ORES
Научно-практический электронный журнал
Оригинальные исследования (ОРИС), *ISSN 2222-0402*
www.ores.su
original-research.ru

УДК 004.6

ОБРАБОТКА СВЕРХ БОЛЬШИХ МАССИВОВ ДАННЫХ И ИХ ВИЗУАЛИЗАЦИЯ

Янаева Марина Викторовна,
кандидат технических наук, доцент,
Кубанский государственный технологический университет,
Россия, г. Краснодар.
yanaevam@mail.ru

Раджабов Азамат Олимжонович,
студент 3 курса,
Кубанский государственный технологический университет,
Россия, г. Краснодар.
azamat.radzhabov00@mail.ru

Зевахин Тимур Родионович,
студент 3 курса,
Кубанский государственный технологический университет,
Россия, г. Краснодар.
zevahintimur1@gmail.com

Аннотация

В данной статье рассматривается решение проблемы обработки сверх больших массивов данных и их визуализация. Для решения задачи были выбрани такие подходы как: параллельная фильтрация, слайсинг, брашинг, которые дают большее представление о больших массивах данных.

Ключевые слова: большие данные, массивы данных, параллельная фильтрация, кластеризация, слайсинг, брашинг, визуализация, тонкий клиент.

PROCESSING SUPER LARGE DATA AND THEIR VISUALIZATION

Marina V. Yanaeva,
candidate of technical sciences, associate professor,
Kuban State Technological University,
Russia, Krasnodar.
yanaevam@mail.ru

Azamat O. Radzhabo,
3rd year student,

Kuban State Technological University,
Russia, Krasnodar.
azamat.radzhabov00@mail.ru

Timur R. Zevakhin,
3rd year student,
Kuban State Technological University,
Russia, Krasnodar.
zevahintimur1@gmail.com

ABSTRACT

This article discusses the solution to the problem of processing oversized data arrays and their visualization. To solve the problem, we chose such approaches as: parallel filtering, slicing, brushing, which give a better idea of large data arrays.

Keywords: big data, data arrays, parallel filtering, clustering, slicing, brushing, visualization, thin client.

Обработка сверхбольших массивов данных

Большие данные — это большие объемы, высокая скорость и/или большое разнообразие наборов данных, которые требуют новых форм обработки, чтобы обеспечить улучшенную оптимизацию процессов, обнаружение идей и принятие решений. Проблемы больших данных заключаются в сборе данных, их хранении, анализе, совместном использовании, поиске и визуализации. Визуализацию можно рассматривать как «внешнюю часть» больших данных. Существуют следующие мифы о визуализации данных:

• Все данные должны быть визуализированы: важно не слишком полагаться на визуализацию; некоторые данные не нуждаются в методах визуализации, чтобы раскрыть их сообщения.

• Следует визуализировать только качественные данные [1]. Простая и быстрая визуализация может выявить ошибки в данных, а также выявить интересные тенденции.

• Визуализация всегда покажет правильное решение или действие: Визуализация не может заменить критическое мышление.

• Визуализация приведет к определенности: визуализация данных не означает, что они показывают точную картину того, что важно. Визуализацией можно управлять с помощью различных эффектов [2].

В настоящее время обработка сверхбольших массивов данных является одной из основных задач развития информационных технологий. Однако, применение данной технологии не так популярно. Это связанно со специфичными требованиями к обработке данных, характерными только для предприятий.

Подходы к визуализации используются для создания таблиц, диаграмм, изображений и других интуитивно понятных способов представления данных. Визуализация больших данных не так проста, как традиционные небольшие наборы данных [3]. Расширение традиционных подходов к визуализации уже появилось, но этого недостаточно. При визуализации крупномасштабных данных многие исследователи

используют извлечение признаков и геометрическое моделирование, чтобы значительно уменьшить размер данных перед их фактическим рендерингом.

Визуализация «больших данных» относится к областям как научной, так и информационной визуализации [4]. 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 [5].

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.

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.

Как уже отмечалось, задачи научной визуализации больших данных близки к задачам, решаемым параллельной визуализацией и распределенными вычислениями. Разделение на параллельные и распределенные вычисления часто основано на том, что в первом случае параллелизм свойственен данным, а во втором — задачам. Это деление весьма условно. Гораздо важнее их сходство с точки зрения используемых технологий программирования [6]. Направление, связанное с визуализацией, сформировалось в параллельных вычислениях, называемых визуальными суперкомпьютерами. Разработана терминология, разработан ряд моделей и прототипов системы. Первоначальные усилия разработчиков программного обеспечения в области суперкомпьютеров были направлены на разработку новых технологий, а обязательными условиями были: поддержка интерактивного режима и возможность обработки больших данных, включая их визуализацию. Облачные вычисления обычно рассматриваются в рамках распределенных вычислений. Анализ результатов в этом случае осуществляется с помощью веб-технологий («тонкий» клиент).

