УДК 004.827

В.В. Пасічник, Н.Б. Шаховська
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

МОДЕЛЮВАННЯ НЕВИЗНАЧЕНОСТЕЙ У СХОВИЩАХ ДАНИХ

Проаналізовано проблеми, що виникають під час опрацювання невизначеності у сховищах даних. Визначено основні відмінності у поданні та опрацюванні невизначеностей у сховищах даних та базах даних. Уведено модель кортежу з невизначеністю та розширено алгебру опрацювання невизначеностей на рівні кортежу.

Main problems in data analisys in datawarehouse are described. The data structures of described system is designed. The uncertainties in input data are described. Main differences in datawarehouse uncertainces and relation database uncertainces are described. The uncertaince tuple model is inputed. The algebra of uncertaince worked is added.

Вступ

Корпоративні інформаційні системи, створені на основі реляційних СУБД, як правило, ефективно розв'язують задачі обліку, контролю і збереження даних. Однак у силу своєї специфіки реляційна структура не дає змоги розв'язувати задачі аналізу наявної інформації з необхідною продуктивністю. Особливо гострою ця проблема є в гетерогенних інформаційних середовищах, якщо в центральному офісі організації й у філіях експлуатуються СУБД різних виробників.

Вирішенням проблем продуктивності є створення спеціалізованої бази даних — сховища даних (Data Warehouse) — одним із застосувань якого є обробка та аналіз інформації.

Сховища даних дають змогу розвантажити оперативні бази даних, завдяки чому користувачі можуть ефективніше і швидше отримувати необхідну інформацію. Вони можуть входити до загальної корпоративної мережі, якою у сховище за заздалегідь визначеним розкладом копіюється накопичена за день чи за тиждень інформація. Оскільки дані змінюються рідко, до сховища даних не висувають твердих вимог, які зазвичай висувають до традиційних баз даних: відсутність аномалій під час виконання операцій чи відновлення видалення і надмірності збереження інформації. Але саме специфічність сховища даних приводить до появи нових проблем, які не зустрічалися у реляційних базах даних. Однією з таких проблем є опрацювання невизначеної інформації, яка має новий якісний характер.

Постановка задачі

Сховище даних визначають як предметно-орієнтований, інтегрований, залежний від часу набір даних, призначений для підтримки прийняття рішень різними групами користувачів. Оскільки сховище має предметно-орієнтований характер, його організація скерована на змістовний аналіз інформації, а не на автоматизацію бізнес-процесів. Ця властивість визначає архітектуру побудови сховища і принципи проектування моделі даних, відмінні від тих, які застосовують в оперативних системах.

Враховуючи специфіку, до проектування сховищ даних зазвичай висувають такі *вимоги*:

- Виділення статичних даних, що регулярно модифікуються.
- Спрощення вимог до запитів з метою виключення запитів, що могли б вимагати множинних запитів SQL у традиційних реляційних СУБД.
- Підтримка складних запитів SQL, що вимагають послідовної обробки великої кількості запитів.

Уведемо формальну модель сховища даних.

Реляційною базою даних називають трійку

$$DB = \langle r, R, Z \rangle,$$

де r – множина відношень бази даних,

R – множина їх схем,

Z – множина обмежень цілісності.

Тоді *сховищем даних*, побудованим на основі реляційної моделі, назвемо трійку

$$DW = \langle DB, rf, Rf, func \rangle,$$

де DB – множина баз даних (або множина відношень, їх схем та обмежень цілісності, які можна вважати окремою базою даних та які містять інформацію про певну частину предметної області – наприклад, дані складського обліку),

rf – відношення, у якому зберігається агрегована інформація і за даними якого приймають рішення (*відношення фактів*); Rf – схема відношення rf; $func$ – множина процедур прийняття рішень.

Тоді нові дані (або рішення) – це результат застосування функцій сховища даних над відношенням фактів:

$$Design = func(rf, user\_param).$$

де $user\_param$ – параметри користувача (або вимоги), які висувають до рішення.

Оскільки відношення rf містить агреговану інформацію з відношень баз даних, то зв'язок між ним і відношеннями баз даних DB приводить до утворення так званого гіперкуба даних (моделі багатовимірного подання даних) [1].

Виміром назвемо універсум відношень бази даних DB_i – $V_i : Universum(DB_i)$. Кожен вимір містить напрямки консолідації даних, що складаються із серії послідовних рівнів узагальнення (рівнів ієрархії).

Відношення між вимірами – деяке відношення, яке є зв'язком між вимірами.

$$V_1, V_2, ..., V_n \rightarrow \mathrm{Re}l.$$

Наприклад, зв'язком між вимірами *Покупець* та *Товар* є відношення *Продаж*. Оскільки не між всіма об'єктами, що зберігаються у вимірах, можна встановити однозначне відношення, то це призводить до виникнення невизначеності відносно вимірів.

