Київська міська державна адміністрація
Науково-виробниче об'єднання "Міськсистемотехніка"

На правах рукопису
УДК 681.3

МАЛЮТА ТЕТЯНА ОЛЕКСІЇВНА

МЕТОДИ І ЗАСОБИ РОБОТИ З НЕПОВНИМИ І НЕЧІТКИМИ ЗНАЧЕННЯМИ В СИСТЕМАХ РЕЛЯЦІЙНИХ БАЗ ДАНИХ

Спеціальність 05.13.17
Теоретичні основи інформатики

Автореферат дисертації на здобуття вченого ступеня кандидата технічних наук

Київ 1993

Робота виконана в НВС "Міськсистемотехніка"

Науковий керівник: чл.-кор. АН України, д.ф.-м.н., професор Стогній Анатолій Олександрович.

Офіційні опоненти: д.т.н.,проф. Б.В. Ігнатенко
к.т.н., с.н.с. Н.Д. Ващенко

Провідна організація - Інститут програмних систем науково-технічного комплексу "Інститут кібернетики ім. В.М. Глушкова" АН України (м.Київ)

Захист відбудеться 10 листопада 1993 р. об 14⁰⁰ год. на засіданні Спеціалізованої Ради Д 166.01.01 в науково-виробничому об'єднанні "Міськсистемотехніка" за адресою: 252004, м. Київ, вул. Червоноармійська 23-б.

З дисертацією можна ознайомитись в бібліотеці НВО "Міськсистемотехніка".

Автореферат розісланий "8" 10 1993 р.

Вчений секретар
Спеціалізованої Ради
Д 166.01.01 Гладун В.П.

Дв-28.311

А Н О Т А Ц І Я

Дисертація присвячена вирішенню однієї з проблем підвищення семантичної виразності реляційних баз даних - представленню і обробці в реляційних базах даних неповної, неточної або відсутньої інформації. Ця проблема має важливе значення в зв'язку з поширенням систем штучного інтелекту і експертних систем, інтелектуалізацією систем обробки даних і інформаційних систем, що в свою чергу приводить до зближення і спільного використання "інтелектуальних" систем і систем обробки даних, зокрема, систем управління базами даних. Серед можливих шляхів зближення систем штучного інтелекту і систем управління базами даних - розширення можливостей традиційних баз даних засобами представлення і обробки знань. Оскільки системи штучного інтелекту, експертні системи вимагають можливостей по обробці неточної і нечіткої інформації, то ці можливості повинні бути реалізовані і в системах управління базами даних. В дисертаційній роботі проаналізовані існуючі підходи до розширення **реляційних** моделей баз даних можливостями обробки неповної і нечіткої інформації. В результаті аналізу дається відповідь на запитання: " В чому ж полягає розширення реляційної моделі баз даних для обробки неповної та неточної інформації? ". Розширення реляційної моделі починається не на рівні **реляційної алгебри**, а полягає у розширенні визначення **домена**, введенні в нього нечітких значень. В результаті введення в домен нечітких значень виникає необхідність розширення поняття відношення і, відповідно, пезного розширення реляційної алгебри. При практичному використанні розширення реляційної моделі полягає у введенні в СУБД **нового типу даних**, який вимагає застосування до себе **нових операцій.**

На захист виносяться наступні результати досліджень:

1. Розширення реляційної моделі баз даних для обробки в них неповної інформації.

2. Методика розширення традиційних реляційних систем управління базами даних введенням в них нового типу даних для

представлення і обробки неповної та неточної інформації.

3. Програмно- алгоритмічна реалізація методики на програмних макетах та моделях в середовищі СУБД реляційного типу, а також її практична апробація та впровадження в інформаційних системах, побудованих на основі реляційних баз даних.

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Розповсюдження систем штучного інтелекту і експертних систем вимагає їх використання у сукупності з системами управління базами даних для збереження в останніх даних і знань. СУБД повинні мати можливості для відображення знань і, зокрема, неповних і неточних даних.

Об'єкт дослідження. Поширеною моделлю баз даних є реляційна модель. Поряд з багатьма перевагами над іншими моделями вона має такий недолік, як обмежена семантична виразність. Зокрема, в реляційних базах даних ускладнена обробка неповних, неточних або неіснуючих даних. **Об'єктом дослідження** є реляційна модель даних і її розширення для представлення неповністю визначених даних.

Предмет дослідження. Аналіз можливостей обробки в реляційних базах даних неповних даних проводиться дослідниками з кінця 70-х років. Більшість запропонованих і проаналізованих підходів давала некоректні розширення реляційної моделі. В роботах не було чітко визначено, в чому саме полягає розширення, і яким чином воно має здійснюватись. В результаті проведених дисертантом досліджень та зроблених узагальнень можна стверджувати, що розширення можливостей реляційної моделі баз даних для обробки неповної інформації полягає у розширенні визначення домена введенням в нього неповних даних. При розширенні можливостей традиційних СУБД реляційного типу це полягає у введенні в СУБД нового типу даних, який дозволяє відображати неповноту інформації, разом з операціями над цим типом даних. Введення неповноти даних на рівні домена вимагає відповідного розширення реляційної алгебри. Розширення реляційної алгебри і методика введення нового типу даних в СУБД реляційного типу для обробки неповної інформації і є **предметом дослідження.**

