ВИЗУАЛИЗАЦИЯ «БОЛЬШИХ ДАННЫХ» (“BIG DATA VISUALIZATION”, “LARGE DATA VISUALIZATION”) ОТНОСИТСЯ К ОБЛАСТЯМ КАК НАУЧНОЙ, ТАК И ИНФОРМАЦИОННОЙ ВИЗУАЛИЗАЦИИ

Анварбекова Хилола Нумановна
андижанский государтвенный университет (аду) им.з.м.бабура,
Факультет Компьютер инжиниринг и информационный техналогии,
Магистрант,
Рузибаев Аскар Жиянбекович
андижанский государтвенный университет (аду) им.з.м.бабура,
Факультет Компьютер инжиниринг и информационный техналогии,
Кафедра ИНФОРМАЦИОННЫЙ ТЕХНОЛОГИИ» , к.т.н.,старший преподаватель
Отакишиева Гулшаной Абдулазиз Кизи
андижанский государтвенный университет (аду) им.з.м.бабура,
Факультет Компьютер инжиниринг и информационный техналогии,
Кафедра КОМПЬЮТЕРНАЯ ИНЖЕНЕРИЯ» ,старший преподаватель

В первом случае «большие данные» возникают в результате сложного компьютерного моделирования различных объектов и процессов. Во втором - имеет место визуальное описание и представление абстрактной информации, получаемой в результате процесса сбора и обработки много категориальных данных, для анализа которых необходимо применение нескольких количественных и качественных мер оценки.

Работы, описывающие результаты по визуализации больших данных, появились сравнительно недавно. Среди них можно выделить «Белую книгу» компании Intel, посвященную визуализации результатов «большого счета» [1], “установочную” публикацию известного специалиста в области компьютерной визуализации и человеко-компьютерного взаимодействия Б.Шнейдермана [2] и введение к спецвыпуску, посвященному визуализации больших данных [3].

Отметим, что один из соавторов работы [3], профессор Калифорнийского Университета в Дэвисе К.-Л. Ма (Kwan-Liu Ma), в настоящее время является ведущим специалистом по визуализации больших данных. Он является соавтором большого числа научных работ, а также организаторм конференций, семинаров и лекций по данному направлению.

Нас в этой работе интересуют, прежде всего, при визуализация результатов при параллельных и распределенных вычислений. В связи с «большими данными» уместно также вспомнить понятие «большого счета» (многочасового, многосуточного или даже многонедельного процесса научных вычислений). Современные вычислительные системы в ходе компьютерного моделирования способны генерировать очень большие (large or huge) по объему файлы с данными. В тоже время оценка размера файла, прежде всего, связана с возможностями «супер вычислений», мощности которых непрерывно возрастают. Усложнение структуры данных связано не только с архитектурой вычислительного комплекса, но и с возможностью решения более сложных в математическом плане задач, например, задач, приводящих к многомерным решениям.

Можно увязать понятие «большие данные» с некоторым предельным (на данный момент) случаем обработки данных, при котором универсальные подходы к анализу и визуализации не работают. Тогда в качестве больших данных могут рассматриваться многомерные и много-категориальные данные, данные большого объема, данные с неполной информацией.Предельный случай формирует вызовы, на которые необходимо ответить, чтобы двигаться дальше. Решение возникающих проблем приводит к тому, что сегодняшние «большие данные» завтра становятся нормой.

Среди задач визуализации "больших данных" рассматриваются следующие:

- визуализация потоков данных;
- визуальный интеллектуальный анализ данных (Visual data mining);
- визуальный поиск и рекомендации (Visual search and recommendation);
- описание ситуаций на основе больших данных с использованием визуализации (Big data storytelling using visualization);
- масштабируемые методы параллельной визуализации

В плане научной визуализации работы по визуализации больших данных затрагивают проблемы визуализации объемных данных, включая параллельный объемный рендеринг и визуализацию объемных данных без прерывания работы системы (In situ volume visualization).

Многие задачи визуализации программного обеспечения, возникающие, например, при рассмотрении трасс выполнения

параллельных программ, также связаны с большими и очень большими объемами данных.

Отметим, что методы визуализации в этом случае, как правило, заимствуются из методов, используемых в информационной визуализации.

Удаленная визуализация позволяет расширить круг пользователей систем «большого счета». Средства онлайн визуализации, во-первых, позволяют, вмешиваясь в работу программы в процессе счета, оперативно оценивать промежуточные результаты и принимать решения поизменению параметров вычисления, а во-вторых, на их базе возможна разработка систем визуальной отладки параллельных программ. Имеет место большое количество публикаций по данным вопросам. Разработки нашего исследовательского коллектива в этом направлении описаны, в частности, в работах [1], [2], [3].

Аппаратные решения - ряд аппаратно-программных комплексов, предоставляющих предконфигурированные решения для обработки больших данных: Aster MapReduce appliance, Oracle Big Data appliance, Greenplum appliance. Эти комплексы поставляются как готовые к установке в центры обработки данных телекоммуникационные шкафы, содержащие кластер серверов и управляющее программное обеспечение для массово-параллельной обработки[4].

Рассмотрим в самых общих чертах анализ эффективности распараллеливания алгоритмов, например, на визуализации системах.