У свою чергу, *Rel* можуть бути параметрами для інших відношень між вимірами і тим самим створювати ієрархію вимірів.

Осями багатовимірної системи координат є основні атрибути аналізованого бізнес-процесу. На перетинах вимірів (dimensions) знаходяться дані, які кількісно характеризують процес - виміри (measures).

Формування відношення rf здійснюється на основі функції агрегування Agg [1]:

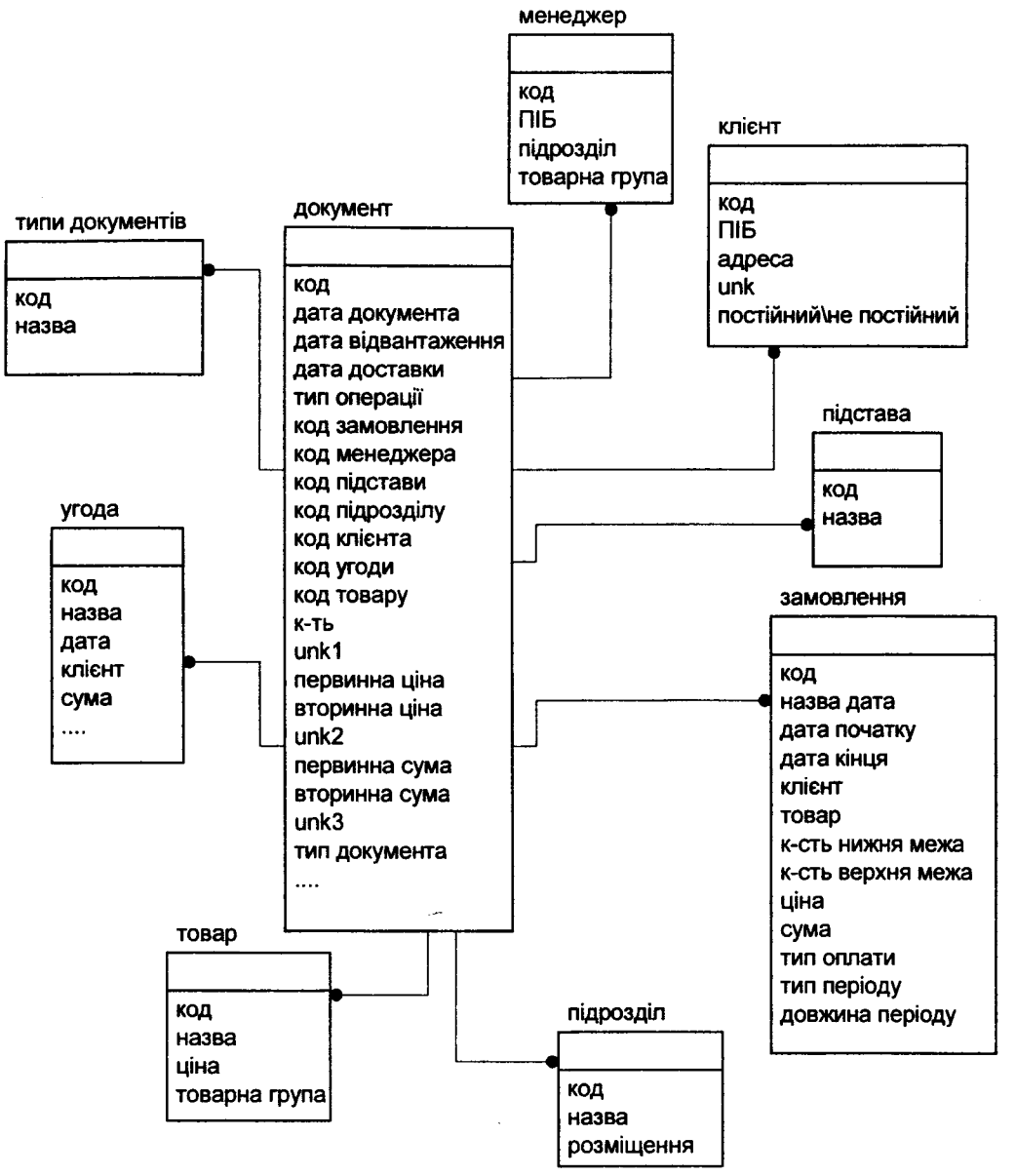
$$rf : Agg(\mathrm{Re}l_1, ..., \mathrm{Re}l_n)$$

Внаслідок встановлення відношень між вимірами та операцій агрегування гіперкуб у переважній більшості випадків є сильно розрідженим, тому проблема опрацювання невизначеності є набагато складнішою, ніж у реляційних базах даних.

Перейдемо до огляду невизначеностей, які виникають у сховищах даних.