Актуальність проблеми та історія проблеми. Перші пропозиції щодо розширення реляційної моделі баз даних для обробки неповних даних були сформульовані Коддом в кінці 70-х років, який ввів в базу **невідоме значення** /null value/. Модель Кодда не була коректною. В подальших дослідженнях Гранта, Ліпського, Біскапа, Заніоло пропонувались різні шляхи по розширенню реляційної моделі для обробки невідомих значень, але практично всі підходи давали некоректне розширення моделі. Тобто, в рамках введеної семантики невизначеного значення /області визначення/ результат операцій був різним в залежності від порядку обробки неповних відношень: інтерпретація неповного відношення відповідно до його семантики і застосування реляційних операцій до повних баз даних, і навпаки, застосування розширених реляційних операцій до неповного відношення і інтерпретація результату.В подальшому з'явились дослідження по введенню в реляційну модель **неповних, неточних** і **нечітких даних** Буклеса, Петрі, Вонга, Прада, Земанкової. Пропонувались різні підходи до моделювання таких даних: теорія ймовірностей, багатозначна логіка, різні емпіричні оцінки. Неповнота даних вводилась в реляційну модель на різних рівнях- на рівні запиту до традиційної реляційної бази даних, на рівні кортежів, коли існує невизначеність щодо цілого кортежу, належить він базі, чи ні. Проведений дисертантом аналіз показав, що найбільші виразні можливості забезпечує введення невизначеності на рівні значень атрибутів бази. Серед засобів моделювання невизначеностей дисертант виділяє емпіричні оцінки, які не вимагають великої попередньої роботи по їх визначенню, і які найбільше відповідають застосуванню баз даних в "інтелектуалізованих" інформаційних системах. Серед емпіричних оцінок найбільш розробленою є теорія можливостей і нечітких множин. В останні роки з'явились роботи по реалізації розширень існуючих систем обробки даних можливостями обробки неповної інформації і по створенню окремих систем, які включають такі можливості. З'явились роботи, присвячені вирішенню конкретних прикладних задач, в яких реалізовано потрібний для вирішення задачі, обмежений набір можливостей по обробці неповних даних. В

даний час актуальною є розробка загальної методики введення нового типу нечітких даних в існуючі СУБД реляційного типу, що дозволяло би реалізувати більшість можливих випадків виникнення неповноти або нечіткості даних.

Методи дослідження. Методом дослідження є об'єктно-орієнтований підхід, який використовується при формуванні нового типу даних і введенні його в існучі СУБД реляційного типу. Для моделювання неповноти і неточності даних використовується теорія можливостей і нечітких множин, яка є добре розробленою і дозволяє моделювати переважну більшість можливих ситуацій з неповнотою інформації.

Ціль дисертації. Дослідження, виконані в дисертації, орієнтовані на розширення можливостей існуючих СУБД реляційного типу засобами обробки нечіткої і неповної інформації. **Ціль дисертації**-аналіз та дослідження можливих варіантів розширення реляційної моделі баз даних для представлення і обробки неповної та нечіткої інформації в базах даних реляційного типу. Розробка методики такого розширення і її реалізація в вигляді програмних макетів та програмно-алгоритмічних моделей, які дозволяють відпрацювати запропоновані дисертантом підходи на практиці. В дисертації автором викладені, обгрунтовані і винесені на захист наступні основні результати:

Теоретичні:

аналіз та дослідження розширень реляційної моделі бази даних для обробки неповної та неточної інформації та введення в СУБД реляційного класу нового типу даних, який дозволяє представляти і обробляти недовизначені дані.

Прикладні:

розробка методики введення нового типу даних для представлення неповної інформації в існуючі СУБД реляційного типу і реалізація методики на програмних макетах та програмно-алгоритмічних моделях в середовищі реляційної СУБД .

Наукова новизна. В дослідженнях, які проводились по розширенню реляційної моделі баз даних для обробки неповних і неточних даних, не був визначений зміст "розширення", що приводи-

до некоректного розширення алгебри. **Наукова новизна** даної дисертаційної роботи полягає в аналізі змісту поняття "розширення" реляційної моделі для обробки неповних даних, розширенні реляційної моделі, яке починається на рівні доменів і розповсюджується на рівень реляційної алгебри, можливістю обробки неповних даних в рамках запропонованої моделі. Це твердження виступає методологічною базою для розробки "розширеної" СУБД реляційного типу. Вперше запропонована методика розширення можливостей існуючої СУБД реляційного типу для обробки неповної і неточної інформації.

Практична цінність. Дисертація була виконана в руслі досліджень, які проводились дисертантом в період з 1987 по 1992 р.р. в Львівському політехнічному інституті в НДЛ- 45 /Обчислювальний центр / і на кафедрі "Автоматизовані системи управління" по замовленню ДКНТ, Міністерства освіти України, Фізико- механічного інституту АН України, виробничих об'єднань "Ювелірпром", "Світоч", Дрогобицького автокранового заводу.

Результати дисертації впроваджені на Львівському ювелірному заводі, Одеському ювелірному заводі, Львівському ВО "Світоч", Дрогобицькому автокрановому заводі, Львівському політехнічному інституті в процесі виконання цільових замовлень цих організацій в рамках господарських договорів та в учбовому процесі під час викладання курсу "Бази та банки даних і знань".

Рекомендації по використанню. Запропонована методика може бути використана для розширення функціональних можливостей існуючих СУБД реляційного типу засобами обробки нечіткої і неповної інформації.

Апробація. Основні результати досліджень доповідались автором на міжнародних, всесоюзних та республіканських конференціях і семінарах:

-міжнародній конференції з математичних основ баз даних /Берлін, 1989/;

- міжнародній науковій конференції з інтелектуальних систем управління /Варна, 1989/;

- всесоюзній конференції "Системи баз даних і знань"

/Калінін, 1989/;

- міжнародній конференції "Системи баз даних і знань" /Львів, 1991/;

- республіканській конференції "Проблемно- орієнтовані діалогові системи" /Батумі, 1988/;

- спеціальних семінарах в Львівському політехнічному інституті, НВО "Міськсистемотехніка", інституті кібернетики ім. В.М. Глушкова АН України.

Матеріали і результати дисертації опубліковані на Україні і за кордоном. Загальна кількість публікацій по темі дисертації- 10.

