

УДК 51.001.57+004.652.4+004.852

Р.Б. Кравець

Національний університет “Львівська політехніка”,
кафедра “Інформаційні системи та мережі”

ОРГАНІЗАЦІЯ БАГАТОВИМІРНОГО ПОДАННЯ ТА АНАЛІЗУ ІНФОРМАЦІЇ У РЕЛЯЦІЙНИХ БАЗАХ ДАНИХ

© Кравець Р.Б., 2006

Author has formalized multidimensional data representation using the theory of relational databases (RDB). Main multidimensional data representation concepts – dimension and data hypercube – are described mathematically. Operations with data hypercube for information analytical processing in RDB are considered in the paper. These operations are realized using relational algebra operations.

Формалізовано багатовимірне подання даних з використанням апарата теорії реляційних баз даних (РБД). Математично описано основні поняття багатовимірною подання даних – вимір та гіперкуб даних. Введено операції над гіперкубами даних, які дозволяють аналітично опрацювати інформацію, що зберігається в РБД. Для реалізації багатовимірного підходу в РБД операції над гіперкубами даних виражені операціями реляційної алгебри.

1. ПОСТАНОВКА ПРОБЛЕМИ

В основі багатьох задач, пов'язаних із плануванням та прийняттям рішень, лежить багатопараметричний аналіз показників, що описують ділянку конкретної ПО. Для проведення такого аналізу необхідні засоби, які б подавали і дозволяли опрацьовувати користувачеві дані у багатовимірному вигляді. Для цього під час проектування БД у ПО виділяються показники, які потрібно аналізувати, та величини, від яких ці показники залежать. Багатовимірні структури, що отримуються у процесі проектування, описують функції багатьох змінних, де показники є значеннями функцій, а величини, від яких залежать ці показники, – змінними-аргументами. У простих випадках, коли показник залежить від однієї чи двох змінних, значення функції у БД задаються одно- чи двовимірною таблицею. Загалом, коли маємо функцію $F(X_1, X_2, \dots, X_n)$ від n змінних, її значення задаються у вигляді n -вимірної таблиці, у рядках якої розміщені значення змінних $X_1, X_2, \dots, X_n$, а у комірках, що стоять на перетині рядків $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$, – значення функції в точці $(x_1, x_2, \dots, x_n)$. Проілюструємо сказане на прикладі.

Приклад 1. Багатовимірне подання даних.

Розглянемо приклад торговельного підприємства, для якого потрібно проаналізувати обсяг продажу товару за регіонами країни поквартально. У цьому випадку ми будемо оперувати функцією *обсяг продажу*, яка залежить від змінних *товар*, *регіон* та *час*. Множину значень змінної *товар* складає перелік товарів, які продає підприємство, множину значень змінної *регіон* – регіони країни, а множину *час* – квартали року. Значенням функції є сума продажу конкретного товару у певному регіоні за певний квартал. Функція задається у вигляді куба, зображеного на рис. 1.

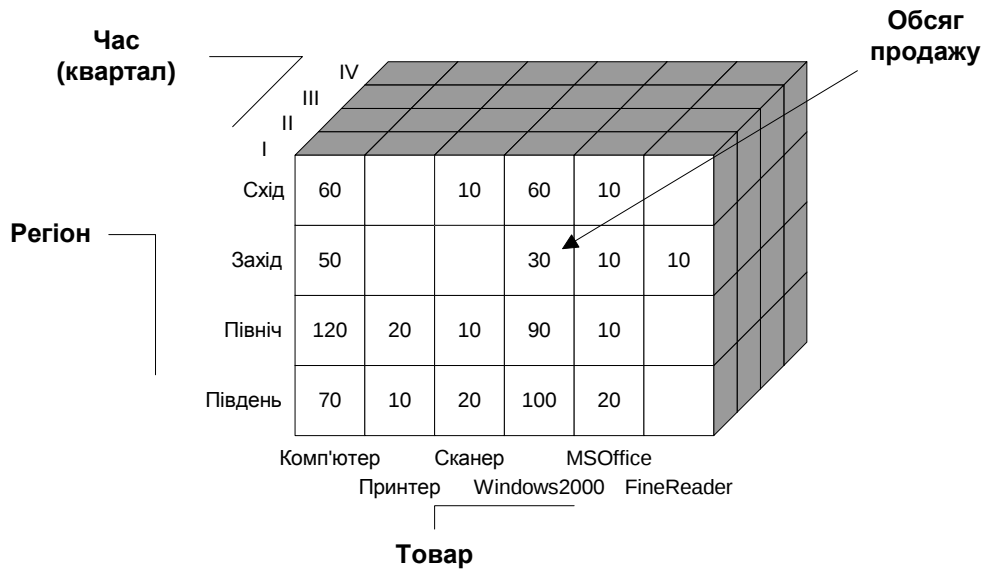


Рис. 1. Куб даних “Обсяг продажу”

Ідею аналітичного опрацювання інформації (OLAP) на основі багатовимірного підходу висунув Код [1]. Відзначимо два ключові моменти, які характеризують багатовимірний підхід до опрацювання інформації, запропонований Кодом:

- у багатовимірному вигляді дані подаються користувачеві на концептуальному рівні;
- на логічному та фізичному рівнях дані організовані у вигляді інформаційних відношень реляційної моделі і зберігаються в РБД.

Це означає, що користувач опрацьовує дані операціями багатовимірного аналізу, які у свою чергу, використовують засоби РБД.

Виходячи з цього, автором пропонується математична формалізація багатовимірного підходу, яка базується на апараті реляційної моделі даних. При цьому структури даних багатовимірного підходу описуються на основі такого ж формалізму, що й структури даних реляційної моделі, а операції багатовимірного аналізу виражаються через операції реляційної алгебри. Такий підхід автора до формалізації багатовимірного підходу дозволяє дотриматись пропозицій, сформульованих Кодом. Крім того, він забезпечує можливість реалізації багатовимірного підходу в рамках реляційних СКБД.

2. АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ

В основі концепції аналітичного опрацювання інформації лежить багатовимірне подання даних. Дані та об'єкти в БД містять інформацію переважно на найнижчому рівні деталізації. Наприклад, відношення “продукція” в БД “збут” може містити атрибути, які деталізують інформацію на низькому рівні абстракції, такі як код продукції, назва продукції, дата виготовлення, ціна тощо. Для аналітичного опрацювання інформації часто необхідно агрегувати великі масиви даних для подання інформації на високому концептуальному рівні. Наприклад, можна агрегувати дані про реалізацію продукції для отримання узагальненого опису збуту. Такий вид аналітичного опрацювання інформації називається узагальненням даних.

Узагальнення даних – це процес абстрагування великих наборів відповідних даних у БД від низького концептуального рівня до високого та подання інформації на різних рівнях

ієрархії класів даних. Відповідні методи узагальнення даних поділяють на дві категорії: атрибутно-орієнтована індукція [2, 3] та багатовимірний підхід [4, 5, 6].

2.1. Атрибутно-орієнтована індукція

У атрибутно-орієнтованому підході завдання на виконання видобування знань формується у вигляді запиту на SQL-подібній мові видобування знань, який формує результат з відповідного набору даних у БД. На основі отриманого результату узагальнюють дані за допомогою таких методів узагальнення даних, як видалення негрупових атрибутів, обчислення агрегатних функцій та ін. Узагальнені дані подаються у вигляді узагальнених відношень, які використовуються для подальшого опрацювання та перетворення узагальнених даних у різного роду знання, або відображаються в інших формах. Наприклад, можуть виконуватися операції згортки та розгортки для подання даних на різних рівнях абстракції; узагальнені відношення можуть відображатися у вигляді підсумкових таблиць, графічних діаграм та інших засобів подання інформації. На основі узагальнених відношень можуть генеруватись характеристичні, дискримінантні та класифікаційні правила.