Оскільки розглядається сховище даних, побудоване на основі реляційної моделі, то у нього можуть входити ті самі типи невизначеності, що й у відношення реляційної бази даних [8]. Проте, враховуючи багатомірність подання даних, виникнення невизначеності слід розглядати окремо для кожного типу відношень сховища даних. Так, якщо невизначеність виникла у відношеннях вимірів, то вона поширюватиметься на всі об'єкти у відношенні фактів rf, які володіють вказаним виміром. Якщо невизначеність виникає у відношенні фактів та стосується вимірів (тобто зовнішніх ключів), то виникає задача визначення приналежності виміру до вказаного об'єкта.

Наведемо приклад схеми відношень для демонстрації причин та місць появи невизначеностей. Нехай є сховище даних предметної області "Торговельне підприємство". З локальних баз даних у центральне сховище даних надходить узагальнена інформація про обсяги продажу товарів та замовлення (плани купівлі).



На цій схемі відношення "*Документ*" – відношення фактів. Інші відношення – відношення вимірів. Відношення вимірів є частково денормалізованими, оскільки містять агреговану інформацію, що надходить з локальних баз даних.

Наведемо приклади ситуацій, які приводять до виникнення невизначеності у цьому відношенні:

1) один із підрозділів передав лише приблизну інформацію про обсяги продажу певного товару. Тоді **атрибут** *к-ть* є **неточним**. Факт неточності моделюють за допомогою атрибута *unk1*, який показує рівень довіри до значення *к-ть*.

2) не вказано *коду менеджера* – виникає невизначеність типу **відсутність.**

3) у відношенні "*Менеджер*" не вказано значення атрибута товарна група, тобто невідомо, яким саме товаром може торгувати певний менеджер. Тому у відношенні фактів виникає невизначеність типу **ненадійність** на рівні кортежу.

4) у відношенні "*Клієнт*" не вказано значення атрибута *постійний\не постійний*. Це призводить до виникнення **невизначеності** типу **нечіткість** за атрибутами відношення фактів

вторинна ціна та *вторинна сума*. Факт нечіткості моделюють за допомогою атрибутів unk2 та unk3 відповідно.

У відношенні *Клієнт* уведено невизначеність на рівні атрибута, що моделюється за допомогою атрибута unk. Така невизначеність не поширюється на відношення фактів, але повинна коректно опрацьовуватись у запитах.

Цілі статті

Статтю присвячено розгляду особливостей представлення невизначеності та методам опрацювання невизначеностей у сховищах даних, побудованих на основі реляційної моделі.

Розглянемо проблеми подання невизначеностей у *відношенні фактів*.

Ключ відношення фактів складається з множини зовнішніх ключів відношень вимірів. Існування факту невизначеності за зовнішнім ключем можна розглядати у таких аспектах:

- Об'єкт, який моделюється кортежем відношення фактів з відсутніми значеннями зовнішніх ключів, не володіє властивостями, описаними у відношеннях розмірності (немає значення за вказаним виміром) – така невизначеність притаманна і відношенням реляційних баз даних [3, 4, 6], і тому у статті не буде розглядатися.
- Відомо, що значення за вказаним виміром існує, але сьогодні 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 можна застосовувати 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 типів [1,2] (виникнення ієрархії об'єктів)
- Є неповна або часткова інформація про значення, для відображення якої використовують 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.

Тепер розглянемо невизначеність, яка виникла у відношенні вимірів. Як і у відношенні фактів, невизначеність може виникати на рівні атрибута, кортежа та відношення. Поява невизначеності на рівні атрибута і кортежа у зв'язку з багатовимірністю представлення інформації призводить до поширення невизначеності на усі кортежі відношення, у яких значення зовнішнього ключа за вказаним виміром не є порожнім. Оскільки відношення фактів містить велику кількість кортежів, то опрацювання невизначеності традиційними засобами (інтервальна математика, багатозначна логіка) [5] стає неефективним через велику кількість операндів.

Отже, специфіка сховищ даних (а саме його багатовимірність) 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 даному випадку поширюється на усе сховище даних. Тому для опрацювання невизначеності у сховищі даних необхідно використати якісно новий підхід, потреба застосування якого не виникала у реляційних базах даних.

У роботі розглянуто задачі:

1) узагальнення методів подання різних типів невизначеностей (інтервальні значення, лінгвістичні змінні, стохастичні оцінки, ступені істинності кортежу або підмножини значень атрибутів кортежу) у сховищах даних та забезпечити коректне опрацювання цих типів невизначеностей під час виконання операцій аналітичного аналізу. Розширити традиційні реляційні оператори з метою опрацювання різних типів невизначеностей, введених у відношення на різних рівнях (стосується цілого кортежу, підмножини значень атрибутів, на рівні значення атрибута);

2) зменшення невизначеності у сховищах даних з використанням методів інтелектуального аналізу;

3) шляхом аналізу зв'язків, які виникають між кортежами відношення, зменшення невизначеності типу "значення невідоме", "нечіткість", "недостовірність", "багатозначність інтерпретацій", "частковість", яке полягає:

- у збільшенні значення ступеня довіри до підмножини атрибутів кортежу;
- у визначенні відсутніх значень;
- у заміні значень атрибутів із низьким значенням ступеня довіри на значення атрибутів.