Структура дисертації. Дисертація складається з вступу, чотирьох глав, заключення і додатків. Об'єм дисертації- стор., тексту- стор., ілюстрацій- . Список використаної дисертантом літератури складає 60 першоджерел. В заключенні сформульовані основні результати, висновки і погляди автора на розвиток досліджень в області моделей база даних і знань реляційного типу. В додатку наводяться програмні макети та програмно-алгоритмічні моделі з реалізації розширення функціональних можливостей СУБД реляційного типу для обробки неповної інформації.

ЗМІСТ РОБОТИ

У вступі проаналізовані підходи до рішення проблеми представлення неповної та неточної інформації в реляційних базах даних. Їх можна умовно розділити на дві групи: ті, що розглядають випадки повної відсутності інформації про значення атрибута, т.з. **невідомі значення** /сюди ж можна віднести інтервали можливих значень/, і підходи, в яких розглядаються способи представлення і обробки **частково невизначених значень**. Цей поділ відображає також хронологічний розвиток досліджень.

Введемо основні позначення і визначення, які використовуються в дисертації. Характеристики, що описують об'єкти реального світу, називаються атрибутами /позначаються A, B,.../, їх множина позначається U. Кожний атрибут приймає значення з деякої множини, яка називається доменом атрибута /позначається D/. Функція DOM: $A_1 \to D_1$ задає відповідність між множиною атрибутів і доменами. Під

атрибутом розуміється пара $\langle A_1, D_1 \rangle$. Відношенням називається властивість, яка виділяє певну підмножину з декартового добутку доменів: $r \subseteq D_1 x D_2 ... x D_n$. Кортежем називається впорядкована послідовність $\langle d_1, d_2, ..., d_n \rangle$, в якій $d_1 \subseteq D_1$. Кожному відношенню можна поставити у відповідність предикат, який приймає значення "істина", коли виконується відношення, і приймає значення "хиба" в інших випадках. Множину предикатів на відношеннях позначимо Σ. Тоді визначаємо реляційну модель як четвірку $\langle U, D, DOM, \Sigma \rangle$, а реляційну базу даних як п'ятірку $\langle U, D, DOM, \Sigma, \Omega \rangle$, де Ω-множина операцій на множині атрибутів і доменів.

Що означає "неповнота" і "нечіткість" інформації в базах даних? Реляційна база даних- це представлення певної реальної предметної області /ПО/. Вважаєм, що ПО описується п'ятіркою $\langle U^0, D^0, DOM^0, \Sigma^0, \Omega^0 \rangle$. Існує наша система уявлень про ПО-представлення ПО /ППО/, яка описується п'ятіркою $\langle U^1, D^1, DOM^1, \Sigma^1, \Omega^1 \rangle$, отриманою з ПО відображенням φ_1. В результаті деякого відображення φ_2 будується реляційна база даних $\langle U^2, D^2, DOM^2, \Sigma^2, \Omega^2 \rangle$, яка використовується, як представлення ПО. Реляційна база даних фактично є результатом суперпозиції двох відображень φ_1 і φ_2. У випадках, коли можна вважати, що ППО співпадає з ПО, ми отримуємо реляційну базу даних, яку називаєм "традиційною". Але коли ППО неповністю відповідає ПО, ми отримуємо реляційну базу даних, яка, в свою чергу, відображає це неповне знання про ПО. В дисертаційній роботі розглядається випадок, коли нема повчих знань про систему відношень Σ, тобто нема повної інформації про властивість ПО, яка формує відношення. Неповнота інформації може зустрічатись на різних рівнях: невідомо, чи властивість притаманна ПО- невизначеність на рівні відношення, відомо, що властивість притаманна ПО, але невідомо, чи притаманна даному об'єкту- невизначеність на рівні кортежів, відомо, що властивість притаманна ПО і об'єкту, але невідомо, як вона на об'єкті проявляється- невизначеність на рівні значень атрибутів. В дисертаційній роботі розглянуто останні два випадки невизначеності.

Тобто, одна з причин виникнення неповноти або неточності

інформації в базі даних /як і в довільному, зв'язаному чи не зв'язаному з комп'ютерними технологіями відображенню ПО/- неточні, суб'єктивні представлення про реальність.

Друга причина виникнення неповноти або неточності інформації- використання розмитих категорій, наприклад, **МОЛОДИЙ, ВЕЛИКИЙ.** Використання таких категорій може бути наслідком першої причини, а може бути викликано необхідністю спрощення моделі для забезпечення її прозорості. Тобто, навіть якщо є повна інформація про предметну область, відображення φ_1 сформує представлення предметної області з неповнотою інформації внаслідок абстрагування і узагальнення.

Ці дві причини /які є взаємозв'язані/ приводять до необхідності використовувати дані, які не можуть бути представлені, як точні, чітко визначені значення.

В **першому розділі** розглянуті підходи до обробки невідомих значень.Вперше розширення реляційної алгебри на відношення з невідомими значеннями було запропоновано Коддом. Для обробки невідомих значень він використовував тризначну логіку з третім значенням істиності "невідомо" /позначається ω/. Підхід грунтується на принципі заміни невизначеностей, який визначає умови, при яких логічний вираз приймає значення ω. Запит до бази має два результати: TRUE- результат /кортежі, на яких умова запиту дає значення "істина", і MAYBE- результат /кортежі, на яких умова запиту дає значення ω/.

В публікаціях по реляційних базах даних наводиться ряд прикладів некоректності запропонованого підходу /деякі теоретико-множинні операції над відношеннями, тавтологія в умові запиту і т.п./ Це викликано тим, що підхід не є достатньо обгрунтованим: в ньому відсутнє формальне визначення семантики невідомого значення, є протиріччя в логічних основах підходу . В подальшому проводились більш детальні проробки розширень реляційної моделі. Всі вони, по-перше, базуються на припущенні, що в базі даних зберігається не само відношення, а множина тверджень про відношення, і, по-друге, відрізняються способами формального визначення семантики невідомого значення.

