УДК 681. 32

ФОРМАЛІЗАЦІЯ ПРЕДМЕТНОЇ ОБЛАСТІ БАГАТОВИМІРНИХ БАЗ ДАНИХ

к.т.н. Г.А. Кучук
(подав проф. А.В. Корольов)

Запропоновано підхід до побудови формалізованої моделі предметної області і специфікацій інформаційних вимог користувачів для багатовимірних баз даних, які використовують OLAP - технологію.

Основна проблема, яка виникає при опрацюванні великих об'ємів інформації, що міститься в розподіленій базі даних (РБД) із структурою, яка постійно ускладнюється – це збереження прийнятного часу реакції системи на запит. Застосування архітектури “клієнт - сервер” дозволяє встановлювати спеціалізовані сервери, оптимізовані для розв‘язання задач специфічного управління даними. При цьому потрібно використовувати запросо - центровані схеми баз даних (БД), орієнтовані на використання багатовимірних масивів (БМ), для яких найбільш ефективною є технологія інтерактивного аналітичного опрацювання даних OLAP (*Online Analytical Processing*), що дозволяє проводити динамічний синтез, аналіз і консолідацію великих об'ємів багатовимірних даних [1]. Консолідація дозволяє виконувати узагальнюючі операції із зв'язаними даними. За рахунок динамічного вибору способу фізичного збереження даних і використання ефективних технологій стиску даних досягається максимальне використання вільного простору як зовнішньої, так і оперативної пам'яті. Це також дозволяє завантажувати більше реальних даних в оперативну пам'ять, що приводить до істотного підвищення продуктивності.

Крім консолідації OLAP - сервери баз даних можуть виконувати ряд аналітичних операцій із даними багатовимірних масивів. Так, досить суттєвою є операція, зворотна консолідації – “спадний аналіз” (*drill - down*), яка дозволяє відобразити більш дрібні деталі розглянутих консолідованих даних. Для аналізу тенденцій і пошуку закономірностей призначена операція “розбивка з поворотом” (*slicing and dicing*).

OLAP - технологія пред'являє ряд вимог до OLAP - інструментарію ([2], 12 правил Кодда), що перераховані нижче.

1. Можливість багатомірного концептуального уявлення даних.
2. Прозорість бази даних, архітектури даних і неоднорідних джерел вхідних даних.
3. Повний доступ до необхідних для аналізу даних.
4. Збереження стійкості функціонування розроблених за OLAP - технологією систем при збільшенні вимірності даних.

5. Ефективність функціонування в середовищі “клієнт – сервер”.
6. Еквівалентність вимірів за структурою та можливостями.
7. Адаптація фізичної схеми розробленої системи до конкретної аналітичної моделі з метою динамічної оптимізації розрідженої матриці.
8. Підтримка багатокористувального інтерфейсу.
9. Можливість розпізнавання ієрархії вимірностей та автоматичного виконання перехресних узагальнюючих обчислень серед вимірностей.
10. Простота аналітичного маніпулювання даними.
11. Гнучкість засобів формування звітів.
12. Відсутність обмежень на кількість вимірів або рівнів узагальнення.

Багатовимірний або реляційний OLAP – інструмент повинен підтримувати процес розробки даних (*data mining*): діставання з великих баз даних достовірної, попередньо невідомої, комплексної та значимої інформації з метою використання її в процесі прийняття рішень [3]. Методи розробки даних звичайно базуються на таких операціях, як прогнозуєме моделювання, сегментування баз даних, аналіз зв‘язків та знаходження відхилень [4].

