УДК 004.827

Н.Б. Шаховська, Ю.О. Сєров
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

ПОДАННЯ ТА ОПРАЦЮВАННЯ НЕВИЗНАЧЕНОСТЕЙ У ГЕОІНФОРМАЦІЙНИХ СИСТЕМАХ

Проаналізовано проблеми, що виникають під час опрацювання даних у геоінформаційних системах. Наведено структури даних геоінформаційної системи та основні задачі, які повинна розв'язувати система. Проаналізовано методи збільшення інформативності геоінформаційних систем. Уведено узагальнений тип невизначеності.

Main problems in data analisys in geosystems are described. The data structure of geoinformation system is designed and main aims are overviewed. Geoinformational system effectiveness increasing methods are analized. The uncertainties in input data are described.

Постановка задачі у загальному вигляді

Головна відмінність ГІС-технологій від технологій сховищ даних, побудованих на основі реляційної моделі, полягає у встановленні зв'язку між картографічною інформацією та тематичними даними у формі реляційних баз данних. Це дозволяє в інтерактивному режимі легко переходити від табличного подання данних до картографічного і навпаки або поєднувати їх. Тому можливість комбінувати геометричні та атрибутивні дані визначає якісно новий підхід до аналізу данних з метою прийняття на його основі обгрунтованого рішення [1].

Важливими особливостями ГІС є можливість забезпечення комплексного введення, контролю, зберігання, відображення та аналізу різної семантичної та картографічної інформації, підготовки картографічних матеріалів для аналітичної обробки, прийняття ефективних рішень на основі аналізу та інтерпретації просторово розподілених даних. Власне різнорідність інформації, якою повинна оперувати система, і призводить до виникнення ряду проблем серед яких, одними із найголовніших є опрацювання невизначеної інформації та проблема пошуку інформації.

Стаття присвячена розгляду вищенаведених проблем та побудові методів їх вирішення.

Актуальність роботи

Головною проблемою, яка постає під час розробки ГІС, є відсутність єдиної системи та політики інтегрування розрізнених геоінформаційних даних з метою їх загального використання. Зрозуміло, що неузгоджене створення, накопичення та актуалізація великого обсягу такої інформації потребує витрачання зайвих коштів і не дає очікуваного результату. Сьогодні на теренах України немає системи, яка б могла використовуватись для комплексного інформаційного забезпечення, формування аналітичних та прогнозних даних та підтримки прийняття рішень стосовно гармонійного розвитку регіону [1].

Розгляд проблеми опрацювання невизначеності дозволить вирішити питання підвищення достовірності та надійності інформації, а, отже, підвищить ефективність рішень, вироблених системою.

Геоінформаційна система призначена для комплексного інформаційного забезпечення, формування аналітичних та прогнозних даних та підтримки прийняття рішень стосовно гармонійного розвитку території.

Система містить такі функції:

1) обліку:

нормативної та законодавчої бази;

структури адміністративно-територіальних одиниць; облік та аналіз картографічної інформації;

кліматичних умов;

моніторинг стану довкілля, аналіз охорони довкілля;

структури населення;

зайнятості та структури зайнятості;

впливу господарською діяльності на довкілля, прогнозування та попередження надзвичайних ситуацій;

обсягів виробництва та розвитку господарської діяльності;

забезпеченості житлово-комунальними послугами;

рекреаційно-оздоровчої сфери та забезпеченості її послугами населення;

культурно-освітнього забезпечення населення;

в галузі капітального будівництва;

транспортної інфраструктури та транспортної мережі;

природних умов проживання населення;

демографічного стану;

виробництва та споживання мінеральних, паливно-енергетичних, водних, біологічних, земельних ресурсів у господарській діяльності;

землеустрою та землекористування для земель сільськогосподарського та промислового використання;

розвитку рекреаційної сфери;

забезпеченості ресурсами житлово-комунальної сфери;

2) аналізу:

структури населення;

зайнятості та структури зайнятості;