В основі атрибутно-орієнтованої індукції лежить метод узагальнення даних, який полягає в наступному: для кожного атрибута обчислюється розподіл даних на основі відповідного набору даних; формуються відповідні рівні абстракції для узагальнення даних атрибута; кожен кортеж у відношенні замінюється відповідним узагальненим кортежем. Процес узагальнення за один прохід набору даних формує агреговані значення атрибутів та розміщує їх в узагальненому відношенні.

Для підтримки операцій аналітичного опрацювання інформації, таких як згортка та розгортка даних, метод атрибутно-орієнтованої індукції узагальнює набір відповідних даних до мінімального рівня абстракції та використовує такі мінімально узагальнені відношення для навігації за рівнями абстракції даних.

2.2. Багатовимірний підхід

Для позначення багатовимірного підходу у різних авторів вживаються декілька різних варіантів назв, такі як “багатовимірні бази даних”, “матеріалізовані перегляди” та “аналітичне опрацювання інформації (OLAP)”. Основна ідея багатовимірного підходу полягає у виконанні певних трудомістких обчислень, які часто здійснюються, особливо таких, які містять функції агрегації (кількість, сума, середнє, максимум та інші), і збереження цих результатів в БД для їх подальшого використання у процесі підтримки прийняття рішень, видобування знань та інших застосуваннях. Обчислення функцій агрегації здійснюються відповідно до групування різних наборів чи піднаборів атрибутів. Значення кожного атрибута можуть об'єднуватися в ієрархічні або розгалужені структури. Наприклад, “дата” може групуватися за днями, місяцями, роками чи тижнями, у результаті отримуємо розгалужену структуру. Узагальнення та деталізація даних виконуються в БД за допомогою операцій згортки та розгортки, причому операції згортки зменшують кількість вимірів в гіперкубі даних або узагальнення значення атрибутів до вищого концептуального рівня, а операції розгортки виконують обернені дії. Оскільки в аналітичному опрацюванні інформації часто доводиться оперувати даними в агрегованому вигляді, збереження наперед обчислених результатів у БД забезпечить значне зменшення часу отримання відповіді на запити до БД.

Багатовимірний підхід придатний до використання в багатьох застосуваннях. Для нього розробляються методи індексації в багатовимірних БД та методи інкрементного

оновлення даних. Зазначимо, що реальна багатовимірна БД може бути дуже розрідженою через те, що далеко не для кожного набору вимірів існує конкретне значення. Це вимагає розробки методів ефективного збереження та опрацювання даних в розріджених багатовимірних БД. Крім того, якщо запит містить значення нижчого рівня ієрархії, ніж значення відповідного виміру, немає жодних методів оптимального використання наперед обчислених результатів, що зберігаються в БД.

Зазначимо, що для того, щоб застосовувати багатовимірний підхід у видобуванні знань, для нього потрібно побудувати формальну модель. У цій роботі пропонується багатовимірна модель подання даних як формалізація багатовимірного підходу.

3. ФОРМУЛЮВАННЯ ЦІЛЕЙ

У статті розглядаються такі питання:

- побудова математичної моделі багатовимірного подання даних в РБД на основі застосування апарата теорії реляційних баз даних;
- математичний опис поняття виміру, гіперкуба даних та відношення між вимірами;
- формалізація основних операцій над гіперкубами даних та дослідження властивостей цих операцій;
- спосіб реалізації структур багатовимірного подання даних засобами РБД;
- реалізація операцій над гіперкубами даних засобами реляційної алгебри та мови запитів SQL.

4. ОСНОВНИЙ МАТЕРІАЛ

4.1. Структури даних багатовимірного подання даних

У багатовимірному підході подання даних структури, у яких зберігаються значення функцій, прийнято називати **гіперкубами даних**, а змінні-аргументи – **вимірами**. Дамо формальне означення поняттям гіперкуба даних та виміру. Нехай $A_1, A_2, \dots, A_m$ – атрибути з універсуму U , визначені на доменах $dom(A_1), dom(A_2), \dots, dom(A_m)$ відповідно.

Означення 1. Виміром D називається підмножина декартового добутку $dom(A_1) \times dom(A_2) \times \dots \times dom(A_m)$. Значенням виміру d називається кортеж виміру D . Кількість всіх кортежів виміру D називається обсягом виміру D і позначається як $|D|$.

Означення 2. Схемою виміру D називається сукупність атрибутів цього виміру $D = \{A_1, A_2, \dots, A_m\}$.

Проілюструємо поняття виміру на прикладі гіперкуба даних, зображеного на рис. 1.

Приклад 2. Виміри гіперкуба даних.

Обсяг продукції торговельного підприємства аналізується за трьома параметрами: товар, географічне розташування покупця (регіон країни) та час купівлі (квартал). Відповідно до цього гіперкуб даних складається з трьох вимірів: *товар*, *регіон* та *час*. Вимір *товар* описується атрибутом *код_товару*, який набуває значень на відповідному домені $dom(код_товару) = \{\text{'Комп'ютер'}, \text{'Принтер'}, \text{'Сканер'}, \text{'Windows2000'}, \text{'MSOffice'}, \text{'FineReader'}\}$. Схеми виміру *регіон* складається з атрибута *код_регіону*, який набуває значень на множині $dom(код_регіону) = \{\text{'Схід'}, \text{'Захід'}, \text{'Північ'}, \text{'Південь'}\}$. І третім виміром є *час*, який описується атрибутом *квартал* з доменом $dom(квартал) = \{\text{'I'}, \text{'II'}, \text{'III'}, \text{'IV'}\}$.

Нехай $D_1, D_2, \dots, D_n$ – виміри зі схемами $D_1, D_2, \dots, D_n$ відповідно.

Означення 3. Гіперкубом даних $H(D_1, D_2, \dots, D_n)$ називається відображення $H: D_1 \times D_2 \times \dots \times D_n \rightarrow V$, де V – деяка множина, на якій задана функція агрегації Agg елементів множини. Координатою гіперкуба даних $H(D_1, D_2, \dots, D_n)$ називається сукупність значень вимірів $(d_1, d_2, \dots, d_n)$. Значенням гіперкуба даних $H(D_1, D_2, \dots, D_n)$ в координаті $(d_1, d_2, \dots, d_n)$ називається елемент $h \in V$, у який відображається координата $(d_1, d_2, \dots, d_n)$.

Отже, гіперкуб даних визначається як множина

$$H(D_1, D_2, \dots, D_n) = \{(d_1, d_2, \dots, d_n, h) \mid \forall i, 1 \leq i \leq n, d_i \in D_i, h \in V, h = h(d_1, d_2, \dots, d_n)\}.$$

Означення 4. Схемою гіперкуба даних H називається сукупність схем вимірів та множина значень гіперкуба даних $H = \{D_1, D_2, \dots, D_n, V\}$.

Кількість вимірів, від яких залежить гіперкуб даних $H(D_1, D_2, \dots, D_n)$, визначає вимірність гіперкуба даних. Отже, гіперкуб даних $H(D_1, D_2, \dots, D_n)$ має вимірність n , або, інакше кажучи, є n -вимірним.

Отже, гіперкуб даних утворюється як багатовимірна структура, вимірами якої є інформаційні відношення. Атрибут гіперкуба, який описує показник, що аналізується, набуває значень на множині, на якій визначена функція агрегації її елементів. Наявність такої функції дає можливість агрегувати дані гіперкуба, що є необхідним для аналітичного опрацювання інформації. На рис. 2 наведено загальний вигляд тривимірного гіперкуба даних.

