Алгоритм репрезентативного сэмплинга для параллельных систем баз данных*

Д.Д. Янцен, М.Л. Цымблер

Южно-Уральский государственный университет (Челябинск)

Сэмплинг

Сэмплинг базы данных представляет собой получение выборки из неё с целью оценить статистические параметры данных или ускорить вычисления в интеллектуальном анализе данных

Последовательные СУБД

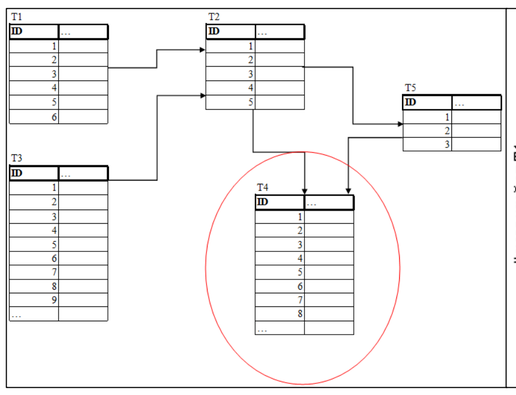
Случайный сэмплинг – выборка производится независимо от значений. В репрезентативном сэмплинге выборка сохраняет статистические особенности данных.

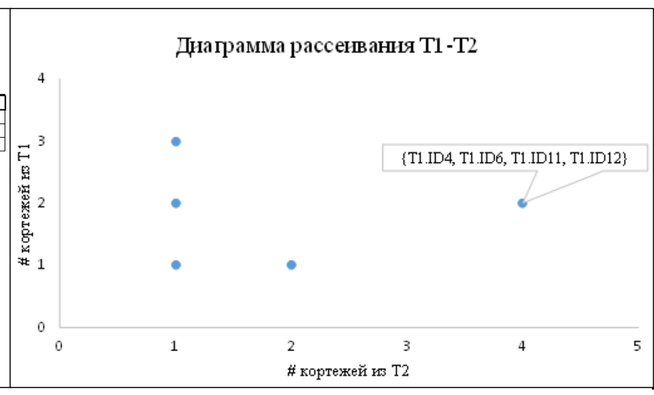
Параллельные СУБД

Основные проблемы выполнения сэмплинга в ПСУБД:

- Сохранение соотношения размеров фрагментов системы;
- Сохранение соотношения кортежей, которые необходимо передавать на другие узлы при выполнении запроса, к кортежам, которые должны быть обработаны на текущем узле.

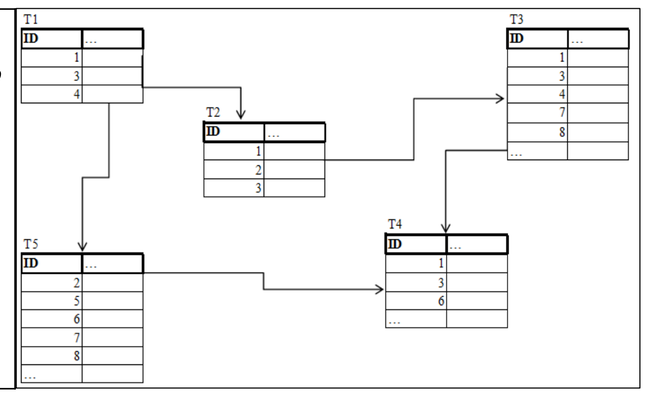
Алгоритм CoDS для последовательных баз данных