впливу господарської діяльності на довкілля, прогнозування та попередження надзвичайних ситуацій;

розвитку господарської діяльності;

забезпеченості житлово-комунальними послугами;

3) аналізу та прогнозування, планування:

обсягів виробництва та розвитку господарської діяльності;

забезпеченості житлово-комунальними послугами;

рекреаційно-оздоровчої сфери та забезпеченості її послугами населення;

культурно-освітнього забезпечення населення;

в галузі капітального будівництва;

транспортної інфраструктури та транспортної мережі;

природних умов проживання населення;

демографічного стану;

виробництва та споживання мінеральних, паливно-енергетичних, водних, біологічних, земельних ресурсів у господарській діяльності;

землеустрою та землекористування для земель сільськогосподарського використання та промислового використання;

розвитку рекреаційної сфери;

забезпеченості ресурсами житлово-комунальної сфери;

функціональної структури території;

умов життєдіяльності;

ефективності та планування економічної діяльності;

типового проектування.

Цілі статті

Як видно із перерахованих задач, геоінформаційна система (ГІС) опрацьовує різнорідну за структурою та формою подання вхідну інформацію: картографічну, числову, текстову, графічну тощо. Виникає необхідність збереження великих чисел (протяжність кордонів адміністративної

одиниці, чисельність населення, обсяг продукції тощо); надзвичайно малих (величина викидів небезпечних речовин, вміст мінеральних речовин у ґрунті тощо); інтервальних (температура повітря); лінгвістичних оцінок (якість) – оскільки проектована система є людино-машинною, і тому згенеровані рішення повинні видаватися у термінах, звичних для людини; неточних даних (заміри, виконані у надзвичайних умовах).

Засобами сучасних СКБД не можна забезпечити коректного зберігання невизначеної інформації. Крім того, низький рівень достовірності (точності даних) призводить до зниження якості рішень, згенерованих системою.

Отже, **виникають такі задачі**:

1. Коректного подання та зберігання відсутніх, нечітких, недостовірних, інтервальних, лінгвістичних та інших невизначених даних;

2. Коректного опрацювання невизначеності – виникає через ряд факторів:

– застосування операцій інтервальної алгебри призводить до постійного розширення інтервалу, а отже, зниження довіри до результату (наприклад, операція множення для формування результату обирає найменший та найбільший з усіх добутків);

– виконання арифметичних операцій над точними даними призводить до зниження точності через заокруглення;

– застосування операцій над значеннями атрибутів, які містять ступені довіри, призводить до постійного зменшення довіри (грунтується на властивостях логічного І для логік будь-якого порядку).

3. Усунення чи зменшення невизначеності (корекції даних у системі РАПІД) з метою підвищення інформативності вхідних даних та покращання якості прийнятих рішень – прийняття рішень на основі видобування нових знань залежить від якості вхідної інформації (чим більша кількість вхідних даних і чим більший ступінь довіри до них, тим точніше можна встановити залежності між даними).

Для розв'язання поставлених задач пропонується:

Для вирішення проблеми збереження – спроектувати логічні моделі для збереження метаданих, які описуватимуть характер невизначеності та позначатимуть атрибути, яких ця невизначеність стосується;

Для вирішення проблеми опрацювання – забезпечити коректність опрацювання інтервальних, надточних та лінгвістичних даних шляхом збереження історії операції; розширити оператори реляційної алгебри (вибірки, проекції та з'єднання) для врахування фактора невизначеності;

Для вирішення проблеми зменшення невизначеності – використовуючи залежності між характеристиками об'єктів, проводити класифікування, результатом якого буде визначення класу, до якого належить об'єкт, а, отже, – усунення невизначеності.

Базуючись на об'єктно-орієнтованому підході до проектування схем даних та зберігаючи зв'язки між об'єктами, використовувати рух за мережею об'єктів, а також їх властивості для зменшення невизначеності шляхом аналізу характеристик об'єктів, зв'язаних між собою.