Розглянемо особливості моделювання невизначеності типу "відсутність" та "нечіткість" у сховищах даних, побудованих на основі реляційної моделі.

Основний матеріал

Проблеми введення невизначеності на рівні кортежу

Одним із методів моделювання неточних, нечітких та часткових даних є уведення до схеми відношення додаткового атрибута, значення якого вказує ступінь довіри до невизначених даних. У результаті цього неможливо[2]:

- вказати, які саме атрибути містять невизначені значення;
- розрізнити, чи значення додаткового атрибута є однією з характеристик об'єкта (наприклад, технічна характеристика надійності приладу), чи воно позначає надійність даних кортежу (тоді йдеться про нечітке відношення, яке складається з невизначених кортежів).

У зв'язку з цим виникають проблеми:

- Збереження цілісності – операції з'єднання з невизначеним кортежем приводять до невизначеності кортежів у відношеннях-операндах операції;
- Істинності запитів (операція проекції) – невідомо, наскільки істинним буде запит, наприклад, на вибірку підмножини атрибутів невизначеного кортежу.

Тому у нашій роботі вважатимемо, що невизначеність може вводитися на рівні підмножини атрибутів відношення, причому до цієї множину не входять первинні та зовнішні ключі.

Покажемо, чому невизначеність не може виникати на рівні значень зовнішніх ключів відношення фактів. Нехай у відношенні фактів з розглянутого вище прикладу потрібно зазначити факт невизначеності на рівні атрибута *менеджер*, тобто показати, що невідомо, який саме менеджер продав товар. Нехай така невизначеність моделюється за допомогою атрибута *unk*.

Тоді на семантичному рівні неможливо розрізнити факти: чи ми не впевнені у значенні цього атрибута, чи в існуванні об'єкта, який описаний у сховищі даних за допомогою коду менеджера , значення якого не визначене (тобто виникла невизначеність на рівні кортежу відношення *менеджер*).

Тому для коректного опрацювання даних невизначеність може виникати на рівні виділених атрибутів:

Документ

код	Дата документа	Дата відвантаження	Дата доставки	Тип операції	Код замовлення	Код менеджера	Код підстави	Код підрозділу	Код клієнта	Код угоди	Код товару	кількість	Unk1	Первинна ціна	Вторинна ціна	Unk2	Первинна сума	Вторинна сума	Unk3	Тип дукумента
01	2.2.05	2.2.05		прихід	04	Менджер1		Склад1	Клієнт1	Угода1	Сист блок	**2**	**0.4**	22	24		44	48		торгівля
02	2.3.05	2.3.05		розхід	01	Менджер1		Склад1	Клієнт1	Угода1	монітор	1		**23**	**26**	**0.8**	23	26		торгівля
03	3.6.05	3.6.05		прихід	03	Менджер1		Склад1	Клієнт2	Угода1	клавіатура	**4**		**2**	**4**		**8**	**16**	**0.6**	торгівля

Тут значення атрибута *unk1* подає ступінь довіри до значення атрибута *кількість*, *unk2* – атрибутів *первинна ціна*, *вторинна ціна*, *unk3* – атрибутів *первинна сума*, *вторинна сума* та перерахованих атрибутів.

Модель кортежу з невизначеністю

Уведемо поняття кортежу з невизначеністю.

Кортежем із невизначеністю t_unk назвемо кортеж, підмножина значень атрибутів якого містить неповні, нечіткі чи недетерміновані дані. Тобто, об'єкт, який моделюється у відношенні сховища даних цим кортежем, існує, але частина інформації про нього відсутня, нечітка, неповна, недетермінована тощо. Крайній випадок незнання про об'єкт відображається у відношенні як існування значення первинного ключа у кортежі з невизначеністю. Такий кортеж не піддається аналізу, оскільки для нього неможливо встановити значення цільових атрибутів.

Кортеж із невизначеністю не порушує обмежень цілісності, оскільки моделює реальний об'єкт. Крім того, інформацію, що у ньому міститься, можна опрацьовувати у запитах, результати яких будуть відбивати сам факт незнання про об'єкт або ступінь довіри до даних про об'єкт.

Значення атрибутів кортежу з невизначеністю поділимо на такі групи:

- Чіткі (відомі) – значення первинного ключа, зовнішніх ключів (можуть бути відсутні). Позначимо їх через A;
- Відсутні – фізично відсутня інформація. Позначимо їх через $\perp$;
- Нечіткі – для підмножин атрибутів (для атрибута) введено додатковий атрибут *UNK*, який указує ступінь істинності значень цих атрибутів. Доменом такого атрибуту є числові дані, що моделюють імовірнісні дані, значення функції приналежності нечітких множин, ступінь істинності багатозначної логіки, процентні відношення, коефіцієнти, різноманітні шкали або лінгвістичні оцінки [4, 8]. За замовчуванням значенню атрибута *UNK* присвоюємо значення, яке означає найвищий ступінь істинності. Детальніше цей атрибут описано нижче. Крайніми випадками введення нечіткості є:
  - додавання атрибутів типу *UNK* до усіх атрибутів, крім чітких;
  - додавання атрибута *UNK* до усіх значень кортежу.