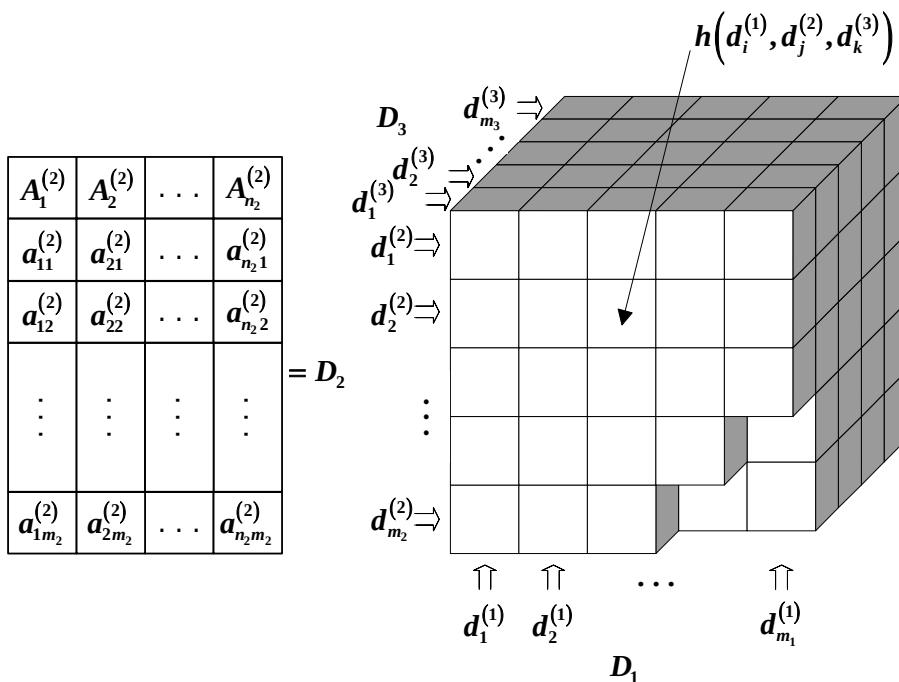


Рис. 2. Багатовимірне подання даних

Важливою умовою для аналітичного опрацювання інформації є можливість вибирати значення гіперкуба на різних рівнях деталізації значень вимірів. У випадку, коли між деякими вимірами існує зв'язок типу “один-до-багатьох”, у якому один вимір є головним, а інший підпорядкованим, можна агрегувати значення гіперкуба даних, замінюючи при цьому під-

порядкований вимір головним. Такий спосіб опрацювання даних гіперкуба дозволяє проводити аналіз на різних рівнях деталізації інформації. Для забезпечення механізму агрегування даних на основі міжвимірних зв'язків введемо поняття **відношення між вимірами**.

Означення 5. Відношенням r між вимірами D_1 та D_2 називається підмножина декартового добутку $D_1 \times D_2$, така, що:

$$\begin{aligned} & \forall d_1 \in D_1 (\exists d_2 \in D_2 \wedge \exists (d_1, d_2) \in r), \\ & \forall d_1^{(1)} \in D_1 \wedge \forall d_1^{(2)} \in D_1 \wedge \forall d_2^{(1)} \in D_2 \wedge \forall d_2^{(2)} \in D_2 \wedge \\ & \exists (d_1^{(1)}, d_2^{(1)}) \in r \wedge \exists (d_1^{(2)}, d_2^{(2)}) \in r \left((d_1^{(1)} = d_1^{(2)}) \rightarrow (d_2^{(1)} = d_2^{(2)}) \right). \end{aligned}$$

Означення 6. Схемою відношення r між вимірами D_1 та D_2 називається сукупність схем вимірів $R = \{D_1, D_2\}$, $K = D_1$ – ключ відношення.

Іншими словами, відношення r між вимірами D_1 та D_2 – це такий набір кортежів типу (d_1, d_2) , де d_1 – кортеж виміру D_1 , d_2 – кортеж виміру D_2 , в яких значення d_1 повністю визначає значення d_2 . При цьому вимір D_1 називається ключовим виміром відношення r . Зауважимо, що у відношенні між вимірами ключ не є властивістю якогось із екземплярів відношення, а радше суттєвою характеристикою співвідношення між вимірами.

Проілюструємо поняття відношення між вимірами на прикладі гіперкуба даних, зображеного на рис. 1.

Приклад 3. Відношення між вимірами.

Крім вимірів, наведених у прикладі 2, введемо ще один вимір *тип_товару*, який описується атрибутом *код_типу_товару*. Атрибут *код_типу_товару* набуває значень на відповідному домені $dom(код_типу_товару) = \{ \text{“Апаратне забезпечення”}, \text{“Програмне забезпечення”} \}$. На рис. 3 зображене відношення між вимірами *тип_товару* та *товар*. Це відношення описує класифікацію товарів за типами.



Рис. 3. Відношення між вимірами

Тепер на основі відношення $(тип_товару, товар)$ ми можемо аналізувати обсяг продажу як у розрізі товарів, так і у розрізі типів товарів, агрегуючи значення гіперкуба даних при переході з рівня товарів на рівень типів товарів.

Опишемо тепер операції багатовимірної підходу до аналітичного опрацювання інформації.

4.2. Операції багатовимірної аналізу даних

Для багатовимірної аналізу інформації автором пропонується використовувати такі операції над гіперкубами даних: об'єднання за виміром, зрізом, проекцією та згортою по відношенню. Опишемо ці операції детальніше.

4.2.1. Операція об'єднання за виміром

Операція об'єднання за виміром потрібна для інтеграції даних, отриманих із різних джерел, в єдиному гіперкубі даних, а також для занесення нових даних у гіперкуб.

Нехай $H_1(D^{(1)}, D_1, D_2, \dots, D_n)$ та $H_2(D^{(2)}, D_1, D_2, \dots, D_n)$ – гіперкуби даних зі схемою $H = \{D, D_1, D_2, \dots, D_n, V\}$, причому значення вимірів $D^{(1)}$ та $D^{(2)}$ є різними, тобто $D^{(1)} \cap D^{(2)} = \emptyset$. У результаті виконання операції об'єднання за виміром D над гіперкубами даних $H_1(D^{(1)}, D_1, D_2, \dots, D_n)$ та $H_2(D^{(2)}, D_1, D_2, \dots, D_n)$ утворюється гіперкуб даних $H(D, D_1, D_2, \dots, D_n)$, який об'єднує у собі дані з обох гіперкубів, тобто

$$H(D, D_1, D_2, \dots, D_n) = H_1(D^{(1)}, D_1, D_2, \dots, D_n) \cup_D H_2(D^{(2)}, D_1, D_2, \dots, D_n) = \\ = \{(d, d_1, d_2, \dots, d_n, h) \mid \forall i, 1 \leq i \leq n, d_i \in D_i, d \in D^{(1)} \cup D^{(2)}, h = h(d, d_1, d_2, \dots, d_n)\}.$$

Проілюструємо цю операцію на прикладі.

Приклад 4. Операція об'єднання за виміром.

Маємо два гіперкуби даних *обсяг_продажу_1* та *обсяг_продажу_2* зі схемою {товар, регіон, час}. Гіперкуб даних *обсяг_продажу_1* містить інформацію про обсяг продажу товару за перше півріччя, а гіперкуб даних *обсяг_продажу_2* – за друге. Об'єднуючи ці гіперкуби даних за виміром *час*, отримуємо гіперкуб даних з інформацією про обсяг продажу товару протягом року (рис. 4).

