Проектування гібридних сховищ даних з врахуванням джерел даних як задача оптимізації

Яцишин А. Ю.
Національний технічний університет України «Київський Політехнічний Інститут», andrew.yatsyshyn@hotmail.com

Building of hybrid data warehouses considering data sources as optimization problem is discussed in this paper. Author analyses existing methods to solve optimization problems and solutions that use these methods. Then author explains his choice of family of genetic algorithms and presents his own adaptation.

Вступ

Проектування сховища даних інформаційної системи є важливим етапом її створення. Від сховища даних суттєво залежить швидкодія такої системи. Крім того, при певній побудові сховища даних (наприклад, з використанням OLAP) забезпечується гнучкість інформаційної системи завдяки ефективному створенню нових представлень даних (для звітів).

Однак у сучасних інформаційних системах часто виникає необхідність поєднати високу швидкодію з гнучкістю системи. Для таких вимог доцільно використовувати гібридне сховище даних, що поєднує переваги реляційної бази даних OLAP і багатовимірної бази даних OLTP.

Однак використанні гібридного сховища даних необхідно вирішувати питання про розподіл даних між базами даних. Це може бути зроблене в ручному режимі, однак такий підхід має суттєві недоліки у ситуації, коли сховище даних змінюється при отриманні нових даних.

У таких випадках доцільно здійснювати автоматичне перепроектування сховища даних відповідно як до даних, що додаються, так і до запитів, які виконуються до цих даних. Постає задача пошуку оптимальної структури сховища даних, що є по суті задачею оптимізації.

Постановка задачі

Запишемо постановку задачі проектування гібридних сховищ даних так, як це зроблено в [4] та [5].

Задані множини атрибутів розділених файлів S, файлів XML X, відношення у реляційній базі даних R, виміри багатовимірної бази даних D та міри M. Крім того, відоме порогове значення частот доступу до даних. Спроектувати гібридне сховище даних, визначивши області сховища даних A, таблиці T та атрибути B.

Знайти такі значення ознак розміщення L_a, індексування I_c, матеріалізації M_c, а також ознаки джерел даних у областях A_s, на яких значення цільової функції

$$z = (1+\sum_{s=1}^{n_s} A_S(\frac{T_{sb}}{\hat{T}_{sb}} - \frac{T_{sq}}{\hat{T}_{sq}}))\times$$
$$\times(\sum_{t=1}^{n} t_i(L_a,\{I_c\},\{M_c\}) + \sum_{j=1}^{n-1} T_j + (T_a^{'} - \hat{T}_a)L_a) \quad (1)$$

буде мінімальним серед всіх можливих наборів значень цих змінних.

Критерієм оптимальності сховища є кількість доступів до даних, тобто операцій читання даних, які необхідно провести для виконання запиту до сховища даних.

Змінні T_{ab} та T_{aq} при цьому позначають часи проектування та виконання запитів відповідно. Вони отримуються з сховища в ході розв'язання задачі.

Існуючі методи та рішення

Згідно [3] та [4] існують такі методи вирішення задач оптимізації, застосовні в умовах, коли аналітичний запис ЦФ не відомий (випадкового пошуку):

1. Метод з поверненням на невдачному кроці
2. Метод найкращої проби
3. Метод випадкового пошуку, що повторюється
4. Метод випадкового пошуку з постійним радіусом пошуку та випадковими напрямками
5. Метод Монте-Карло

З цих методів доцільно використовувати сімейство генетичних алгоритмів, оскільки воно дозволяє швидше досягнути оптимуму за рахунок відсіювання неоптимальних рішень (непридатних особин). Наведені алгоритми не дозволяють відсіювати особини настільки, щоб можна було говорити про їх ефективність. Крім того, відомо, що генетичні алгоритми дозволяють знайти те ж рішення, що й метод Монте-Карло за меншу кількість кроків

Генетичні алгоритми використовуються в низці рішень проектування та оптимізації сховищ даних. Прикладом такого рішення є наведене в [1]. Такі рішення можна використати використати для оптимізації реляційних чи багатовимірних баз даних, але їх недостатньо для задачі проектування гібридних сховищ даних.

У зв'язку з цим я пропоную розв'язувати задачу проектування гібридних сховищ даних з врахуванням сховищ даних за допомогою адаптивного генетичного алгоритму. Адаптивність алгоритму полягає в тому, що для кожного наступного покоління знаходимо пари "батько-нащадок", які мають найбільшу кількість спільних і найменшу кількість змінених генів, співставляємо ті гени, які змінилися від їх батьків і якщо маємо значення ЦФ гірше, то ці гени фіксуються у значенні, протилежному тому, на яке вони змінилися.

Проілюструємо це за допомогою малюнку 1..

Рисунок 1. Ілюстрація адаптивного генетичного алгоритму

На рисунку 1 показані по вертикалі популяції особин у порядку розвитку, по горизонталі – їх особини. Особини популяцій співставлені за батьківством. Синім кольором виділені активність гена (ген є) в особині, білим – пасивість (гену немає). Використання такого фіксування виключає подальшу появу «хворобливих» генів в особинах популяції. У випадку нашої задачі це особливо актуально, бо гени впливають в тому числі і на області сховища, а при певних генах може бути суттєво погіршене значення як часу виконання запитів, так і часу перепроектування сховища.

Висновки

У даній статті розглянуто існуючі підходи та рішення стосовно проектування гібридних сховищ даних. Крім того, запропоновано використовувати адаптивний генетичний алгоритм, що дозволяє суттєво підвищити ефективність проектування гібридних сховищ даних.

Література

[1] Wen-Yang Lin. A Genetic Selection Algorithm for OLAP Data Cubes [Текст], Knowledge and information systems, vol. 6 / Wen-Yang Lin,I-Chung Kuo – 2004

[2] Корнеенко В. П. Методы оптимизации [Текст] / Корнеенко В. П. – М.: Высшая школа, 2007. – 664 с.-
[3] Методы оптимизации (базовый курс) / Режим доступу: http://bigor.bmstu.ru/?cnt/?doc=MO/base.cou
[4] Томашевський В.М.Математична модель задачі проектування гібридних сховищ даних з врахуванням структур джерел даних [Текст]. Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка: Зб. наук. пр. /Томашевський В.М.,Яцишин А.Ю. – К.:Век+, – 2011. – № 53. – 211
[5] Томашевський В.М. Особливості проектування гібридних сховищ даних з врахуванням джерел даних [Текст]. Подано до друку у Вісник Національного університету „Львівська політехніка”, секція "Інформаційні системи та мережі" / Томашевський В.М.,Яцишин А.Ю. - 2012