УДК 51.001.57+004.652.4+004.827

Н. Шаховська
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

МОДЕЛЬ КОНСОЛІДОВАНИХ ДАНИХ ТА ЇХ ОПРАЦЮВАННЯ ЗА УМОВ НЕВИЗНАЧЕНОСТІ

© Шаховська Н., 2013

Проаналізовано проблеми опрацювання розрізнених даних. Побудовано формальну модель сховища консолідованих даних та розроблено метод зменшення невизначеності даних.

Ключові слова: невизначеність даних, модель сховища, консолідовані дані.

Problems which arise up during work with separate sources with depositories information using and databases are analyzed. There is formalized model of consolidation datawarehouse and there is built the methods for uncertainty elimination .

Key words: uncertainty data, warehouse model, consolidated data.

Вступ

Інформаційні об’єкти, що описують певну предметну область, консолідовані дані та зв’язки між об’єктами, становитимуть простір даних. Однією зі задач, яка виникатиме у процесі консолідації, є невизначеність даних, що є результатом дублювання, неточності, відсутності, суперечливості даних. Також невизначеність виникатиме внаслідок встановлення неправильних зв’язків між об’єктами. Тому виникає задача зменшення невизначеності для підвищення якості даних.

Оскільки дані надходять з різних джерел, 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 атрибута, кортежа та відношення (невизначеність у схемі опису). Поява невизначеності на рівні атрибута і кортежа у зв’язку з багатовимірністю відображення інформації призводить до поширення невизначеності на всі примірники опису певного концепту. Оскільки простір даних об’єднує мільйони даних про об’єкти проблемної області, то опрацювання невизначеності традиційними засобами (інтервальна математика, багатозначна логіка) стає неефективним через велику кількість операндів.

Отже, специфіка простору даних (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 весь простір даних. Тому для опрацювання невизначеності у просторі даних необхідно використати якісно новий підхід, потреба застосування якого не виникала у реляційних базах даних та сховищах даних.

Аналіз літературних джерел і постановка задачі

Прокласифікуємо типи невизначеності за характером їх появи у просторі даних. Однією з перших робіт у цьому напрямі є робота Л. Заде [1]. Г. Цельмер [2] підкреслює, що невизначеність, будучи об'єктивною формою існування реального світу, обумовлена, з одного боку, об'єктивним існуванням випадковості як форми прояву необхідності, а з іншого – неповнотою кожного акту відображення реальних явищ у людській свідомості. Причому неповнота відображення принципово

непереборна через загальний зв'язок всіх об'єктів реального світу і нескінченність їх розвитку. Полягає невизначеність у різноманітті перетворення можливостей на дійсність, в існуванні множини (як правило, нескінченної кількості) станів, в яких об'єкт, що змінюється в динаміці, може перебувати в майбутній момент часу.

Ф. Найт під ситуацією невизначеності розуміє недостатню обізнаність та необхідність діяти, спираючись на думку, а не на знання [3]. В.В. Черкасов трактує невизначеність як постійну мінливість умов, швидку й гнучку переорієнтацію виробництва, дії конкурентів, зміну ринку і т.п. Невизначеність він називає найхарактернішою причиною ризику в управлінській діяльності. Він виділяє два класи невизначеності: "доброякісна" невизначеність, коли для невідомих чинників є статистичні або імовірнісні характеристики, і "погана" невизначеність, коли таких характеристик в принципі не можна отримати, причому існують методи означення обох видів невизначеності, що виникають в реальних завданнях (Вентцель, 1980).

У (Моісеєв, 1975) наводиться така класифікація невизначеностей [4]:

– 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. [5] наведено детальнішу класифікацію невизначеностей у сучасних економічних системах.

У [6] визначено типи невизначеностей, природою яких є:

1) значення невідоме (відсутнє);
2) неповнота інформації;
3) нечіткість (стохастичність) – використання розподілу для встановлення істинності знань;
4) неточність (стосується числових даних);
5) недетермінованість процедур виведення рішень (випадковість);
6) ненадійність даних;
7) багатозначність інтерпретацій;
8) лінгвістична невизначеність: невизначеність значення слова, невизначеність змісту речення.