Грант вводить поняття області визначення невідомого значення, яка задає множину правильних підстановок невідомого значення. Вводиться третє значення істинності- "невідомо". Особливість підходу Гранта заключається в тому, що він залишається в рамках двозначної логіки. Для цього кожному предикату P ставиться у відповідність два предиката- P_T і P_M /TRUE- і MAYBE- предикати відповідно/, де P_T приймає значення "істина" тоді і тільки тоді, коли P є істиним для всіх вірних підстановок для невизначеностей в P, а P_M є істиним тоді і тільки тоді, коли P є істиним хоча б для однієї вірної підстановки для невизначеностей в P. На підставі цих загальних положень вираховуються P_T і P_M для різних предикатів на значеннях атрибутів, атрибутах, таблицях, а також даються визначення розширених реляційних операцій. Отримана модель досліджується на коректність. Виявилось, що результати, які отримуються в результаті застосування введених операцій, в загальному випадку не можуть бути отримані при знаходженні спочатку всіх можливих розширень тих таблиць, що обробляються, а потім застосуванням звичайних операторів. Розширена модель Гранта також не є коректною.

Велике дослідження можливостей коректного розширення реляційної моделі на бази даних з невідомими значеннями проведене Ліпським та Імелінським. Як критерій коректності розширеної моделі було запропоновано поняття репрезентативної системи. Репрезентативною системою називається трійка < T, Rep, Ω >, де T- множина неповних відношень, Rep- відображення, яке ставить у відповідність кожному неповному відношенню деяку множину відношень, Ω- множина реляційних операторів, які використовуються. Для Ω- виразів f слід було б чекати, щоб виконувалось \Rep(f(T)) = f(Rep(T)). Ця вимога виявилась дуже сильною і автори вводять більш слабу Ω- еквівалентність таким чином, щоб Rep(f(T)) апроксимувало f(Rep(T)). Тоді приведена вище трійка буде репрезентативною системою, якщо для любого неповного відношення R буде виконуватись $Rep(f(R)) =_\Omega f(Rep(T))$. Показано, що для неповних відношень Кодда репрезентативна система може включати тільки дві операції- проекцію і об'єднання. Однією з причин того, що на неповних

відношеннях Кодда не може коректно виконуватись з'єднання, є те, що неможливо відобразити нерівність двох невідомих значень. Показано, що введення в відношення додаткової інформації про нерівність невідомих значень між собою /вибір невідомих значень з деякої нескінченної множини невідомих значень/ може суттєво розширити набір коректно виконуваних розширених алгебраїчних операцій. В репрезентативну систему для таких відношень включаються операції проекції, вибору з додатковою умовою, об'єднання і з'єднання.

Біскап також використовує в своїх роботах поняття репрезентативної системи, але вводить його відмінним від Ліпського способом. Фактично, підхід Бискапа можна вважати обгрунтуванням підходу Кодда і MAYBE- результатів. В цьому випадку допускається можливість і достовірність цілих кортежів, для чого вводиться додатковий атрибут STATUS, який може приймати два значення m і d /m- кортеж можливо належить результату, d- кортеж достовірно належить результату/. В рамках розширення реляційної алгебри Біскапа всі окремі реляційні операції виконуються коректно, але підхід не розповсюджується на реляційні вирази, як у Ліпського.

В роботах Заніоло було показано, що ряд логічних і теоретико-множинних проблем може бути вирішений за допомогою більш примітивної інтерпретації невідомого значення. Невідоме значення трактується, як повна відсутність інформації, що не дозволяє зробити висновок чи є це значення "невідомим", чи "неіснуючим". Розширення реляційної алгебри по Заніоло не вимагає суттєвого ускладнення механізмів обробки запитів. Однак, на думку дисертанта, така інтерпретація є досить штучною і не дозволяє побачити ряд проблем. Крім того, в роботі Кеплера показано, що Заніоло не вдалось вирішити теоретико-множинні проблеми розширеної моделі, і, крім того, отримано більш загальний результат, який полягає в слідуючому: не можна побудувати розширення реляційної алгебри, яке грунтується на ідеї поповнення невідомого значення, для якого виконувались би всі теоретико- множинні властивості.

В роботі Вассілю розглядається моделювання неповноти інформації за допомогою денотаційної семантики, причому

враховується також інтерпретація відсутнього значення, як "неіснуючого". Множина значень істинності розширюється слідуючим чином: $T^0 = T \cup \{ bot, top \}$, де $T = \{ "істина", "хибя" \}$. Елемент bot апроксимує будь- який елемент T і його інтерпретують, як "значення невідоме", елемент top апроксимується будь- яким елементом з T і інтерпретується, як "значення не існує". Підхід відрізняється від використання чотиризначної логіки. Коректно обробляються запити, в яких умова є тавтологією.

Проаналізовані підходи до обробки невідомих значень в реляційних базах даних відрізняються визначенням семантики невідомого значення, але всі базуються на ідеї поповнення невідомого значення з області його визначення. Запропоновані розширення реляційних операцій, які не є коректними. З точки зору дисертанта недоліки запропонованих моделей витікають з того, що при виконанні реляційних операцій до невідомих значень застосовуються ті ж оператори, що і до значень бази, які визначаються традиційними типами даних /числа, символи і т.п./. Невідомі значення треба розглядати, як дані іншого типу, і застосовувати до них відповідні, визначені для них оператори.

В **другому розділі** проаналізовані підходи по представленню в базах даних неповної або частково визначеної інформації. Для моделювання частково визначеної інформації використовуються теорія ймовірностей, нечіткі множини, багатозначні логіки.