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

Отже, кортеж із невизначеністю *t_unk* – це множина значень характеристик об'єкта сутності, описана трійкою

$$\mathbf{t\_unk = <dom(A),\ dom(A\_unk),\ dom(UNK)>},$$

де A – підмножина атрибутів із чіткими значеннями,

A_unk – підмножина атрибутів із нечіткими та недетермінованими значеннями,

UNK – підмножина атрибутів із ступенями істинності значень атрибутів *A_unk*.

Прикладом відношення, що містить кортежі з невизначеністю, є відношення фактів, показане нижче. Тут обведені атрибути належать підмножині *A*, атрибути, виділені жирним – множині A_*unk*, решта атрибутів (із назвами unk) – множині *UNK*.

Документ

код	Дата документа	Дата відвантаження	Дата доставки	Тип операції	Код замовлення	Код менеджера	Код підстави	Код підрозділу	Код клієнта	Код угоди	Код товару	кількість	Unk1	Первинна ціна	Вторинна ціна	Unk2	Первинна сума	Вторинна сума	Unk3
01	2.2.05	2.2.05		прихід	04	Менеджер1		Склад1	Клієнт1	Угода1	Сист блок	2	0.4	**22**	**24**		**44**	**48**	
02	2.3.05	2.3.05		розхід	01	Менеджер1		Склад1	Клієнт1	Угода1	моніторр	1		**23**	**26**	0.8	**23**	**26**	
03	3.6.05	3.6.05		прихід	03	Менеджер1		Склад1	Клієнт2	Угода1	клавіатура	4		**2**	**4**		**8**	**16**	0.6

Зрозуміло, що вказання невизначеності для підмножини атрибутів відношення ускладнює розуміння схеми сховища даних і повинне бути документоване. Тому введемо відношення *attr* із схемою *Attr*, у якому зберігатиметься залежність між чіткими та нечіткими атрибутами відношень сховища даних:

attr

Id	Первинний ключ
Rel_name	Назва відношення
Attr_name	Назва атрибута
UNK_name	Назва атрибута з невизначеністю
Prior_id	Зовнішній ключ відношення Attr

Алгебра опрацювання невизначеності у сховищах даних

Оскільки сумарне відношення складається з множини зовнішніх ключів сторонніх відношень, то при виникненні невизначеності на рівні зовнішніх ключів розглядають випадки, коли відношення є повністю з'єднувальними або не повністю з'єднувальними [4].

Для *повністю з'єднувальних* відношень введення атрибута *UNK* не впливає на операцію з'єднання (природне з'єднання або еквіз'єднання), оскільки його значення не змінюватимуться [5].

У випадку *неповної з'єднувальності* значення атрибута *UNK* для кортежів підлеглої таблиці, які не потрапляють у відношення, дорівнюватимуть найнижчому ступеню довіри (за вже згаданим постулатом замкненості світу):

$$r \underset{unk}{\triangleright\!\triangleleft} s' = \pi_{(R,B,NVL(UNK,\min)UNK)}(r \underset{\bullet}{\triangleright\!\triangleleft} s'), \qquad (2)$$

де r – традиційне відношення зі схемою *R*, s' – відношення, до схеми якого входить атрибут *UNK* ($S' = S \bigcup UNK$), *B* – множина тих атрибутів з *S*, які не належать схемі *R* $B \subset S$, ($B \not\subset S \bigcap R$), *min* – значення, яке означає найнижчий ступінь довіри (нуль); *NVL*(<величина1>,<величина2>) – операція, яка у випадку відсутності значення <величини1> присвоює їй значення <величини2>; $NVL(UNK,\min)UNK$ – операція, яка присвоює *min* усім значенням атрибута UNK для нез'єднувальних кортежів відношення s', $\underset{\bullet}{\triangleright\!\triangleleft}$ – ліве з'єднання (вводяться усі кортежі відношення r і лише ті кортежі відношення s', у яких значення за з'єднувальними атрибутами збігаються). Спочатку виконується операція лівого з'єднання для відношень зі схемами S' і R. Потім над отриманим за попередньою операцією відношенням здійснюють операцію проекції, за якою утвореним у результаті з'єднання порожнім значенням атрибута *UNK* присвоюється значення *min*.