Для сховища даних вважалося, що невизначеність може виникати на рівні значення та на рівні кортежу [7]. Оскільки для простору даних притаманним є встановлення факту довіри до джерела, то рівні уведення невизначеності необхідно розширити.

Змінимо рівні уведення типів невизначеностей у просторі даних (рис. 1).

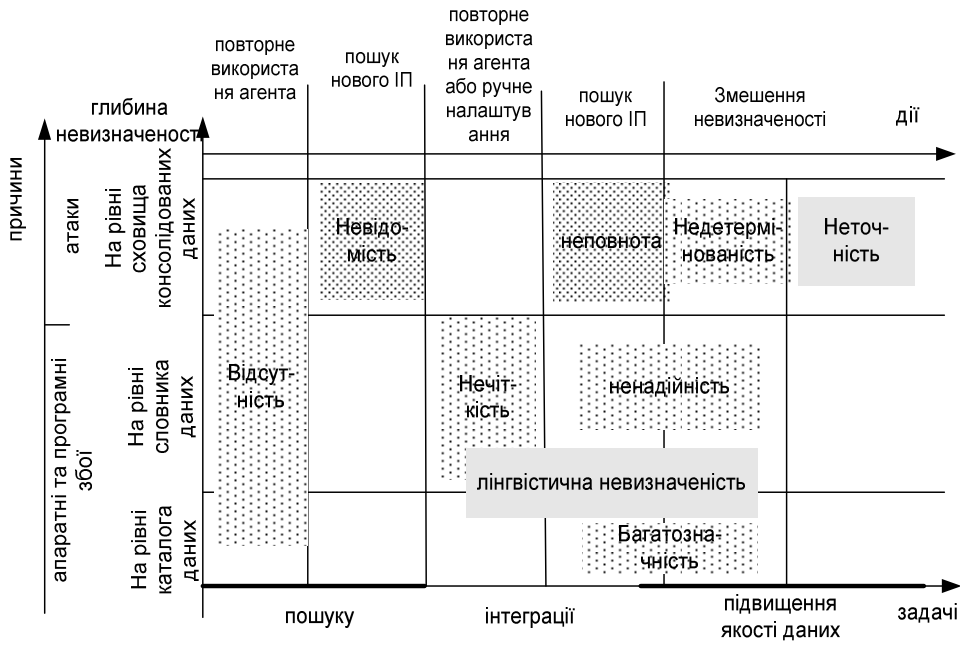

Рис. 1. Типи невизначеності у сховищі консолідованих даних простору даних та рівні їх введення

Невизначеності на рівні сховища даних виникають на основі атак – блокування даних у джерелі, приховуванням частини інформації тощо.

Невизначеності на рівні словника та каталогу даних виникають насамперед на основі програмних збоїв, а також через наявність атак на рівні джерел даних.

Розглянемо детальніше вказані типи невизначеностей та виявимо місця їх появи у сховищах та просторах даних [7]. 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 каталозі даних та словнику. Відсутність може виникнути або через те, що необхідної характеристики не знайдено у інформаційних продуктах простору даних, або вона не включена до каталогу чи словника через недостатній рівень довіри. Для усунення цього виду невизначеності необхідно повторне використання агента, можливо зі зменшеним рівнем довіри до даних.

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, $unk \cup \boldsymbol{A} = s, unk \cap \boldsymbol{A} = \varnothing$.

Подання цього типу невизначеності є ідентичним до подання у сховищах даних. Поява невідомості на рівні каталогу даних призводить до зашумлення всієї інформації, що отримується від джерела даних з невідомим атрибутом.

Неповнота є станом об'єкта, у якому є підмножина відсутніх значень характеристик. Якщо ця підмножина є порожня і ми говоримо 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 словнику даних як результат збоїв роботи інтелектуального агента визначення структури джерела:

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

Моделюється так само, як і в сховищах даних, але, на відміну від сховища даних, виникає і на рівні відношення (каталогу даних).