Серед підходів, які використовують ймовірнісне представлення неповноти даних, особливе місце займає підхід Вонга, який розглядає невизначеність ніби "над" базою даних. В базу заносяться тільки точні дані. Крім того, існують апріорні ймовірнісні розподіли, які зв'язують атрибути з іншими характеристиками предметної області /які теж можуть зберігатись в базі/. На підставі відомих апріорних розподілів ймовірностей відбувається попередня статистична обробка запитів, потім перетворений запит адресується до основної бази даних. Підхід можна використовувати в тих випадках, коли неповні дані моделюються за допомогою ймовірнісних розподілів і при проектуванні бази даних можна розділити точні і ймовірнісні дані.

В ряді робіт ймовірнісні розподіли вводяться на рівень кортежів- в базу додається атрибут, значення якого визначають ймовірність того, що відповідний кортеж належить до відношення.

Використання ймовірнісних методів ускладнюється необхідністю обробляти великі об'єми даних при визначенні розподілів. Крім того, теорія ймовірностей є дуже нормативною для того, щоб використовуватись для моделювання невизначеностей деякої природи /наприклад, лінгвістичних невизначеностей/. Більше розповсюдження отримали методи моделювання неповної інформації за допомогою емпіричних оцінок, в першу чергу- за допомогою нечіткостей. Існують підходи, які зберігають однорідність бази даних. Найпростішим способом введення нечіткостей в базу даних є використання їх на рівні кортежів аналогічно ймовірнісним розподілам. Це не вимагає суттєвого ускладнення механізмів обробки даних. Такий підхід можна використовувати і у випадку, коли невизначеність зустрічається тільки в запитах при використанні нечітких понять або нечітких відношень /наприклад, вибрати з бази всіх молодих осіб, або вибрати всіх осіб з близькими інтересами/. Тоді дані зберігаються в звичайному вигляді, а окремі відношення визначають нечіткі поняття /наприклад, **МОЛОДИЙ**/ або матрицю близькості між значеннями атрибута /наприклад, **ІНТЕРЕСИ**/. В ряді робіт детально розроблено використання нечіткого відношення "близькості" або "подібності" замість відношення рівності.

Найбільші можливості для представлення неповноти даних мають неоднорідні бази даних, коли невизначеність вводиться на рівень значень у відношеннях. Такий підхід дає можливість відображати в базах наступні випадки неповноти даних:

- значення знаходиться в інтервалі або є одним з дискретної множини значень, в тому числі сюди відноситься невідоме значення;
- значення не існує;
- є неповна або часткова інформація про значення, яка представляється за допомогою розподілу емпіричних оцінок або нечіткого поняття.

Детально розроблено використання для представлення неповноти даних теорії можливостей і апарата нечітких множин Заде. Значення

атрибутів мають двоїсту оцінку- можливість і необхідність виникнення саме цього значення для даного кортежа. Результат обробки кожного запиту буде включати два відношення: кортежі, які "можливо" належать результату, і кортежі, які "необхідно" належать результату.В ряді робіт для моделювання неповноти використовується багатозначна логіка.

Класифікацію проаналізованих підходів можна представити схематично:

Виникнення неповноти інформації на рівні запиту або кортежу не вимагає суттєвого ускладнення обробки запиту до бази, але не може охопити всі випадки неповноти даних. Найбільші виразні можливості забезпечує введення неповноти інформації на рівень значень атрибутів бази. Серед проаналізованих підходів до моделювання неповноти інформації дисертант виділяє емпіричні оцінки.

В **третьому розділі** дається обгрунтування вибору інструментарію для моделювання неповноти і нечіткості даних і формулюється суть поняття "розширення" реляційної моделі для частково визначених даних.На думку дисертанта одним з найбільш цікавих з точки зору практичного використання може бути підхід, який спирається на апарат теорії можливостей . Він дозволяє змоделювати більшість випадків неповноти даних, включаючи неіснуючі значення, і нечіткі запити до бази. Серед робіт, в яких описується реалізація використання в реляційних базах даних неповних даних і обробки

нечітких запитів, переважають підходи, які спираються на апарат теорії можливостей або на його аналоги. Такого підходу вимагали і розробки, викликані реальними задачами і потребами. З нашої точки зору саме такий підхід може бути корисним при використанні реляційних баз даних в "інтелектуалізованих" інформаційних системах або при стикуванні їх з цими системами, оскільки з його допомогою добре відображаються саме емпіричні, експертні оцінки і судження, і він природньо інтегрується з методами представлення знань в "інтелектуалізованих" інформаційних системах. Крім того, в теорії можливостей люба подія отримує **двоїсту** оцінку: **необхідність** цієї події і її **можливість**, що дозволяє більш адекватно і повно оцінювати нечіткі дані. Саме ця особливість теорії можливостей не використовувалась в проаналізованих нами публікаціях по розширенню реляційних моделей баз даних.

Мірою можливості Π на множині U буде функція $P(U \to [0,1]$, де $P(U)$ - множина підмножин W, така, що $\Pi(\varnothing)=0$, $\Pi(U)=1$, $\forall A,B \in P(U)$, $\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$. Як наслідок $\forall A \in P(W)$, $\max(\Pi(A), \Pi(\bar{A}))=1$.

Коли заданий розподіл можливостей $\pi: W \to [0,1]$, міра можливості визначається наступним чином: $\forall A \in P(W)$, $\Pi(A) = \sup_{W \in A} \pi(W)$.

На основі можливості будується міра необхідності $\forall A \in P(W)$, $N(A)=1-\Pi(\bar{A})$.

Для N виконується $\forall A \in P(W)$, $\min (N(A), N(\bar{A}))=0$, $\forall A \in P(W)$, $N(A) = \inf_{W \in \bar{W}} [1-\pi(W)]$.

