезпечує основу для розуміння впливу брудних даних на аналіз даних, а також допомагають

обрати методи роботи з брудними даними та показники для вимірювання якості даних. В такій класифікації важливо враховувати не тільки текстові та числові дані, а й всі інші типи даних, що використовуються та обробляються в певному ресурсі – мультимедійні дані, потокові дані з різного обладнання і частково структуровану інформацію (наприклад, метадані без повної стандартизації або формати представлення знань). У цій роботі ми пропонуємо онтологічну модель НКД, що орієнтована на модель подання інформації у семантичних Wiki-ресурсах. Ця модель містить таксономію даних та формально описує ті методи й засоби , які дозволяють знаходити та перетворювати ці дані у форми, більш придатні для семантичної обробки. Запропонована розширена класифікація спрямована на визначення джерел брудних даних в цьому технологічному середовищі і дозволяє формалізувати їх типи для більш коректної обробки та по змозі способи запобігання їх появі.

Таксономія НКД

У даній статті ми пропонуємо використовувати більш розширену класифікацію для визначення джерел брудних даних, формалізації їхніх типів для коректнішої обробки засобами м'яких обчислень та за можливості шляхів запобігання їх виникненню. Цей підхід орієнтований на підтримку технологій створення та використання семантичних інформаційних ресурсів великого обсягу та складною структурою. Важливо розуміти, що той самий фрагмент даних може бути віднесений одночасно до різних підкласів, якщо він містить одночасно кілька різних некоректностей. Основна мета створення такої таксономії – забезпечити однозначну ідентифікацію типу НКД, щоб дати відповідь на питання щодо можливості та шляхів їх перетворення. Це дозволяє ефективніше перетворювати сирі дані на Smart data без безпосередньої участі експертів зі знань на всіх етапах – необхідні рекомендації здобуваються з цієї таксономії (рис.1).

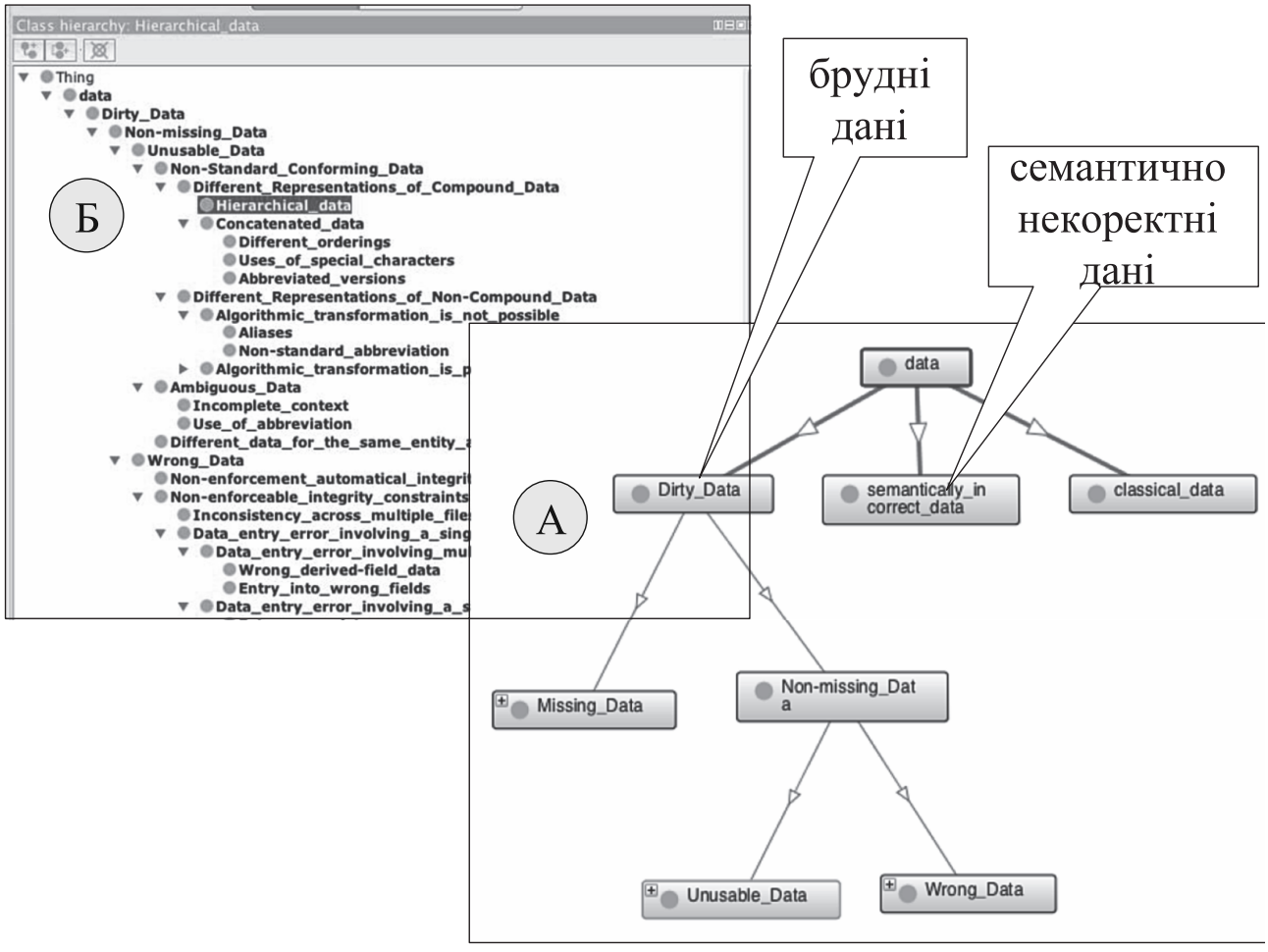

Рис.1. Таксономія класичних та некласичних даних (верхній рівень)

Запропонована класифікація НКД доповнена додатковим класом – *семантично некоректні дані* (рис.1-А). На відміну від брудних даних, різницю таких даних і класичних можна виявити тільки на етапі їх семантичної інтерпретації, у разі якщо збережені значення не відповідають обмеженням ПрО. Тому таке виявлення потребує аналізу знань щодо цієї ПрО із зовнішніх джерел. Наприклад, семантична некоректність може стосуватися віку співробітника, якщо мінімальне або максимальне значення визначаються певними характеристиками його професії та вимогами законодавства певної країни. Семантично некоректні дані можуть мати ті самі джерела, що й брудні дані, але, крім того, існують підкласи семантично некоректних даних, що є такими через некоректність вибору понять, пов'язаних з тими атрибутами, значення яких є некоректними, з неповною семантичною подібністю обраних атрибутів та з вибором області значень цих атрибутів.

До прикладу, якщо замість поняття "працівник" обрано поняття "людина", то це не дозволяє коректно відобразити дані щодо службових собак (таких, як пес Патрон), які входять до складу певного підрозділу .