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 в істинності підмножини значень неключових атрибутів.

$$s=\{\boldsymbol{A}, unk_1, unk_2,\ldots, unk_n\},\ \boldsymbol{A}\in K, \boldsymbol{A}',\ \boldsymbol{1}\le n\le|\boldsymbol{A}'|,\ unk_{attr}=P^{attr}(i,j),\ |\boldsymbol{A}|\ge\{unk_1, unk_2,\ldots, unk_n\}$$

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 інтервальними величинами). Цей тип невизначеності моделюється за допомогою додаткового атрибута і може виникати через нечіткість в словнику даних.

На відміну від сховищ даних, у просторах даних 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 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 сховищі консолідованих даних:

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

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. 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 у каталозі 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 просторів даних виникає внаслідок опрацювання напівструктурованої інформації (тексти, веб-сторінки тощо).

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

Для задачі *зменшення невизначеностей* удосконалено метод, який використано для зменшення невизначеності у сховищах даних реляційного типу – усунення невизначеності на основі методу видобування знань. Невідоме значення атрибута розглядається як позначка класу, а сама задача усунення невизначеності трансформується у задачу віднесення до класу. Використання цього методу дає змогу усувати невизначеності типу "невідомий" та "неповний" на рівні значення атрибута та підмножини атрибутів. Проте, на відміну від сховищ даних, необхідно ще враховувати рівень довіри до джерела даних, тобто працювати з невизначеністю на рівні відношення.

Задача усунення (зменшення) невизначеностей – це побудова гомоморфного відображення множини даних, що зберігаються у сховищі консолідованих даних, у множину даних, що використовуються для підтримання прийняття рішень (рис. 2) з метою підвищення якості консолідованих даних та прийняття на їх основі ефективних керівних стратегічних рішень, враховуючи ймовірність появи атак. Атака – додавання до простору даних джерела, структура даних яких викликає багатозначність інтерпретацій у словнику синонімів *Dic*.

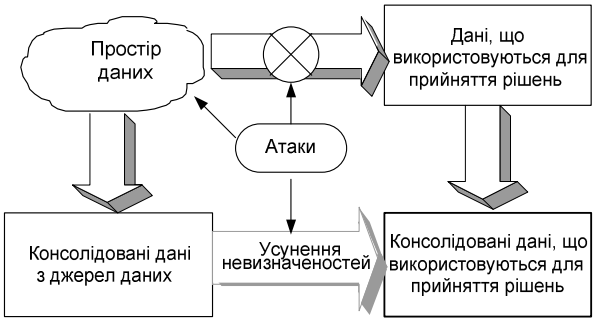

Рис. 2. Постановка задачі усунення невизначеності у сховищі консолідованих даних простору даних

Одним із методів моделювання неточних, нечітких та часткових даних є введення у каталог джерела додаткового атрибута, значення якого визначає ступінь довіри до невизначених даних.

Основний матеріал

1. Модель сховища консолідованих даних

Схема сховища консолідованих даних 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; множину 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; схему словника синонімів *Dic* та схему каталогу даних *Cg*:

$$Cg' = <\{C_1, C_2, ..., C_n\}, \{C\_unk_1, C\_unk_2, C\_unk_p\}, \{Unk_1, Unk_2, \ldots, Unk_m\}, Dic, Cg>, \quad (3.1)$$

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

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

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.

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:

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

де $\boldsymbol{C}$ – підмножина значень атрибутів із чіткими значеннями, $\boldsymbol{C} = t_{rel} \cup t_{dw} \cup t_{text}$, ***C_unk*** – підмножина значень атрибутів із нечіткими та недетермінованими значеннями, ***Unk*** – підмножина значень атрибутів із ступенями істинності значень атрибутів ***C_unk***, $\{dic\}$ – множина значень словника даних, $\{cg\}$ – множина значень каталогу даних.

Сховище консолідованих даних cg' – множина відношень зі схемою Cg' та множиною кортежів консолідованих даних ***dc***

Модель сховища консолідованих даних містить дані з усіх типів джерел простору даних.

2. 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 ***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 не відомо). Такий метод подання ступеня істинності за замовчуванням обираємо за принципом замкненості.