Применительно к параллельным и распределенным вычислениям мы рассмотрим ряд программных технологий, применимых также к задачам обработки, анализа и визуализации «больших данных».

Фильтрация данных включена в стандартный графический конвейер, состоящий из фильтрации данных, геометрической обработки и растеризации (рендеринга). Любая из этих трех частей может быть реализована параллельно или последовательно. Однако все же важно разобраться, где заканчивается параллелизм, а также выяснить, какой тип визуализации данных поддерживает клиент. В случае параллельной фильтрации данных выполняется только параллельная фильтрация данных, поэтому отфильтрованные математические данные передаются клиенту. Параллельная фильтрация востребована для двух видов задач: визуализации больших объемов данных и многопараметрических задач, требующих активного взаимодействия пользователя и системы в процессе визуального анализа для уточнения этих параметров или рефакторинга. В частности, разработанный конструктор отображения поддерживает применение фильтров и изменение параметров. Эта последняя возможность определяется как императивный подход [7].

Мы собираемся дать ряд определений, связанных с фильтрацией данных.

Фильтр — это операция с данными, которая определяет объем данных. Это исключает случаи фрагментации. Удаление отрывков является обязательным условием для создания предварительного заказа: формально мы можем сказать, что они выполнены идеально. В соответствии с этим назначением сжатие и обработка изображений могут быть отфильтрованы, как в некоторых случаях с данными.

Таким образом, фильтрация данных предполагает получение необходимых (интересно, каких) данных в кратчайшие сроки. Это также позволяет вам создать еще одну минимальную задачу, в которой данные должны быть отфильтрованы для получения более полезной информации с минимальными затратами (например, базовое обучение во время работы над оценкой, сотрудничеством, заявлениями и т.д.).). Большие данные определяют данные, которые могут быть просмотрены полностью или в течение разумного времени. Таким образом, важным вопросом в вопросе обработки данных является вопрос о рассчитанных показателях эффективности.

Фильтрация данных — это процесс получения необходимой (интересной) информации в кратчайшие сроки. Возможна еще одна минимальная формулировка проблемы, в которой целью фильтрации данных является получение максимального объема полезной информации с минимальными затратами (например, расчет, взаимодействие, интерпретация и т. Большие объемы данных определяются как данные, которые не отображаются полностью и могут быть в течение приемлемого периода времени. Таким образом, важной задачей является управление данными является вопрос показателей оценки эффективности.

Слайсинг – построение срезов. Частный случай фильтрации данных, когда функция от данных равна константе. Кроме сечения плоскостью – это такие стандартные виды отображения, как изолинии и изоповерхности. Условие – равенство константе фактически сокращает размерность данных на единицу. Изменение константы приводит к построению фазового пространства. В качестве константы может выступать идентификатор функции или максимальная длинна графа. В качестве такого примера приведем работу, в которой параметры статического анализа (графа) уточняются после проведения динамического анализа. Данный подход можно определить, как робастная модель статического анализа с рефакторингом.

Общепринятый визуальный подход закрашивание “brushing” интерактивно выделяет подмножества данных, чаще всего с помощью цвета. Он поддерживает визуальную связь разнородных объектов, тем самым решая проблему визуальной фрагментации. В отличие от фильтрации, которая удаляет данные с дисплея, закрашивание дает добавочную информацию, накладывая уточненное изображение на существующую структуру. Однако для закрашивания нужно собственное кодирование, то есть графический атрибут, который используется для выделения выступающих точек, должен быть сохранен независимо от того, является ли он цветом, формой или текстурой. Эти графические атрибуты уже являются стандартными в визуализации с большим количеством параметров (множественные визуализации). В частности, цвет подходит для категоризации. Таким образом, брашинг – это категоризация или выделение цветом. Это не уменьшает объем данных, а лишь уточняет их структуру [8].

Группировка – это тематическая категоризация или классификация. Данные принадлежат кластеру, расстояние до которого минимально. С помощью кластеризации также возможно сжатие данных, поскольку элемент с наибольшей емкостью остается в кластере. Один из стереотипов группировки - "разделяй и властвуй". В связи с развитием интернет-технологий направление вектора формализации сместилось с методов линейного программирования на нечеткие множества, которые, с точки зрения программиста,

представляют собой ассоциативные массивы. В этом случае основой кластеризации является алгоритм Map-Reduce.

Эффективность в модели потока данных достигается за счет возможности конвейерной обработки данных (один из методов распараллеливания). В работе, например, описана система программирования, основанная на бинарной коммуникативной модели.

Считается, что решением проблемы растущей административной сложности вычислительных инфраструктур являются автономные вычисления. Автономные вычисления относятся к компьютерным системам, обладающим способностью к самопознанию и самоуправлению [9]. Такая система может характеризоваться одним или несколькими из следующих признаков:

Автоконфигурация: система может комбинировать новые и существующие компоненты без вмешательства администратора;

Самооптимизация: система может постоянно пытаться изменить конфигурацию, чтобы определить, является ли текущая оптимальной;

Самовосстановление (self-healing) — система может отслеживать ошибки и восстанавливаться после неисправных аппаратных или программных компонентов;

Самозащита: система может отслеживать попытки взлома и реагировать соответствующим образом.