Всі класи даних (як класичні дані і НКД, так і семантично некоректні дані) мають багато підкласів (рис.1-Б), ієрархія та деталізованість яких залежать від цілі класифікації. В даній роботі основна увага звертається на те, як ці особливості впливають на обробку інформації на семантичному рівні. Запропонована таксономія НСД реалізована як окремий випадок онтології з єдиним типом відношення «клас-підклас» та формалізована засобами Protege. Класи цієї онтології пов'язані ієрархічним відношенням "клас-підклас", що можуть бути візуалізовані за допомогою плагіну OntoGraf (рис.3), а екземплярами є різноманітні приклади НКД та тих класичних даних, в які вони можуть бути перетворені – вручну або автоматизовано.

Методи обробки НКД

Кожний тип НКД потребує різних методів детектування та обробки. Іноді спочатку потрібно виявити, який саме тип некоректності присутній у даних, тому що це може бути незрозуміло із власно даних. Тільки після цього можна знайти відповіді на наступні питання: чи потребують такі дані виправлення, чи можна перетворити їх на класичні дані, і, якщо можливо, чи виконується таке перетворення автоматизовано чи з допомогою людини та чи потребує воно застосування додаткових джерел знань або інструментів аналізу.

У запропонованій таксономії на верхньому рівні всі брудні дані поділяються на два класи відповідно до того, наявні хоч якісь дані чи вони пропущені (missing) – їх значення взагалі відсутні на певний момент часу. Третього варіанту не може бути. Дані вважаються відсутніми, якщо у певне поле не введено жодного значення. В іншому випадку дані введено, і вони вважаються брудними з інших причин.

Дані можуть бути пропущеними з різних причин: 1. коли це дозволено відповідно до їх змісту (нульові дані) – значенні невідомі або неважливі, або 2. коли пропущене введення даних не дозволено.

У першому випадку дані можуть бути відсутні через те, що вони ще невідомі, але вже існують (наприклад, відомо, що людина має електронну пошту, але адреса невідома), через те, що вони поки що відсутні (людина ще не завела пошту, але збирається це зробити) або ж через те, що їх значення відсутнє в принципі (людина померла за багато століть до появи Інтернету). У другому випадку це є помилкою введення і потребує знаходження відсутніх значень (наприклад, у переліку виконавців проєкту відсутнє прізвище керівника).

Зрозуміло, що це різні види нульових даних, і тому у багатьох системах м'яких обчислень з ними пов'язують різні спеціальні значення ("не відомо", "не існує", "не визначено"). Логічне виведення на основі таких даних базується на багатозначних логіках, де для кожного спеціального значення існують правила виведення та аксіоми. Найпростіші з них – значення "не відомо" замінюється на набір усіх припустимих значень, а "не існує"

– на значення, що не співпадає з жодним існуючим. Визначення типу пропущених даних потребує знань щодо ПрО, які отримуються від експерта або із зовнішніх джерел знань.

Таку інформацію доцільно пов'язати безпосередньо із класами та підкласами таксономії НСД. Для цього пропонується розширена онтологічна модель НКД, до якої надаються наступні класи, значення яких використовуються як об'єктні властивості підкласів НКД: 1. метод детектування НСД (рис.2-А); 2. метод перетворення на НКД; 3. зовнішні джерела інформації (рис.2-Б).

До зовнішніх джерел інформації належать різноманітні бази знань та онтології різної виразності, такі як тезауруси, контрольовані словники тощо

Екземпляри цих класів – це конкретні методи Data Mining, машинного навчання, логічного виведення, а також посилання на зовнішні онтології ПрО.

Ці властивості даних можуть розглядатися як логічні змінні зі значеннями "Так" і "Ні", або як нечіткі логічні змінні з ймовірнісними значеннями в діапазоні від 0 до 1. Такі властивості даних певним чином дублюють інформацію, що неявно представлена у таксономії НСД (тому що саме ці параметри і є основою для таксономічного поділу на верхньому рівні), але їх використання значно полегшує обробку інформації – ці дані можуть використовуватися в умовах для запитів щодо пошуку засобів обробки.

Популяція онтологічної моделі екземплярами є складним процесом, що потребує детального аналізу актуальних досліджень із різних напрямків аналізу даних та є поза межами даного дослідження. Але запропонована онтологічна модель задає структуру того, як відповідна інформація може бути представлена та пов'язана з іншими елементами (рис.2-В).

