Алгоритм репрезентативного сэмплинга для параллельных систем баз данных*

Д.Д. Янцен, М.Л. Цымблер

Южно-Уральский государственный университет

Сэмплинг баз данных в настоящее время является одним из наиболее часто используемых подходов к интеллектуальному анализу сверхбольших объемов данных, который позволяет сократить объем анализируемых данных и время выполнения аналитических операций ценой снижения точности результатов. При случайном сэмплинге данные из исходной базы данных отбираются в сэмпл вне зависимости от их значений. Репрезентативный сэмплинг должен обеспечивать такой отбор данных, при котором сэмпл максимально точно сохраняет статистические особенности исходной базы данных. На сегодня параллельные системы баз данных признаются научным сообществом как единственное эффективное средство для организации хранения и обработки сверхбольших баз данных. При применении технологий сэмплинга в параллельных СУБД необходимо обеспечить не только репрезентативность сэмпла каждого фрагмента базы данных, но и сохранение в сэмпле ряда важных особенностей параллельной базы данных в целом: перекос данных по узлам кластерной системы, и соотношение кортежей, которые при выполнении запроса необходимо передавать на другие узлы («чужих»), к кортежам, которые должны быть обработаны на текущем узле («своих»).

Целью данной работы является создание алгоритма репрезентативного сэмплинга для параллельных систем баз данных, сохраняющего вышеуказанные характеристики. Данный алгоритм основан на алгоритме CoDS [2] для последовательных реляционных баз данных, который обеспечивает высокую степень репрезентативности при высоком быстродействии.

Идея алгоритма заключается в объединении кортежей выбранной (стартовой) таблицы, имеющих одинаковые свойства, в группы, и создании сэмпла на их основе. Для сохранения соотношения «своих» и «чужих» кортежей этап объединения кортежей стартовой таблицы в группы модифицируется таким образом, чтобы кортежи, для которых необходима пересылка, попадали в отдельную группу. Затем из каждой группы данных в сэмпл поочередно до достижения заданного размера переносятся кортежи с текущим минимальным значением функции фактора влияния на репрезентативность сэмпла, которая вычисляется следующим образом:

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

где T^* – стартовая таблица в исходной базе данных; TS^* – стартовая таблица сэмпла; dp – обрабатываемая группа кортежей; $RDP(dp)$ – список групп, имеющих общие кортежи; α – доля сэмпла от исходного размера; t – текущий кортеж; φ – функция фрагментации стартовой таблицы; $k \in [0, 1]$ – коэффициент, определяющий степень важности сохранения перекоса данных.

Литература

1. Seshadri S., Naughton J.F. Sampling Issues in Parallel Database Systems. // 3rd International Conference on Extending Database Technology, Vienna, Austria, March 23-27, 1992. Lecture Notes in Computer Science, 1992. P. 328-343.
2. Buda T.S., Cerqueus T., Murphy J., Kristiansen M. CoDS: A Representative Sampling Method for Relational Databases // 24th International Conference on Database and Expert Systems Applications, Prague, Czech Republic, August 26-30, 2013. Lecture Notes in Computer Science, 2013. P. 342-356.

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