При проектуванні баз даних, які працюють з багатовимірними масивами за OLAP - технологією, необхідна попередня формалізація предметної області прикладного програмного забезпечення. Нехай $\mathbf{F = \{f_i \mid i = \overline{1,I}\}}$ - множина функцій, які підлягають автоматизації; $\mathbf{Z = \{z_j \mid j = \overline{1,J}\}}$ - множина плануємих задач обробки даних; $\mathbf{U = \{u_k \mid k = \overline{1,K}\}}$ - множина користувачів системи; $\mathbf{P = \{p_m \mid m = \overline{1,M}\}}$ - множина процесів та об‘єктів автоматизації; $\mathbf{V = \{v_l \mid l = \overline{1,L}\}}$ - множина інформаційних елементів (вхідних і вихідних даних) предметної області системи. На парах вищеперелічених множин визначимо реляційні відношення $\mathbf{r_n}$, які описуються булевими матрицями суміжності та складають множину реляційних відношень $\mathbf{R = \{r_n \mid n = \overline{1,N}\}}$. Виділимо такі типи реляційних відношень предметної області БД [5]: $\mathbf{FZ = (fz_{ij})}$; $\mathbf{FU = (fu_{ik})}$; $\mathbf{FP = (fp_{im})}$; $\mathbf{FV = (fv_{il})}$; $\mathbf{ZU = (zu_{jk})}$; $\mathbf{ZP = (zp_{jm})}$; $\mathbf{ZV = (zv_{jl})}$; $\mathbf{PV = (pv_{ml})}$. Модель предметної області – це кортеж $\mathbf{M_{пр} = <F, Z, U, P, V, R>}$, що описується відповідними множинами та реляційними відношеннями.

Визначимо множину інформаційних елементів системи $\mathbf{H = P \cup V = \{h_s \mid s = 1,S\}}$. Нехай $\mathbf{H_k = \{h_{ks} \mid s = \overline{1,S}\}}$ - множина інформаційних елементів, що описують предметну область **k** - го користувача. Тоді його відношення до множини **H** визначається у вигляді бінарного вектора реляційного відношення $\mathbf{UH = (uh_{ks})}$. Подамо специфікацію інформаційних вимог користувача **k** у вигляді бінарної моделі $\mathbf{M^k_{сп} = < h_{ks_1}, r_n, h_{ks_2} >}$, де $\mathbf{h_{ks_1}, h_{ks_2} \in H_k}$ -

структурні елементи моделі, пов‘язані відношенням $\mathbf{r_n}$. Для побудови $\mathbf{M^k_{сп}}$, по-перше, необхідно провести аналіз матриць **FU** та **ZU** для визначення повного переліку функцій і задач обробки, плануємих для використання. Далі за матрицею **FP** вибираються потрібні процеси та об‘єкти і визначаються відношення між ними $<\mathbf{p_1,r^{(p)},p_2}>$. Аналіз матриці **PV** дозволяє встановити належність вхідних та вихідних інформаційних елементів кожному із процесів і об‘єктів, що розглядаються.

Для забезпечення ефективного функціонування багатовимірної бази даних в середовищі “клієнт -сервер” необхідно вже на перших стадіях розробки визначити варіанти розбиття користувачів на групи на основі аналізу спільності їх предметних областей. Для цього розглянемо множину інформаційних елементів системи, що містить елементи, які повторюються, з різних множин $\mathbf{H_k}$: $\mathbf{H^{(o)}=\bigcup_{k=1}^{K} H_k}$. Для кожного $\mathbf{H_k}$ визначимо множини $\mathbf{H_k^{(\alpha)} = H^{(o)} - H_k}$, $\mathbf{H_k^{(p)} = H_k^{(\alpha)} \cap H_k, k \in \overline{1,K}}$. Якщо $\mathbf{H_k^{(p)} \neq \varnothing}$, $\mathbf{card\,(H_k^{(\alpha)}) > N_{дост}}$, де $\mathbf{N_{дост}}$ - задана величина, то предметна область користувача **k** дозволяє розглядати його у якості представника однієї з груп користувачів РБД. Чисельне значення $\mathbf{N_{дост}}$ визначимо за допомогою міри подібності $\mathbf{\mu_k : (H_k, H_k^{(\alpha)}) \to [0,1]}$, яка враховує загальні ($\mathbf{p_{11}^{(k)}}$), специфічні для $\mathbf{H_k}$ ($\mathbf{p_{10}^{(k)}}$) та для $\mathbf{H_k^{(d)}}$ ($\mathbf{p_{01}^{(k)}}$) інформаційні елементи [6]:

$$\mu_k = p_{11}^{(k)} / \left(p_{11}^{(k)} + 2(p_{10}^{(k)} + p_{01}^{(k)})\right) = p_{11}^{(k)} / \left(n_0^{(k)} + p_{10}^{(k)} + p_{01}^{(k)}\right), \qquad (1)$$

де $\mathbf{n_0^{(k)} = p_{11}^{(k)} + p_{10}^{(k)} + p_{01}^{(k)}}$ - загальна кількість інформаційних елементів.