Существует достаточно широкий спектр работ по адаптивному интерфейсу, который можно рассматривать как частный случай автономных вычислений, связанный с самооптимизацией под конкретного пользователя. Примерами из области параллельных вычислений являются планирование очереди задач и оптимизация краудсорсинга.

Технология интеллектуального анализа данных представляет собой процесс поиска ранее неизвестных, нетривиальных, но практически полезных корреляций, закономерностей и тенденций в массивах необработанных данных. Интеллектуальный анализ данных использует различные статистические и математические методы и алгоритмы [10].

Визуализации могут быть статическими или динамическими. Интерактивная визуализация часто приводит к открытиям и работает лучше, чем инструменты статических данных. Интерактивные визуализации могут помочь лучше понять большие данные. Интерактивная очистка и связь между подходами к визуализации и сетями или веб-инструментами могут облегчить научный процесс. Веб-визуализация помогает своевременно получать динамические данные и поддерживать визуализацию в актуальном состоянии.

Расширения некоторых традиционных подходов к визуализации для обработки больших данных далеко не достаточно в функциях. Необходимо разработать больше новых методов и инструментов визуализации больших данных для различных приложений больших данных. В этой статье представлены достижения в области визуализации больших данных, а также проведен SWOT-анализ существующих программных средств визуализации для визуализации больших данных. Это поможет разработать новые методы и инструменты для визуализации больших данных. Аналитика и визуализация больших данных могут быть тесно интегрированы, чтобы лучше всего работать с приложениями больших данных [11]. Иммерсивная виртуальная реальность (VR) — это новый и мощный метод работы с высокой размерностью и абстракцией. Это значительно облегчит визуализацию больших данных.

Список литературы:

1. Хан М., Хан С. С., Методы визуализации данных и информации и интерактивные механизмы: обзор, Международный журнал компьютерных приложений, 34 (1), 2011, С. 1-14.
2. Сухарита В., Субаш С. Р. и Пракаш П., Визуализация больших данных: ее инструменты и проблемы, Международный журнал прикладных инженерных исследований, 9 (18), 2014, С. 5277-5290.
3. Саймон П., Визуальная организация: визуализация данных, большие данные и поиск лучших решений, Harvard Business Review, 2014, С. 1–8.
4. Портер Б., Визуализация больших данных в Drupal: использование визуализации данных для открытия знаний, отчет, Вашингтонский университет, 2012, С. 1-38.
5. Фокс П. и Хендлер Дж., Изменение уравнения визуализации научных данных, Наука, 331 (11), 2011, С. 705-708.
6. Китчин Р. Революция данных: большие данные, открытые данные. Инфраструктуры данных и их последствия. Публикации SAGE; 2014.
7. Акеркар Р. Вычисления на больших данных. Бока-Ратон, 2013.
8. Руссом П. Управление большими данными. Отчет о передовой практике TDWI, исследование TDWI; 2013.
9. Бейер М. А., Лейни Д. Важность "больших данных": определение. Стэмфорд: Gartner; 2012.
10. Майер-Шенбергер В., Кукьер К. Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим. С.240, 2014.
11. Черняк Л. Большие данные – новая теория и практика //Открытые системы. СУБД– 2011. – №10. – С.18-25.

References:

1. Khan M., Khan S. S., Data and information visualization methods and interactive mechanisms: a review, International Journal of Computer Applications, 34 (1), 2011, pp. 1-14.
2. Suharita V., Subash S. R. and Prakash P., Big Data Visualization: Its Tools and Challenges, International Journal of Applied Engineering Research, 9 (18), 2014, pp. 5277-5290.
3. Simon P., Visual Organization: Data Visualization, Big Data, and Finding Better Solutions, Harvard Business Review, 2014, pp. 1–8.
4. Porter, B., Big Data Visualization in Drupal: Using Data Visualization for Knowledge Discovery, Report, University of Washington, 2012, pp. 1-38.
5. Fox, P. and Handler, J., Changing the Science Data Visualization Equation, Nauka, 331(11), 2011, pp. 705-708.
6. Kitchin R. Data revolution: big data, open data. Data infrastructures and their implications. SAGE Publications; 2014.
7. Akerkar R. Big Data Computing. Boca Raton, 2013.
8. Russom P. Big data management. TDWI Best Practice Report, TDWI Study; 2013.
9. Beyer M. A., Laney D. The importance of "big data": a definition. Stamford: Gartner; 2012.

10. Mayer-Schenberger V., Kukier K. Big data: A revolution that will change the way we live, work and think. P.240, 2014.
11. Chernyak L. Big data - new theory and practice // Open systems. DBMS - 2011. - No. 10. - P.18-25.