ТЕХНИКА РЕПЛИКАЦИИ ДЛЯ СБАЛАНСИРОВАННОЙ КЛАСТЕРИЗАЦИИ МОДЕЛИ РАСПРЕДЕЛЁННОГО ГРАФА

Станкевич Е.В.

Россия, 150000, г. Ярославль, ул. Советская, д.14

Рост объёма хранимых и обрабатываемых данных в конечном итоге приводит к необходимости увеличения вычислительной мощности систем оперирующих этими данными. Наиболее предпочтительным способом повышения производительности является горизонтальное маштабирование, при котором прирост производительности системы осуществляется за счёт увеличения числа вычислительных узлов. Для систем управления базами данных требуется разработка специальных подходов, позволяющих производить горизонтальное маштабирование в рамках используемой в системе модели данных. В работе расматривается подход к горизонтальному маштабированию графовой базы данных, моделью данных которой является граф.

Формально задача горизонтального маштабирования графовой БД может быть представлена как поиск распределённого графа, вершины которого разбиты на заданное число кластеров. Каждый такой кластер представлен на отдельном вычислительном узле. С целью обеспечения равномерной нагрузки на вычислительные узлы размеры полученных кластеров должны быть сбалансированы, кроме того величины разрезов графа на кластеры должны быть минимальны для снижения объёма трафика данных, передаваемых между узлами. Задача сбалансированной кластеризации графа известна как NP-полная, предлагаемые схемы разбиения представляют собой некоторые эвристики. Одним из таких подходов является алгоритм Ja-be-Ja. Алгоритм представляет собой процедуру локального поиска оптимальной раскраски графа в N цветов. В начале работы алгоритма всем вершинам равновероятно приписывается один из цветов. Вводится понятие энергии вершины, как число смежных вершин отличного от её цвета. На каждой итерации путём обмена цветов вершин алгоритм производит уменьшение общей энергии, для предотвращения сходимости к локальному минимуму моделируется отжиг. В проведённых авторами экспериментах алгоритм демонстрирует высокое качество кластеризации.

Основным недостатком данного алгоритма является зависимость качества кластеризации от изначально выбранного числа кластеров. Так, при разбиении графа на 2 кластера, при фактическом наличии в графе 3-х кластеров одинакового размера, из-за жёсткого выполнения требования сбалансированности алгоритм оказывается неспособным найти близкое к оптимальному разбиение. Целью работы стала разработка модификации алгоритма Ja-be-Ja, устраняющей недостатки оригинальной версии. Основная идея предлагаемого подхода заключается в использовании техники репликации вершин графа на нескольких вычислительных узлах таким образом, что локальные реплики представляют собой кеш наиболее используемых вершин соседних кластеров. Для изучения характеристик предложенного метода нечёткой кластеризации графа была разработана и реализована имитационная модель распределённого статичного графа. Анализ полученных данных показал, что предложенный подход может быть применён для более качественной кластеризации по сравнению с оригинальной версией алгоритма.