Виходячи із заданих реляційних відношень, визначимо:

$$p_{11}^{(k)} = \sum_{n=1}^{n_0} \alpha_n^{(k)};\ \ \alpha_n^{(k)} \in \{0;1\};\ \ \alpha_n^{(k)} = 1 \Leftrightarrow \exists n : \left(h_n \in H_k\right) \& \left(\sum_{k=1}^{K} uh_{kn} \geq 1 \right); \qquad (2)$$

$$p_{10}^{(k)} = \sum_{n=1}^{n_0} \beta_n^{(k)};\ \ \beta_n^{(k)} \in \{0;1\};\ \ \beta_n^{(k)} = 1 \Leftrightarrow \exists n : \left(h_n \in H_k\right) \& \left(\sum_{k=1}^{K} uh_{kn} = 1 \right); \qquad (3)$$

$$p_{01}^{(k)} = \sum_{n=1}^{n_0} \gamma_n^{(k)};\ \ \gamma_n^{(k)} \in \{0;1\};\ \ \gamma_n^{(k)} = 1 \Leftrightarrow \exists n : \left(h_n \notin H_k\right) \& \left(\sum_{k=1}^{K} uh_{kn} \geq 1 \right). \qquad (4)$$

З врахуванням (2 - 4) функція подібності (1) прийме вигляд

$$\mu_k = \sum_{n=1}^{n_0^{(k)}} \alpha_n^{(k)} / \left(n_0^{(k)} + \sum_{n=1}^{n_0} (\beta_n^{(k)} + \gamma_n^{(k)})\right).$$

Задамо на множині користувачів $\mathbf{U = \{u_k\}}$ відношення належності

$\mathbf{R_п}$ наступним чином: $\mathbf{u_k \in U_p \overset{R_п}{\Leftrightarrow} \mu_k \geq \mu_0^*}$, де $\mathbf{U_p \subset U}$ - множина розподілених користувачів; $\mathbf{\mu_0^*}$ - критична міра подібності. Якщо $\mathbf{u_k \notin U_p}$, то користувач у своїй предметній області працює з локальною базою даних.

Для застосування мережної архітектури "клієнт – сервер" отримані результати використовуються при визначенні інформаційного складу мережних БД, розміщених на сервері. При цьому для всіх $\mathbf{u_k \in U_p}$ визначається безнадмірна підмножина інформаційних елементів $\mathbf{H_p \subseteq H}$: $\mathbf{H_p = \bigcup\limits_{k:\, u_k \in U_p} H_k}$. Потім на основі варіювання величини міри подібності визначаються кластери, що виділяються на основі узагальнення їх предметних областей [7]: для $\mathbf{\delta = (1-\mu_0^*)/N_\delta}$, де $\mathbf{N_\delta}$ - число інтервалів варіювання подібності, визначається група користувачів $\mathbf{U_p^{(i)}}$, для якої $\mathbf{R_п(U_p^{(i)}) \in [\mu_i^*, \mu_i^* + \delta)}$. Якщо $\mathbf{card\,(H_p^{(i)}) \geq N_{дост}}$, то відповідний кластер виділяється і починається формування наступного кластера. У результаті формується ієрархічна структура мінімально інформаційно пов'язаних між собою кластерів користувачів (як з $\mathbf{U_p}$, так і з $\mathbf{U \backslash U_p}$), для кожного з яких у подальшому визначається інформаційний склад бази даних.

Розглянемо інформаційні вимоги користувачів $\mathbf{\nu}$ - го кластеру $\mathbf{K_\nu}$, що складають множину $\mathbf{T^{(\nu)} = \{ t_k \mid k = \overline{1, K^{(\nu)}} \}}$. Нехай до складу вимоги $\mathbf{t_k}$ входять структурні елементи предметної області, що складають множину $\mathbf{H_k^{(\nu)} \subset \bigcup\limits_{u \in K_\nu} H_u}$. Тоді матриця семантичної суміжності вимоги $\mathbf{k}$ кластеру $\mathbf{K_\nu}$, означена як $\mathbf{B_k^{(\nu)} = (b_{ij}^{(k,\nu)})}$, має вимірність $\mathbf{card(H_k^{(\nu)})}$, а $\mathbf{b_{ij}^{(k,\nu)} = 1}$ тоді і тільки тоді, якщо між структурними елементами $\mathbf{h_i}$ і $\mathbf{h_j}$ існує відношення $\mathbf{r_{ij}}$, таке, що елемент $\mathbf{h_i}$ розширює елемент $\mathbf{h_j}$.

