УДК 004.

Шаховська Н. Б.[1], Болюбаш Ю. Я.[2]

[1]Д-р техн. наук, професор, Національний університет «Львівська політехніка», Україна, E-mail: nataly233@gmail.com
[2]Здобувач, Національний університет «Львівська політехніка», Україна

ОПРАЦЮВАННЯ НЕВИЗНАЧЕНОСТІ У ВЕЛИКИХ ДАНИХ

У статті уведено поняття терміну «великі дані» та проаналізовано причину їх появи. Показано рівні виникнення невизначеності у «великих даних». Сформовано модель сховища даних з невизначеністю та розроблено операції над ним. Подано метод формування агрегату з врахуванням невизначеності.

Ключові слова: великі дані, інформаційний продукт, невизначеність.

ВСТУП

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 структуризацію 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 роботи 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, взаємонезалежних 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.

Проаналізуємо інформаційні технології для організації різнотипних інформаційних об'єктів та налагодження обміну інформації між ними.

1 ОГЛЯД ЛІТЕРАТУРНИХ ДЖЕРЕЛ ТА ПОСТАНОВКА ЗАДАЧ

Опрацюванням різнотипних неузгоджених даних дослідники займалися з 70-х років XX 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, видобуванні 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-застосувань (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

 DOI 10.15588/1607-3274-2014-1-14

зв'язку з формуванням нових вимог до джерел даних та їх динамічному додаванні.

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

Великі дані (Big Data) в інформаційних технологіях – набір методів та засобів опрацювання структурованих і неструктурованих різнотипних динамічних даних великих обсягів з метою їх аналізу та використання для підтримки прийняття рішень. Є альтернативою традиційним системам управління базами даних і рішеннями класу Business Intelligence. До цього класу відносять засоби паралельного опрацювання даних (NoSQL, алгоритми MapReduce, Hadoop) [1].

Визначальними характеристиками для великих даних є обсяг (volume, в сенсі величини фізичного обсягу), швидкість (velocity, в сенсах як швидкості приросту, так і необхідності високошвидкісної обробки та отримання результатів), різноманіття (variety, в сенсі можливості одночасної обробки різних типів структурованих і напівструктурованих даних).

З одного боку, через свою неоднорідність і постійне зростання Big Data вимагають до себе нестандартних підходів у зберіганні та опрацюванні. Для ефективної роботи необхідні комплексні рішення моніторингу, фільтрації, структурування та пошуку ієрархічних зв'язків. З іншого – використовуючи Big Data, можна спостерігати за величезною множиною змінних, і на основі наданої інформації виявляти глобальні тренди і висновки, розглядаючи певну ситуацію в перспективі.

Однією з технологій, що доцільно використовувати для роботи з Великими даними, є простір даних.

Простір даних – це блоковий вектор, що містить множину інформаційних продуктів предметної області, поділену на три блоки: структуровані дані (бази, сховища даних), напівструктуровані дані (XML, електронні таблиці) та неструктуровані дані (текст). Над цим вектором та його окремими елементами визначено операції та предикати, які забезпечують [1]: перетворення різних елементів вектора один в одного; об'єднання елементів одного типу; пошук в елементах за ключовим словом.

ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

Як було зазначено вище, розглядається задача опрацювання даних, що надійшли з різних, наперед неузгоджених джерел. Ідеалізована схема опрацювання різнотипних даних подана на рис. 1.

Як бачимо, певна множина даних може бути відсутня у джерелах даних, а інша може перекриватися у різних інформаційних продуктах. Тому виникає проблема дублювання, відсутності, неповноти та нечіткості даних.

Невизначеність може виникати на рівні атрибуту, кортежу та відношення (невизначеність у схемі опису). По-

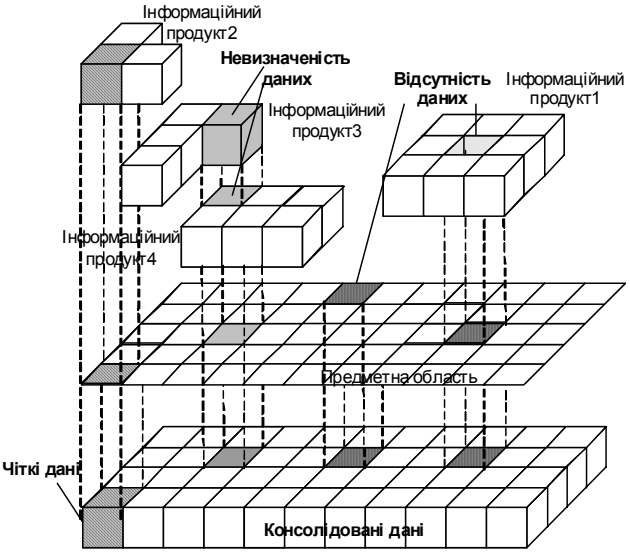

Рис. 1. Схема консолідації даних

ява невизначеності на рівні атрибута і кортежа у зв'язку з багатовимірністю відображення інформації призводить до поширення невизначеності на всі примірники опису певного концепту. Оскільки об'єднуються мільйони даних про об'єкти проблемної області, то опрацювання невизначеності традиційними засобами (інтервальна математика, багатозначна логіка) стає неефективним через велику кількість операндів.

Розглянемо проблеми подання невизначеностей у Big Data. Вважатимемо, що для тимчасового зберігання інформація потрапляє у федеративне сховище даних.

Об'єкт, який моделюється кортежем відношення фактів з відсутніми значеннями зовнішніх ключів, не має властивостей, описаних у відношеннях метаданих – така невизначеність притаманна і відношенням реляційних баз даних.

Відомо, що значення за вказаним атрибутом існує, але на певний час воно невідоме, що викликає необхідність застосовувати алгоритми видобування даних для усунення невизначеності – така невизначеність також існує у реляційних базах даних, але методи її опрацювання не можуть застосовуватись у сховищах даних, оскільки сховищам даних притаманні не тільки зв'язки між об'єктами різних типів, але й між об'єктами одного й того ж типу (виникнення ієрархії об'єктів).

