Аналитика и визуализация «Больших Данных»: почему «Большие Данные» являются Большой проблемой?
Analysis and Visualization «Big Data»: why «Big Data» is a «Big Problem»?

Малярова М. В.

Малярова Мария Виталиевна / Malyarova Maria Vitalievna – студент,
кафедра математической кибернетики и компьютерных наук,
факультет компьютерных наук и информационных технологий,
НИУ Саратовский государственный университет, г. Саратов

Аннотация: *в статье описывается проблема анализа и визуализации результатов при обработке больших данных. Приводится некоторая классификация типов данных и сравнительно быстродействующие методики анализа больших данных.*
Abstract: *the article describes the problem of analysis and visualization of results in the processing of big data. Provided are some of the classification of types of data and relatively fast method of analysis of big data.*

Ключевые слова: *анализ, большие данные, визуализация, кластеризация, данные, массово-параллельная обработка, пространственный анализ, k-деревья.*
Keywords: *analysis, big data, visualization, clustering, massively parallel processing data, spatial analysis, k-tree.*

Информация сейчас — это собирательный поток из вещей повседневной жизни. Поток, состоящий полностью из кредитных карт, телефонов, инфраструктуры городов и датчиков оборудованных зданий. Объемы данных настолько быстро растут, что общее удвоение мощности роста данных происходит каждые 18 месяцев. «Большие Данные», как правило, возникают в результате сложного компьютерного моделирования объектов или различных процессов. Также имеет место представление абстрактной информации, которая получается в ходе исследований и обработки данных, для анализа которых необходимо применение нескольких видов оценки.

Классификацию объемов данных можно представить так:

Огромные наборы данных: от 1000 гигабайт до нескольких терабайт. «Большие данные»: от нескольких терабайт до сотен терабайт и экстремально «Большие Данные»: от 1000 до 10000 терабайт.

Почему же Большие данные оказались проблемой? С момента появления термина «Big Data» прошло не более восьми лет. Люди сталкиваются с большими потоками данных, которые нужно проанализировать, но чтобы это сделать, необходимы машины со соизмеримыми с данными мощностями. Машины, которые используются для распределенных вычислений, способны генерировать большие по объему выходные данные. Нередко проблему «Больших Данных» упрощают, связывая её с законом Мура, с разницей, что в случае с потоком данных используется феномен удвоения роста данных за год. Случилось так, что способность порождать данные оказалась масштабнее, чем способность анализировать. Данные обрабатываются исключительно для получения информации, которой должно быть такое количество, из которого можно было получить знание о предмете исследования. В современных исследованиях, которые производятся за счет параллельных и распределенных вычислений, необходима визуализация и оценка результатов.

Информационная визуализация «Больших Данных» тесно переплетается с методикой «Data Mining», то есть с методиками обнаружения значимых корреляций, зависимостей в результате анализа хранимой информации, выявления отношений между данными различного типа, таких как: ассоциации, последовательности аналогии и кластеры. Применяются различные методы выделения и извлечения информации, которые позволяют выявить систематизированные структуры данных и вывести из них правила для принятия решений и прогнозирования их последствий.

В свою очередь научная визуализация связана с анализом научных данных, которая включает в себя такие операции, как идентификация, локализация, категоризация, кластеризация, ассоциация, валидация и корреляция.

Компьютерные приложения приближаются все ближе к росту объемов входных данных для анализа. Тенденции к потребности аналитики данных вкупе с большими скоростями обработки данных привели к возникновению направления «аналитика Больших Данных». Объемы в сочетании с высокой скоростью требуют машин, которые выдержат эту нагрузку. На сегодняшний день производители предлагают специализированные программно-аппаратные системы, такие как: SAP HANA, Oracle Big Data Appliance, Oracle Exadata Database Machine и Oracle Exalytics Business Intelligence Machine, Teradata Extreme Performance Appliance, NetApp E-Series Storage Technology, IBM Netezza Data Appliance, EMC Greenplum, Cloudera, DataStax, Northscale, Splunk, Palantir, Factual, Kognitio, Datameer, TellApart, Paraccel, Hortonworks.

Существуют три вариации задач, которые тесно связаны с проблемой «Больших Данных». Первая задача состоит в хранении и управлении. Объем данных в сотни терабайт или петабайт не позволяет легко хранить и управлять данными с помощью традиционных реляционных баз данных. Вторая задача заключается в вопросе: «Как можно организовать неструктурированные данные?». Большинство входных данных могут оказаться неструктурированными медиа-объектами: видео, изображения и так далее. Последняя задача заключается в самом анализе «Больших Данных». Как анализировать неструктурированную информацию? Как на основе Больших Данных составлять простые отчеты, строить и внедрять углубленные прогностические модели?

Общность параллельных и распределенных вычислений, с точки зрения применяемых технологий программирования, очень важна для анализа данных. Рассмотрим некоторые технологии, которые позволяют обрабатывать и анализировать «Большие Данные».

Кластеризация. Кластеризация — это задача разбиения множества объектов на группы, называемые кластерами. Основные алгоритмы кластеризации основаны на динамическом поиске ближайших по дистанции кластеров. В связи с развитием технологий направление вектора формализации сместилось от методов линейного программирования. В данном случае фундаментом кластеризации является алгоритм Map-Reduce. MapReduce — модель распределенных вычислений, представленная компанией Google.

Алгоритмы с внешней памятью. С точки зрения научной визуализации, технологии, которые применяют алгоритмы работы с внешней памятью, минимизирующие накладные расходы ввода-вывода, очень важны [1]. Двумя наиболее часто используемыми подходами к визуализации является организация данных со многими разрешениями. Для этих случаев обычно применяют k-дерево при реструктуризации данных. Реструктуризация данных по k-дереву обеспечивает возможность быстрого поиска и извлечения данных.

Фильтрация данных. Фильтрация данных включается в стандартный графический конвейер, состоящий из фильтрации данных и их геометрической обработки. Работа фильтрации над большими данными происходит параллельно [2]. Параллельная фильтрация данных востребована для класса многопараметрических задач, требующих активного взаимодействия исследователя и системы в процессе визуального анализа.

Анализ «Больших Данных» требует использования новых технологий компьютерной графики, сред виртуальной и расширенной реальности. Возникает необходимость проведения комплексных исследований не только с точки зрения компьютерных наук и математики, но и с точки зрения когнитивной психологии. С долей вероятности, завтра проблему не будут порождать сегодняшние «Большие Данные», но проявится проблема поступательной завтрашней информации. Проблема останется вечно неразрешенной, пока не появятся автоматизированные системы, которые могли бы приспособиться к постоянному росту объема «Больших Данных».

Литература

1. *Бахтерев М. О., Васёв П. А., Казанцев А. Ю., Альбрехт И. А.* Методика распределенных вычислений RiDE // Параллельные вычислительные технологии (ПаВТ’2011). Челябинск: Издательский центр ЮУрГУ, 2011.
2. *Brodlie K., Brooke J., Chen M., Chisnall D., Fewings A., Riding M.* Visual Supercomputing - Technologies, Applications and Challenges // Eurographics 2004, STAR Reports. Pp. 37-68.