Для з'ясування взаємозв'язків між структурними елементами з метою визначення складу інформаційних груп побудуємо матрицю семантичної досяжності $\mathbf{A_k^{(\nu)} = (a_{ij}^{(k,\nu)})}$, одиничний елемент якої визначає наявність шляху на орграфі $\mathbf{G(B_k^{(\nu)})}$ з $\mathbf{h_i}$ до $\mathbf{h_j}$ із певним змістом, що має сенс. На підставі матриці $\mathbf{A_k^{(\nu)}}$ будуються попередня множина $\mathbf{C(h_i)}$ - їй відповідають одиничні елементи $\mathbf{i}$ - го стовпця, та множина досяжності $\mathbf{D(h_j)}$ - їй відповідають одиничні елементи $\mathbf{j}$ - го рядка. Аналіз цих множин дозволяє побудувати інформаційну структуру

предметної області кластеру $\mathbf{K_v}$. Одна з її складових - це окремі інформаційні елементи (висячі вершини орграфа $\mathbf{G(B_k^{(v)})}$), що складають множину $\mathbf{H_k^{(v,0)}}$ - нульовий рівень ієрархії $\mathbf{P_0}$. Інші елементи складають окремі структурні групи більш низьких рівнів ієрархії – множини $\mathbf{H_k^{(v,m)}}$, де індекс $\mathbf{m}$ позначає належність до рівня ієрархії $\mathbf{P_m}$, яка визначається за допомогою матриці зв'язків між відповідними групами $\mathbf{A_k^{(v,g)} = (a_{ij}^{(k,v,g)})}$. До рівня $\mathbf{P_1}$ віднесемо всі групи, для яких $\mathbf{C(h_i^{(g)}) \cap D(h_i^{(g)}) = D(h_i^{(g)})}$. Належність до інших рівнів ієрархії визначається ітеративно із співвідношення

$$\mathbf{P_m = \{h_i^{(g)} \in H_k^{(v)} \setminus P_1 \setminus \ldots \setminus P_{m-1} \;\Big|\; C_{m-1}(h_i^{(g)}) \cap D_{m-1}(h_i^{(g)}) = D_{m-1}(h_i^{(g)})\}},$$

де $\mathbf{C_{m-1}(h_i^{(g)})}$ и $\mathbf{D_{m-1}(h_i^{(g)})}$ - відповідні множини, які визначені на підмножині $\mathbf{H_k^{(v)} \setminus P_1 \setminus \ldots \setminus P_{m-1}}$.

Упорядкування інформаційних груп за рівнями ієрархії дозволяє виділити для кожного кластера кореневі і проміжні групи й остаточно визначити як структуру предметної області багатовимірної бази даних, так і її зв'язок із зовнішнім середовищем.

ЛІТЕРАТУРА

1. Berson A., Smith S. Data Warehousing, Data Mining & OLAP. – McGraw Hill Comp.Inc., 1997. – 586 p.

2. Codd E., Codd S., Salley C. Providing OLAP to User - Analysts. – Arbor Software Corp., 1993. – *http://www.arborsoft.com/essbase/wht_ppr/coldToc.html.*

3. Simoudis E. Reality Check for Data Mining. – IEEE Expert. – 1996. – Oct. – P. 26 - 33.

4. Cabena P., Hadjinian P., Stadler R. Discovering Data Mining from Concept to Implementation. – New Jersey, USA: Prentice - Hall Ptr, 1997. – 344 p.

5. Сахаров А.А. Концепция построения и реализации информационных систем, ориентированных на анализ данных // СУБД. – 1996. – № 2. – С. 55 - 70.

6. Коллинз Г., Блей Дж. Структурные методы разработки систем. От стратегического планирования до тестирования. – М.: Финансы и статистика, 1986. – 264 с.

7. Мандель И. Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с.

Подана до редколегії 13.11.2000