Визначення можливості та необхідності можуть бути розширені на використання нечітких множин.

Інформація про значення атрибута A буде визначатись розподілами можливостей $\pi_{t(A)}$ для кортежа t на $D(A) \cup \{e\}$, де e - додатковий елемент, який відповідає випадку, коли для даного об'єкта значення не існує, тобто $\pi_{t(A)} : D(A) \cup \{e\} \to [0,1]$.

Результат операцій на розширеній базі даних буде складатися з кортежів, обов'язково задовільняючих результату операції та можливо задовільняючих результату /відповідає TRUE і

MAYBE-результату/, де їх ступінь належності буде відповідати мірі необхідності та можливості відповідно. Визначені формули для оцінювання атомарних і складних нечітких умов при обробці запитів.

Для демонстрації приведених положень в якості прикладу використаємо відношення **ОСОБИ**:

ОСОБИ	ІМ'Я	ВІК	РІСТ	КОЛІР ВОЛОССЯ
	ІВАНЧУК	25	180	0.9/блондин + 1/русявий
	ПЕТРЕНКО	[20-25]	високий	-
	СИДІР	молодий	190	русявий

Для охоплення всіх можливих ситуацій виникнення нечіткостей при роботі з РБД треба забезпечити можливість представлення і обробки:

- нечітких і неповних даних /невідомі значення, інтервали значень, нечіткі підмножини, неіснуючі значення/;
- нечітких операторів /наприклад, оператори подібності, близькості/;
- лінгвістичних змінних, тобто ідентифікаторів нечітких підмножин, на доменах атрибутів /наприклад, **МОЛОДИЙ**/;
- модифікаторів нечітких операторів і лінгвістичних змінних /наприклад, **ДУЖЕ МОЛОДИЙ, НАБАГАТО СТАРШИЙ**/;
- матриць близькості на значеннях домена /наприклад, близькість кольорів волосся, яка використовувалась би в запитах з умовою типу **ПОДІБНИЙ ДО** блондина/.

Крім того, бажано було б дати можливість користувачу формувати власні нечіткі поняття, нечіткі оператори і нечіткі модифікатори. Ні в одній з розглянутих нами робіт ці можливості не були розроблені повністю.

В дисертаційній роботі запропонована методика розширення традиційної реляційної моделі даних для представлення і обробки неповних даних і нечітких запитів, яка спирається на апарат теорії можливостей для представлення нечіткостей і дозволяє змоделювати всі перераховані вище випадки нечіткостей і невідомих та неіснуючих значень.

В чому полягає "розширення" традиційної моделі реляційної бази даних при спробі реалізувати роботу з нечіткостями? Можна стверджувати, що основні положення реляційної алгебри, визначення реляційних операцій достатньо загальні і не накладають обмежень на зміст бази даних і запити до неї. Традиційно використовувався обмежений набір типів елементів доменів атрибутів бази, обмеженим був набір операторів, які використовувались при формуванні умов для реляційних операцій. Фактично розширення реляційної моделі полягає у розширенні визначення домена, включенні в нього неповністю визначених значень. На рівні СУБД це означатиме введення нового типу даних, який дозволяє задавати в базі даних неповністю визначені дані, і операцій над цим типом даних. Цей тип даних будем називати **нечітким**.

Введення неточних даних в реляційну базу даних на рівні значень атрибутів тягне за собою необхідність введення неповноти інформації на рівні кортежів, тобто, необхідність подальшого розширення визначення відношення, і, відповідно, розширення реляційної алгебри.

Тобто, розширення реляційної моделі почалося з розширення визначення домена і розповсюдилось на визначення відношення і, відповідно, на реляційну алгебру над розширеними відношеннями. Крім множини атрибутів, які входять в схему відношення, в розширеному відношенні розглядаються два додаткові атрибута POS і NEC, домени яких є одиничним інтервалом, і які відповідають можливості та необхідності включення кортежів у відношення:

$$U^* = U \cup POS \cup NEC$$

$$D_{POS} = D_{NEC} = [0, 1]$$

В розширеному відношенні R^* виділяємо дві частини: інформаційну, яка визначається схемою відношення, і характеристичну, яка визначається додатковими атрибутами POS і NEC.

Підхід до обробки неповних даних, як до даних нового типу, дозволяє уникнути некоректності при "розширенні" реляційної моделі баз даних. Використання теорії можливостей, як інструмента для моделювання невизначеностей, дозволяє охопити більшість випадків

невизначеності даних, а використання двоїстої міри невизначеності забезпечує найбільш адекватну оцінку результатів обробки запитів до бази.

В **четвертому розділі** описується методика введення в реляційну модель бази даних нового типу даних для обробки неповних і неточних даних.

Ми пропонуємо методику формування такого типу даних, який в подальшому іменується **НЕЧІТКІСТЬ**, засобами СУБД реляційного типу. При включенні в систему нового типу даних пропонується скористатись можливостями об'єктно- орієнтованого підходу, якщо вони передбачені в системі, або змоделювати його засобами інструментальної системи. Введення нового типу даних в систему включає наступні етапи.

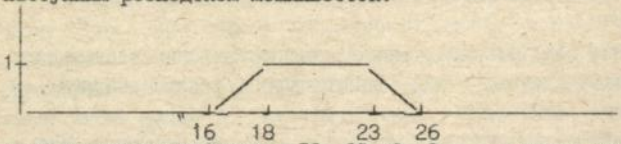
Реєстрація нового типу даних. Задання нового типу даних **НЕЧІТКІСТЬ** полягає в реєстрації його в системі, а також у визначенні набору операторів і функцій, які можуть застосовуватись до даних нового типу. Для систем, в яких не передбачено включення нових типів даних , реєстрація може проводитись штучно. Наприклад, створюється відношення **НЕЧІТКІСТЬ:**

НЕЧІТКІСТЬ. АТРИБУТ	ДОМЕН	ДОВЖИНА	ТОЧНІСТЬ
ВІК	N	3	0
РІСТ	N	3	0
КОЛ.ВОЛ.	C	IO	-

В **АТРИБУТ** заносяться назви атрибутів, домени яких будуть визначатись на новому типі даних, в **ДОМЕН**- базовий домен, на якому будуть визначатись нечіткі дані /N- числовий, C- символьний/.