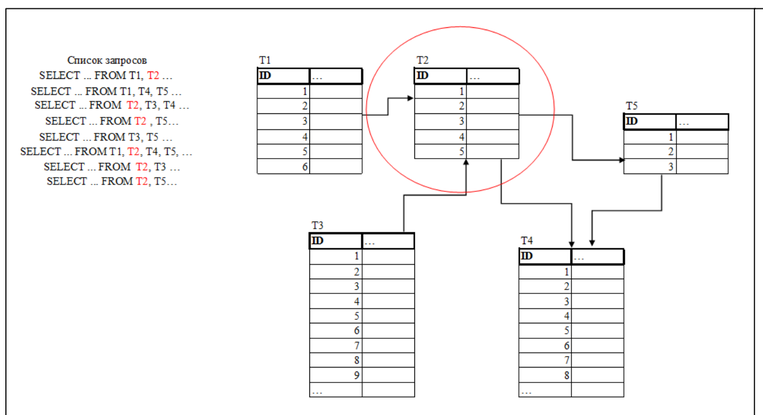


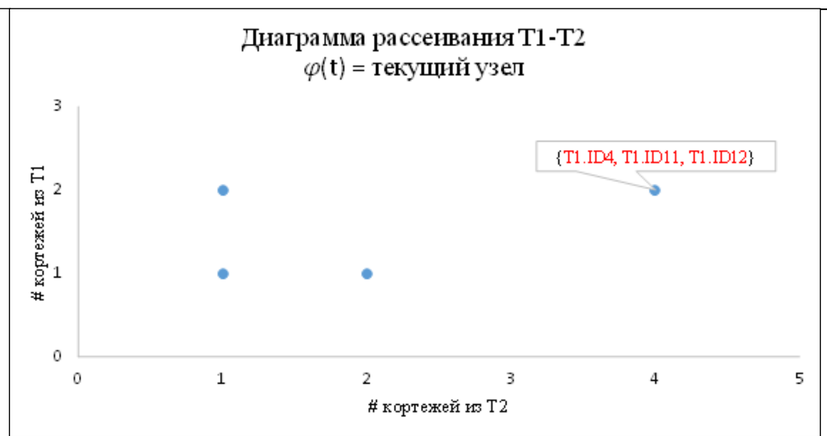
$$IF(T^*.t) = \sum_{dp' \in RDP(dp)} \frac{\|dp' \cap TS^*\|}{\alpha * \|dp'\|},$$

где
T^* - стартовая таблица,
TS^* – стартовая таблица сэмпла,
t – текущий кортеж,
dp – группа данных (точка на диаграмме рассеивания),
$RDP(dp)$ – группы данных, имеющие общие кортежи с dp.



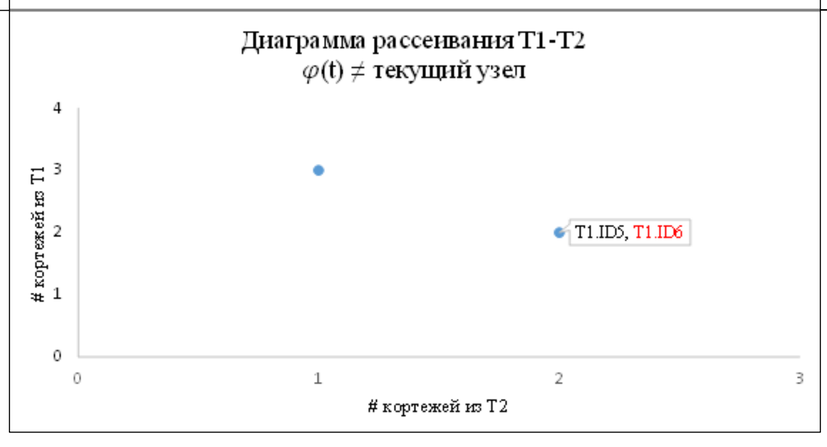
Модификация алгоритма CoDS для ПСУБД





$$IF(T^*.t) = \sum_{dp' \in RDP(dp)} \frac{\|dp' \cap TS^*\|}{\alpha * \|dp'\|} + k * \frac{\left\|TS^*_{\varphi(t)}\right\| / \|TS^*\|}{\left\|T^*_{\varphi(t)}\right\| / \|T^*\|},$$

где
φ – функция фрагментации стартовой таблицы,
$k \in [0,1]$ – коэффициент, определяющий степень важности сохранения перекоса данных.



* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 12-07-00443-а.