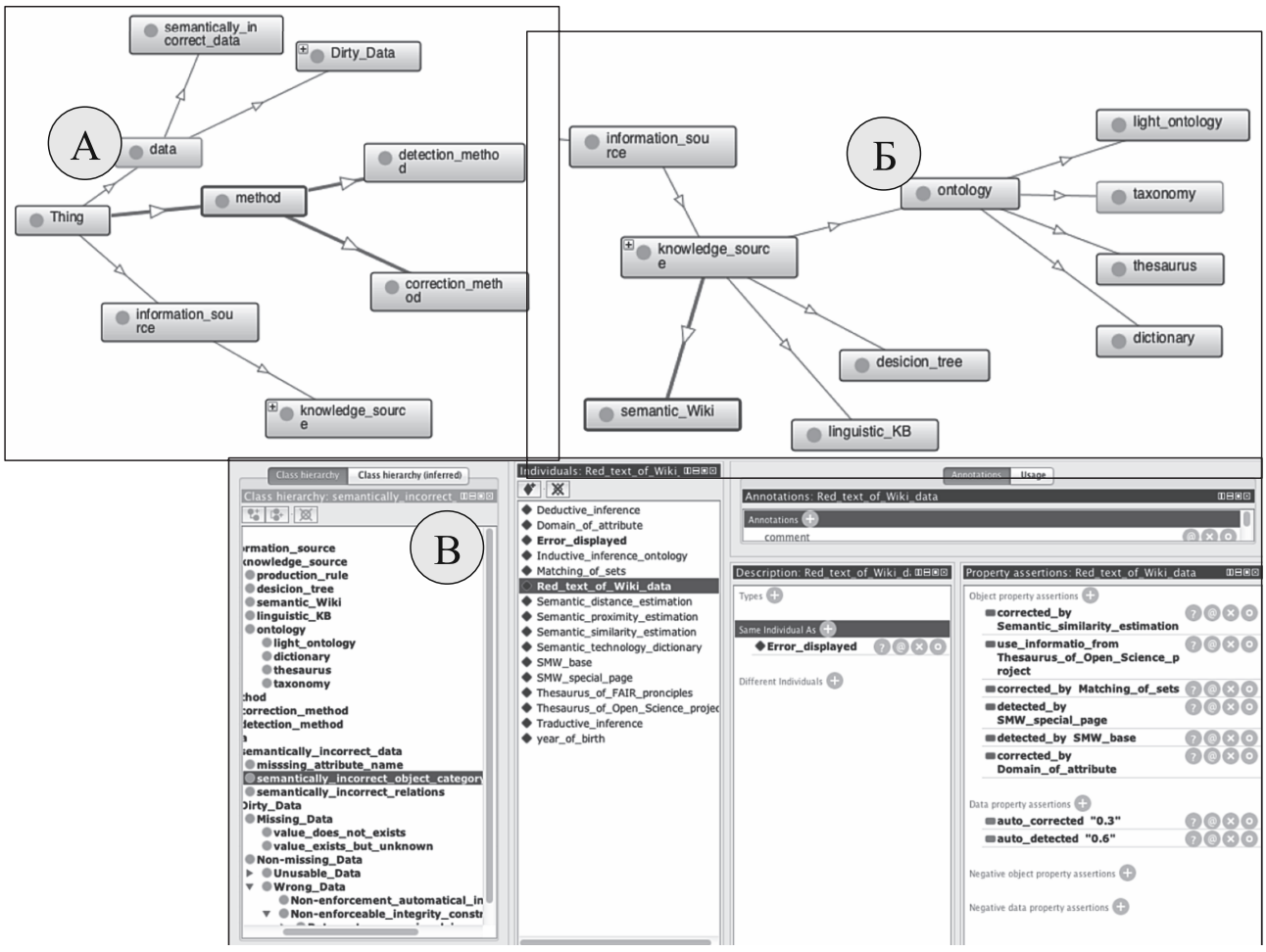

Рис.2. Методи, що використовуються для знаходження та корекції НКД

Брудні дані та семантичні Wiki

Прояви брудних та некоректних даних багато в чому залежать від того технологічного середовища, в якому ці дані створюються, зберігаються та обробляються. Найбільш важливим фактором є виразність, яку середовище забезпечує як для подання інформації (даних та метаданих), так і для їх пошуку: спотворене може бути лише те, що можна відобразити. Крім того, різноманітні інструментальні засоби містять різні види автоматизованого узгодження даних. Потрібно враховувати, що джерелом брудних даних може бути невдалий вибір моделі подання даних, яка не пертинентна вирішуваній проблемі, – наприклад, використання текстового типу замість числового призводить до неправильного впорядкування даних.

Надалі ми пропонуємо конкретизацію таксономії НКД для технологічного середовища семантичних Wiki (а саме – MediaWiki та його розширення Semantic MediaWiki), в якому присутній широкий спектр брудних даних із різних джерел. Семантизація Wiki-ресурсів значно розширює виразність цієї технології, але викликає необхідність аналізувати також і семантичну коректність інформації.

Wiki-технології забезпечують розподілену обробку інформації у відкритому середовищі Web. Особливістю цієї технології є те, що користувачі самостійно створюють та редагують контент сторінок. З одного боку, це забезпечує швидкий розвиток Wiki-ресурсів та збільшення їх обсягу, а з іншого – викликає велику кількість різноманітних помилок та неоднозначностей у даних. Це потребує розробки додаткових моделей та методів перевірки сирих даних, визначення джерел того, що вони стають НКД. Важливо розуміти, що причини нечіткості та неповноти даних у Wiki-ресурсах не завжди є наслідком помилок або відсутності достовірної інформації. Саме тому Wiki-середовище стає цікавим прикладом для класифікації брудних даних та пошуку шляхів їх трансформації у класичні дані.

У багатьох випадках контент Wiki-ресурсу перетворюється на НКД через невдалий вибір моделі та структури контенту, що не пертинентні реальному світу. На жаль, у багатьох випадках така ситуація визначається тільки в процесі накопичення гетерогенного контенту, в якому потрібно подавати складні інформаційні об'єкти, які не відповідають попередньо обраним моделям. Основним елементом контенту Wiki-ресурсу є Wiki-сторінка, що має унікальне ім'я та набір властивостей, котрі можуть розглядатися як метадані.

Найбільш уживане зараз програмне забезпечення для Wiki-систем – MediaWiki [8], яку використовують проекти такі відомі проєкти, як Wikipedia, Wikidata та Wikibooks. Введення даних у MediaWiki підтримується зручним редактором контенту і забезпечує наступні елементи структурування даних: 1. *категорії,* які дозволяють групувати сторінки (сторінка може належати до довільної кількості категорій, а відношення часткового впорядкування дозволяє створювати набори ієрархій цих категорій); 2. *посиланнями* між Wiki-сторінками; 3. *простори імен* сторінок; 3. *шаблони,* які уніфікують окремі елементи контенту сторінок.

MediaWiki не містить засобів перевірки узгодженості використання цих елементів та не відображає семантику зв'язків між ними. Для вирішення цих проблем використовують семантичні плагіни, що розширюють виразність Wiki-ресурсу за допомогою семантичної розмітки, що дозволяє пов'язувати певні елементи контенту з поняттями ПрО. Така розмітка допомагає у структуруванні інформації і робить дані доступнішими для автоматичного аналізу. Наприклад, плагін *Semantic MediaWiki* (SMW) [9] дозволяє пов'язувати зв'язки між Wiki-сторінками та даними з поняттями довільної ПрО та підтримує пошук за цими зв'язками, щоб інтегрувати інформацію з різних Wiki-сторінок, та генерувати за Wiki-сторінками онтологічні структури [10], які можуть використовувати інші системи [11]. Приклад Wiki-ресурсу на основі SMW – портал Великої української енциклопедії – е-ВУЕ [12] .

SMW дозволяє доповнювати контент Wiki-ресурсі: 1. *семантичними властивостями* Wiki-сторінок; 2. *шаблонами*

типових інформаційних об'єктів, які забезпечують уніфіковану семантичну розмітку та спрощують введення значень властивостей; 3. *семантичними запитами.* Саме для цих даних можлива поява семантичної некоректності різних типів. Це викликано тим, що SMW дозволяє безпосередньо формалізувати семантику класів та екземплярів ІО, але не містить достатньо розвинутих засобів для контролю їх несуперечності та узгодженості. Для семантичних Wiki-ресурсів виникає необхідність перевірки семантичної узгодженості введених даних з правилами ПрО, які відображені у зовнішніх джерелах знань, таких як онтології.

Одним із інструментів для цього є використання метрик семантичної подібності та семантичної близькості між елементами онтології, які дозволяють кількісно оцінити пертинентність використання як теги розмітки для семантичних властивостей цих класів, відношень та екземплярів онтології відповідної ПрО.

Семантична близькість є окремим випадком семантичної спорідненості ІО, що стосується спільних властивостей ІО, тоді як семантична спорідненість відображає ймовірність використання ІО в спільному контексті. Семантично близькі поняття ПрО – це нечітка множина, яка включає набір понять, для яких кількісне значення семантичної близькості з обраним поняттям вище заданого порогу. Міри визначення семантичної близькості понять на основі онтологій використовують їхні властивості (атрибути і відношення з іншими поняттями) [13] та взаємне положення в онтологічних ієрархіях [14, 15].

Використовуючи поняття певної ПрО як теги семантичної розмітки Wiki-ресурсу, потрібно переконатися, що обране ім'я властивості відповідає відношенню саме цієї області і не використовується в іншому значенні в іншій ПрО (в такому випадку виникає потреба у додатковому уточненні). Тому одним із важливих кроків перевірки семантичної узгодженості Wiki-даних є визначення кількісної оцінки семантичної спорідненості. Така спорідненість між тегами певної Wiki-сторінки оцінюється як функція від семантичної відстані між відповідними поняттями онтології ПрО та може використовувати довільну підмножину зв'язків між поняттями ПрО, що відповідає цілям оцінювання [16].