Є неповна або часткова інформація про значення, для відображення якого використовується додатковий атрибут, що характеризує рівень істинності даних та містить значення функцій розподілу, лінгвістичних змінних, ступенів істинності багатозначних логік (може вводитися на рівні значення атрибута, підмножини значень атрибутів або кортежа). Існування такої невизначеності приводить до появи нечіткого відношення, яке може містити суперечливу інформацію.

Крім того, невизначеність може виникати внаслідок отримання агрегованої інформації, коли необхідно знати детальні дані, наприклад, невідомі обсяги продажу у вказаному регіоні за вказаним товаром.

Отже, специфіка Big Data:

– наявність множини різнотипних джерел;

– дублювання даних;

– неоднозначність опису джерел даних,

приводить до того, що невизначеність, яка у традиційних реляційних базах даних розглядалася у межах одного відношення і могла виникати на рівні атрибута, кортежа та на рівні відношення, в цьому випадку поширюється через сприйняття користувачем інформації на все федеративне сховище даних (гіперкуб даних). Тому для опрацювання невизначеності у гіперкубі даних необхідно використати якісно новий підхід, потреба застосування якого не виникала у реляційних базах даних.

У федеративному сховищі даних невизначеність може виникати і у відношеннях метаданих.

Проаналізуємо місця виникнення невизначеностей у сховищах даних.

1. Невизначеність у схемі посередника (медіатора).

Посередник (mediator) – програмний компонент, що, з одного боку, взаємодіє з користувачем інтегруючої системи, та, з іншого боку, з інформаційними джерелами. Він надає єдину «точку входу» (програмний інтерфейс) для запитів користувачів та виконує основні стадії опрацювання запиту:

– визначення джерел, які можуть містити результат запиту;

– декомпозицію на запити до конкретних джерел (на основі їхніх описів);

– оптимізацію плану виконання.

Схема посередника – це множина схем термінів, що зустрічаються у запитах. У термінах сховища даних посередником є метод визначення структури джерела. Схемою посередника є множина таблиць метаданих. Він не обов'язково охоплює усі атрибути будь-якого з джерел, але містить інформацію про домени джерела даних. Невизначеність у схемі посередника може виникнути з кількох причин. По-перше, якщо схеми посередника автоматично визначаються з даних джерел під час запуску, виникає невизначеність з приводу результатів запиту. По-друге, коли домени є широкими, виникає невизначеність стосовно відповідності схем даних чи їх перекриття.

Іншими словами, невизначеність у схемі посередника виникає внаслідок порівняння структур даних джерел для завантаження з них інформації. Така невизначеність призводить до неточного відображення схеми джерела і є джерелом для інших невизначеностей. Причинами невизначеності зазначеного типу є зовнішні (атаки), програмні, апаратні збурення в процесі відбору, опрацювання та завантаження даних.

2. Невизначеність у схемі відображення.

Зазвичай виникає у словнику синонімів (відношеннях метаданих місця). Вказаний тип є частковим випадком невизначеності у схемі посередника. Оскільки словник синонімів визначає семантичні відношення між термінами в джерелах даних, які є повністю незалежними, а багато первинних відображень схем будуть автоматично отримані, то отримані відображення можуть бути неточними. Прикладом такої невизначеності може бути випадок, коли одним терміном ідентифікують різні об'єкти (полісемія).

3. Невизначеність даних сховища консолідованих даних.

Зрозуміло, що через неструктурованість даних, а також через автоматичність завантаження даних частина з них може бути невизначеною. Крім того, системи, які включають багато джерел, можуть містити недостовірні або суперечливі дані. Невизначеність може виникати навіть у тому випадку, коли первинні дані були точними, оскільки для відображення одної характеристики можуть використовуватись різні домени.

Прикладом предметної області, яка яскраво демонструє такий тип невизначеності, є система перевірки достовірності подій. У цьому випадку важливу роль відіграє ступінь довіри до джерела даних.

4. Невизначеність запитів.

Невизначеність запитів виникає у зв'язку з наявністю різних моделей даних та їх виразної потужності, оскільки система сама трансформує запит, отриманий від користувача, наприклад, на основі ключових слів. Під час перетворення цього типу запиту у SQL-запит до структурованого джерела може виникнути невизначеність з результатами запиту.

Невизначеність запитів яскраво демонструють пошукові системи, де за запитом користувачеві надається надто багато результатів пошуку і лише частина з них насправді задовольняє користувача.

ПОДАННЯ НЕВИЗНАЧЕНОСТІ У СХОВИЩАХ ДАНИХ

Прокласифікуємо типи невизначеності за характером їх появи у просторі даних. Однією з перших робіт у цьому напрямі є робота Л. Заде [2]. Г. Цельмер підкреслює, що невизначеність, будучи об'єктивною формою існування оточуючого нас реального світу, обумовлена, з одного боку, об'єктивним існуванням випадковості як форми прояву необхідності, а з іншого – неповнотою кожного акту відображення реальних явищ в людській свідомості. Причому неповнота відображення принципово непереборна через загальний зв'язок всіх об'єктів реального світу і нескінченності їх розвитку. Виражається невизначеність в різноманітті перетворення можливостей у дійсність, в існуванні множини (як правило, нескінченної кількості) станів, в яких об'єкт, що змінюється в динаміці, може перебувати в майбутній момент часу (Цельмер, [3]).

У (Моісеєв, 1975) наводиться така класифікація невизначеностей [5]:

– за ступенем невизначеності: імовірнісна, лінгвістична, інтервальна, повна невизначеність;

– за характером невизначеності: параметрична, структурна, ситуаційна;

– за використанням одержаної в ході керування інформації: переборна і невиправна.