Основний матеріал

Структури даних геоінформаційної системи

Мінімальною одиницею, яка використовується при проектуванні геоінформаційних систем, є район [1]. Тому наведемо схему інформаційних потоків району (рис. 1).

Наведемо основні об'єкти та їх характеристики, що використовуються для подання геоінформаційної інформації.

Вхідною інформацією для комплексної ГІС є [2,4]:

Картографічні дані,

Природно-кліматичні умови

Дані соціальної сфери

Дані людської діяльності.

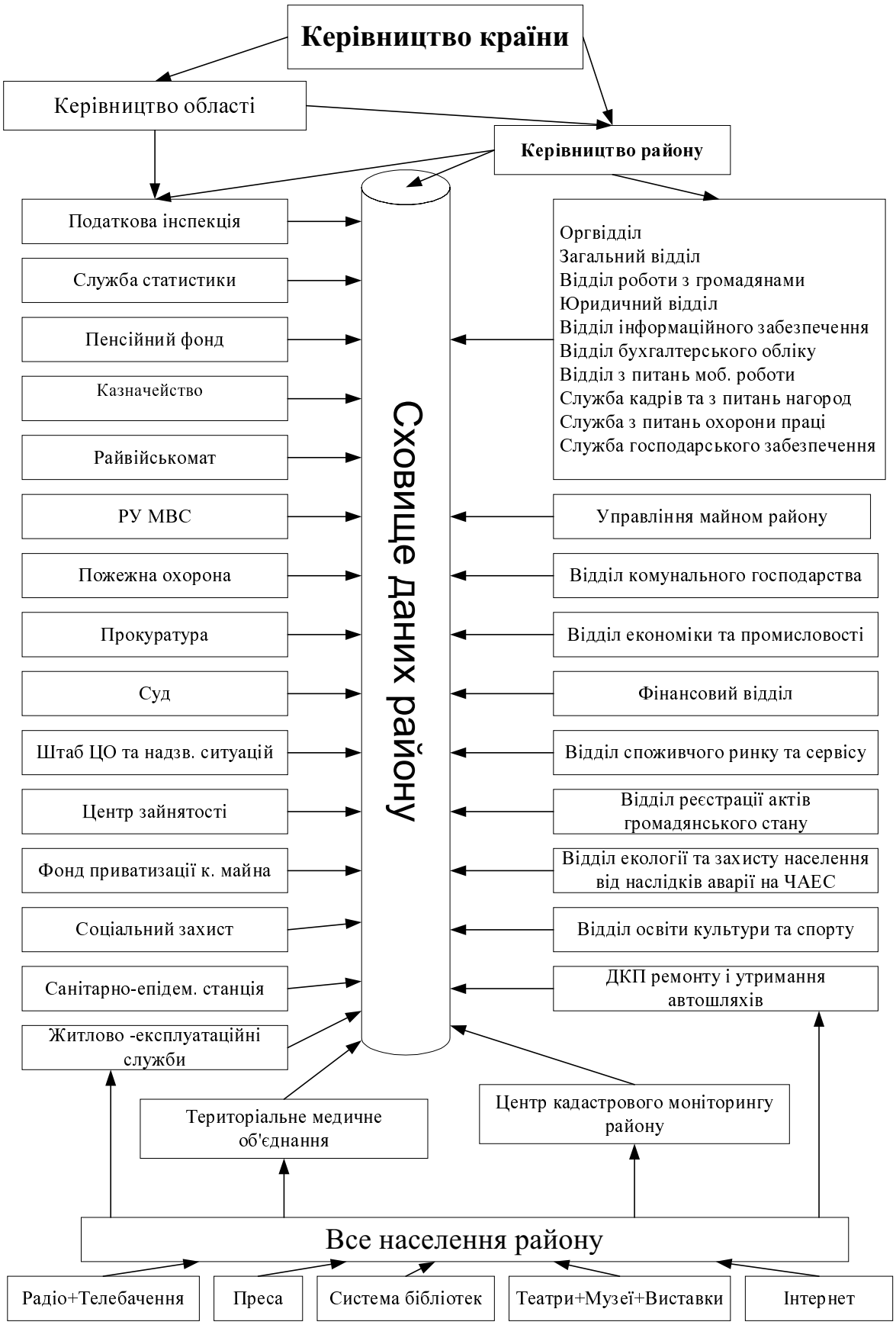

Рис. 1. Блок-схема інформаційних потоків району

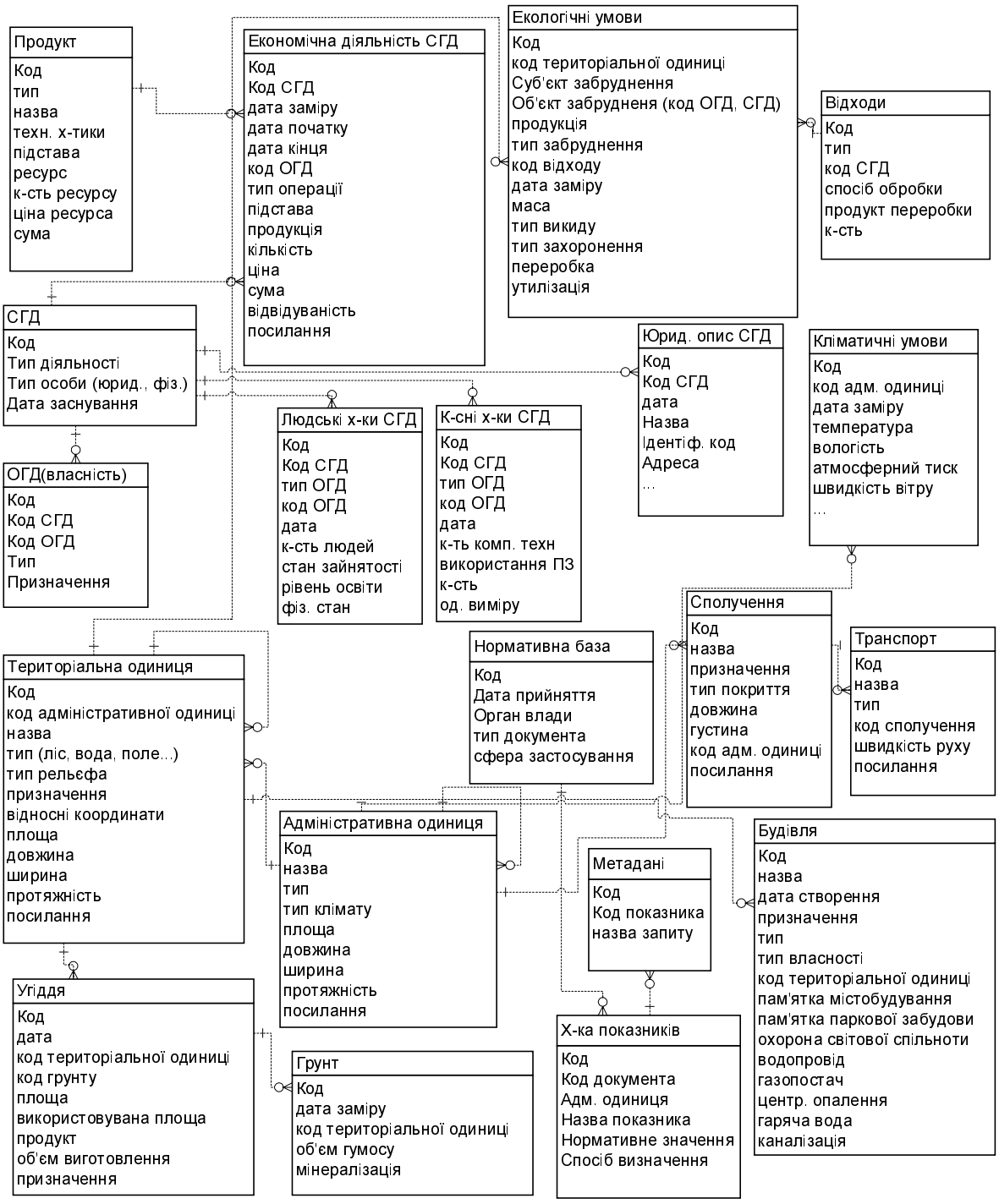

Рис. 2. ERD геоінформаційної системи: СГД – суб'єкт господарської діяльності; ОГД – об'єкт господарської діяльності; Коди ОГД формуються на основі кодів з таблиць Грунт, Будівля, Угіддя, Сполучення; Продукція – назва продуктів праці, послуг, історичних цінностей, рекреаційних цінностей

Як видно із опису поданих об’єктів ГІС-системи, необхідність опрацювання невизначеності зустрічається майже для кожного об’єкта. Тому проаналізуємо проблему подання невизначеної інформації.

Подання та опрацювання невизначених даних

У геоінформаційній області невизначеності виникають через обробку великої кількості числової інформації. Типовими об’єктами, які містять невизначеність, є: кліматичні умови (температура, тиск, відносна вологість), характеристики території (протяжність кордонів, величина