Для визначення подібності між тегами семантичного Wiki-ресурсу виникає потреба у вимірюванні подібності тих слів, що використовуються як імена тегів (тобто імен семантичних властивостей сторінки), а не понять, яким відповідають ці теги. Така подібність дозволяє відокремити досить схожі імена різних понять від різних імен близьких за змістом або тотожних понять. Це дозволяє розв'язувати семантичну некоректність, що виникає внаслідок колективної паралельної роботи спеціалістів різних галузей зі вдосконалення структури Wiki-ресурсу: досить часто створюються семантичні властивості зі схожими іменами, які мають різне значення у різних галузях знань.

Е-ВУЕ та брудні дані

Розглянемо детальніше використання онтологічної моделі брудних даних семантичного Wiki-ресурсу на прикладах, пов'язаних із розробкою та поповненням е-ВУЕ (vue.gov.ua). Ми обрали цей інформаційний ресурс, тому що він побудований на MediaWiki та SMW, має складну структуру, великий обсяг та відображає зв'язки між поняттями різних областей. В ньому представлені екземпляри різних ІО, що описані даними різних типів – текст, числа, посилання, мультимедіа тощо.

У створенні та оновленні гасел е-ВУЕ бере участь велика кількість співробітників установи, тоді як сам контекст створюють незалежні експерти різних ПрО. Внаслідок цього виникають неоднозначні тлумачення вимог щодо форми подання даних, щодо структури та правил застосування шаблонів типових ІО та їхніх атрибутів. Використання таксономії НКД дозволяє більш точно виявляти причини виникнення неточності й некоректності даних у Wiki-ресурсі та у разі можливості рекомендувати шляхи їх перетворення. Залежно від типу некоректності, необхідно змінювати дані або вносити доповнення та зміни в їхню модель.

Для інтеграції набору семантичних властивостей виділяються типові ІО – групи гасел, віднесених до визначеного набору категорій та які мають фіксований набір характеристик.

Відсутні дані в е-ВУЕ. Відсутні дані допустимі у шаблонах ІО, якщо для деяких екземплярів певні властивості невідомі (на даний момент або взагалі). Цю ситуацію потрібно враховувати в процесі створення шаблону (тобто обов'язково проводити перевірку, що значення не є порожнім), тому що в іншому випадку спроба вивести неіснуючу інформацію призведе до помилок. Така перевірка потребує додаткових обчислень, тому потрібно консультуватися зі спеціалістами ПрО щодо її необхідності.

Наприклад, в е-ВУЕ шаблон ІО «Персоналія» містить параметр, що відповідає семантичній властивості «Псевдонім». Але не всі видатні особи, представлені в енциклопедії, мали псевдоніми. Тому код шаблону містить перевірку, за результатами якої значенні властивості виводиться тільки в тому випадку, якщо воно не є порожнім:

{{if||{{{Псевдонім|}}}| ‘’’Псевдоніми’’’ {{#arraymap:
{{{Псевдонім|}}}|;|x|[[Псевдонім::x]]}}}}

Інформація про можливість таких ситуацій має бути отримана від експерта ПрО до початку використання відповідного шаблону.

Для складніших ситуацій (неточна інформація, відсутність частини значення) можна ввести різні значення для різних типів відсутніх даних та виконувати перевірку для кожного з варіантів:

{{if||{{{Псевдонім|::Null value}}}|
‘’’Значення не визначено’’’
{{if||{{{Псевдонім|:: Other type}}}|
‘’’Значення не представлене в придатній формі}}