У Дієва В. С. і Трухачева Р.І. [4, 6] наводиться детальніша класифікація невизначеностей в сучасних економічних системах (Диев, 2001; Трухачов, 1981). У [7] визначено типи невизначеностей, природою яких є:

– значення невідоме (відсутнє);

– неповнота інформації;

– нечіткість (стохастичність) – використання розподілу для встановлення істинності знань;

– неточність (стосується числових даних);

– недетермінованість процедур виведення рішень (випадковість);

– ненадійність даних;

– багатозначність інтерпретацій;

– лінгвістична невизначеність: невизначеність значення слова, невизначеність змісту речення.

На рис. 2 подано рівні уведення типів невизначеностей у сховищі даних. Невизначеності на рівні агрегованих даних виникають на основі атак – блокування даних у джерелі, приховуванням частини інформації тощо. Невизначеності на рівні метаданих виникають, в першу чергу, на основі програмних збоїв, а також через наявність атак на рівні джерел даних (змін структур даних джерел).

Розглянемо детальніше вказані типи невизначеностей та виявимо місця їх появи у сховищах даних [8]. Аналізуватимемо невизначеності, що виникають у результаті консолідації даних у єдине джерело (локальне чи віртуальне), а, отже, матимемо справу зі структурованими даними. Для подання єдиного джерела використовуватимемо реляційну модель.

Відсутність даних виникає внаслідок відсутності опису необхідної характеристики у метаданих. Відсутність може виникнути або через те, що необхідної характеристики не знайдено у інформаційних продуктах, що є джерелом для сховища даних, або вона не включена до метаданих через недостатній рівень довіри.

Невідомість даних зустрічається на рівні значення характеристики (атрибуту у реляційних базах даних) і означає, що значення притаманне об'єкту, але невідоме:

$$s=\{\boldsymbol{A}, unk\},$$

де s – об'єкт, який описується кортежем характеристик консолідованих даних, unk – відсутнє значення, $\boldsymbol{A}$ – решту значень атрибутів характеристик кортежу консолідованих даних, $unk \cup \boldsymbol{A} = s, unk \cap \boldsymbol{A} = \varnothing$.

У випадку появи невідомості на рівні метаданих призводять до зашумлення всієї інформації, що отримується від джерела даних з невідомим атрибутом.

Неповнота є станом об'єкту, у якому є підмножина відсутніх значень характеристик. Якщо ця підмножина є порожня і ми говорило про реляційне подання даних, то отримаємо традиційний кортеж. Відсутність інформації є також частковим випадком неповноти інформації, коли кількість невідомих значень атрибутів кортежу дорівнює 1. Неповнота може з'являтися як і у відношенні, у яке інтегруються дані, так і у метаданих як результат збоїв роботи методу визначення структури джерела.

$$s=\{\boldsymbol{A}, \{unk\}\}, |unk|<|\boldsymbol{A}|.$$

Невизначеності типів 3–8 класифікують як неоднозначність даних, що переважно виникають на рівні об'єкта або підмножини значень характеристик, із яких формується кортеж. Вони виникають як результат атак на рівні джерел даних (інформаційних продуктів).

Нечіткість виникає через неповне вивчення або неоднозначне відображення характеристик сутності. Моделюється за допомогою доповнення схеми відношення додатковим атрибутом (атрибутами), значення яких містять рівень впевненості у істинності підмножини значень неключових атрибутів. Також вона подає рівень довіри до характеристики $P^{attr}(i, j)$.

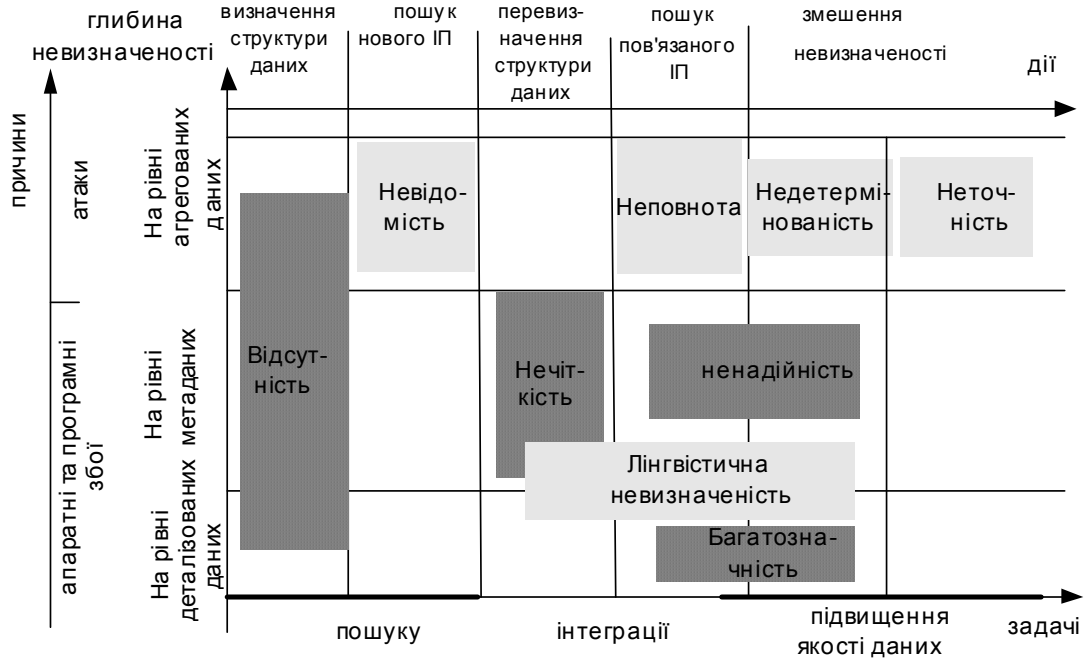

Рис. 2. Типи невизначеності у федеративному сховищі даних та рівні їх введення

$$s=\{\boldsymbol{A}, unk_1, unk_2,\ldots, unk_n\},$$

$$\boldsymbol{A}\in K, \boldsymbol{A}', \boldsymbol{1}\le n\le|\boldsymbol{A}'|,$$

$$unk_{attr}=P^{attr}(i,j),\ |\boldsymbol{A}|\ge\{unk_1, unk_2,\ldots, unk_n\}$$

де K – множина значень ключів, $\boldsymbol{A}'$ – підмножина значень неключових атрибутів. Рівень впевненості може позначатися за допомогою числової шкали, лінгвістичних оцінок, нечіткої величини тощо.