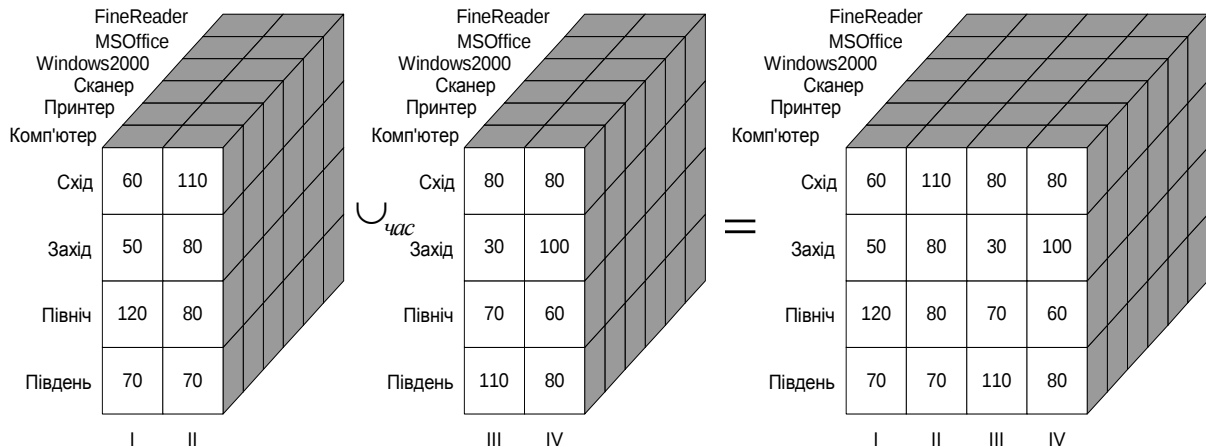


Рис. 4. Операція об'єднання за виміром

4.2.2. Операція зрізу

Гіперкуб даних не аналізується одразу за всіма вимірами, від яких він залежить. Переважно вибірка даних із гіперкуба здійснюється для конкретних значень певного набору вимірів, і вільними залишаються, як правило, один чи два виміри, за якими проводиться подальший аналіз. Таке опрацювання інформації виконується за допомогою операції зрізу. Ця операція дозволяє з $(n+1)$ -вимірного гіперкуба даних отримати n -вимірний, зафіксувавши значення одного з вимірів.

Нехай $H_1(D, D_1, D_2, \dots, D_n)$ – гіперкуб даних зі схемою $\mathbf{H} = \{D, D_1, D_2, \dots, D_n, V\}$.

Операція зрізу за значенням $D = d^*$ над гіперкубом даних $H_1(D, D_1, D_2, \dots, D_n)$ вибирає із цього гіперкуба ту частину, для якої вимір D набуває значення d^* . У результаті отримуємо такий гіперкуб даних:

$$\begin{aligned} H(D_1, D_2, \dots, D_n) &= \text{Slice}(D = d^*; H_1(D, D_1, D_2, \dots, D_n)) = \\ &= \{(d_1, d_2, \dots, d_n, h) \mid \forall i, 1 \leq i \leq n, d_i \in D_i, h = h_1(d^*, d_1, d_2, \dots, d_n)\}. \end{aligned}$$

Розглянемо приклад, який ілюструє операцію зрізу.

Приклад 5. Операція зрізу.

Маємо гіперкуб даних *обсяг_продажу* зі схемою $\{\text{товар}, \text{регіон}, \text{час}\}$. Для того, щоб отримати інформацію про обсяг продажу товару за регіонами країни за перший квартал, потрібно виконати операцію зрізу із значенням виміру $\text{час} = \text{'I'}$. У результаті виконання операції отримаємо гіперкуб даних зі схемою $\{\text{товар}, \text{регіон}\}$. Вихідний та результуючий гіперкуби даних зображені на рис. 5.

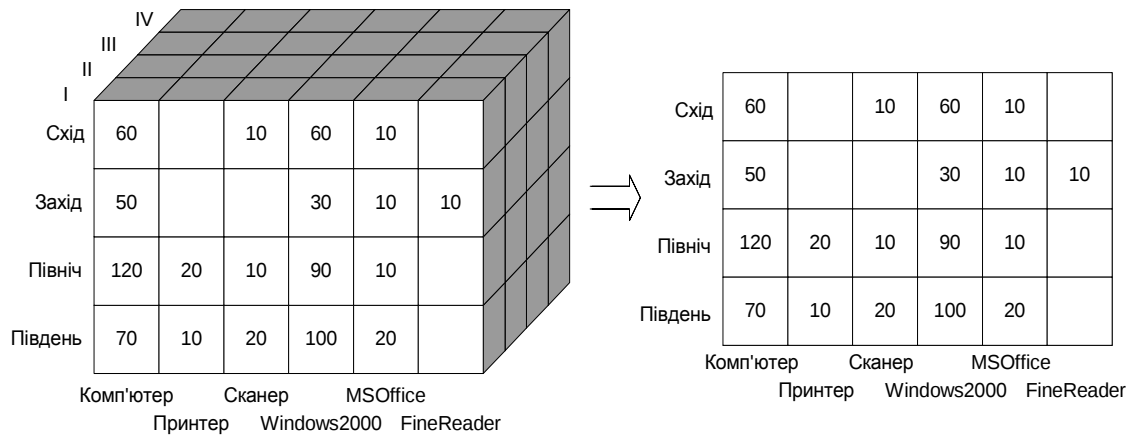


Рис. 5. Операція зрізу

Для операції зрізу виконується властивість комутативності.

Твердження 7. Нехай маємо гіперкуб даних H над вимірами $D_1, D_2, \dots, D_n$. Справедлива така властивість:

$$\begin{aligned} \text{Slice}(D_1 = d_1; \text{Slice}(D_2 = d_2; H(D_1, D_2, \dots, D_n))) &= \\ = \text{Slice}(D_2 = d_2; \text{Slice}(D_1 = d_1; H(D_1, D_2, \dots, D_n))) &. \end{aligned}$$

Використовуючи властивість комутативності, можемо узагальнити операцію зрізу, фіксуючи значення одразу k вимірів

$$\text{Slice}(D_1 = d_1, D_2 = d_2, \dots, D_k = d_k; H(D_1, D_2, \dots, D_n)),$$

де $k \leq n$.

Зауважимо, що за допомогою операції зрізу можна робити вибірку конкретного значення гіперкуба даних:

$$h(d_1, d_2, \dots, d_n) = \text{Slice}(D_1 = d_1, D_2 = d_2, \dots, D_n = d_n; H(D_1, D_2, \dots, D_n)).$$

4.2.3. Операція проєкції

Ще однією операцією, яка дозволяє зменшити кількість вимірів гіперкуба даних, що аналізується, є операція проєкції. Для виконання цієї операції вибираємо вимір, за яким буде здійснюватися проєкція. Цей вимір не входить у схему результуючого гіперкуба даних. Значення гіперкуба даних агрегуються за усіма іншими вимірами, які складатимуть схему гіперкуба-проєкції.

Нехай $H_1(D, D_1, D_2, \dots, D_n)$ – гіперкуб даних зі схемою $\mathbf{H} = \{D, D_1, D_2, \dots, D_n, V\}$. Результатом виконання операції проєкції за виміром D над гіперкубом даних $H_1(D, D_1, D_2, \dots, D_n)$ є гіперкуб даних $H(D_1, D_2, \dots, D_n)$, значення якого агрегуються за вимірами $D_1, D_2, \dots, D_n$, тобто:

$$H(D_1, D_2, \dots, D_n) = \text{Project}(\text{Agg}; D; H_1(D, D_1, D_2, \dots, D_n)) = \\ = \{(d_1, d_2, \dots, d_n, h) \mid \forall i, 1 \leq i \leq n, d_i \in D_i, h = \text{Agg}(d \in D; h_1(d, d_1, d_2, \dots, d_n))\},$$

де Agg – деяка функція агрегації над множиною значень гіперкуба даних.

Наведемо приклад виконання операції проєкції.