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

$$slice: \sigma^{cons}{}_{\left(\begin{array}{l}Unk\Theta unk)\cup(C\_unk\Theta c\_unk)\cup\\ \cup\sigma_C(Dic)\cup\sigma_C(Cg)\end{array}\right)}(cg') = \left\{\begin{array}{l} t\in dc \,|\, t(Unk)\Theta unk, t(C\_unk)\Theta c\_unk, meta_{Unk,C\_unk}=1,\\ \sigma_C(Dic)\text{ Is Not NULL}, \sigma_C(Cg)\text{ Is Not NULL}, unk = P(cg')\end{array}\right\},$$

де 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*) – операція, введена у стандарті SQL 92. 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 відно-

шення. У випадку неповної з'єднувальності значення атрибута *Unk* для кортежів підлеглої таблиці, які не потрапляють у відношення, вважатимуться такими, що мають найвищий ступінь довіри.

$$across: r \underset{cons}{\triangleright\!\triangleleft} cg' = \text{IIF}\left(\sigma_{C\cup C\_Unk=X}(Dic); \pi_{\sigma_{C\cup C\_Unk=X}(Dic)}(r \underset{\bullet}{\triangleright\!\triangleleft} cg'); \pi_{(R,B,NVL(Unk,\min))}(r \underset{\bullet}{\triangleright\!\triangleleft} cg')\right),$$

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. В 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 дорівнює 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 $t_1, t_2, ..., t_m \in S'$.

3. Зменшення невизначеності консолідованих даних

Аналіз великих обсягів даних потребує виявлення груп атрибутів, що утворюють функціональні залежності. Проте в реальних умовах значно частіше зустрічаються набори даних, в яких важливі залежності, визначені тільки на деякій підмножині значень групи ключових атрибутів; називатимемо такі залежності частковими функціональними залежностями (ЧФЗ).

Тобто, часткова функціональна залежність – це ФЗ, визначена в деякій селекції основного відношення.

$$F_p: \; K = \{a_i\}, a_i \in A, D = \{a_j\}, a_j \in A, R' \subset R: K \to D \big| R'. \qquad (1)$$

Багато залежностей мають не чітко детермінований характер; називатимемо їх імовірнісними продукційними залежностями (ІПЗ).

Імовірнісна продукційна залежність (ІПЗ) – це продукційне правило в селекції основного відношення, яке справджується для значущої кількості об'єктів цієї селекції. Поріг значущості повинен визначатись експертним шляхом, або виходячи з розрахунків імовірності помилкового виділення цієї залежності.

$$F_I: \; K = \{a_i\}, a_i \in A, D = \{a_j\}, a_j \in A, : P\left(k \in K \to d \in D\right) = p. \qquad (2)$$

Тут k та d – кортежі значень деяких груп атрибутів K та D відповідно.

Основним показником достовірності такої залежності є відношення кількості об'єктів, для яких має місце така ІПЗ до кількості об'єктів в селекції:

$$P(F_I) = \frac{\left|\sigma_{k\in K \wedge d\in D}(R)\right|}{\left|\sigma_{k\in K}(R)\right|}. \tag{3}$$

Класифікаційним правилом назвемо імовірнісну продукційну залежність між підмножинами атрибутів *X* та *Y* у сховищі консолідованих даних cg', яка зустрічається у тестовому наборі cg' зі ступенем відповідності (довіри) *s*, де $(X = x) \to (Y = y)$.

Будується класифікаційне правило на основі навчального набору даних у cg', де значення позначок класу (значення підмножини атрибутів *Y*) відомі. Класифікаційне правило будується для схеми Cg' і тому не залежатиме від нових кортежів, що надходитимуть у відношення сховища консолідованих даних (незалежність від тестового набору).

Позначкою класу назвемо лінгвістичну змінну або звичну характеристику об'єктів, яка є значеннями підмножини атрибутів *Y* і позначає об'єкти зі спільними (подібними зі ступенем *s*) значеннями підмножини атрибутів *X*. Домени атрибутів, що належать до підмножини *Y*, $y \in dom(Y) = \pi_Y(Cg')$, обов'язково повинні містити скінченну та наперед відому множину значень.

Позначки класу обираються з наперед відомої множини значень (у межах досліджуваної області є фіксованими), а віднесення до класу об'єктів, інформація про які щойно надійшла у сховище консолідованих даних, здійснюється на основі класифікаційних правил.

Додавання позначок здійснюватиметься автоматично, оскільки надходження нових джерел даних у простір даних – також динамічне.