Внутрішнє і зовнішнє представлення даних нового типу. Зовнішнє представлення даних типу **НЕЧІТКІСТЬ** є символьним. Треба визначити внутрішнє представлення даних цього типу, яке б забезпечувало можливість їх обробки.В основу вибраного підходу покладено функцію розподілу можливостей над доменом D атрибута: $\pi(d)$, $d \in D$, за допомогою якої моделюються нечіткі поняття і нечіткі оператори. Для практичних застосувань її зручно представляти функцією трапецієдальної форми, наприклад, нечітке поняття **МОЛОДИЙ** над доменом атрибуту **ВІК** буде характеризуватись

наступним розподілом можливостей:



I буде записуватись четвіркою < I8, 23, 2, 3 >, де перше число відповідає початку інтервалу, на якому функція приймає значення I, друге число- кінцю цього інтервалу, третє і четверте- відхилення від першої і другої точок відповідно, на яких функція приймає ненульове значення. Таке представлення функції розподілу можливостей є цілком достатнім для більшості застосувань, оскільки невеликі зміни форми функції /яка представляє, як правило, суб'єктивну оцінку події або явища і не може вважатись абсолютно точною/ не впливають сильно на результат запиту. Невідоме значення представляється спеціальною функцією розподілу $\pi_1(d) = 1 \ \forall d \in D$. Неіснуюче значення представляється функцією $\pi_0(d) = 0 \ \forall d \in D$. Розроблені програмно- алгоритмічніпроцедури перетворення даних з внутрішнього представлення в зовнішнє і навпаки. Для атрибутів, для яких базовим є домен скалярів, використовується явне задання розподілу можливостей.

Константи нового типу даних. У випадку визначення на новому типі даних констант, останні теж мають бути зареєстровані в системі. Для нечітких даних константами виступають виділені назви нечітких множин, наприклад, **МОЛОДИЙ**. Вони заносяться у відношення **КОНСТАНТИ**, яке має наступну структуру:

КОНСТАНТИ	АТРИБУТ	НАЗВА	T1	T2	T3	T4
	ВІК	МОЛОДИЙ	I8	23	2	3

де в колонку НАЗВА заносяться зафіксовані назви нечітких множин відповідного атрибуту, а T1- T4 трапецієдальний розподіл можливостей внутрішнього представлення константи.

Оператори. Для нового типу даних задаються оператори, які можуть застосовуватись до них,зокрема, арифметичні оператори і оператори порівняння /для нечітких даних над базовими числовими

доменами/, оператори близькості /для нечітких даних над скалярними базовими доменами/, і інші, які користувач захоче визначити над нечіткими даними. Арифметичні оператори реалізовані у вигляді стандартних процедур і можуть бути включені в систему. Оператори порівняння і близькості, як і нечіткі константи, моделюються за допомогою трапецієдальних розподілів, але розподіли залежать від аргументів операторів. Оператори визначаються за допомогою таблиці **ОПЕРАТОРИ:**

ОПЕРАТОРИ НАЗВА	Т1	Т2	Т3	Т4	АРГУМ	ВИРАЗ
БІЛЬШЕ	0	2	0	1.5	2	|a1 - a2|

де ОПЕРАТОР- назва оператору, Т1- Т4- четвірка розподілу, АРГУМ- кількість аргументів оператора, ВИРАЗ- результуючий вираз від аргументів, від якого залежить розподіл.

Модифікатори операторів і нечітких множин. Інколи виникає необхідність використовувати складні оператори або константи, наприклад, ДУЖЕ МОЛОДИЙ, НАБАГАТО СТАРШИЙ, і т.п. Їх можна задавати звичайним чином. У випадку, коли одна частина оператора або константи виступає модифікатором другої, тобто змінює функціонально її розподіл можливостей, то модифікатор для зручності задаємо окремо у відношенні **МОДИФІКАТОРИ:**

МОДИФІКАТОРИ НАЗВА	ФУНКЦІЯ
ДУЖЕ	f

де НАЗВА визначає назву модифікатора, а ФУНКЦІЯ- назву функції від розподілу оператора або константи, якій передує модифікатор.

Матриці близькості. Для обробки запитів, в умови яких включений атрибут над нечітким доменом з базовим скалярним доменом, потрібна інформація про близькість значень домена між собою. Матриця близькості на значеннях атрибута задається відношенням **МАТРИЦЯ:**

МАТРИЦЯ АТРИБУТ	ЗНАЧЕННЯ 1	ЗНАЧЕННЯ 2	БЛИЗЬКІСТЬ
КОЛІР ВОЛ.	БЛОНДИН	РУСЯВИЙ	0.9

де АТРИБУТ визначає атрибут на якому задається матриця близькості, ЗНАЧЕННЯ 1 і ЗНАЧЕННЯ 2- значення з домену цього атрибуту, між якими визначається близькість.