Приклад 6. Операція проєкції.

Маємо гіперкуб даних *обсяг_продажу* зі схемою $\{\text{товар}, \text{регіон}, \text{час}\}$ з інформацією про квартальні обсяги продажу товарів за регіонами країни. Потрібно отримати інформацію про річний обсяг продажу товарів у розрізі регіонів. Для цього потрібно виконати операцію проєкції над гіперкубом *обсяг_продажу* за виміром *час*. При цьому отримаємо новий гіперкуб даних зі схемою $\{\text{товар}, \text{регіон}\}$, значеннями якого є суми обсягів продажу за усі квартали року окремо для кожного товару по кожному регіону. Результат виконання цієї операції зображений на рис. 6.

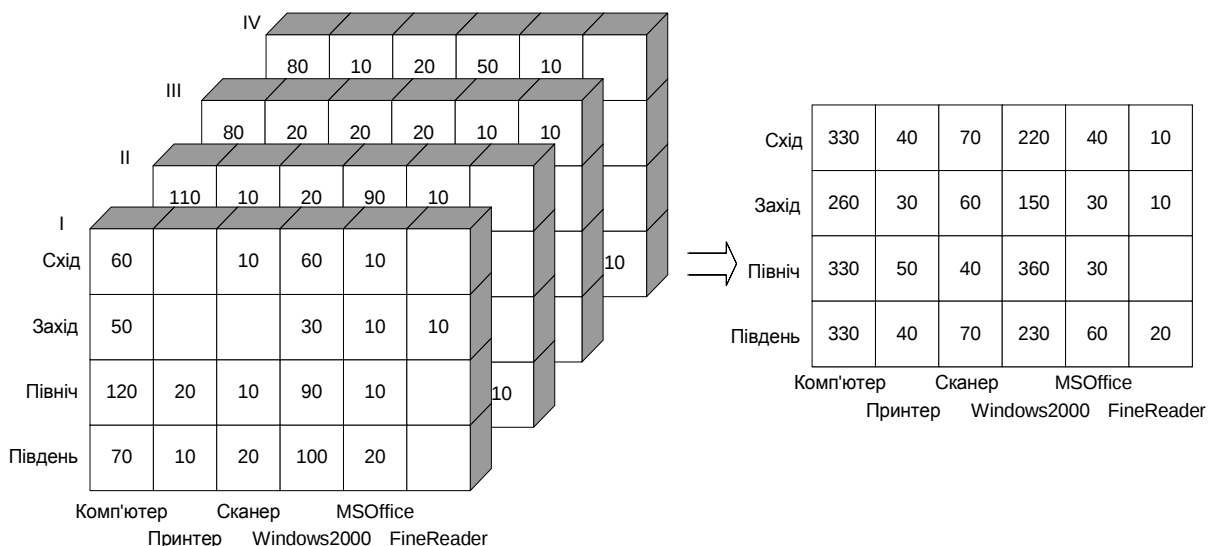


Рис. 6. Операція проєкції

Зазначимо, що операції проєкції та зрізу є комутативними.

Твердження 8. Нехай маємо гіперкуб даних H над вимірами $D_1, D_2, \dots, D_n$. Існує така властивість:

$$\begin{aligned} & Project(Agg; D_1; Slice(D_2 = d_2; H(D_1, D_2, \dots, D_n))) = \\ & = Slice(D_2 = d_2; Project(Agg; D_1; H(D_1, D_2, \dots, D_n))). \end{aligned}$$

Крім того, для операції проєкції виконується властивість комутативності за умови, що функція агрегації є комутативною.

Твердження 9. Нехай маємо гіперкуб даних H над вимірами $D_1, D_2, \dots, D_n$. Якщо функція агрегації Agg значень гіперкуба даних H є комутативною, то існує така властивість:

$$\begin{aligned} & Project(Agg; D_1; Project(Agg; D_2; H(D_1, D_2, \dots, D_n))) = \\ & = Project(Agg; D_2; Project(Agg; D_1; H(D_1, D_2, \dots, D_n))). \end{aligned}$$

У випадку, якщо функція агрегації комутативна, можемо узагальнити операцію проєкції, виконуючи її одразу за k вимірами:

$$Project(Agg; D_1, D_2, \dots, D_k; H(D_1, D_2, \dots, D_n)),$$

де $k \leq n$.

4.2.4. Операція згортки за відношенням

Операція згортки за відношенням дозволяє агрегувати значення гіперкуба даних на основі зв'язку між вимірами, надаючи можливість користувачеві аналізувати інформацію на різних рівнях її деталізації.

Нехай $H_1(D_1, D_2, \dots, D_n)$ – гіперкуб зі схемою $\mathbf{H}_1 = \{\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_n, \mathbf{V}\}$, D – вимір зі схемою $\mathbf{D}$ і $r(D, D_1)$ – відношення між вимірами D та D_1 зі схемою $R = \{\mathbf{D}, \mathbf{D}_1\}$, $K = \mathbf{D}$. Результатом виконання операції згортки за відношенням $r(D, D_1)$ над гіперкубом даних $H_1(D_1, D_2, \dots, D_n)$ є гіперкуб даних $H(D, D_2, \dots, D_n)$ зі схемою $\mathbf{H} = \{\mathbf{D}, \mathbf{D}_2, \dots, \mathbf{D}_n, \mathbf{V}\}$ такий, що

$$\begin{aligned} & H(D, D_2, \dots, D_n) = Rollup(Agg; r(D, D_1); H_1(D_1, D_2, \dots, D_n)) = \\ & = \{(d, d_2, \dots, d_n, h) \mid \forall i, 1 \leq i \leq n, d_i \in D_i, h = Agg(d \in D \wedge d_1 \in D_1 \wedge (d, d_1) \in r; h_1(d_1, d_2, \dots, d_n))\}, \end{aligned}$$

де Agg – деяка функція агрегації над множиною значень гіперкуба даних.

Проілюструємо операцію згортки за відношенням на прикладі.

Приклад 7. Операція згортки за відношенням.

Маємо гіперкуб даних *обсяг_продажу* зі схемою $\{\text{товар}, \text{регіон}, \text{час}\}$ і вимір *тип_товару*. Між вимірами *товар* та *тип_товару* встановлений зв'язок, зображений на рис. 3. Цей зв'язок описується відповідним відношенням зі схемою $\{\text{тип_товару}, \text{товар}\}$. Результат виконання операції згортки гіперкуба *обсяг_продажу* за відношенням між вимірами *товар* та *тип_товару* зображений на рис. 7.

Зазначимо, що операції згортки за відношенням та зрізу є комутативними.

Твердження 10. Нехай маємо гіперкуб даних H над вимірами $D_1, D_2, \dots, D_n$, відношення між вимірами $r(D, D_1)$. Існує така властивість:

$$\begin{aligned} & Rollup(Agg; r(D, D_1); Slice(D_2 = d_2; H(D_1, D_2, \dots, D_n))) = \\ & = Slice(D_2 = d_2; Rollup(Agg; r(D, D_1); H(D_1, D_2, \dots, D_n))). \end{aligned}$$

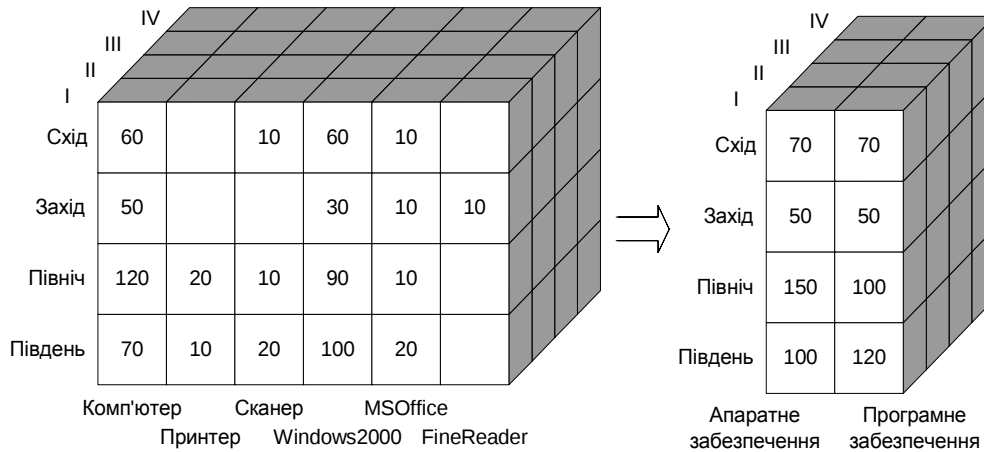


Рис. 7. Операція згортки за відношенням

Як і у випадку операції проєкції, для операції згортки за відношенням виконується властивість комутативності за умови, що функція агрегації є комутативною.