Неточність отримується внаслідок застосування математичних операцій над числовими даними (цього типу є також невизначеність, яка виникає внаслідок роботи з інтервальнимим величинами). Цей тип невизначеності моделюється за допомогою додаткового атрибута і може виникати через нечіткість в метаданих. Виникає доволі часто у зв'язку з опрацюванням даних, що зберігались на різних платформах, використовувались для вирішення різного класу задач.

$$s=\{\boldsymbol{A},\{unk\}\},\ \{unk\}\subset\boldsymbol{A},\ Design(\boldsymbol{A})\in\{unk\}.$$

Недетермінованість процедур виведення рішень (випадковість) виникає у тому випадку, коли необхідно зберігати проміжні або кінцеві результати процедур виведення або прийняття рішень, а також – у відношенні фактів на рівні значень агрегованих атрибутів. Моделюється за допомогою розширення схеми даних та виникає винятково у агрегованих даних (гіперкубі):

$$s=s\cup\{unk\},\ \{unk\}\notin\boldsymbol{A},\ Design(s)\in\{unk\}.$$

Ненадійність є типом невизначеності, який вважається однією із характеристик об'єкта. Хоча сама природа цієї характеристики є невизначеною, у відношенні як її домен використовують традиційну числову шкалу та застосовують до її значень традиційні математичні операції. Виникає внаслідок визначення довіри звернення до джерела даних $P(j)$. Моделюється за допомогою доповнення схеми каталогу даних додатковим атрибутом. Значення цього атрибута змінюється у результаті роботи простору даних. Представляється як характеристика, обернена до значення довіри до джерела даних.

$$s=s\cup[unk_j],\ unk_j\notin\boldsymbol{A},\ unk_j=\frac{1}{P(j)}.$$

Багатозначність інтерпретації є одним із джерел виникнення суперечностей. Такий тип невизначеності виникає найчастіше на рівні детальних даних через отримання інформації із різних джерел і неможливість визначення істинності даних. Для відображення цього типу невизначеності схему відношення доповнюють додатковим атрибутом, який містить ступінь впевненості у істинності даних кортежу. Від типу нечіткість відрізняється тим, що вводиться на рівні відношення.

Лінгвістична невизначеність пов'язана з використанням природної мови в інформаційних ресурсах (у текстових файлах та веб-ресурсах), які мають якісний характер, і може виникати внаслідок нерозуміння (незнання) значення слова або нерозуміння змісту речення. Такий тип невизначеності зустрічається у системах опрацювання текстової інформації (системи автоматизованого перекладу, системи для самонавчання тощо). У контексті сховищ даних виникає внаслідок опрацювання напівструктурованої інформації (тексти, веб-сторінки тощо).

Розглянуті типи невизначеностей можуть накладатись або бути джерелом появи одна одної.

МОДЕЛЬ СХОВИЩА ДАНИХ З НЕВИЗНАЧЕНІСТЮ

Схема сховища даних з невизначеністю Cg' – скінченна множина імен атрибутів $\{A_1, A_2, ..., A_n\}$, значення яких є чіткими; $\{A\_unk_1, A\_unk_2, A\_unk_p\}$ з нечіткими або недермінованими значеннями; множина імен атрибутів $\{Unk_1, Unk_2, \ldots, Unk_m\}$, доменами яких є числові дані, що моделюють імовірнісні дані, значення функції приналежності нечітких множин, ступінь істинності багатозначної логіки, процентні відношення, коефіцієнти, різноманітні шкали або лінгвістичні оцінки; схеми словника синонімів *Dic* та схему метаданих *Cg*:

$$Cg'=<\{C_1, C_2,...,C_n\}, \{C\_unk_1, C\_unk_2, C\_unk_p\},\ \{Unk_1, Unk_2, \ldots, Unk_m\}, Dic, Cg>.$$

Невизначеними вважаються значення атрибутів множини ***C_unk***, а рівень довіри до них зберігається у значеннях атрибутів множини ***Unk.***

Для відображення зв'язків між атрибутами множин ***C_unk*** та ***Unk*** використано бінарне відношення **Meta,** значення якого визначаються на основі вибірки представлення джерела *i* в каталозі даних *Cg*:

$$Meta=\left|meta_{ij}\cdot\sigma_{\arg(i)}(Cg)\right|,\ \forall i=\overline{1,p},\forall j=\overline{1,m},$$

$$meta_{ij}=\begin{cases}1, Unk_j\Leftrightarrow C\_unk_i\wedge\sigma_{\arg(j)}(Dic)\\ 0, \text{в іншому випадку}\end{cases}.$$

Сума по рядках бінарного відношення рівна 1, оскільки вважатимемо, що ступінь довіри до атрибута не вказуватиметься двома і більше атрибутами із множини **Unk**:

$$\forall i=\overline{1,p},\sum_1^n meta_{ij}=1.$$

Введення відношення **Meta** дозволить моделювати будь-які типи невизначеностей, не розширюючи доменів атрибутів.

Кортеж консолідованих даних dc – інформаційний опис об'єкта *t* джерела даних *S*, поданий у вигляді множини (кортежу) значень характеристик (атрибутів), підмножина значень атрибутів якого містить дані про об'єкт, джерело даних та синонімічні назви об'єкта, при-

чому ці дані можуть бути неповні, нечіткі чи недетерміновані. Тобто, об'єкт, який моделюється у джерелі даних цим кортежем, існує, але частина інформації про нього відсутня, нечітка, неповна, недетермінована тощо.

Наведемо приклади кортежу консолідованих даних для різних типів інформаційних ресурсів.