Обробка даних типу НЕЧІТКІСТЬ. Методика була реалізована на СУБД Clipper, в якій не передбачено введення нових типів даних. Для обробки запитів до бази були написані процедури, які реалізують основні реляційні операції над нечіткими даними. Для зручності цим операціям були дані окремі назви: FPROJECTION, FUNION, FJOIN, FSELECT /префікс F відповідає англійському терміну нечіткий- fuzzy/.Ці процедури обробляються препроцесором Clipper. Для обробки умови, яка може зустрічатись при заданні операції, був написаний синтаксичний аналізатор. Умова визначається наступною граматикою:

<умова> -> <терм> <з'єднувач> <умова>
<умова> -> <терм>
<терм> -> <умова>
<терм> -> <нечітка множина>
<терм> -> <ім'я атрибуту> <операція> <константа> |
<ім'я атрибуту> <операція> <ім'я атрибуту>
<нечітка множина> -> <модифікатор> <нечітка константа>
<нечітка множина> -> <нечітка константа>
<з'єднувач> -> NOT |AND | OR
<модифікатор> -> з відношення **МОДИФІКАТОРИ**
<нечітка константа> -> з відношення **КОНСТАНТИ**
<операція> -> <звичайна операція> | <нечітка операція>
<звичайна операція> -> = | < | > | <= | >= | ><
<нечітка операція> -> з відношення **ОПЕРАТОРИ.**

Після того, як користувач задасть всі необхідні для обробки нечіткостей дані /для чого створені відповідні засоби/, він може формувати запити до розширеної реляційної бази даних.

Розроблені програмні макети та програмно- алгоритмічні моделі розширення можливостей реляційної СУБД можливостями обробки неповної та неточної інформації і методика формування в СУБД нового типу даних **НЕЧІТКІСТЬ** забезпечують гнучкість при настройці на предметну область і на користувача. Розробка була впроваджена

на ряді підприємств в програмних пакетах та АРМах по планово-економічній та бухгалтерській діяльності, де користувач в інтерактивному режимі має можливість формувати нечіткі запити до бази даних з оперативною, нормативно- довідковою та архівною інформацією по обліку виробничої діяльності підприємства.

ЗАКЛЮЧЕННЯ

Аналіз досліджень в області розширення реляційної моделі баз даних для обробки неповної інформації виявляє відсутність чіткого формулювання змісту відповідного "розширення" моделі. Робота націлена на обгрунтування тези, що розширення моделі відбувається на рівні доменів і розповсюджується на рівень реляційної алгебри. При розширенні можливостей СУБД обробкою неповної інформації це приводить до введення нового типу даних в СУБД.

Пропонується методика розширення реляційної моделі існуючих СУБД реляційного типу новим типом даних, який відповідає неповним, неточним, нечітким даним. Результати роботи дозволять використовувати існуючі СУБД реляційного типу в системах баз даних і знань, експертних системах і інших "інтелектуалізованих" інформаційних системах.

Результати дисертації опубліковані в основних роботах:

1. Малюта Т.А., Пасичник В.В. Реализация диалогового проектирования реляционных баз данных с использованием реляционной БД // Материалы респ. конф. "Проблемно- ориентированные диалоговые системы", 25- 28 окт. 1988г. - Батуми, 1988. - с.140-145.

2. Малюта Т.А., Пасичник В.В. Расширение средств реляционной СУБД для обработки размытых значений данных // Тез. докл. Всесоюз. науч.- практ. школы семинара "Программное обеспечение ЭВМ: индивидуальная технология, интеллектуализация разработки и применение", 5- 10 дек. 1988 г. - Ростов- на- Дону, 1988. - с. 74-76.

3. Брона И.И., Малюта Т.А., Пасичник В.В. Реляционные базы данных с нечеткими значениями // Реляционные базы данных с нечеткими значениями. - Новосибирск, 1989.- с. 1- 53. - (Препр./АН СССР. ВЦ

CO; 846.
4. Stogniy A.A., Malyuta T.A., Pasitschnik V.V. Means for management of relation fuzzy data bases- way to merging of systems of data bases and knowledge bases // Lecture Notes in Computer science. - Berlin: Springer; Tokyo: Verlag, 1989. - N 364. - p.337- 347.
5. Малюта Т.А., Пасичник В.В. Интеллектуализация реляционных баз данных путем введения в них возможностей работы с неполной и неточной информацией // Conference on intelligent management systems / Bulgarian academy of science. - Varna, 1989. - p.96-103.
6. Pasitschnik V.V., Malyuta T.A. Use of abstract data types to manipulate the incomplete and imprecise information in relational databases // 12-th International Seminar on Database Management Systems. - Suzdal, 1989.- Books 2.- p.132- 139.
7. Брона И.И., Пасичник В.В., Малюта Т.А. Неопределенные и неполные значения в реляционных базах данных // Материалы 4-ой Всесоюз. конф. "Банки данных и знаний", Калинин, нояб. 1989. - с.I-I8.
8. Малюта Т.А. Использование в базах данных нечетких данных- средство слияния баз данных и баз знаний // Вестник Львовского политехнического института, N 248, из- во "Свит", I990. - с.8I-85.
9. Васиков В.Л., Копчак О.И., Малюта Т.А., Пасичник В.В.Разработка автоматизированной системы по научному направлению "Однородные вычислительные среды и систолические структуры" // Отчет по НИР, ЛПИ, I989, рег.N OI860053898.
IO. Копчак О.И., Малюта Т.А., Пасичник В.В. Исследование и разработка системы обработки данных на базе ПЭВМ профессионального класса с ориентацией на СУБД реляционного типа // Отчет по НИР, ЛПИ, I99I, рег.N OI890040I4I.

Підп. до друку 11.03 93 . Формат 60x84 1/16.
Папір друк. № 2. Друк. офс. Умовн.друк.арк. 1.5
Умовн.фарб.-відб. 1.5 Обл.-вид.арк. 1.37
Тираж 100 прим. Зам. 79 . Безплатно
ЛПІ 290646 Львів-I3, Ст.Бандери, I2
Дільниця оперативного друку ЛПІ
Львів, вул. Городоцька, 286