Для опрацювання та аналізу невизначеностей за допомогою запиту в реляційних операторах здійснюють селекцію кортежів за значеннями атрибута *UNK*. Додавши відношення, у якому би містилися "розшифровки" значень атрибута *UNK*, або використовуючи лінгвістичні оцінки, ми отримуємо можливість побудови нечітких запитів (наприклад, "Вибрати усіх працівників, рівень інформованості про яких є високий"). Варто зазначити, що у разі стовідсоткової довіри до кожного кортежу ми отримуємо традиційне реляційне відношення та застосовуємо традиційні операції над ним.

Нехай r та s – відношення зі схемою *R*, r' та s' – відношення зі схемою $R \cup UNK$. 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 $R \cup UNK$.

Доповнення до відношення r' працюватиме коректно у разі присвоєння усім значенням атрибута *UNK* найнижчого ступеня довіри (апріорі вважається, що інформація, яку заносять у відношення, є правдивою та повною, а про решту інформації нам нічого не відомо). Такий метод подання ступеня істинності за замовчуванням вибрано за принципом замкненості світу та вибору апарату багатозначної логіки Лукасевича для опрацювання даних (розглянуто нижче).

Оператор вибірки передбачає аналіз нечіткого значення за атрибутом *UNK*.

$$\sigma_{(UNK\ \Theta\ unk)\cup(A\ \Theta\ a)}\left(r'\right)=\left\{t\in r'\middle|\ t\left(UNK\right)\ \Theta\ unk,\ t\left(A\right)\ \Theta\ a\right\},\qquad(3)$$

де Θ – множина символів (знаків) бінарних відношень над парами доменів. Вважається, що кожний атрибут A порівняний за рівністю й за нерівністю. Як правило, будуть вживатися тільки такі знаки порівняння над одним доменом: $=, \neq, <, \leq, \geq, >$ [6].

Зазначимо, що розширений оператор вибірки зберігає властивості комутативності та дистрибутивності відносно булевих операцій.

Здійснюючи проекцію відношення з кортежами з невизначеністю, відслідковують зв'язок атрибута *UNK* із підмножиною атрибутів. Тому розширимо оператор проекції:

$$\pi_X^{UNK}(r)=$$
$$=\text{IIF}(\neg\ \text{ISNULL}(\sigma_{rel\_name=R\cup attr\_name=X}(attr));\ ,\qquad(4)$$
$$\pi_{X\cup\pi UNK\_name(\sigma rel\_name=R(attr))}(r);\pi_X(r))$$

де IIF(*умова*; *дія 1*; *дія 2*) – операція, введена у стандарті SQL 92 [4]. 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.

Розширений оператор проекції зберігає властивості традиційного оператора проекції.

Для решти операцій введемо *узагальнений оператор* над відношеннями з невизначеними даними:

$$R=\gamma(r,A,lingvistic,\beta),\qquad(5)$$

де $\boldsymbol{r}=\{r_1,...,r_n\}$ – множина відношень зі схемою $\boldsymbol{R}$ (які можуть бути об'єднані у єдине універсальне відношення), $\boldsymbol{A}$ – множина цільових атрибутів, ($\boldsymbol{A}\in\boldsymbol{R}$); *lingvistic* – множина невизначених змінних, з якими порівнюються значення цільових атрибутів у $\boldsymbol{r}$, β – множина операторів реляційної алгебри над $\boldsymbol{r}$. Результатом виконання оператора γ буде множина відношень $\Re$, яку будують на основі застосування до $\boldsymbol{r}$ операторів з β за атрибутами з множини $\boldsymbol{A}$ з врахуванням значення нечіткої змінної з множини *lingvistic* (або множини нечітких змінних) та додаванням до отриманого відношення атрибута *UNK* ($\Re = R\cup UNK$), який характеризує ступінь відповідності значень цільових атрибутів у вихідних кортежах до значення змінних з *lingvistic*. Значення цього атрибута прямо пропорційно залежить від:

- кількості атрибутів у множині $\boldsymbol{A}$ (чим більше використовувати атрибутів для аналізу, тим точнішим буде отриманий результат);
- способу задавання змінних з *lingvistic*;
- обраного правила порівняння нечітких величин.

Застосування γ *-оператора* дасть змогу:

- здійснити кластеризацію за ступенями відповідності та визначити ступінь істинності результату аналізу фрагмента відношення стосовно результатів аналізу усього відношення: якщо існує залежність між атрибутами відношення, то результати аналізу за усіма атрибутами та за їхньою підмножиною будуть різними, і власне значення атрибута *UNK* характеризуватиме ступінь відповідності значень критичних атрибутів цього кортежу до результатів аналізу усього відношення;
- проаналізувати, наскільки повними є дані і наскільки точно вдасться довизначити порожні значення. γ-оператор можна розглядати як класифікаційне правило, де лівою його частиною будуть цільові атрибути, а правою – критичні. Таке класифікаційне правило має ступінь довіри багатозначної логіки, виражений значенням *UNK*. Як зазначалось, до множини цільових та критичних атрибутів належать атрибути, які входять до складу функціональних залежностей, та додаткові атрибути, виділені для аналізу з врахуванням потреб певної предметної області. Тому

з'являється можливість заповнення за класифікаційними правилами порожніх значень критичних атрибутів аналізованого кортежу значеннями критичних атрибутів із тих кортежів відношення, у яких значення атрибутів у лівій частині дорівнює значенням атрибутів лівої частини даного кортежу із ступенем довіри *UNK*.

