

розміру відеопамяті, потрібної для роботи ПЗ, *mincpurate* - мінімальне значення тактової частоти процесора, потрібної для роботи ПЗ.

Решта найменувань взяті з бази даних і використовують для вказування зв'язку між відношеннями (так звані ключі).

### Висновок

У роботі поставлено задачу пошуку комп'ютерів з заданими вимогами до потужності їхнього апаратного забезпечення, розроблено теоретичні співвідношення процедури пошуку. Метод базується на алгебрі реляційних відношень та спеціальному алгоритмі порівняння апаратних характеристик. Реалізація цього методу дасть змогу пришвидшити пошук підходящих для встановлення деякого ПЗ комп'ютерів, що підвищить продуктивність роботи користувачів.

### Список використаних джерел

1. [Електронний ресурс]. — Режим доступу <http://www.w3schools.com>
2. [Електронний ресурс]. — Режим доступу [http://uk.wikipedia.org/wiki/%D0%A0%D0%B5%D0%BB%D1%8F%D1%86%D1%96%D0%B9%D0%BD%D0%B0\\_%D0%B0%D0%BB%D0%B3%D0%B5%D0%B1%D1%80%D0%B0](http://uk.wikipedia.org/wiki/%D0%A0%D0%B5%D0%BB%D1%8F%D1%86%D1%96%D0%B9%D0%BD%D0%B0_%D0%B0%D0%BB%D0%B3%D0%B5%D0%B1%D1%80%D0%B0)
3. [Електронний ресурс]. — Режим доступу <http://sqld.info/uk/10/54.html>

УДК 004.6

## ВРАХУВАННЯ СТРУКТУРОВАНОСТІ ДАНИХ ПРИ ПРОЕКТУВАННІ ГІБРИДНИХ СХОВИЩ ДАНИХ

Яцишин А.Ю.

*Національний технічний університет України «Київський політехнічний інститут», аспірант*

### I. Постановка проблеми

Питання проектування сховищ даних завжди було важливим етапом створення інформаційних систем. В зв'язку з необхідністю поєднання швидкодії реляційних баз даних автором були запропоновані гібридні сховища даних. Зокрема, була побудована математична модель та описані особливості проектування гібридних сховищ даних з врахуванням джерел даних.

В існуючих на сьогодні наукових публікаціях розглядаються різні аспекти, пов'язані з роботою з структурованими даними. Однак питання врахування структурованості даних, що містяться в джерелах даних, з яких проектується сховище даних, зокрема гібридне, не було розглянуто.

### II. Аналіз останніх досліджень та публікацій

У роботі [1] розглядаються схеми інтеграції структурованих та напівструктурованих даних, у роботі [2] – інтеграція структурованих та текстових даних. У роботі [3] – робота з деревоподібними структурами. У [5] – адаптивна обробка структурованих даних. У роботах [4],[8],[10] – інтеграція неструктурованих, напівструктурованих та структурованих даних. У роботах [6], [7] – обробка напівструктурованих та структурованих даних. У роботі [9] – виконання запитів до напівструктурованих даних

У статтях [10] і [11] розглядаються питання проектування гібридних сховищ даних з врахуванням джерел даних.

### III. Мета роботи

Ціллю роботи є описати особливості процесу проектування гібридними сховищами даних, пов'язані з структурованістю даних та розширити концепцію гібридних сховищ даних для врахування структурованості даних.

### IV. Виклад основного матеріалу

У роботі [12] представлена математична модель проектування задачі гібридних сховищ даних з врахуванням джерел даних. Вона сформульована наступним чином:

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. Спроекувати гібридне сховище даних, визначивши області сховища даних  $A$ , таблиці  $T$  та атрибути  $B$ .

Знайти такі значення ознак розміщення  $L_a$ , індексування  $I_c$ , матеріалізації  $M_c$ , на яких значення цільової функції часу виконання запиту буде мінімальним серед всіх можливих наборів значень цих змінних.

При цьому джерела можна поділити на 3 основні типи за структурованістю даних:

1. Реляційні бази даних, особливо нормалізовані (знаходяться у 3НФ), та розділені файли є структурованими, оскільки атрибутами відношень чітко задається структура даних.
2. Файли XML та багатовимірні бази даних є напів-структурованими, оскільки можуть містити дані не по всіх з певної множини атрибутів, однак у сукупності можна виділити їх загальну структуру.

3. Нерозділені файли є неструктурованими, оскільки не мають метаданих.