мінералізації, величина викидів небезпечних речовин), якість прийнятих рішень (лінгвістична оцінка, ступінь довіри) тощо. Структура об'єктів подана на рис. 2.

Для опрацювання невизначеностей скористаємося апаратом інтервальної алгебри. Для цього числові значення будемо подавати у вигляді числа та двох значень довіри. Розглянемо для прикладу одне із відношень проектованої системи:

Метеорологічна карта
Код
Регіон
Дата
Температура повітря
Відносна вологість
К-сть викидів
Довіра_нижня_межа
Довіра_верхня_межа

Тут атрибут *Довіра* містить ступінь істинності замірів метеорологічних умов регіонів (тобто відображає невизначеність атрибутів *Температура повітря* та *Відносна вологість*).

Введемо узагальнений тип невизначеності, використання якого не призводило б до порушення цілісності даних, а також дозволяло б подавати усі типи невизначеностей. Таку узагальнену невизначеність можна подавати за допомогою двох складових:

$$Indeterminate=<Value, Trust>,$$

де *Value* – значення невизначеної величини (число, лінгвістична змінна, рядок тощо), *Trust* – довіра до значення (подається інтервалом).

Операції над значеннями, які мають тип *Indeterminate*, виконуються окремо для значення *Value* та окремо для значення *Trust*. Оскільки операції над величинами *Value* не відрізняються від операцій, що виконуються над стандартними типами даних, то зупинимося детальніше на інтервальних операціях, що застосовуються для частини *Trust*.