Твердження 11. Нехай маємо гіперкуб даних H над вимірами $D_1, D_2, \dots, D_n$, відношення між вимірами $r(D^{(1)}, D_1)$ та $r(D^{(2)}, D_2)$. Якщо функція агрегації Agg значень гіперкуба даних H є комутативною, то існує така властивість:

$$\begin{aligned} & Rollup\left(Agg; r(D^{(1)}, D_1); Rollup\left(Agg; r(D^{(2)}, D_2); H(D_1, D_2, \dots, D_n)\right)\right) = \\ & = Rollup\left(Agg; r(D^{(2)}, D_2); Rollup\left(Agg; r(D^{(1)}, D_1); H(D_1, D_2, \dots, D_n)\right)\right). \end{aligned}$$

Крім того, для операцій проєкції та згортки за відношенням виконується властивість комутативності, якщо комутативною є функція агрегації.

Твердження 12. Нехай маємо гіперкуб даних H над вимірами $D_1, D_2, \dots, D_n$, відношення між вимірами $r(D, D_1)$. Якщо функція агрегації Agg значень гіперкуба даних H є комутативною, то існує така властивість:

$$\begin{aligned} & Rollup\left(Agg; r(D, D_1); Project\left(Agg; D_2; H(D_1, D_2, \dots, D_n)\right)\right) = \\ & = Project\left(Agg; D_2; Rollup\left(Agg; r(D, D_1); H(D_1, D_2, \dots, D_n)\right)\right). \end{aligned}$$

Описані операції призначені для виконання аналітичного опрацювання інформації користувачем на концептуальному рівні багатовимірного подання даних. Як зазначалося вище, в РБД виміри та гіперкуби даних організовані у вигляді інформаційних відношень. При цьому операції над гіперкубами даних виражаються за допомогою операцій реляційної алгебри над відповідними інформаційними відношеннями.

Перейдемо тепер до опису запропонованої автором реалізації багатовимірного подання даних засобами РБД.

Опишемо спочатку як реалізуються виміри, гіперкуби даних та відношення між вимірами в РБД. Після цього виразимо операції багатовимірного аналізу даних через операції реляційної алгебри.

4.3. Реалізація структур багатовимірних подання даних в РБД

Із означення виміру випливає, що виміру у багатовимірному поданні даних відповідає інформаційне відношення реляційної моделі. Отже, в РБД вимір реалізується як інформаційне відношення і, відповідно, схема виміру – як схема інформаційного відношення.

Для реалізації гіперкуба даних в РБД нам потрібно подати відображення вимірів гіперкуба на деяку множину із визначеною на ній функцією агрегації. Таке відображення будемо подавати у вигляді інформаційного відношення, схема якого складається з атрибутів, що є первинними ключами вимірів гіперкуба даних, та атрибута, що відповідає множині значень гіперкуба даних. При цьому сукупність ключових атрибутів вимірів утворює ключ інформаційного відношення гіперкуба даних. За цієї умови кожен кортеж інформаційного відношення гіперкуба даних містить деякий унікальний набір значень ключових атрибутів вимірів, якому відповідає значення гіперкуба даних. Зазначимо, що при цьому набори значень ключових атрибутів вимірів, для яких не задано конкретного значення гіперкуба даних, не зберігаються в інформаційному відношенні гіперкуба.

Для уникнення появи помилкових даних, коли в інформаційному відношенні гіперкуба даних присутні кортежі із значеннями ключових полів вимірів, яких немає серед значень ключових полів в інформаційних відношеннях самих вимірів, будемо використовувати обмеження цілісності зовнішнього ключа. Ці обмеження накладемо на кожне ключове поле виміру в інформаційному відношенні гіперкуба даних.

Зазначимо, що якщо схема виміру складається лише з одного атрибута, то для цього виміру не створюється інформаційне відношення, а відповідний атрибут включається в інформаційне відношення гіперкуба даних.

Для реалізації відношення між вимірами будемо використовувати принцип подання в РБД зв'язку типу “один-до-багатьох” між інформаційними відношеннями вимірів. Це означає, що у схему інформаційного відношення підпорядкованого виміру включається ключовий атрибут головного виміру. Для уникнення появи помилкових значень ключового атрибута в кортежах інформаційного відношення підпорядкованого виміру на цей атрибут накладається обмеження цілісності зовнішнього ключа.

На рис. 8 зображені узагальнені схеми інформаційних відношень вимірів та гіперкубів даних, а також показано спосіб реалізації відношень між вимірами.

Зауважимо, що для спрощення опису реалізації багатовимірних подання даних в РБД ми розглядали виміри з одним ключовим атрибутом. Описаний спосіб легко узагальнюється на випадок, коли виміри мають ключі, які складаються з кількох атрибутів.

Розглянемо реалізацію структур багатовимірних подання даних в РБД на прикладі гіперкуба даних *обсяг_продажу*, зображеного на рис. 1.

Приклад 8. Схема гіперкуба даних.

Маємо гіперкуб даних *обсяг_продажу* зі схемою {товар, регіон, час} та вимір *тип_товару*. У свою чергу, схему виміру *товар* утворює множина атрибутів {код_товару, код_типу_товару, назва_товару, одиниця_виміру, ціна} з ключовим атрибутом *код_товару*, схему виміру *регіон* – {код_регіону, назва_регіону, кількість_ мешканців} з ключовим атрибутом *код_регіону*, а схема виміру *час* складається з одного атрибута *квартал*. Вимір *тип_товару* має схему {код_типу_товару, назва_типу_товару}. Крім того, між вимірами *тип_товару* та *товар* існує відношення, яке реалізується за допомогою атрибута *код_типу_товару*.

Виходячи із описаних схем вимірів, з'ясуємо, що схему інформаційного відношення гіперкуба даних *обсяг_виміру* утворює множина {квартал, код_товару, код_регіону, сума_продажу}.