При тому структурованість даних суттєво впливають на дисковий простір, який займає сховище даних, бо у випадку неструктурованих та напівструктурованих даних вони можуть зберігатися розріджено. Крім того, з цього випливає деяке падіння швидкодії сховища даних, яке є істотно залежним від операцій введення – виведення.

Прикладом структурованих даних база паспортів, прикладом напів-структурованих – база медичних карток, неструктурованих – текстові дані.

У зв'язку з цим я пропоную при проектуванні сховища даних визначати структурованість сховища даних та розміщувати дані у сховищі відповідно до їх структурованості

Перш за все, структурованість джерела  $S$  можна визначати наступним співвідношенням:

$$ST_s = \sum_{i=1}^n N_{A_i} / nN_s \quad (1),$$

де  $n$ - кількість записів джерела даних,  $N_A$  - кількість атрибуту запису,  $N_S$  - загальна кількість сховища. При цьому значення змінюються від 0 (неструктуровані дані) до 1 (структуровані дані)

Для розрахунку структурованості приймаються наступні евристики:

1. Для реляційних баз даних в загальному випадку  $ST=1$ , при врахуванні пустих значень («null»), значення можна обрахувати для кожної таблиці за формулою (1), аналогічно для плоских розподілених файлів;
2. Для багатовимірних баз даних структурованість може бути визначена для кожного розрізу куба за формулою (1), при цьому сильно структуровані дані підлягають перерозподілу у реляційну базу даних, а неструктуровані – у базу даних XML.
3. Нерозподілені файли (якщо такі є) мають  $ST=0$ .

На основі інформації про структурованість даних можна покращити алгоритм ініціалізації сховища, представлений у [12], розподіливши дані у сховище наступним чином:

1. Дані з високою структурованістю розміщуються в реляційній базі даних, побудованій у ЗНФ. При такому розміщенні ці дані адаптуються під частий доступ та типові запити.
2. Дані з низькою структурованістю розміщуються в багатовимірній базі даних. При такому розміщенні ці дані адаптуються під аналіз та гнучкі запити.
3. Неструктуровані дані чи дані, структурованість яких не може бути визначена, розміщуються в базі даних XML.

При цьому вищезгаданий розподіл корисний у випадках, коли інформація про запити, що будуть оброблятися системою, невідома.

### Список використаних джерел

4. Catriel Beeri. Schemas for Integration and Translation of Structured and Semi-Structured Data [Текст]. Database Theory — ICDT'99/ Catriel Beeri.,Tova Milo - 1999
5. David A. Grossman. Integrating Structured Data and Text [Текст]. Journal of the American Society of Information Science/ David A. Grossman - 1997
6. Frank Neven. Expressive and efficient pattern languages for tree-structured data. [Текст]. PODS '00 Proceedings of the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems/ Frank Neven, Thomas Schwentick New York - 2000
7. Georg Gottlob. Monadic Queries over Tree-Structured Data [Текст]. Proceedings of 17th Annual IEEE Symposium on Logic in Computer Science/ Georg Gottlob, Christoph Koch New York- 2002
8. 5.Guoliang Li. - EASE: An Effective 3-in-1 Keyword Search Method for Unstructured, Semi-structured and Structured Data. [Текст] SIGMOD '08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data / Guoliang Li.,Beng Chin Ooi, Jianhua Feng, Jianyong Wang, Lizhu Zhou – New York - 2008
9. M. Hagenbuchner- A self-organizing map for adaptive processing of structured data [Текст]. IEEE Transactions on Neural Networks/ M. Hagenbuchner, A. Sperduti, A. C. Tsoi - 2003
10. Peter Buneman. Path Constraints on Semistructured and Structured Data. [Текст]. PODS '98 Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems / Peter Buneman, Wenfei Fan, Scott Weinstein – New York. 1998
11. S. Bergamaschi. Semantic Integration of Semistructured and Structured Data Sources. [Текст]. ACM SIGMOD Record - Volume 28 Issue 1, March 1999 / S. Bergamaschi, S. Castano, M. Vincini - 1999
12. Serge Abiteboul - Querying Semi-Structured Data [Текст]. ICDT '97 Proceedings of the 6th International Conference on Database Theory / Serge Abiteboul - 1997
13. Tirthankar Lahiri.- Ozone: Integrating Structured and Semistructured Data[Текст] DBPL '99 Revised Papers from the 7th International Workshop on Database Programming Languages: Research Issues in Structured and Semistructured Database Programming/ Tirthankar Lahiri, Serge Abiteboul, Jennifer Widom – London - 2000
14. 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
15. 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