У загальному випадку точність інтервального результату визначається такими чотирма факторами [5]:

1. Невизначеністю у заданні вихідних даних.
2. Заокругленнями при виконанні операцій, що змінюють або породжують інтервальні об'єкти.
3. Наближеним характером використовуваного чисельного методу.

Ступенем обліку залежностей між інтервальними об'єктами (змінними і константами), що беруть участь в обчисленні.

Для того, щоб інтервал не збільшувався при виконанні над ним математичних операцій (тобто, щоб не збільшувалась невизначеність), подаватимемо межі інтервалу двома атрибутами з доменами дійсних чисел та використовуватимемо для опрацювання інтервалів оператори інтервальної алгебри.

Спочатку введемо поняття інтервалу.

Інтервалом $[a,b], a<b$ назвемо замкнену обмежену підмножину R дійсних чисел виду [5]

$$[a,b]=\{x \mid x\in R \wedge a\le x\le b\}.$$

Множину всіх інтервалів позначимо через $I(R)$. Якщо А – елемент $I(R)$, $A\in I(R)$, то його лівий і правий кінці будемо позначати як $a,\overline{a}: A=[a,\overline{a}]$. Елементи $I(R)$ називаються *інтервальними числами.*

Арифметичні операції над інтервальними числами визначаються так. Нехай $*\in\{+,-,\cdot,/\}$, $A,B\in I(R)$.