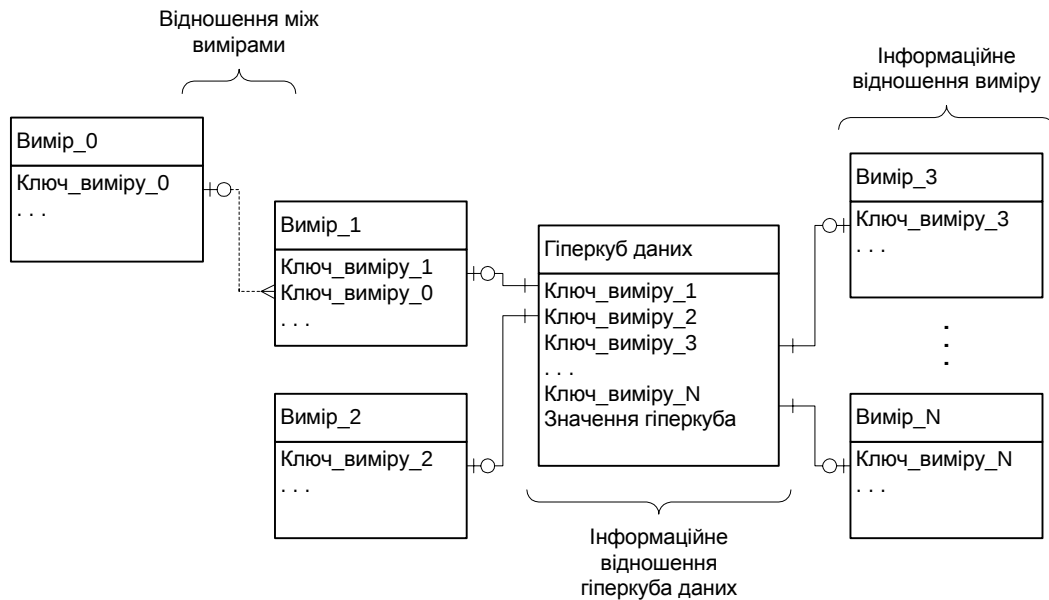


Рис. 8. Схема гіперкуба даних

Схема БД, що реалізує описані у прикладі структури багатовимірної подання даних, зображена на рис. 9.

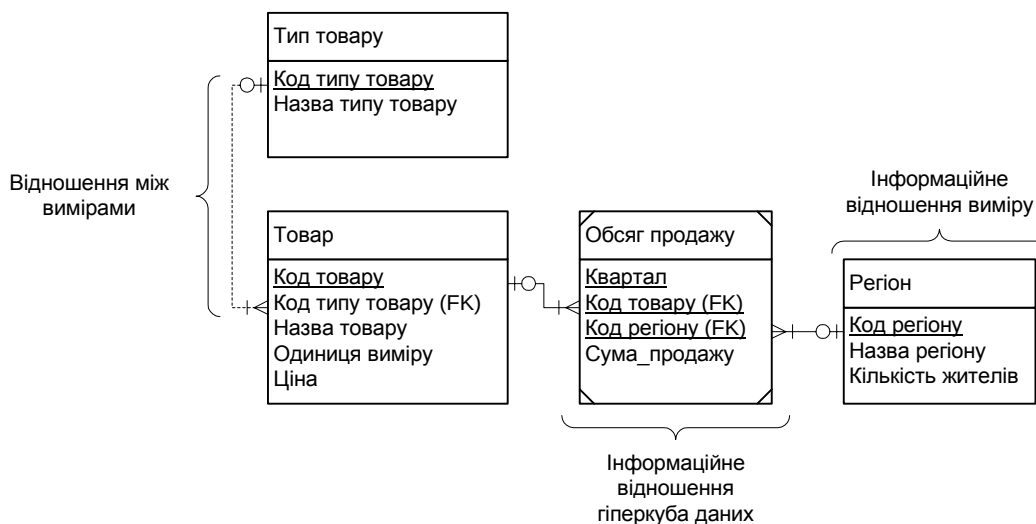


Рис. 9. Схема гіперкуба даних обсяг_продажу

Перейдемо тепер до опису реалізації операцій над гіперкубами даних засобами РБД.

4.4. Реалізація операцій багатовимірної аналізу даних у РБД

Для реалізації запропонованих операцій багатовимірної аналізу даних в РБД буде використувати апарат реляційної алгебри. З цією метою виразимо операції над гіперкубами даних через реляційні операції над інформаційними відношеннями. Опишемо

також реалізацію операцій багатовимірної аналізу даних засобами реляційних СКБД, тобто за допомогою мови запитів SQL.

Наведемо позначення інформаційних відношень, які відповідають вимірам та гіперкубам даних. Інформаційне відношення, яке відповідає виміру D зі схемою $\mathbf{D} = \{K, A_1, A_2, \dots, A_m\}$ та ключем K , будемо позначати як r_D , а схему цього інформаційного відношення – як R_D , причому $R_D = \{K, A_1, \dots, A_m\}$. Нехай тепер маємо гіперкуб даних $H(D_1, D_2, \dots, D_n)$ зі схемою $\mathbf{H} = \{\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_n, \mathbf{V}\}$; виміри цього гіперкуба даних $D_1, D_2, \dots, D_n$ мають схеми $D_1 = \{K_1, A_1^{(1)}, A_2^{(1)}, \dots, A_{m_1}^{(1)}\}$, $D_2 = \{K_2, A_1^{(2)}, A_2^{(2)}, \dots, A_{m_2}^{(2)}\}$, ..., $D_n = \{K_n, A_1^{(n)}, A_2^{(n)}, \dots, A_{m_n}^{(n)}\}$ відповідно. Вимірам $D_1, D_2, \dots, D_n$ відповідають інформаційні відношення $r_{D_1}, r_{D_2}, \dots, r_{D_n}$ зі схемами $R_{D_1} = \{K_1, A_1^{(1)}, A_2^{(1)}, \dots, A_{m_1}^{(1)}\}$, $R_{D_2} = \{K_2, A_1^{(2)}, A_2^{(2)}, \dots, A_{m_2}^{(2)}\}$, ..., $R_{D_n} = \{K_n, A_1^{(n)}, A_2^{(n)}, \dots, A_{m_n}^{(n)}\}$, а гіперкубу даних $H(D_1, D_2, \dots, D_n)$ – інформаційне відношення r_H зі схемою $R_H = \{K_1, K_2, \dots, K_n, V\}$, де $\text{dom}(V) = \mathbf{V}$.

4.4.1. Операція об'єднання за виміром

Нехай $H_1(D^{(1)}, D_1, D_2, \dots, D_n)$ та $H_2(D^{(2)}, D_1, D_2, \dots, D_n)$ – гіперкуби даних зі схемою $\mathbf{H} = \{\mathbf{D}, \mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_n, \mathbf{V}\}$. Гіперкубам даних H_1 та H_2 відповідають інформаційні відношення r_{H_1} та r_{H_2} зі схемою $R_H = \{K, K_1, K_2, \dots, K_n, V\}$. Інформаційні відношення $r_{D^{(1)}}$ та $r_{D^{(2)}}$ зі схемою $R_D = \{K, A_1, A_2, \dots, A_m\}$, які відповідають вимірам $D^{(1)}$ та $D^{(2)}$, задовольняють умову $r_{D^{(1)}} \cap r_{D^{(2)}} = \emptyset$.

У результаті виконання операції об'єднання за виміром отримуємо гіперкуб даних $H(D, D_1, D_2, \dots, D_n)$ зі схемою $\mathbf{H} = \{\mathbf{D}, \mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_n, \mathbf{V}\}$. При цьому інформаційне відношення r_D , яке відповідає виміру D , утворюється як результат об'єднання інформаційних відношень $r_{D^{(1)}}$ та $r_{D^{(2)}}$, тобто:

$$r_D = r_{D^{(1)}} \cup r_{D^{(2)}}.$$

Гіперкубу даних $H(D, D_1, D_2, \dots, D_n)$ відповідає інформаційне відношення r_H зі схемою R_H , яке є результатом об'єднання інформаційних відношень r_{H_1} та r_{H_2} , тобто:

$$r_H = r_{H_1} \cup r_{H_2}.$$

У РБД інформаційне відношення r_D утворюється за допомогою такого SQL-запиту:

```
SELECT  $r_{D^{(1)}}.K, r_{D^{(1)}}.A_1, r_{D^{(1)}}.A_2, \dots, r_{D^{(1)}}.A_m$ 
FROM  $r_{D^{(1)}}$  UNION
SELECT  $r_{D^{(2)}}.K, r_{D^{(2)}}.A_1, r_{D^{(2)}}.A_2, \dots, r_{D^{(2)}}.A_m$ 
FROM  $r_{D^{(2)}}$ .
```

Відповідний SQL-запит, який утворює інформаційне відношення r_H , записується так:

```
SELECT  $r_{H_1}.K, r_{H_1}.K_1, r_{H_1}.K_2, \dots, r_{H_1}.K_n, r_{H_1}.V$ 
```

```

FROM  $r_{H_1}$  UNION
SELECT  $r_{H_2}.K, r_{H_2}.K_1, r_{H_2}.K_2, \dots, r_{H_2}.K_n, r_{H_2}.V$ 
FROM  $r_{H_2}$ .