Обчислення достовірності виконання такої залежності ґрунтується на можливості розкладу такої залежності на складові ІПЗ:

$$P(s\in S \to t\in T) = \sum_{t_i\in T} P(s\in S \to t = t_i) = \sum_{t_i\in T} \frac{\sum_j \left|s = s_j \wedge t = t_i\right|}{\sum_j \left|s = s_j\right|} \tag{4}$$

Як і у випадку з F-залежностями (функціональними залежностями), множину класифікаційних правил, що існують у заданому відношенні, можна представити деякою їх підмножиною, з якої за допомогою правил виведення можна отримати усі класифікаційні правила даного відношення. Оскільки класифікаційні правила є розширенням F-залежностей, то варто розглянути трансформації аксіом для функціональних залежностей для класифікаційних правил.

Рефлексивність. $P(s\in S \to s\in S) = 1$ для будь-якого відношення *r*(*R*).

Доведення: $P(s\in S \to s\in S) = \frac{\left|\sigma_{s\in S \wedge s\in S}\right|}{\left|\sigma_{s\in S}\right|} = \frac{\left|\sigma_{s\in S}\right|}{\left|\sigma_{s\in S}\right|} = 1$

Поповнення. Якщо $P(s\in S \to t\in T) = p$, то $P(s\in S \wedge w\in D(W) \to t\in T) = p$.

Доведення: $P(s\in S \wedge w\in D(W) \to t\in T) = \frac{\left|\sigma_{s\in S \wedge w\in D(W) \wedge t\in T}(R)\right|}{\left|\sigma_{s\in S \wedge w\in D(W)}(R)\right|} = \left|\forall x\in r : q = \pi_{W=w}(x) \in D(W) \Rightarrow w\in D(W)\right| =$

$$= \frac{\left|\sigma_{s\in S \wedge t\in T}(R)\right|}{\left|\sigma_{s\in S}(R)\right|} = P(s\in S \to t\in T) = p$$

Адитивність. Якщо $P(s\in S \to t\in T) = p$ і $P(s\in S \to w\in W) = 1$, то $P(s\in S \to t\in T \wedge w\in W) = p$.

Доведення: $P(s\in S \to t\in T \wedge w\in W) = \frac{\left|\sigma_{s\in S \wedge t\in T \wedge w\in W}\right|}{\left|\sigma_{s\in S}\right|} = \left|s\in s \to w\in W\right| = \frac{\left|\sigma_{s\in S \wedge t\in T}\right|}{\left|\sigma_{s\in S}\right|} = P(s\in S \to t\in T) = p$

Усунення невизначеностей, які зустрічаються серед значень атрибута *Y* відношення *r*, є класифікуванням із використанням модифікованого алгоритму *chase*.

Суть методу:

1) пошук кортежів, у яких є однакові значення по множині атрибутів *X*;
2) пошук кортежів, у яких є однакові значення по множині синонімів атрибутів *X*;
3) розрахунок рівня довіри до джерела кортежів, отриманих на кроках (1) та (2);
4) розрахунок рівня довіри до атрибутів джерел кортежів, отриманих на кроках (1) та (2).
5) визначення кортежів з найбільшими рівнями довіри.

Щоб мати можливість класифікувати об'єкти, необхідно побудувати функції класифікації. Взагалі, у просторі даних може зберігатися інформація про декілька типів класів, і для кожного типу класу є своя підмножина функцій. Одна й та ж функція може застосовуватись для визначення кількох типів класів.

Функціями класифікації називатимемо модифіковані функціональні залежності, які виконуються для певної підмножини кортежів відношення сховища консолідованих даних.

Алгоритм віднесення до класу:

1. **Якщо** $\sigma(cg')$={dc$_1$ (X_1) ↓, ..., dc$_1$ (X_n) ↓} **і** {dc$_2$ (X_1) ↓, ..., dc$_2$ (X_n) ↓,}

   і {dc$_1$ (X_1) ↓, ..., dc$_1$ (X_n) ↓ = dc$_2$ (X_1) ↓, ..., dc$_2$ (X_n) ↓}

   і { dc$_1$ (Y) ↓} і {dc$_2$ (Y) = ⊥} **і Якщо** $\sigma_{X_1}(Dic) = \varnothing$

   то замінюємо ⊥ на dc$_1$ (Y) **і** $dc_1(P) = dc_1(P)/(\sum_i m_{1i}/n)$.

