Исследование и реализация методов Data Mining в BOINC-грид

Головин А. С., Ивашко Е.Е.

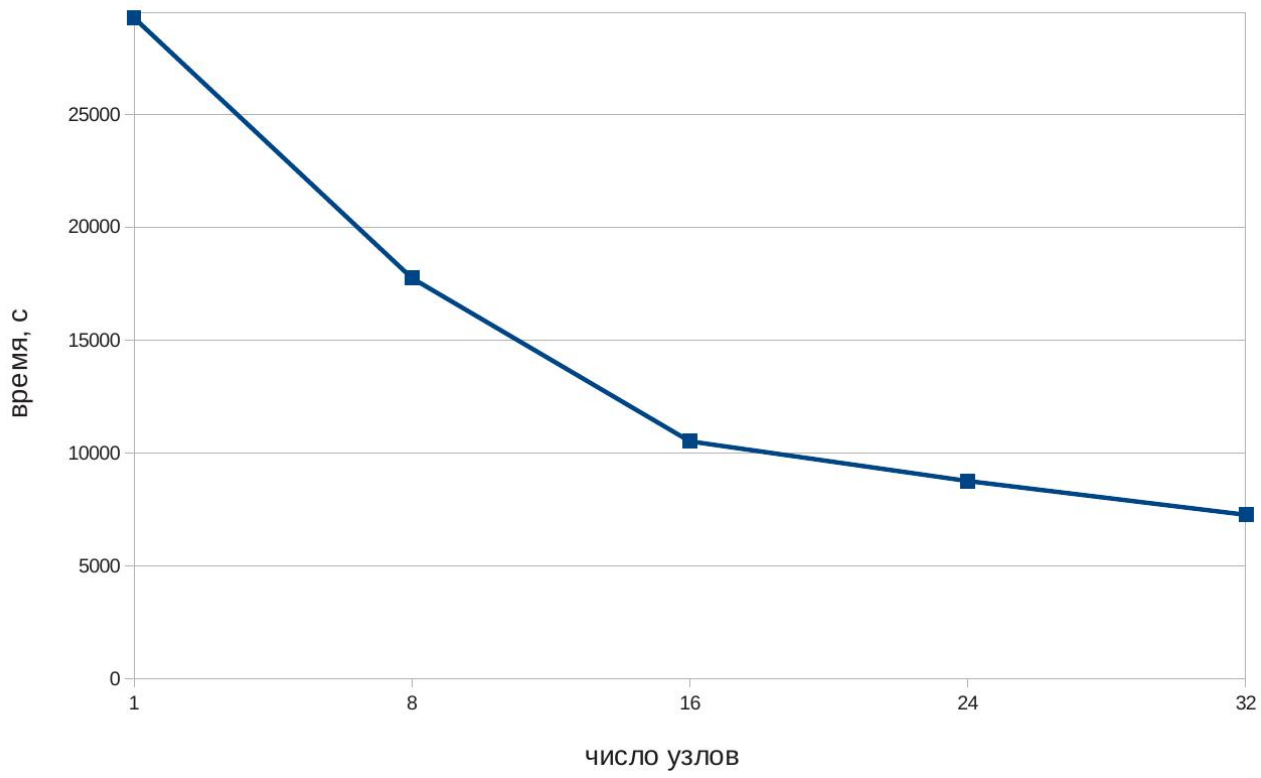
20 августа 2013

Согласно исследованию, проведенному International Data Corporation, во всем мире объемы данных удваиваются каждые два года [1]. Большие объемы данных стали одной из движущих сил фундаментальных изменений в социальной жизни, технологиях, науке и экономике. Для автоматической обработки таких данных используются методы Data Mining, которые помогают находить ранее неизвестные, нетривиальные, практически полезные и доступные для интерпретации зависимости. Одним из популярных методов Data Maning являются алгоритмы поиска ассоциативных правил, которые позволяют находить закономерности между связанными событиями. Например, в медицине ассоциативные правила могут использоваться для нахождения закономерностей между наличием определенных генов и предрасположенностью к заболеваниям.

Большие и сверхбольшие объемы данных для своей обработки требуют значительных вычислительных ресурсов. Использование для этих целей суперкомпьютеров и кластеров, состоящих из серверов, достаточно дорого и доступно не всем. Достаточно эффективной альтернативой таким решениям являются Desktop Grid. Например, открытая программная платформа BOINC, позволяет проекту FreeRainbowTables (цель — доказать отсутствие безопасности при использовании простых хэш-функций для защиты важных паролей) использовать около 40 тысяч компьютеров, обеспечивая производительность в 4,97 петафлопс (по состоянию на 20 августа 2013г.) [2]. Для сравнения самый мощный суперкомпьютер на сегодняшний день, Tianhe-2, имеет пиковую производительность в 54,9 петафлопс [3].

Для повышения эффективности поиска ассоциативных правил в рамках представленной работы был реализован алгоритм Partition на платформе BOINC, основная идея которого — разбиение исходного набора данных на n непересекающихся частей, каждая из которых считается на отдельном узле в грид-сети. На данных, взятых с ресурса Frequent Itemset Mining Implementations Repository, и данных, полученных с помощью программы «IBM generator», были проведены эксперименты (в качестве хостов грид-сети выступали вычислительные узлы кластера КарНЦ РАН). На протестированных наборах данных с

увеличением количества узлов наблюдался прирост производительности, но при слишком большом разбиении исходного набора данных прирост производительности замедлялся.



Проведенные эксперименты также показали, что значительную часть временных затрат на вычисления занимает передача данных клиентам и необходимо адаптировать программу для решения данной проблемы.

1. THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadow s, and Biggest Grow th in the Far East , December 2012, By John Gantz and David Reinsel
2. http://www.allprojectstats.com/po.php?projekt=86
3. http://www.top500.org/lists/2013/06/