Тоді

$$A * B = \{a * b \mid a \in A, b \in B\}, \quad (1)$$

причому у випадку розподілу $0 \notin B$.

Визначення (1) еквівалентне до співвідношень

$$A + B = [a, \bar{a}] + [b, \bar{b}] = [a + b, \bar{a} + \bar{b}], \quad (2)$$

$$A - B = [a, \bar{a}] - [b, \bar{b}] = [a - b, \bar{a} - \bar{b}], \quad (3)$$

$$A \cdot B = [a, \bar{a}] \cdot [b, \bar{b}] = [\min\{ab, \bar{a}b, a\bar{b}, \bar{a}\bar{b}\}, \max\{ab, \bar{a}b, a\bar{b}, \bar{a}\bar{b}\}], \quad (4)$$

$$A / B = [a, \bar{a}] / [b, \bar{b}] = [a, \bar{a}] \cdot [1/\bar{b}, 1/b]. \quad (5)$$

Якщо A і B – вироджені інтервали, то рівності (2)–(5) збігаються зі звичайними арифметичними операціями над дійсними числами. Отже, інтервальне число є узагальненням дійсного числа, а інтервальна арифметика – узагальненням дійсної.

Інтервальні додавання і множення асоціативні і комутативні, інакше кажучи, для $A, B, C \in I(R)$ існують рівності

$$A + (B + C) = (A + B) + C, A + B = B + A$$

$$A \cdot (B \cdot C) = (A \cdot B) \cdot C, A \cdot B = B \cdot A$$

Рівність (1), як і (2) – (5) показує, що якщо один з операндів є невиродженим інтервалом, то результатом арифметичної операції є також невироджений інтервал.

Якщо для подання інтервалу у сховищі даних використовувати два атрибути, то операції додавання та віднімання не будуть відрізнятись від традиційних операцій додавання та віднімання, які використовуються при опрацюванні даних у сховищах даних. Виключення становлять операція множення та пов'язана з нею операція ділення.

Для моделювання невизначеності типу *Indeterminate* у сховищі даних використаємо 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_type> :=<Тип невизначеності атрибута Attr_name >
<Prior_id> :=<Зовнішній ключ відношення Attr>

Відношення attr – це відношення, що містить метадані (тобто, описує дані). Інформація у ньому вважається апріорі чіткою.

У відношенні attr інформація про нечіткість відношення *Метеорологічна карта* записується так (типи невизначеності подані у [3]):

Id	Rel_name	Attr_name	UNK_name	Prior_id
1	Метеорологічна карта	Довіра	нечіткість	
2	Метеорологічна карта	Температура повітря	неточність	
3	Метеорологічна карта	Відносна вологість	неточність	
4	Метеорологічна карта	К-сть викидів	ненадійність	

Переваги використання метаданих для відображення залежностей між атрибутами:

Не потрібно використовувати додаткові відношення для моделювання різних типів невизначеностей.

Не потрібно здійснювати надбудову над реляційною алгеброю (уведення нових операторів), а лише врахувати додаткові умови під час використання традиційних операторів та розробити додатковий інструментарій обробки даних (операції множення-ділення інтервальних значень).

Усунення проблеми подання невизначеності на рівні кортежу дозволить застосувати методи для її зменшення.

Усунення невизначеності

Одним із методів усунення невизначеності є аналіз зв'язків між об'єктами проектованої предметної галузі.

Для зменшення невизначеності шляхом аналізу зв'язків між кортежами відношень необхідно змоделювати структуру, яка б дозволяла зберігати зв'язки довільної арності та забезпечувати доволі простий доступ до інформації на будь-якому рівні ієрархії.