Існує декілька методів визначення значення *UNK*, але вони у цій статті не розглядатимуться. Для простоти вважатимемо, що значення атрибута *UNK* – це відношення визначених цільових атрибутів до усіх цільових атрибутів.

Наведемо приклад γ -оператора – оператор *лінгвістичної вибірки*:

$$\sigma^{unk}_{A\Theta U(r),B} = \{t \in r \mid ([t(A) \geq u1] \wedge [t(A) \leq u2]) \cup UNK\}, \quad (6)$$

де r – традиційне реляційне відношення зі схемою *R*, *A* – атрибут (множина цільових атрибутів) в *R*, за яким проходить вибірка; *unk* – відношення, яке містить значення лінгвістичних змінних, *unk*(*lingvistic_variable*, *infimum*, *supremum*) з атрибутами, які позначають назву лінгвістичної змінної та її нижнє та верхнє значення (одна з меж може бути відсутня); $A\Theta U$ – вираз, результатом якого є порівняння за обраним правилом значення *a* атрибута *A* із значеннями [*u*1, *u*2] лінгвістичної змінної *U*; *B* – послідовність атрибутів, значення яких обов'язково повинні відповідати параметрам вибірки ($B \subset A$); σ^{unk} – оператор лінгвістичної вибірки; *unk* – ступінь істинності значень критичних атрибутів відношення, який визначається у результаті застосування σ^{unk} .

Оператор σ^{unk} , як і аналогічний у традиційній реляційній алгебрі оператор вибірки σ володіє властивістю дистрибутивності відносно бінарних булевих операцій, але втрачає властивість комутативності, оскільки для різних послідовностей *B* отримуються різні результати аналізу.

Проаналізуємо, як працюватиме оператор лінгвістичної вибірки, якщо до відношення застосовувати чіткі параметри. У результаті виконання селекції отримуємо перегляд, у який увійдуть кортежі, які повністю відповідають заданим чітким параметрам *a*, тобто згідно з (6)

$$\sigma^{unk}_{A\Theta U(r),B} = \sigma_{A=a} \cup unk_{lingvictic\_variable='повна\ відповідність'}$$

Продемонструємо застосування оператора лінгвістичної вибірки на прикладі.

Приклад

Вибрати усі виконані протягом останнього місяця замовлення, які задовольняють умови

Критичний атрибут	**Умова вибірки**
Дата початку	01.01.02
Дата закінчення	31.01.02
Тип документа	Банк
Підрозділ	Dept1
Угода	Contract1
1. Інтервал виконання плану та інтервал, зазначений у вибірці, перетинаються 2. Дата документа є у межах плану 3. Враховувати збіг за угодами	

На підприємстві є такі замовлення, що відповідають заданим умовам:

Таблиця 1

Відношення Замовлення

Код	Дата початку	Дата кінця	Тип документа	Підрозділ	Угода	Сума
01	12.12.01	10.01.02	Банк	Dep1	Contract1	120
02	3.01.02	1.02.02	Банк		Contract1	100
05	4.01.02	12.01.02	Банк	Dept1	Contract1	200

З ними пов'язані такі документи.

Таблиця 2

Відношення Документи

Код док.	Код плану	Дата документа	Тип документа	Підрозділ	Угода	Сума
01	01	12.12.01	Банк	Dep1	Contract1	100
02	02	3.01.02	Банк	Dept1	Contract1	100
03		4.01.02	Банк	Dept1	Contract1	200

У результаті застосування γ-оператора (тобто селекції та групування за датами та угодами) отримано таке відношення:

Таблиця 3

Відношення Зіставлення замовлень та документів

Код документа	Код плану	Дата документа	Тип документа	Підрозділ	Угода	Ступінь відповідності
01	01	12.12.01	Банк	Dep1	Contract1	Повна невідповідність
02	02	3.01.02	Банк	Dept1	Contract1	Повна відповідність
03	03	4.01.02	Банк	Dept1	Contract1	Повна відповідність

Крім того, з'явилась можливість довизначити значення атрибута *Підрозділ* у відношенні *Плани* (кортеж з кодом плану 02) та атрибута *Код плану* у відношенні *Документи* (код документа 03).

Отже, традиційну вибірку можна вважати частковим випадком оператора σ^{unk}. Тоді значенням атрибута *UNK* буде максимальний ступінь відповідності.