2. **Якщо** {dc$_1$ (X_1) ↓, ..., dc$_1$ (X_n)↓}

і {в dc$_2$ m з n значень атрибутів – ↓, $n - m$ значень атрибутів – ⊥, $m \le n$ }

і { $P \ge 1 - m/n$ } **і** {по визначених значеннях dc$_1$(X^m) ↓= dc$_2$(X^m) ↓}

і { dc$_1$ (Y) ↓} **і** { dc$_2$ (Y) = ⊥},

то замінюємо ⊥ у r на dc$_1$ (Y) **і** $dc_2(P) = dc_2(P)/(\sum_i m_{2i}/n)$.

3. **Якщо** {в dc$_i$ m_i з n значень атрибутів – ↓, $m_i \le n$ }

і {в dc$_j$ m_j з n значень атрибутів – ↓, $m_j \le n$ }

і {по визначених значеннях dc$_i$(X^m) ↓= dc$_2$(X^m) ↓}

і {по визначених значеннях dc$_j$(X^m) ↓= dc$_2$(X^m) ↓}

і { $m_i/n \le m_j/n$ } **і** { $P \ge 1 - m_i/n$ }

і { dc$_i$ (Y) ↓} **і** { dc$_j$ (Y) ↓} **і** { dc$_2$ (Y) = ⊥},

то замінюємо ⊥ на dc$_j$ (Y) **і** $dc_2(P) = dc_2(P)/(\sum_i m_{2i}/n)$.

Приклад

Нехай маємо базу даних науково-дослідної частини та наукові звіти двох кафедр. Зрозуміло, що між вказаними джерелами існує залежність: потрапляння даних у базу даних науково-дослідної частини через наукові звіти кафедр. Нехай у сховище консолідованих даних у результаті інтеграції потрапило 2 кортежі з таким вмістом:

ID	Author	Title	Publisher	Source	Trust
1	Автор1, Автор2	Назва1	Журнал1	DB1	0,7
2	Автор1, Автор2	Назва1	Журнал2	Text32	0,4

Отримали невизначеність виду "багатозначність інтерпретацій" (відмінності значення атрибута **Publisher**).

У каталозі даних вказано, що *Text32* є джерелом даних для *DB1*. Тоді, не дивлячись, що рівень довіри до першого кортежу є вищим, ніж до другого, у результуючу вибірку потрапить другий кортеж.

Висновки

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

2. Удосконалено операції над відношенням з невизначеністю з метою їх застосування до сховища консолідованих даних, що дозволило реалізувати унарні операції алгебричної системи “сигнатури простір даних”.

3. Розроблено метод зменшення невизначеності даних, що розміщені у сховищі консолідованих даних як основу для подальшого оцінювання якості консолідованих даних.

4. Розглянуті методи доцільно застосовувати також і для прийняття рішень, так як це забезпечує пошук прихованих залежностей між характеристиками сховища консолідованих даних. Такі залежності доцільно враховувати під час прийняття рішень на основі консолідованих даних.

1. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений / Заде Л. – М.: Мир, 1976. – 166 с. 2. Цельмер Г. Учет риска при принятии управленческих решений / Г. Цельмер // Проблемы МСНТИ. – № 3. – С. 94–105, 1980. 3. Найт Ф.Х. Риск, неопределенность и прибыль / Ф.Х. Найт. – М.: Дело, 2003. – 358 с. 4. Моисеев Н.Н. Элементы теории оптимальных систем / Н.Н. Моисеев. – М.: Наука, 1975. – 528 с.5. Трухачев Р.И. Модели принятия решений в условиях неопределенности / Р.И. Трухачев. – М.: Наука, 1981. – 151 с. 6. Згуровський М.З. Основи системного аналізу / Згуровський М.З., Панкратова Н.Д. – К.: Видавнича група BHV, 2007. – 544 с. 7. Шаховська Н.Б. Моделювання невизначеностей у сховищах даних реляційного типу. – Львів, 2007. – автореф. дис. … канд. техн. наук.