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

$$dc = t_{rel} \cup Unk,$$

$$t_{rel} = \{c_1,...,c_n\} \cup \{c\_unk_1,...,c\_unk_m\},$$

де $\{c_1,...,c_n\}$ – значення чітких атрибутів, $\{c\_unk_1,...,c\_unk_m\}$ – значення атрибутів з невизначеністю.

2. Сховище даних – поєднує дані з відношень фактів та вимірів. Множину значень вимірів та характеристик фактів подано як кортеж t_{dw}:

$$dc = t_{dw} \cup Unk,$$

$$t_{dw} = \{c_{11},...,c_{1n}\} \cup ... \cup \{c_{k1},...,c_{kn}\} \cup \{c_{rf_1},...,c_{rf_l}\} \cup$$
$$\cup \{c\_unk_{11},...,c\_unk_{1m}\} \cup ... \cup \{c\_unk_{k1},...,c\_unk_{ks}\} \cup$$
$$\cup \{c\_unk_{rf_1},...,c\_unk_{rf_t}\},$$

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 атрибутів з невизначеністю i-го виміру, $c\_unk_{rf_j}$ – значення j-ї характеристики з невизначеністю відношення фактів.

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

$$dc = t_{text} \cup Unk,$$

$$t_{text} = \{c_1,...,c_n\} \cup \{c\_unk_1,...,c\_unk_m\}.$$

Значення атрибутів кортежу консолідованих даних поділимо на групи.

1.Чіткі (відомі) – значення первинного ключа, зовнішніх ключів (можуть бути відсутні). Позначимо їх через C.

2. Відсутні – фізично відсутня інформація. Позначимо їх через $\perp$.

3. Невизначені – для підмножин атрибутів введена множина атрибутів *Unk*, які вказують ступінь істинності значень цих атрибутів. За замовчуванням значенню атрибута *Unk* присвоюємо значення, яке означає найвищий ступінь істинності. Крайніми випадками введення невизначеності є:

– додавання атрибутів типу *Unk* до усіх атрибутів, крім чітких;

– додавання атрибута *Unk* до усіх значень кортежу.

Зауважимо, що, у випадку стовідсоткової довіри до кожного значення кортежу, ми отримуємо традиційний реляційний кортеж та застосовуємо традиційні операції над ним.

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

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

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

Сховище даних з невизначеністю cg' – множина відношень зі схемою Cg' та множиною кортежів консолідованих даних $\boldsymbol{dc}$.

РОЗРОБЛЕННЯ ОПЕРАЦІЙ НАД КОНСОЛІДОВАНИМИ ДАНИМИ МОДЕЛІ СХОВИЩА

Оскільки сховище консолідованих даних є розширенням сховища даних, побудованого на основі реляційної моделі, то далі удосконалимо операції для роботи з ним.

Для опрацювання та аналізу невизначеностей за допомогою запиту в реляційних операторах слід здійснювати селекцію кортежів за значеннями множини атрибутів $\boldsymbol{Unk}$. У сховищі даних аналогічною до неї є *операція зрізу*. Нехай r та s – відношення зі схемою R, r' та s' – відношення зі схемою $R \cup Unk \cup Dic \cup Cg$. Тоді $r \cap s$, $r \cup s$ і $r - s$ є відношеннями зі схемою R, а $r' \cap s'$, $r' \cup s'$ і $r' - s'$ – відношеннями зі схемою $R \cup Unk \cup Dic \cup Cg$.

Враховуючи ймовірність атак (невизначеність типу «багатозначність»), вибираємо ті джерела даних, рівень довіри до яких вищий за аналогічні:

$$r' = r \cup \sigma_{\max(P(\pi(Cg)))}(Dic) \cup Cg.$$

Доповнення до відношення r' працюватиме коректно у разі присвоєння всім значенням атрибута *Unk* найнижчого ступеня довіри (апріорі вважається, яка ця інформація, що заноситься у відношення є правдивою та повною, а про решту інформації нам нічого невідомо). Обрання такого методу подання ступеня істинності за замовчуванням здійснено, виходячи з принципу замкненості.

Оператор зрізу передбачає аналіз нечіткого значення за множиною значень атрибутів *Unk*.

$$slice: \sigma^{cons\left(\begin{matrix}Unk\Theta unk)\cup(C\_unk\Theta c\_unk)\cup \\ \cup\sigma_C(Dic)\cup\sigma_{\bar{C}}(Cg)\end{matrix}\right)}(cg') =$$

$$= \left\{ \begin{matrix} t \in dc | t(Unk)\Theta unk, t(C\_unk)\Theta c\_unk, meta_{Unk,C\_unk} = 1, \\ \sigma_C(Dic) \text{ Is Not NULL}, \sigma_C(Cg) \text{ Is Not NULL}, unk = P(cg') \end{matrix} \right\},$$

где Θ – множина символів (знаків) бінарних відношень над парами значень доменів. Вважається, що до кожного

атрибута *C_unk* застосовуються операції порівняння. Як правило, будуть вживатися лише такі знаки порівняння над одним доменом: $=, \neq, <, \leq, \geq, >$.

Твердження: Удосконалений оператор зрізу, як і оператор вибірки, зберігає властивості комутативності та дистрибутивності відносно булевих операцій.

Доведення

Нехай $r'(R')$ – відношення, $R' \leftarrow R \cup Unk \cup Dic \cup Cg$, A і B – атрибути в R', і нехай $a \in dom(A)$, $b \in dom(B)$. Тоді має місце рівність: $\sigma^{cons}{}_{A=a}(\sigma_{B=b}(r')) = \sigma_{cons\,B=b}(\sigma_{A=a}(r'))$.

Удосконалений оператор зрізу дистрибутивний відносно бінарних булевих операцій:

$$\sigma^{cons}{}_{A=a}(r' \gamma s') = \sigma^{cons}{}_{A=a}(r') \gamma \sigma_{A=a}(s'),$$

де $\gamma = \cap,\ \cup$ або $-$, а r' і s' – відношення над однією і тією ж схемою.