Неоднозначні дані у е-ВУЕ. Такі дані викликані використанням неоднозначних

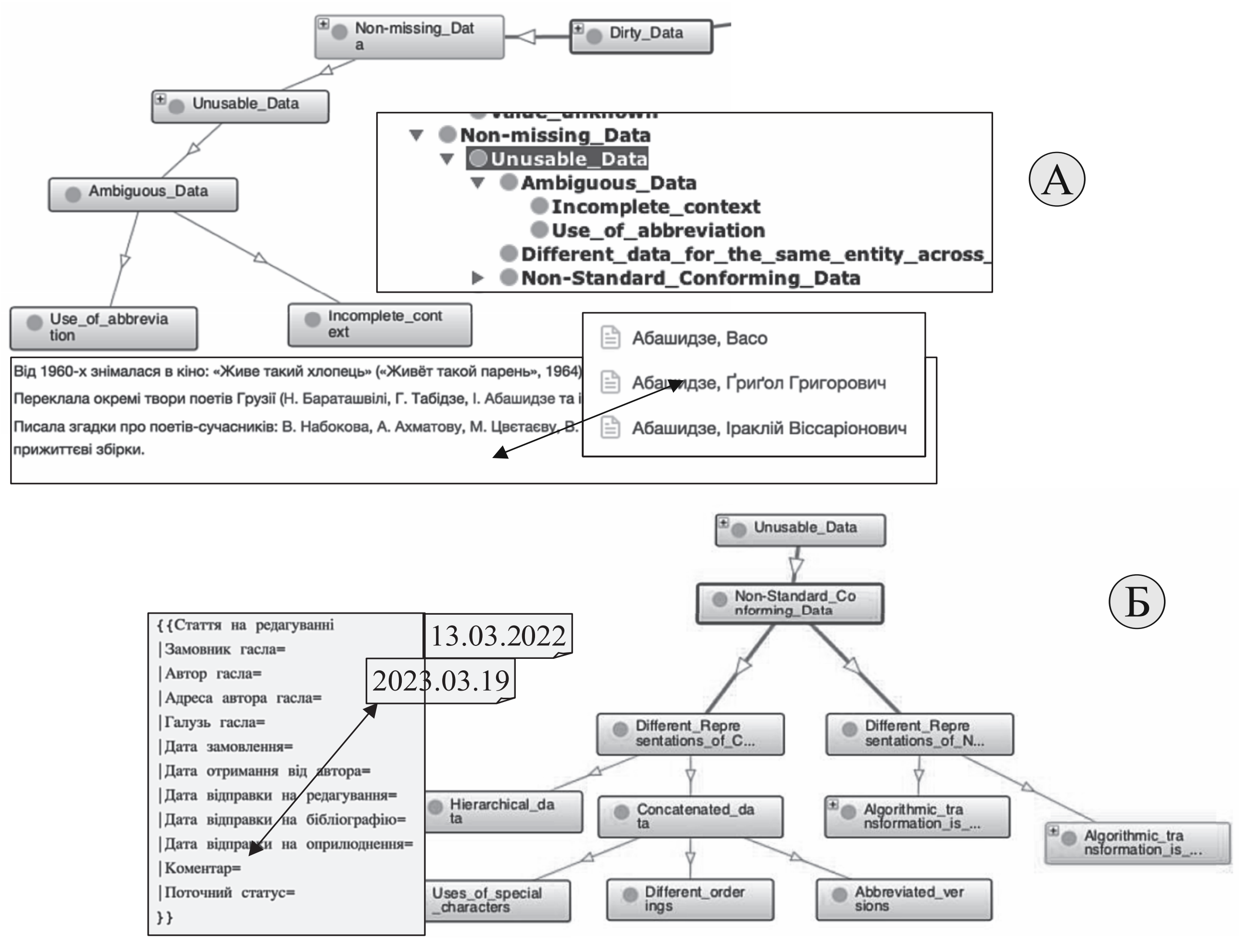

Рис.3. Приклади неоднозначних та непридатих даних у е-ВУЕ

начних абревіатур та скорочень або неповнотою даних. Приклад неповноти даних – у тексті гасла є посилання на інше гасло, але інформація про ім'я цього гасла є неповною та допускає кілька варіантів доповнення (наприклад, наведено тільки прізвище без ініціалів або ініціали без повного імені – рис.3-А). В такому випадку проблема не може бути розв'язана автоматично засобами SMW та потребує або участі експерта, або застосування зовнішніх джерел знань та мір семантичної близькості. Приміром, якщо кілька гасел мають імена, що можуть бути скорочені до використаного у посиланні, то доцільно обрати те, що має більше спільних категорій з гаслом, в якому міститься посилання.

Дані, непридатні для обробки в е-ВУЕ через нестандартне подання. Такі дані не можуть бути ефективно використані (знайдені за запитами, об'єднані тощо) через те, що подання інформації не відповідає прийнятим правилам та стандартам. Причини такої невідповідності можуть бути різними. Частина з них пов'язана із окремими елементами даних, а інша – саме зі зв'язками між елементами даних, присутніми у даних зі складною структурою. Наприклад, службовий шаблон «Стаття на редагуванні» (рис.3-Б) дозволяє визначити поточний статус ще не оприлюдненої статті. Обов'язковими параметрами є тільки «Автор гасла», «Галузь гасла», «Дата замовлення» та «Поточний статус» (їх відсутність є неприпустимою помилкою та належить до НКД «Неприпустимі відсутні дані». Атрибут «Поточний статус» визначає, що саме зараз роблять зі статтею, і його значення мають приймати значення з обмеженої множини {підготовка автором, рецензування, перевірка науковим редактором, літературне редагування, узгодження з автором, інше}. Всі інші значення цього параметру не розпізнаються SMW як помилка, але вони не будуть коректно оброблятися у запитах (рис.3-Б).

Дані, непридатні для обробки в е-ВУЕ через неструктуроване нестандартне по-

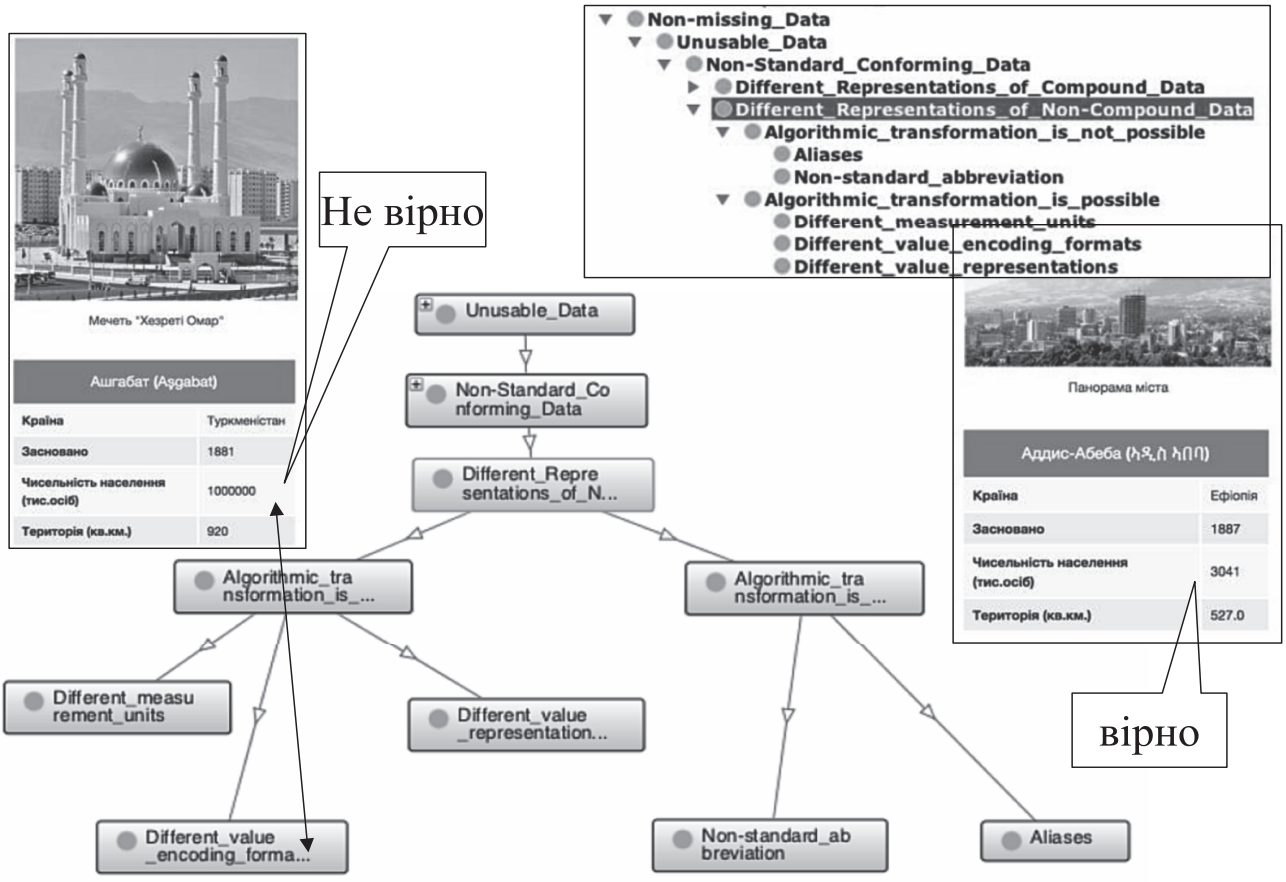

Рис.4. Приклад неоднозначних даних е-ВУЕ, що непридатні для обробки в е-ВУЕ через нестандартне подання

дання – це підтип нестандартного подання даних (рис.4). В цьому випадку дані належать до НКД не через зв'язки з іншими елементами даних, а через властивості окремих значень. Найпоширеніші ситуації в е-ВУЕ з такими даними – це некоректний вибір одиниць виміру. Зокрема, якщо вказано, що чисельність населення наведена у тисячах осіб, то потрібно вводити число в 1000 разів менше реального значення. Якщо для міста Ашгабат введено, що його чисельність – 1000000 тис. осіб, то це є помилкою, яка потребує виправлення. Такі помилки практично неможливо знаходити автоматично (тільки шляхом співставлення із зовнішніми базами даних, інформація в яких може відрізнятися через різний час оприлюднення), але досить легко виконується експертом відповідної ПрО. Слід відзначити, що в деяких випадках помітити такі помилки досить складно, тому що, наприклад, швидкість вітру в метрах на секунду або у кілометрах на годину відрізняються менш ніж на порядок.