При проектировании новых многопроцессорных операционных систем (МП ОС) весьма остро стоит проблема уменьшения накладных расходов, возникающих при планировании процессов. Частью планировщика ОС является функция диспетчеризации задач при их назначении на обработку центральным процессорам (ЦП) [5]. Наименее затратный путь решения проблемы – провести математическое моделирование различных характеристик диспетчера задач.

Ниже анализируется многопроцессорной вычислительного комплексах (МВК) с диспетчеризацией, при задачи визуализации больших данных

В [1] рассматриваются подходы к распараллеливанию задач численного анализа. Качества параллельного алгоритма обычно оцениваются по следующей система критериев [2]: $S_к=T1/T_к$

Где S- коэффициент ускорения вычислений;

В [3] исследуется зависимость коэффициента ускорения от загрузки и интенсивности системы, причем анализ проводится не по алгоритмам, а по времени диспетчирования.

Рассматривается модель МВК с единственной управляющей машиной (УМ) и n однотипными процессорами. Все задачи сначала поступают в УМ и в течение некоторого интервала времени τ анализируются и разделяются (распараллеливаются с вероятностью q_k на (к‹n) независимых частей. Параллельная задача содержит такие подзадачи для обработки информации, которые могут функционировать одновременно (параллельно), поэтому их одновременно предают к процессорам [4].

Предполагается, что все процессоры однотипны и в систему поступает однородный поток задач с интенсивностью λ . Если задача на одно-процессорной машина требует времени решения -Т . то в МВК за счет распараллеливания она может решаться намного быстрее.

Остановимся в кратца на процесса прохождения задач через МВК. Задача с вероятностью -q_k распараллеливается на n частей с помощью УМ за время - τ. Суммарное время решения распараллеленных задач обозначим через Т . Тогда $T=T1+T2+T3+\ldots\ldots+Tn=\sum T_i$

Где T_i (i= 1,2,3,…, k)-случайные величины.

Для удобства предположим, что T_i является одинаково распарал-леленными временими и тогда поступившие задача в МВК требует среднего времени Mτ+MTi (1)

Где (Mτ и MTi - средние значения τ и Ti). Понятно, что в МВК будет эффективнее однопроцессорного условия, тогда- когда выполняется

Tτ‹(к-1)MTi. (2)

Введем понятие коэффициента ускорения прохождения задач в комплекса : K_Y ==MT/(Mτ+MTi) (3)

При KY>1 происходить ускорение прохожденая задач, в противном случае – замедление.

Обозначим ν=1/Mτ и μ=1/MTi , тогда интенсивность обслуживания рассмат-риваемого комплекса ($\mu_{инт.компл.}$) определятся в вида:

$\mu_{инт.компл}=1/(M\tau+MTi)=(к+\mu)/(\nu+к\mu)=\nu/(1+\nu/к\mu)$

Если Mτ→0, то ν→∞ и $\mu_{инт.компл}\rightarrow к\mu$.

Отсюда вытекают следующие выводы:

Если на анализ задач УМ уходит малая доля времена.

То интенсивность комплекса стремится к кμ [5];

Результаты экспериментов на модели позволили выявить , что ответ на поставленный вопрос является положительным. В качестве иллюстрации рассмотрим пример.

При значении μ=0.1÷0.2. λ = 0.3÷1.5 , КУ>1 распралливание дает эффект, а при μ=0.22÷0.5 нет.

Теперь рассмотрим одна из характеристик комклекса -μинт.компл . При значении λ= 0.3÷1.5; μ=0.1÷0.5; $_{\text{μинт.компл}}$→1 , рис.1.3 – загрузка МВК. Если λ<v , то комплекс будет работать незагруженным. При v→∞ загрузка асимптотически приближается к λ/кμ.

Таким образом , при большой загрузке комплекса, т.е. $_{\text{ρзагр.компл}}$→1 распараллеливание не дает эффекта и система работает в режиме пакетной обработки. Распараллеливание задачи дает эффект лишь тогда , когда загрузка комплекса меньше 1 и суммарное время простоя процессоров не учитывается за время функционирования системы.

Тогда приближенно можно ввести оценки затрат времени на визуализация системе.

ЛИТЕРАТУРА:

1. Keim D. Qu H., Ma K.-L. Big-Data Visualization // IEEE Computer Graphics and Applications.July/August 2013. Pp. 50-51.

2. Авербух В.Л., Байдалин А.Ю., Васев П.А., Исмагилов Д.Р., Зенков А.И., Манаков Д.В., Перевалов Д.С., Шагубаков М.Р. Задачи визуализации параллельныхвычислений. // Вопросы атомной науки и техники. Сер. Математическое моделирование физических процессов. 2002. Вып. 3. Стр. 40-

3. Бахтерев М.О., Васёв П.А., Казанцев А.Ю., Манаков Д.В. Система удалённой визуализации для инженерных и суперкомпьютерных вычислений // Вестник ЮжУрГУ, N17 (150), 2009, серия «Математическое моделирование и программирование», Выпуск 3.Стр. 4-11.

4. Наиболее полный список инструментов для анализа данных и машинного обучения // DATASIDE. [2018–2018]. Дата обновления: 26.11.2018. URL: http://ru.datasides.com/big-data-analytic-tools/ (дата обращения: 23.12.2018)

5. Клейнрок Л. Теория массового обслуживания. –М.: Машиностроение, 1979 .-432с.