Аналогом операції згортання у сховищі даних, побудованому на основі реляційній моделі, є *операція проекції*. Здійснюючи проекцію відношення з кортежами консолідованих даних, слід відслідковувати зв'язок підмножини атрибутів *Unk* із підмножиною атрибутів *C_unk*, а також перевіряти, чи для назви атрибута *C_unk* є синонім у словнику синонімів *Dic*. Тому удосконалений оператор згортання подано так:

$$\text{drill-down}: \pi_X{}^{cons}(cg') = \text{IIF}\left(\begin{array}{l} \neg\,\text{ISNULL}\left(\sigma_{Cg=R\cup C\_unk=X}(c\_unk)\right); \pi_{X\cup\pi_{Unk}(\sigma_{Cg=meta(C\_unk,Unk)=1}(c\_unk))}(dc); \\ \text{IIF}\left(\sigma_{C\cup C\_Unk=X}(Dic); \pi_{\sigma_{C\cup C\_Unk=X}(Dic)}(r); \pi_X(dc)\right) \end{array}\right),$$

де IIF(*умова*; *дія 1*; *дія 2*) – оператор умови. У разі виконання *умови* виконується *дія 1*, інакше *дія 2*; ISNULL(r) – логічний оператор, результатом якого є *істина*, якщо відношення-операнд r не містить кортежів, та хиба – у іншому випадку. Також здійснюється пошук синоніма атрибута у словнику синонімів *Dic* ($\sigma_{C\cup C\_Unk=X}(Dic)$) та заміна за потреби ($\pi_{\sigma_{C\cup C\_Unk=X}(Dic)}(r)$).

Твердження: Удосконалений оператор згортання зберігає властивості традиційного оператора проекції.

Доведення: Якщо $X_1 \subseteq X_2 \subseteq .... \subseteq X_m \subseteq R'$, то

$$\pi_{X_1}{}^{cons}\left(\pi_{X_2}{}^{cons}\left(...\left(\pi_{X_m}{}^{cons}(cg')\right)..\right)\right) = \pi_{X_1}{}^{cons}(cg').$$

Оператор з'єднання використовується для зв'язування відношення фактів та відношень вимірів у сховищі консолідованих даних, оскільки воно будується на основі реляційної моделі.

Традиційний оператор з'єднання не може використовуватися для сховищ та просторів даних з консолідованими даними, оскільки для статистичного аналізу необхідне з'єднання відношення фактів з відношеннями вимірів, а за наявності непорожньої підмножини атрибутів ***Unk*** у відношеннях фактів та вимірів таке з'єднання буде некоректним. Також на оператор з'єднання впливає той факт, що виникає необхідність з'єднання не лише за тими атрибутами, що вказані як вхідні параметри, але й перевіряти наявність синонімів у словнику синонімів *Dic*.

Для удосконалення оператора з'єднання слід розглянути випадки, коли відношення є повністю з'єднувальними або не повністю з'єднувальними. Для повністю з'єднувальних відношень уведення множини атрибутів ***Unk*** не впливає на операцію з'єднання. Якщо значення множини атрибутів ***Unk*** містять міру невизначеності зовнішнього ключа відношення, з яким відбувається з'єднання, то ця міра невизначеності переноситься на всі решту значень атрибутів цього відношення. У випадку неповної з'єднувальності значення атрибута *Unk* для кортежів підлеглої таблиці, які не потрапляють у відношення, будуть вважатися рівними найвищому ступеня довіри.

$$across: \underset{cons}{r \triangleright\!\!\triangleleft cg'} = \text{IIF}\Big(\sigma_{C\cup C\_Unk=X}(Dic);$$

$$\pi_{\sigma_{C\cup C\_Unk=X}(Dic)}(r \underset{\bullet}{\triangleright\!\!\triangleleft} cg'); \pi_{(R,B,NVL(Unk,\min))}(r \underset{\bullet}{\triangleright\!\!\triangleleft} cg')\Big),$$

де r – традиційне відношення, cg' – відношення з консолідованими даними, R – множина атрибутів відношення r, S – множина атрибутів відношення cg', не включаючи підмножини атрибутів ***Unk*** ($Cg' = Cg \cup Unk$), B – множина тих атрибутів з S, яких нема у відношенні r ($B \subset Cg$, $B \not\subset Cg \cap R$), min – значення, яке означає найнижчий ступінь довіри, $NVL(Unk, \min)$ – операція, яка присвоює min усім значенням *Unk* для нез'єднувальних кортежів відношення cg', $\underset{\bullet}{\triangleright\!\!\triangleleft}$ – ліве з'єднання.

Спочатку перевіряється, чи необхідно здійснювати з'єднання не за заданими атрибутами, а за синонімами ($\sigma_{C\cup C\_Unk=X}(Dic)$). Якщо ні, то виконується операція лівого з'єднання для відношень з схемами S' і R та проекція за атрибутами-синонімами. У іншому випадку виконується операція лівого з'єднання за спільними атрибутами, потім над отриманим з попередньої операції відношенням здійснюється операція проекції, за якою утвореним у результаті з'єднання порожнім значенням підмножини атрибутів ***Unk*** присвоюється значення min.

Слід зазначити, що коли словник синонімів порожній ($Dic = \varnothing$) і ймовірність звернення до джерел даних загалом та до їх характеристик рівні одиниці ($Unk = 1$), то отримаємо традиційне реляційне з'єднання.

Твердження: Удосконалений оператор з'єднання комутативний та асоціативний.

Доведення

Для даних відношень q', r' і s'

$$\left(q' \triangleright\!\triangleleft r'\right) \triangleright\!\triangleleft s' = q' \triangleright\!\triangleleft \left(r' \triangleright\!\triangleleft s'\right).$$

Введемо позначення для деяких багаторазових з'єднань. Нехай $s_1'\left(S_1'\right), s_2'\left(S_2'\right), ..., s_m'\left(S_m'\right)$ – відношення, $R' = S_1' \cup S_2' \cup ... \cup S_m'$ і S' – послідовність $S_1', S_2', ..., S_m'$. Далі, нехай $t_1, t_2, ..., t_m$ – послідовність кортежів, в якій $t_i \in s_i', 1 \le i \le m$. Кортежі з'єднувальні на S', якщо існує кортеж $t \in R'$, такий, що $t_i = t\left(S_i'\right), 1 \le i \le m$. Кортеж t є результатом з'єднання кортежів $t_1, t_2, ..., t_m \in S'$.