Семантично некоректні дані в е-ВУЕ. Розпізнавання семантично некоректних даних для Wiki-середовища включає наступні ситуації (рис.5):

- використовується ім'я атрибута, яке не існує;
- використовується посилання на значення атрибута, якого не існує;
- категорія значення за змістом не відповідає атрибута;
- введене значення атрибута не релевантне, або містить непотрібні елементи;
- за допомогою семантичних властивостей сформовано складний інформаційних об'єкт, який не може існувати в реальному світі.

Важливо підкреслити, що значеннями даних, оброблюваних в середовищі SMW, можуть бути не тільки текст та числа, але й мультимедійні дані – зображення, аудіо та відео. Розпізнавання їх семантики (наприклад, розпізнавання тексту у зображеннях або розпізнавання мовлення

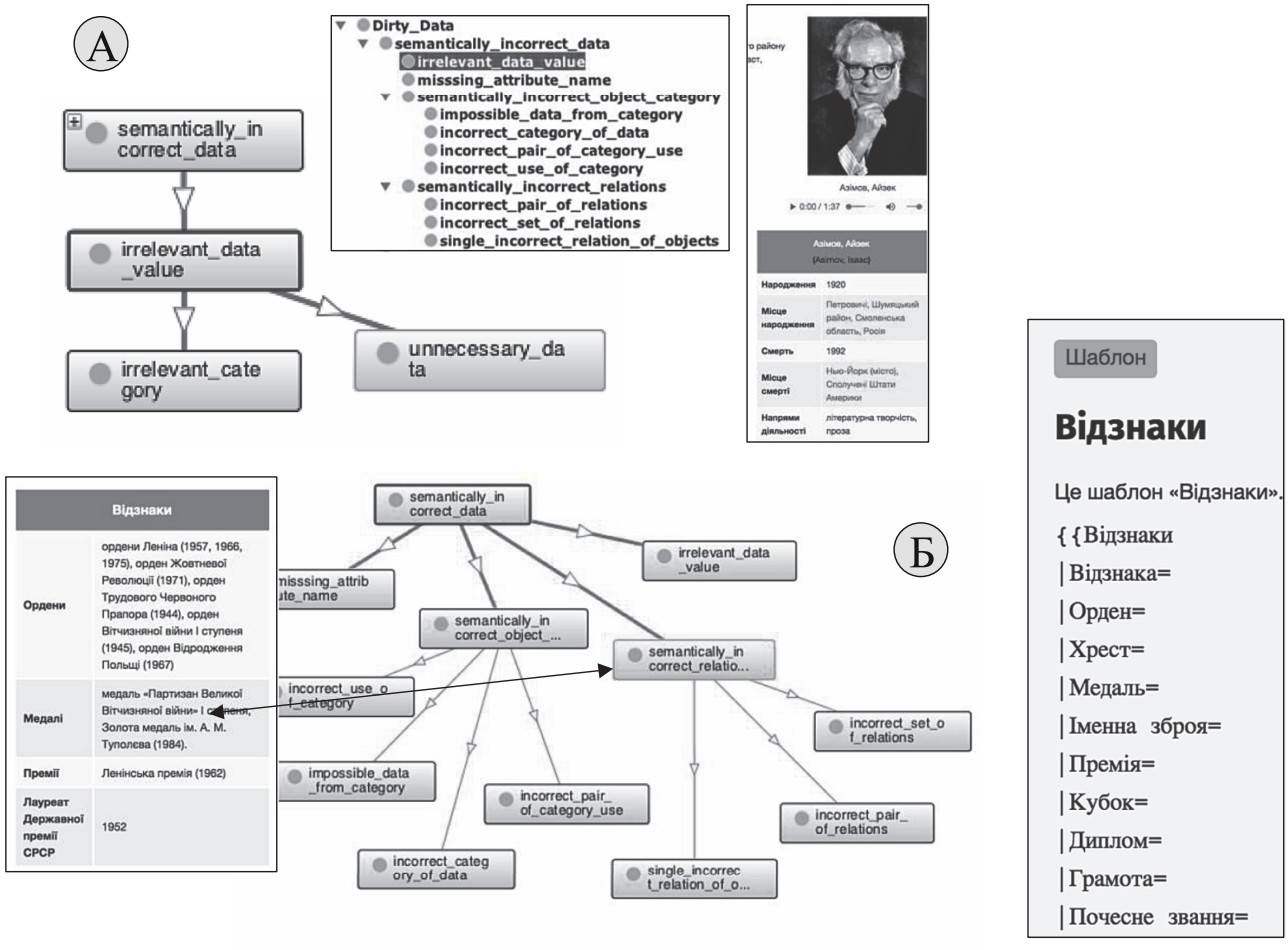

Рис.5. Таксономія семантично некоректних даних (фрагмент)

в аудіофайлах) знаходиться поза сферою дослідження даної роботи, а семантика таких даних визначається на основі аналізу їхніх метаописів.

Пошук семантичних некоректностей вручну з допомогою експерта потребує багато часу для великих обсягів даних та не є надійним. Тому його доцільно застосовувати тільки на початкових етапах створення бази знань Wiki-ресурсу, коли експерт ще тільки шукає коректні відповідності між моделями даних та реальним світом.

Один із поширених варіантів семантично некоректних даних в е-ВУЕ – посилання на сторінку гасла, якої не існує (рис.5-Б). Цей тип НКД контролюється засобами середовища MediaWiki – посилання виводиться червоним кольором. У такому разі доцільно обирати одне з можливих рішень – створювати відповідну відсильну сторінку, змінювати значення посилання на існуюче (якщо була припущена помилка в імені) або перетворювати тип даних такого атрибута на текстовий замість посилання. Останнє рішення обирають для тих атрибутів, які мають багато різних значень, що не використовуються в інших гаслах і не є достатньо значущими для створення окремої сторінки. Наприклад, вказуючи місце народження певної особи, достатньо дати посилання на країну та найближчий регіональний центр, а назву невеликого селища залишити тільки в природномовному контенті сторінки.

Інший варіант семантичної некоректності – використовується значення атрибута, що не є релевантним, хоча відповідає узагальненим вимогам щодо типу даних. Однією з найпоширеніших груп НКД в е-ВУЕ є повторювання у значенні семантичної властивості імені властивості або одиниць вимірювання (рис.5-Б). Так, якщо у значенні властивості «Орден» введено «ордени Леніна», то слово «ордени» є зайвим та заважатиме у семантичному пошуку осіб, нагороджених цим орденом. Найбільш ефективний метод виявлення таких помилок – напівавтоматичний: спочатку створюється семантичний запит, результатами якого є значення властивості, що перевіряється. Потім ці результати впорядковуються, і експерт проглядає їх, переконуючись у правильному виборі значень та коректності впорядкування.