Довизначенням атрибута *код плану* надалі використовували отримане значення для аналізу обсягів виконання планів. Без застосування γ-оператора план з кодом 03 вважався б недовиконаним (або невиконаним), оскільки за ним не було знайдено документів, що насправді було не так.

Висновки

Багатомірність подання даних у сховищах даних спричинила фактори, які суттєво міняють подання та опрацювання невизначеності у сховищах даних та реляційних базах даних. Особливостями роботи із невизначеністю у сховищах даних є необхідність опрацювання невизначеності окремо на рівні відношення фактів та окремо на рівні відношень вимірів. Оскільки у сховищі даних зберігається інформація про усю предметну галузь і основна увага звертається на аналітичні запити, то застосування традиційних реляційних операторів до опрацювання невизначеності призводить до її збільшення (розширення інтервалів, зменшення точності тощо).

Тому у статті розглядалися оператори коректного опрацювання невизначеності у запитах, а також пропонувалось використовувати засоби інтелектуального аналізу для зменшення невизначеності.

Об'єкт дослідження (сховище даних) та методи опрацювання у ньому невизначеності вимагають подальших досліджень.

Наукова новизна одержаних результатів полягає в тому, що на рівні формальної моделі показано причину виникнення невизначеності у сховищі даних та особливість її опрацювання порівняно з реляційною базою даних. Уведено модель кортежу з невизначеністю.

Практична цінність досліджень полягає в уведеннні *узагальненого оператора* над відношеннями з невизначеними даними, який дозволяє коректно опрацьовувати невизначеності.

Подальші дослідження авторів будуть стосуватися таких питань:

1. Побудова алгоритмів для зменшення невизначеності шляхом інтелектуального аналізу даних.
2. Розробка методів коректної роботи з інтервальними зниченнями.

1. Кравець Р.Б. Організація багатовимірного подання та аналізу інформації у реляційній базі даних //Вісник НУ "Львівська політехніка". – 2003. – № 489. 2. Netz A., Chaudhuri S. Integration of Data Mining and Relational Databases. Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000 3. G. Grsrfe, U. Fayyad. On the Efficient Gathering of Sufficient Statistics for Classification from Large SQL Databases. – 1998. – www.aaai.org 4. Huhtala Y., Karkainen J. Tane: An Efficient Algoritm for discovering Functional and Approximate Dependencies// The Computer Journal. 1999. Vol. 42. – № 2. 5. Дюбуа А., Прад А. Теория вероятностей. Приложение к представлению знаний в информатике. – М.: Радио и связь, 1990. 6. Мейер Д. Теория реляционных баз данных. – М.: Мир. – 1987 7. Шаховська Н.Б. Застосування апарату багатозначної логіки у системах баз даних //Вісник НУ "Львівська політехніка" . – 2001. – № 438. 8. Шаховська Н.Б. Методи усунення невизначеностей у базах знань, побудованих на основі реляційного підходу //Вісник НУ "Львівська політехніка". – 2001. – № 438.

УДК 681.3

А.М. Пелещишин
Національний університет "Львівська Політехніка",
кафедра інформаційних систем та мереж

ОПТИМІЗАЦІЯ ФОРУМІВ ТА ІНШИХ ФОРМ СПІЛЬНОТ КОРИСТУВАЧІВ WWW

Розглянуто проблеми формування та оптимізації тематики і наповнення сайтів інтернет-спільнот. Детально досліджено найпопулярнішу форму спільноти – інтернет-форум.

This paper considers problems of thematics and content building and optimization for sites of internet communities. Most popular form of community – bulletin boards are detailed researched.

Постановка проблеми та її актуальність

Розглянуто особливості визначення та оптимізації тематики сайтів таких спеціальних видів, як форумів, блогів, інших форм інтернет-спільнот. Принциповою відмінністю сайтів інтернет-спільнот від традиційних сайтів, що представляють своїх власників, є високий ступінь залежності інформаційного наповнення сайту від його відвідувачів. Така безпосередня залежність відсутня у традиційних представницьких чи інформаційних сайтах.

Ця особливість сайтів інтернет-спільнот різко виділяє їх від решти сайтів під час розв'язання ряду специфічних задач з позиціонування сайту в глобальному середовищі. У першу чергу це стосується задачі визначення тематики та її ефективного подання.

Аналіз досліджень

Як і для традиційних сайтів, для сайтів інтернет-спільнот тематика породжується інформаційним наповненням сайту, а знаходить своє відображення в аудиторії WWW, яка знаходить інформацію, зацікавлюється нею та використовує її [0].

Відмінністю ж є те, що сама аудиторія сайту безпосередньо наповнює сайт інформацією, отже, самостійно формує, уточнює та змінює тематику сайту інтернет-спільноти. Крім прямого інформаційного зв'язку "сайт–користувач", існує й зворотний зв'язок "користувач–сайт", який для деяких типів інтернет-спільнот взагалі є домінуючим (фактично сайт стає похідним від спільноти).