Зв'язок як між кортежами одного відношення, так і між окремими кортежами різних відношень можна встановити за допомогою відношення *dc_link*:

Dc_link

Id	код
Evdate	Дата занесення зв'язку
TableType	Назва або префікс відношення
Table_id	Код кортежу відношення
Prior_id	Посилання на код

Атрибут *Prior_id* є зовнішнім ключем відношення *dc_link* та застосовується для встановлення зв'язку між кортежами відношень, назва яких вказана у *TableType*.

Описана структура (*dc_link* та використання атрибута *prior_id*) дозволяє моделювати складні об'єкти, наприклад, шляхом перерахування їхніх властивостей, та враховувати невизначеність як їхню характеристику, вказавши для кожної властивості ступінь її впливу (відповідності) на об'єкт-предок.

Коли ми говоримо про невизначеності, які є характеристиками об'єкта, доцільно *застосувати об'єктно-орієнтований підхід до проектування схеми бази даних*. Мова йде про об'єкти, які описуються

– через перелічення їхніх складових або властивостей (наприклад, географічні об'єкти, виробничі процеси тощо);

– через вказання зв'язків з іншими об'єктами.

Як бачимо, об'єкти моделюються у вигляді мережевої структури. На рівнях цієї мережі можуть знаходитись стани об'єкта, його властивості тощо.

У такому випадку фактор невизначеності та зв'язок між об'єктами має бути передбачений вже у процесі проектування схеми БД (на відміну від попереднього випадку, де ми говоримо про усунення невизначеності у відношеннях, наповнених тестовим набором [1], тобто існуючих).

Якщо вважати, що за допомогою кортежу відношення подають характеристику об'єкта, його стан у часі або властивість, то моделюванням об'єкта у сховищі даних є встановлення зв'язку між окремими кортежами. А зважаючи на те, що кожен об'єкт є системою певної складності, то за допомогою властивостей складових об'єкта можна визначати властивості самого об'єкта або ж, навпаки, переносити властивості об'єкта на його складові. Тобто, моделювання об'єкта за допомогою перелічення його складових або властивостей та перенесення властивостей з вищого рівня ієрархії на нижчий та навпаки є одним із методів усунення невизначеності його характеристик.

Якщо на рівні описів об'єкта передбачити атрибут для подання ступеня істинності чи відповідності, то моделювання ієрархічної структури об'єктів та їх властивостей у сховищах даних дозволить також позбуватися нечіткостей та неоднозначностей шляхом руху за ієрархією та аналізом ступенів відповідності.

Уведемо ряд операторів для руху мережею записів.

Оператором визначення предка Up назвемо оператор виду

$$Up_{X=x_1}(r) = \sigma_{X=x_1}(\sigma_{X=Y}(r)), \tag{6}$$

де r – відношення (універсальне відношення), у якому зберігається інформація про зв'язки між об'єктами; X – первинний ключ відношення; Y – зовнішній ключ відношення r (вказує на підпорядкування записів, тобто формує мережу); x_1 -– значення первинного ключа запису, для якого здійснюється пошук предка; x_2 – значення зовнішнього ключа, на який здійснює посилання нащадок (код предка).

Оператором визначення нащадка Down назвемо оператор виду

$$Down_{X=x_2}(r) = \sigma_{Y=x_2}(r), \tag{7}$$

X	Y
x_1	x_2
x_2	

Операції визначення предка та нащадка мають такі властивості:

$$Up_{X=x_1}(Down_{X=x_1}(r)) = \sigma_{X=x_1}(r)$$

$$Down_{X=x_1}(Up_{X=x_1}(r)) = \sigma_{X=x_1}(r).$$

Тобто, у результаті послідовного застосування операторів визначення предка та нащадка ми отримаємо традиційний оператор вибірки.

Рухаючись по мережі записів у відношеннях сховища даних, побудованих на основі реляційної моделі, необхідно вказувати спосіб залежностей між записами (визначати предка та нащадка навіть у тому випадку, коли зв'язок між записами є двостороннім). Крім того, потрібно передбачити описи, як саме усувають невизначеність – на основі даних предка чи нащадка.