Семантична некоректність вибору атрибутів даних в е-ВУЕ – це ще один тип семантичної некоректності, викликаний неоднозначністю слів природної мови, які використовуються для назв шаблонів та властивостей. Це пов'язано з тим, що в цьому мультидисциплінарному ІР виникає потреба для ідентифікації понять та відношень різних ПрО, які мають різний зміст (тобто їх необхідно відрізняти у пошуку), але терміносистеми цих ПрО перетинаються й потребують додаткового уточнення. Наприклад, властивість "Відзнака" для персоналій, що стосується нагород за результати діяльності, має зовсім іншу область значення та набір можливих значень, аніж властивість "Особлива відзнака" в біології, яка характеризує відмінності між тваринами та рослинами. Для семантичних властивостей у SMW така перевірка частково автоматизована (рис.12). На жаль, такий пошук надає тільки статистичні оцінки подібності та є основою для подальшого аналізу: потрібно перевірити, чи використовуються в ПрО обидва знайдені терміни (і тоді ситуація не є помилковою), чи використовується тільки один з них (тоді інший є некоректним), чи немає взагалі таких або подібних понять, і потрібно видалити обидва. Для цього доцільно застосовувати запити до онтології ПрО. В першому випадку введені дані за замовчанням будуть віднесені до типу "посилання". Через те, що ймовірність існування Wiki-сторінки з таким ім'ям дуже мала, такі дані будуть виводитися червоним кольором, який у цьому технологічному середовищі вказує на помилку. Обробка такої семантичної некоректності досить проста: користувач має обрати один з варіантів – створити відповідну властивість або замінити використане ім'я властивості на ім'я існуючої.

Семантична некоректність вибору області значення атрибутів в е-ВУЕ пов'язана з тим, що в SMW відсутня можливість конкретизувати область значення властивості типу "Посилання" через набір категорій або обмеження значень

семантичних властивостей. Наприклад, якщо для семантичної властивості "Місце народження" обрано значення "Дніпро (річка)" замість "Дніпро (місто)", то таку некоректність можуть розпізнати тільки експерти, тому що тільки вони можуть відокремити неправильно використані значення від особливих ситуацій (наприклад, людина дійсно народилася у океані на кораблі). На відміну від онтологічних моделей, де можна явно вказати область значення та область визначення відношень, середовище SMW безпосередньо не підтримує такі функцій, але дозволяє створювати онтологічну модель фрагменту ІР (в форматі RDF) [17], яку можна обробляти зовнішніми інструментами для аналізу онтологій.

Семантично некоректні відношення між екземплярами типових ІО в е-ВУЕ. В деяких випадках неприпустимі ситуації певної ПрО легко описати логічними правилами, але SMW не містить відповідного формального апарату. Зокрема, якщо сторінка особи А посилається на сторінку особи Б як на попередника у дослідженнях, але сторінка особи Б посилається на сторінку особи А як на попередника у дослідженнях, то така ситуація є семантично некоректною. Інший приклад – сторінка особи А посилається на сторінку особи Б як на батька, але сторінка особи Б посилається на сторінку особи А як на брата. Автоматизована перевірка таких ситуацій у Semantic MediaWiki неможлива через те, що виразна здатність середовища не дозволяє визначати формально такі характеристики властивостей, як транзитивність, симетричність, антисиметричність тощо. Більш складні поєднання потребують виведення у багатозначній логіці. Такі семантичні неузгодженості можна знаходити на основі логічного виведення, що знаходиться поза можливостями Semantic MediaWiki, але може підтримуватися зовнішніми засобами онтологічного аналізу. Тому може бути запропоноване наступне рішення: 1. Згенерувати RDF-файл за набором Wiki-сторінок, для яких потрібно виконати перевірку на НКД, 2. Виконати перевірку цієї згенерованої сукупності даних.

Семантична некоректність категорії даних в е-ВУЕ. Для багатозначних гасел ця проблема вирішується безпосереднім додаванням назви ПрО або категорії поняття до назви гасла, що використовують багатозначні терміни, – як-от, "Болід (астрономія)" замість "Болід" та "Бетховен (кратер)" замість "Бетховен". Це знімає семантичну неоднозначність, але може спричинити некоректні посилання з інших гасел: в них можуть бути використані імена без цих уточнень, особливо у випадку, якщо ці гасла створювалися раніше, ніж ті, на які вони посилаються. Для перевірки цього доцільно створювати онтологічне представлення певного фрагменту ІР, що перевіряється. SMW надає можливість генерувати результати семантичних запитів у форматі RDF. Після цього згенеровану онтологію можна співставити із зовнішніми онтологічними моделями або продемонструвати її структуру експерту. Слід відзначити, що графічне подання онтологічної інформації значно спрощує її сприйняття людиною та є потужним інструментом для виявлення семантичних некоректностей різних типів.

Найпростіший випадок такої некоректності – помилковий вибір категорії сторінки (безпосередньо або через використання нерелевантного шаблону). Перший випадок, причиною якого зазвичай є копіювання контенту іншої сторінки для подальшого редагування, легко відстежити за допомогою оцінок семантичної подібності. Але це потребує написання спеціалізованого програмного коду. Другий випадок значно менш наочний, тому що категорія може бути надана сторінці одним із багатьох вкладених шаблонів. Детектування таких ситуацій може базуватися на виявленні групи сторінок з однаковими некоректно обраними категоріями та їх порівняння для виявлення спільно використаних шаблонів. Таке дослідження потребує глибокого аналізу бази даних ІР. Запобігти таким ситуаціям дозволяє створення онтологічної моделі ІР (не автоматизоване), в якому формалізовано фіксуються всі відношення між семантичними властивостями,

категоріями та шаблонами безпосередньо у момент їх створення у ресурсі [18]. Онтологічна модель має значно більшу виразність порівняно з SMW і дозволяє відображати характеристики цих елементів бази знань та обмеження щодо їх використання. Важливо, що запити до такої моделі виконуються автоматично, тобто можна визначити для сторінки з некоректною категорією усі ті шаблони, що містять таку категорію. Крім того, для її аналізу можна застосовувати різноманітні спеціалізовані аналітичні інструменти. Зараз існує велика кількість інструментальних засобів для перевірки різних аспектів якості онтологій. Наприклад, OOPS! (OntOlogy Pitfall Scanner!) (http://oops.linkeddata.es/) – відкрите програмне забезпечення, яке дозволяє виявляти транзитивні та симетричні властивості об'єктів. Вибір засобів перевірки залежить від того, які саме семантичні некоректності потрібно перевірити. Після цього можливо прийняти одне з двох можливих рішень – внести зміни у відповідний шаблон або згенерувати інший шаблон з іншим набором категорій.

Потрібно враховувати, що всі перевірки семантики у Wiki-середовищі, що стосуються використання категорій, потребують написання додаткового програмного коду, на відміну від обробки семантичних властивостей. Тому доцільно дублювати інформацію щодо категорій за допомогою апарату семантичних властивостей, які надає Semantic MediaWiki.