АГРЕГАЦІЯ РОЗРІДЖЕНОГО ГІПЕРКУБА ДАНИХ

Практика розроблення і впровадження реляційних систем збирання даних показала, що через різні причини збір первинних даних здійснюється лише частково, а тому не завжди може бути оптимальним для використання. Це приводить до необхідності застосування багатовимірних баз даних з частковою або слабкою заповненістю. При цьому створені багатовимірні куби даних мають низьку щільність заповнення даними, а тому є *розрідженими*. Тому виникають такі проблеми:

– низька ефективність пошуку і витягання інформації з розрідженого гіперкуба даних;

– некоректність використання отриманих при агрегації значень у розріджених гіперкубах даних.

Разом з тим, розріджені гіперкуби даних містять потенційно цінну інформацію, ефективне використання якої може відіграти значну роль при прийнятті рішення.

Отже, основними проблемами, які виникають у задачах аналізу розрідження гіперкуба є зниження якості рішень та погіршення агрегації розріджених гіперкубів даних.

У більшості випадків при створенні інформаційних систем, орієнтованих на аналіз даних, питання подання інформації в розріджених гіперкубах даних обходяться стороною. Та методи роботи зі щільними і розрідженими гіперкубами даних повинні істотно різнитися. Тому, розроблення альтернативних методів пошуку і агрегації даних, що дозволяють вирішити вищезгадані проблеми, є актуальним завданням.

Покажемо, яким чином вплине відсутність показника нижнього рівня на формування агрегату.

Перш за все наведемо методи обчислення агрегату. Існує декілька традиційних методів агрегації (табл. 1).

Вибір того або іншого методу агрегації даних залежить від конкретного вирішуваного завдання. Технологічно процедура підрахунку агрегатів виконується з використанням т.з. карт агрегації, що включають стандартні

Таблиця 1. Стандартні методи агрегації

SUM	Додавання деталізованих даних	$P = \sum_{i=1}^{N} x_i$
WSUM	Зважена сума	$P = \sum_{i=1}^{N} p_i x_i$
MIN (MAX)	Мінімальне (максимальне) значення	$P = \min_{l \in N}(x_i)$
AVERAGE	Середнє значення	$P = \frac{\sum_{i=1}^{N} X_i}{N}$
WAVERAGE	Зважене середнє	$P = \frac{\sum_{i=1}^{N} p_i x_i}{\sum_{i=1}^{N} p_i}$

методи агрегації, вказані в табл. 1. У багатьох популярних OLAP-системах як метод агрегації «за замовчуванням» використовується метод додавання, що передбачає наявність первинних даних на нижньому рівні ієрархії. Проте, виникає питання про застосовність цих методів при агрегації даних в розріджених гіперкубах.

При вирішенні серйозних аналітичних завдань аналітику важливо знати не тільки значення показника, але і те, наскільки він може довіряти набутому значенню. Обчислення агрегата за методом середнього значення за наявності первинних даних за усіма значеннями нижнього рівня в ієрархії дає 100 %-ву достовірність, оскільки немає причин вважати, що це середнє значення могло бути чим-небудь спотворене.

Очевидно, що в стандартних методах агрегації не враховується ситуація невизначеності первинних даних, що відповідають деяким міткам нижнього рівня ієрархічного виміру. Але ж саме таку ситуацію являє собою агрегація даних у розрідженому гіперкубі.

Під час виконання агрегації в розрідженому гіперкубі за методом обчислення середнього необхідне введення додаткового параметра, що характеризує рівень вірогідності отриманого результату. Технологічно ця операція може здійснюватися шляхом створення додаткової карти агрегації, що включає розрахунок рівня вірогідності для кожного, отриманого в ході агрегації, значення.

Обчислення агрегату на першому рівні ієрархії (l=1) здійснюється за формулою:

$$Ag_j^1 = \frac{\sum_{i=1}^{V_{jj}} ag_i^0}{V_j},$$

де V_j – кількість фактів, які відповідають атрибутам, що є дочірніми по відношенню до атрибута j.

Узагальнюючи, одержимо формули обчислення агрегатів на решті рівнів ієрархії:

$$Ag_j^l = \frac{\sum_{i=1}^{V_{jj}} ag_i^{l-1}}{V_j}, l = 1,...,N \cdot$$

Розглянутий метод може бути застосований при побудові карт агрегації в розріджених гіперкубах даних і дає можливість оцінити рівень достовірності одержаних результатів на етапі аналізу.

Опишемо пропонований метод формально. Нехай ми маємо ієрархічний вимір з N рівнями. Первинні дані відповідають нижньому рівню ієрархії (l=0). Поставимо у відповідність кожній i-ій мітці нижнього рівня ієрархії величину, що характеризує міру достовірності фактів так (рис. 3): $t_i^0 = 1$ у випадку, якщо існує факт, що відповідає цій мітці, і $t_i^0 = 0$, якщо такого факту не існує.

Обчислення агрегату на певному рівні ієрархії (l=1) здійснюється за формулою:

$$X_j^1 = \frac{\sum_{i=1}^{M_j} x_i^0}{M_j},$$

де M_j – кількість фактів, які відповідають міткам, що є дочірніми по відношенню до мітки j.

Обчислення рівня достовірності відповідного агрегату здійснюється за формулою:

$$T_j^1 = \frac{\sum_{i=1}^{K_j} t_i^0}{K_j},$$

де T_j – кількість міток, що є дочірніми по відношенню до мітки j.