```

4.4.2. Операція зрізу

Нехай маємо гіперкуб даних $H_1(D, D_1, D_2, \dots, D_n)$ зі схемою $\mathbf{H}_1 = \{\mathbf{D}, \mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_n, \mathbf{V}\}$. Виміру D відповідає інформаційне відношення r_D зі схемою $R_D = \{K, A_1, A_2, \dots, A_m\}$, а гіперкубу даних H_1 – інформаційне відношення r_{H_1} зі схемою $R_{H_1} = \{K, K_1, K_2, \dots, K_n, V\}$. Для виміру D зафіксуємо значення d^* . Цьому значенню відповідає кортеж $(k^*, a_1^*, a_2^*, \dots, a_m^*)$ інформаційного відношення r_D , який, у свою чергу, повністю визначається ключовим значенням k^* . Отже, операція зрізу за значенням $D = d^*$ над гіперкубом даних H_1 реалізується за допомогою операції вибірки за значенням $K = k^*$ з інформаційного відношення r_{H_1} . У результаті отримуємо гіперкуб даних $H(D_1, D_2, \dots, D_n)$ зі схемою $\mathbf{H} = \{\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_n, \mathbf{V}\}$, якому відповідає інформаційне відношення r_H зі схемою $R_H = \{K_1, K_2, \dots, K_n, V\}$.

Інформаційне відношення r_H утворюється як результат виразу:

$$r_H = \pi_{R_H}(\sigma_{K=k^*}(r_{H_1})).$$

Відповідний SQL-запит, який реалізує операцію зрізу гіперкуба даних в РБД, записується так:

```

SELECT  $r_{H_1}.K_1, r_{H_1}.K_2, \dots, r_{H_1}.K_n, r_{H_1}.V$ 
FROM  $r_{H_1}$ 
WHERE  $r_{H_1}.K = k^*$ .

```

4.4.3. Операція проєкції

Нехай маємо гіперкуб даних $H_1(D, D_1, D_2, \dots, D_n)$ зі схемою $\mathbf{H}_1 = \{\mathbf{D}, \mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_n, \mathbf{V}\}$. Йому відповідає інформаційне відношення r_{H_1} зі схемою $R_{H_1} = \{K, K_1, K_2, \dots, K_n, V\}$. Результатом виконання операції проєкції є гіперкуб даних $H(D_1, D_2, \dots, D_n)$ зі схемою $\mathbf{H} = \{\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_n, \mathbf{V}\}$, якому відповідає інформаційне відношення r_H зі схемою $R_H = \{K_1, K_2, \dots, K_n, V\}$. Інформаційне відношення r_H отримується внаслідок агрегування значень атрибута V інформаційного відношення r_{H_1} з групуванням за атрибутами $K_1, K_2, \dots, K_n$, тобто:

$$r_H = A_{K_1 K_2 \dots K_n}(r_{H_1}; \text{Agg}(V)).$$

Операція проєкції реалізується за допомогою такого SQL-запиту:

```

SELECT  $r_{H_1}.K_1, r_{H_1}.K_2, \dots, r_{H_1}.K_n, \text{Agg}(r_{H_1}.V)$ 
FROM  $r_{H_1}$ 
GROUP BY  $r_{H_1}.K_1, r_{H_1}.K_2, \dots, r_{H_1}.K_n$ .

```

4.4.4. Операція згортки за відношенням

Нехай маємо гіперкуб даних $H_1(D_1, D_2, \dots, D_n)$ зі схемою $H_1 = \{D_1, D_2, \dots, D_n, V\}$, вимір D зі схемою D та відношення $r(D, D_1)$ між вимірами D та D_1 . Виміру D відповідає інформаційне відношення r_D зі схемою $R_D = \{K, A_1, A_2, \dots, A_m\}$, виміру D_1 – інформаційне відношення r_{D_1} зі схемою $R_{D_1} = \{K_1, K, A_1^{(1)}, A_2^{(1)}, \dots, A_m^{(1)}\}$, а гіперкубу даних H_1 – інформаційне відношення r_{H_1} зі схемою $R_{H_1} = \{K_1, K_2, \dots, K_n, V\}$. У результаті виконання операції згортки гіперкуба даних $H_1(D_1, D_2, \dots, D_n)$ за відношенням між вимірами D та D_1 отримаємо гіперкуб даних $H(D, D_2, \dots, D_n)$ зі схемою $H = \{D, D_2, \dots, D_n, V\}$, якому відповідає інформаційне відношення r_H зі схемою $R_H = \{K, K_2, \dots, K_n, V\}$. Інформаційне відношення r_H отримується так:

$$r_H = A_{KK_1 \dots K_n} (r_{H_1} \triangleright \triangleleft r_{D_1}; \text{Agg}(V)).$$

Відповідний SQL-запит, який реалізує цю операцію, записується так:

```
SELECT  $r_{D_1}.K, r_{H_1}.K_2, \dots, r_{H_1}.K_n, \text{Agg}(V)$ 
FROM  $r_{D_1}, r_{H_1}$ 
WHERE  $r_{D_1}.K_1 = r_{H_1}.K_1$ 
GROUP BY  $r_{D_1}.K, r_{H_1}.K_2, \dots, r_{H_1}.K_n$ .
```

5. ВИСНОВКИ

У статті відображені такі результати:

- побудовано математичну модель багатовимірного подання даних в РБД на основі застосування апарата теорії реляційних баз даних. Математично описано поняття виміру, гіперкуба даних та відношення між вимірами. Формалізовано основні операції над гіперкубами даних та досліджено властивості цих операцій;
- показано спосіб реалізації структур багатовимірного подання даних засобами РБД. Реалізовано операції над гіперкубами даних засобами реляційної алгебри та мови запитів SQL.

1. Codd E.F., Codd S.B., Salley C.T. *Providing OLAP (on-line analytical processing) to user-analysts: an IT mandate*. – E.F. Codd & Associates, 1993. 2. Han J., Cai Y., Cercone N. *Knowledge discovery in databases: an attribute-oriented approach* // *Proc. Int. Conf. Very Large Databases*. – Vancouver (Canada). – 1992. – P. 547–559. 3. Han J., Fu Y. *Exploration of the power of attribute-oriented induction in data mining* / Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. (editors). *Advances in knowledge discovery and data mining*. – AAAI/MIT. – 1996. – P. 399–421. 4. Gupta A., Harinarayan V., Quass D. *Aggregate-query processing in data warehousing environment* // *Proc. Int. Conf. Very Large Databases*. – Zurich (Switzerland). – 1995. – P. 358–369. 5. Harinarayan V., Ullman J.D., Rajaraman A. *Implementing data cubes efficiently* // *Proc. ACM SIGMOD Int. Conf. Management of Data*. – Montreal (Canada). – 1996. 6. Кравець Р.Б. *Багатовимірна модель даних у системах аналітичної обробки інформації* // *Вісн. Нац. ун-ту "Львівська політехніка"*. – 1998. – № 330. – С. 147–153.