Наведені приклади НСД, що можливі у технологічному середовищі SMW та виникали в процесі створення e-BYE, не вичерпують усі ситуації, які проаналізовано в запропонованій вище онтологічній моделі. Крім того, в процесі розвитку інструментарію для розробки семантичних Wiki-ресурсів, збільшується набір ситуацій, що виявляються та розв'язуються вбудованими засобами. З іншого боку, зростання обсягу та ускладнення структури інформаційних ресурсів на основі цієї технології призводить до появи нових прикладів НСД, які можуть бути класифіковані на основі цієї моделі, але потребують спеціалізованих засобів обробки.

Висновки

Запропонована у дослідженні онтологічна модель призначена для класифікації різних типів брудних та семантично некоректних даних, що уможливлює ефективніший пошук методів виявлення таких даних та засобів їх обробки. Така обробка, що може розглядатися як одна зі складових Smart data, має зробити дані придатними для автоматичного аналізу та використання в інших інформаційних системах. Онтологічний підхід забезпечує інтеграцію запропонованої моделі з іншими зовнішніми онтологіями, що описують різноманітні методи та програмні засоби аналізу даних (наприклад, онтологія індуктивних методів [19] та онтологія Data Mining [20]) та можуть бути застосовані для пошуку некоректностей у даних та їх очищення, а також із онтологіями ПрО, в яких представлені більш коректні, точні та актуальні відомості.

У роботі використано досвід розробки бази знань портальної версії Великої української енциклопедії e-BYE, великою за обсягом, із складною структурою та великою кількістю різноманітних гетерогенних інформаційних об'єктів. Через те, що в створенні цього інформаційного ресурсу бере участь велика кількість спеціалістів різних наукових напрямків із різною областю експертизи та різною кваліфікацією щодо застосування знанняорієнтованих інформаційних технологій, виникає багато розбіжностей у розумінні правил подання та структурування даних. Тому виникає необхідність у формалізованих та масштабованих рішеннях для знаходження та опрацювання різноманітних типів нечіткості, неповноти та семантичної некоректності контенту.

Слід підкреслити, що наведені приклади некласичних даних, що можливі у технологічному середовищі SMW та виникали в процесі створення e-BYE, не вичерпують усі ситуації, які проаналізовано в запропонованій вище онтологічній моделі. Крім того, в процесі розвитку інструментарію для розробки семантичних Wiki-ресурсів, збільшується набір ситуацій, що виявляються та розв'язуються вбудованими засобами. З іншого боку,

зростання обсягу та ускладнення структури інформаційних ресурсів на основі цієї технології призводить до появи нових прикладів НСД, які можуть бути класифіковані на основі цієї моделі, але потребують спеціалізованих засобів обробки.

Запропонований підхід може бути корисним для створення інших великомасштабних ресурсів як на основі технології семантичних Wiki, так і інших технологічних платформ колаборативної обробки розподілених даних та знань.

References

1. Zadeh L. A. Fuzzy sets and information granularity. Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers, 1979, pp.433-454.
2. Motro, A., Smets, P. Uncertainty Management in Information Systems: From Needs to Solutions. Springer, 1997. 464 p. DOI: http://dx.doi.org/10.1007/978-1-4615-6245-0.
3. Codd E. F. Missing information (applicable and inapplicable) in relational databases. ACM Sigmod Record, 15(4), 1986, pp.53-53.
4. Parsons S. Current Approaches to Handling Imperfect Information in Data and Knowledge Bases // Knowledge and Data Engineering IEEE, Vol.8, №3, 1996. pp. 483-488.
5. Zadeh L. A. The concept of a linguistic variable and its application to approximate reasoning. Information sciences, 8(3), 1975pp.199-249, DOI: http://dx.doi.org/10.1016/0020-0255(75)90036-5.
6. Kim W., Choi, B. J., Hong E. K., Kim S. K., Lee D. A taxonomy of dirty data. Data mining and knowledge discovery, 7, 2003, pp.81-99.
7. Kim W., Chae K. J., Cho D. S., Choi B., Jeong A., Kim M., Yong H. S. The Chamois component-based knowledge engineering framework. Computer, 35(5), 2002, pp.45-54.
8. Koren Y. Working with MediaWiki. San Bernardino, CA, USA: WikiWorks Press. 157-159(2012). URL: uplooder.net.
9. Semantic MediaWiki. https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki.
10. Guarino N. Formal Ontology in Information Systems. Formal Ontology in Information Systems. // Proc. of FOIS'98, 3-15, 1998.
11. Rogushina J.V., Grishanova I.J. Ontological methods and tools for semantic extension of the media WIKI. Problems in programming, № 2-3, 2020. pp.61-73. DOI:10.15407/pp2020.02-03.061.
12. Andon P.I., Rogushina J.V., Grishanova I.Y., Reznichenko V.A., Kyrydon A.M., Aristova A.V., Tyschenko A.O. Experience of Semantic Technologies Use for Development of Intelligent Web Encyclopedia. Proc. of the 12th International Scientific and Practical Conference of Programming (UkrPROG 2020),CEUR Workshoop Proceedings, 2021, Vol-2866, P.246-259. http://ceur-ws.org/Vol-2866/ceur_246-259andon24.pdf
13. Tversky A. Features of similarity. Psychological review, 84(4), 1977, pp.327-341.
14. Rada R., Mili H., Bicknell E., Blettner M. Development and application of a metric on semantic nets. IEEE transactions on systems, man, and cybernetics, 19(1), 1989, pp.17-30.
15. Resnik P. Semantic Similarity in a Taxonomy: An Information-Based
Measure and its Application to Problems of Ambiguity in Natural Language. In: Journal of Artificial Intelligence Research 11, 1999, pp.95-130..
16. Rogushina J. Use of Semantic Similarity Estimates for Unstructured Data Analysis. Selected Papers of ITS 2019. CEUR Vol-2577, pp.246-258. URL: http://ceur-ws.org/Vol-2577/paper20.pdf [last accessed 2023/02/122].
17. RDF Web Ontology Language. Overview, W3C, 2012. https://www.w3.org/RDF/ [last accessed 2023/02/15].
18. Rogushina J., Grishanova I. Ontological methods and tools for semantic extension of the media WIKI technology. Problems in Programming, № 2-3, 2020, pp.61-73.
19. Pidnebesna H., Stepashko V. Ontology Application to Constructing the GMDH-Based Inductive Modeling Tools. Semantic Web Technologies, 2022, pp. 263-292.
20. Panov P., Dzeroski S., Soldatova L. OntoDM: An ontology of data mining. In: 2008 IEEE International Conference on Data Mining Workshops, IEEE, 2008, pp. 752-760.

Одержано: 05.04.2023

Про автора:

Рогушина Юлія Віталіївна,
канд.фіз.-мат.наук, с.н.с.
Публікації в українських виданнях – 200,
публікації в іноземних журналах – 40.
Індекс Хірша: Scopus – 5, Google Scholar – 20.
ORCID http://orcid.org/0000-0001-7958-2557.

Місце роботи автора:

Інститут програмних систем НАН України,
03181, Київ-187, проспект Академіка Глушкова, 40,
e-mail: ladamandraka2010@gmail.com,
066 550 1999.