Формування мережі та встановлення зв'язків між об'єктами дозволяє використовувати властивість об'єктно-орієнтовано підходу – наслідування.

Уведемо оператори усунення невизначеності у мережевій структурі.

Оператор усунення невизначеності за даними предка Heir:

$$\sigma_{X=x_1, Value=val}(r) = Heir(\sigma_{X=x_1, Value=val}(r), Down_{X=x_1, Value=Null}(r)). \tag{8}$$

Приклад: якщо у відношенні є кортеж з відомостями про величину атмосферного тиску певного дня у Львівській області, а відсутня інформація про величину атмосферного тиску у Львові, то за допомогою оператора *Heir* ця інформація може бути отримана.

У випадку усунення невизначеності кортежа-предка на основі аналізу множини кортежів-нащадків необхідно враховувати можливість наявності суперечливої інформації. Тому кортежі-нащадки вимагають попереднього групування за значеннями *Trust.* Всередині групи до значень Value застосовується логічна операція *АБО*, а до груп – логічна операція *І*.

Для прикладу розглянемо фрагмент відношення:

Admin_unit

Регіон	Клімат	Довіра	Посилання
Галичина			
Львів	П	0.9	Галичина
Тернопіль	ПК	0.8	Галичина
Івано-Франківськ	П	0.8	Галичина
Поділля			
Хмельницький	ПК	0.5	Поділля
Вінниця	ПК	0.9	Поділля
Чернівці	П	0.6	Буковина

У цьому відношенні символом “П” позначено помірний клімат, а “ПК” – помірноконтинентальний.

Як бачимо, за атрибутом *Клімат* у нас є невідомі значення. Об’єкти *Галичина, Буковина* та *Поділля* є предками. Ставиться задача усунення невизначеностей типу відсутність, неточність, частковість, багатозначність інтерпретацій та ненадійність даних на основі значень нащадків.

Після виконання кроків:

– аналіз значень атрибутів *Довіра*, *Посилання*,

– визначення кількості появ значень атрибута *Клімат* для об’єктів, у яких значення атрибута *Посилання* відсутнє

– вдалося визначити значення атрибута *Клімат* для об’єктів *Галичина* та *Поділля* (*помірний* та *помірноконтинентальний відповідно*).

Отже, усунення невизначеності кортежів-предків може відбуватися шляхом аналізу кортежів-нащадків.

Висновки

Використання метаданих для опису невизначеностей дозволяє усувати ряд проблем, пов’язаних із виникненням невизначеності на рівні кортежу. Науковою новизною є використання алгоритмів аналізу зв’язків між кортежами відношення для зменшення невизначеності. Уведено узагальнений тип невизначеності, який дозволяє моделювати усі типи невизначеностей, які підтримуються у реляційному відношенні . Практична цінність полягає у розробленні логічних схем даних для збереження зв’язків між атрибутами відношень, правил, метаданих тощо, а також побудови алгоритмів для усунення відсутніх, неоднозначних, неповних та нечітких даних шляхом аналізу мережевої структури.

1. ГІС Форум 2000// Матеріали конференції. – К., 2000. 2. Габрель М.М. Методологічні основи просторової організації містобудівних систем. – Автореф. дис. … д-ра техн. наук. – К., 2002. 3. Шаховська Н.Б. Матоди усунення невизначеностей у базах знань, побудованих на основі реляційного підходу // Вісник Національного університету “Львівська політехніка”. 2003. № 489. 4. Ковтун В.Я., Москаленко Л.В. Средства извлечения геологических знаний из электронных хранилищ данных в геологических фондах / /Труды международного семинара Диалог 2001 по компьютерной лингвистике и её приложениям. – Таруса, 2001. – Т. 2. Прикладные проблемы – www.dialog-2001.asp.ru 5. Алтунин А.Е., Семухин М.В. Модели и алгоритмы принятия решений в нечетких условиях: Монография. – Тюмень: Издательство Тюменского государственного университета, 2000. – 352 с.