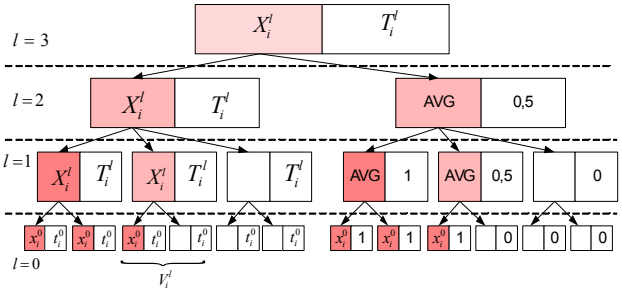

Рис. 3. Агрегація розрідженого гіперкуба даних

Узагальнюючи, одержимо формули обчислення агрегатів на решті рівнів ієрархії:

$$X_j^l = \frac{\sum_{i=1}^{M_j} x_i^{l-1}}{M_j}, l = 1,...,N \cdot$$

Розглянутий метод може застосовуватися для побудови карт агрегації в розріджених гіперкубах даних і дає можливість оцінки рівня достовірності одержаних результатів на етапі аналізу.

ВИСНОВКИ

Визначено поняття терміну Big Data і проаналізовано причини їх появи. Також визначено одну з проблем Big Data – появу невизначеності.

Побудовано модель сховища консолідованих даних, яка є розширенням моделі відношення з невизначеністю.

Удосконалено операції над відношенням з невизначеністю з метою їх застосування до сховища консолідованих даних.

Побудовано процедуру попереднього формування агрегату з врахуванням невизначеності.

СПИСОК ЛІТЕРАТУРИ

1. *Шаховська, Н. Б.* Аналіз методів опрацювання показників соціо-еколого-економічного розвитку регіону / Н. Б. Шаховська, Ю. Я. Болюбаш // Східно-європейський журнал передових технологій. – 2013. – Том 5, № 2(65). – С. 4–8.
2. *Заде, Л.* Понятие лингвистической переменной и его применение к принятию приближенных решений / Л. Заде. – М. : Мир, 1976. – 166 с.
3. *Цельмер, Г.* Учет риска при принятии управленческих решений/ Г. Цельмер // Проблемы МСНТИ. – 1980. – № 3. – С. 94–105.
4. *Найт, Ф. Х.* Риск, неопределенность и прибыль / Ф. Х. Найт. – М. : Дело, 2003. – 358 с.
5. *Моисеев, Н. Н.* Элементы теории оптимальных систем / Н. Н. Моисеев. – М. : Наука, 1975. – 528 с.
6. *Трухачев, Р. И.* Модели принятия решений в условиях неопределенности / Р. И. Трухачев. – М. : Наука, 1981. – 151 с.
7. *Згуровський, М. З.* Основи системного аналізу / М. З. Згуровський, Н. Д. Панкратова. – К. : Видавнича група BHV, 2007. – 544 с.
8. *Шаховська, Н. Б.* Моделювання невизначеностей у сховищах даних реляційного типу. – Львів, 2007. – автореферат дис. на здобуття канд. техн. наук

Стаття надійшла до редакції 27.12.2013.

Шаховска Н. Б.[1], Болюбаш Ю. Я.[2]
[1]Д-р техн. наук, профессор, Национальный университет «Львивська политехника», Украина
[2]Соискатель, Национальный университет «Львивська политехника», Украина

ОБРАБОТКИ НЕОПРЕДЕЛЕННОСТИ В БОЛЬШИХ ДАННЫХ

В статье введено понятие «Большие данные» и проанализированы причины их появления. Показано уровни возникновения неопределенности в «Больших данных». Сформирована модель хранилища данных с неопределенностью и разработаны операции над ним. Приведен метод формирования агрегата с учетом неопределенности.

Ключевые слова: большие данные, информационный продукт, неопределенность.

Shakhovska N.[1], Bolubash Yu.[2]
[1]Doctor of Science, Professor, National University «Lviv Polytechnic», Ukraine
[2]Candidate for degree, National University «Lviv Polytechnic», Ukraine

INDECISION PROCESSING IN BIG DATA

This paper introduced the concept of the term Big Data and analyzes the cause of their appearance. Thus, the specificity of Big data (the presence of diverse set of sources, data doubling, ambiguity describing data sources) leads to the fact that the indeterminacy in traditional relational databases considered within a relationship and could occur at the level of attribute and tuple-level attitude in this case extends through the perception of the user information on the entire data space. Therefore, for processing indeterminacy in the Big data must use a different approach, the need for the use of which has not had in relational databases and data warehouses. There is the level of uncertainty in the Big Data show. There are formed data warehouse model with uncertainty and developed operations on it. There is posted forming unit method, taking into account uncertainty.

Keywords: big data, information product, uncertainness.

REFERENCES

1. Shakhovska N. B., Bolubash Yu. Ja. Analis metodiv opratsuvannia pokaznykiv sotsio-ekologo-ekonomichnogo rozvytku regionu, *Shidno-yevropeyskij zhurnal peredovyh tehnologij*, 2013,Vol. 5, No. 2(65), pp. 4–8
2. Zade L. Ponyatie lingvisticheskoj peremennoj i ego primenenie k priniatiyu reshenij. Moscow, Mir, 1976, 166 p.
3. Tselmer G. Utchet riska pri priniatii upravlencheskih reshenij, *Problemu MSNTI*, 1980, No. 3, pp. 94–105.
4. Nait Ph., Risk H., Neopredelennost i pribyl. Moscow, Delo, 2003, 358 p.
5. Moiseev N. N. Elementy teorii optialnyh sistem. Moscow, Nauka, 1975, 528 p.
6. Trukhachov R. I. Modeli priniatija reshenij v uslovijah neopredelennosti. Moscow, Nauka, 1981, 151 p.
7. Zgurovskyj M. Z., Pankratova N. D. Osnivy systemnogo analizu. Kiev, BHV, 2007, 544 p.
8. Shakhovska N. B. Modeluvannja nevyznachenostej u chovyshhah danyh reliatsijnogo typu. Lviv, 2007. avtoreferat dys. Na zdobuttia kand. tehn. nauk