УДК 004.041

БОЛЬШИЕ ДАННЫЕ: ПРОБЛЕМЫ ОБРАБОТКИ

Лобанов А.А., к.т.н., доцент, МГТУ МИРЭА, E-mail: cvdisser@list.ru
Москва, Россия

Аннотация. Дается анализ проблемы «больших данных». Раскрываются причины появления проблемы и факторы, которые ведут к ее появлению. Дается сравнение больших данных и обычных данных. Показано, что проблема больших данных состоит не только в больших объемах коллекций данных. Важными факторами являются: ограничения на время обработки и анализа данных, а также в рост сложности информационных моделей и информационных коллекций. Описан методический и алгоритмический инструментарий, применяемый при обработке больших данных.
Ключевые слова: Информация, данные, большие данные, информационные объемы, методы обработки, сложность, информационные технологии, анализ.

BIG DATA: PROCESSING PROBLEMS

Lobanov A.A., PhD., associate professor, MSTU MIREA, E-mail: cvdisser@list.ru
Moscow, Russia

Abstact. The analysis of the problem of "big data." Reveal the causes of the problem and the factors that lead to its emergence. Provides a comparison of large data and normal data. It has been shown that the problem of large data is not only large amounts of data collection. Important factors are: restrictions on the processing and analysis of data, as well as to the growth of the complexity of information models and information collections. A methodical and algorithmic tools used in the processing of large data
Keywords: Information, data, big data, information volumes, processing methods, the complexity of information technology, analysis

Введение. В последние годы много говориться о проблеме «больших данных» (Big Data) [1, 2, 3, 4]. Чаще всего эту проблему связывают с необходимостью обработки структурированных и неструктурированных данных больших объёмов. Для характеристики «больших данных» используют критерий «три *V*»: объём (*volume*), скорость (*velocity*), многообразие (*variety*) , однако такой критерий является существенным упрощением ситуации. Появление термина соотносят с 2008 годом [5]. Введение термина «большие данные» связывают с Клиффордом Линчем – редактором журнала Nature [5], подготовившему серию работ на эту тему. Это обозначает признание проблемы в некомпьютерных сферах.

Проблему больших данных обнаружили специалисты в области дистанционного зондирования Земли более 50 лет назад [6, 7], затем ее отметили программисты 40-50 лет назад. Затем ее зафиксировали аналитики 20-30 лет назад. И только в последние десять лет

она открылась для бизнес -аналитиков и журналистов, что и привело к их повышенному вниманию к такому явлению и появлению термина.

В процессе развития человеческого общества происходит наблюдение человека за объектами, явлениями и процессами окружающего мира. Как результат наблюдения происходит получение информации в информационном поле [8, 9], накопление опыта и формирование описаний объектов, явлений и процессов. Первичное описание объектов окружающего мира состояло в формировании количественных и качественных свойств, характеристик, признаков и отношений между ними. Это описание представляет собой информационные коллекции. Вторичное описание состояло в формировании моделей и систем, формируемых на основе анализа первичных коллекций данных. Чем сложнее объект исследования, тем большего количества информации требует его описание и тем объемнее и сложнее информационные коллекции, составляющие такое описание.

Рост объемов собираемой информации и требование ее обработки и хранения делают актуальным исследования в области методов и алгоритмов анализа больших и сверхбольших наборов данных. В работе [10] высказано предположение, что выявление закономерностей в больших массивах данных становится основным инструментом исследования и получения новых знаний. Рост объемов данных характеризует не только IT-компании, но и научную сферу [11], а также широкий спектр организаций в самых различных областях [12]. В современной науке возникло новое направление, связанное с анализом больших и сверхбольших наборов данных, Big Data [2].

Описание больших данных. Описания больших данных применяемых в разных сферах являются аргументом в пользу проведения исследований и разработок, направленных на создание масштабируемых аппаратных и программных решений проблем. Пока пределом возможностей приложений, ориентированных на обработку больших объемов данных, являются петабайтные наборы и гигабайтные потоки данных. Но в соответствии с тенденцией ожидаются еще большие масштабы и объемы данных

При создании приложений, работающих с большими данными, приходится сталкиваться со следующими проблемами: большие объемы данных [1], интенсифицированные потоки данных [12], существенное сокращение допустимого времени анализа данных [2], предел времени принятия решений при любом количестве данных [4], возрастание морфологической сложности моделей, возрастание структурной сложности [13] моделей и систем, возрастание вычислительной сложности [3], относительный рост слабоструктурированной исходной информации, относительный рост нечеткой информации, рост потребностей в параллельных вычислениях [5] и т.д.

Упрощенно проблемы работы с данными большого объема приведены в таблице 1.

Таблица 1. **Сравнительные характеристики обычных и больших данных.**

Характеристика	**Обычные данные**	**Большие данные**
Формат	Однородный	Неоднородный
Объем	Мегабайты гигабайты	Петабайты
Распределенность данных	нет	есть
Тип задачи	Первого рода	Второго рода
Тип моделей решателей	Алгоритмические	Статистические
Тип моделирования	Имитационное моделирование	Стохастическое
Топологическая сложность	Приемлемая	Высокая
Вычислительные ресурсы	Обычные	Повышенной мощности

Приложения, ориентированные исключительно на обработку больших объемов данных, имеют дело с наборами данных объемом от нескольких терабайт до петабайта. Как правило, эти данные поступают в нескольких разных форматах и часто распределены между несколькими местоположениями. Обработка подобных наборов данных обычно происходит в режиме многошагового аналитического конвейера, включающего стадии преобразования и интеграции данных.

Требования к вычислениям обычно почти линейно возрастают при росте объема данных, и вычисления часто поддаются простому распараллеливанию. К основным исследовательским проблемам относятся управление данными, методы фильтрации и интеграции данных, эффективная поддержка запросов и распределенности данных.

Особо следует подчеркнуть распределенность данных, которая сама по себе создает проблемы даже при не очень большом объеме. Это мотивирует разработку специальных пространственных моделей данных [14], которые часто отображают свойства информационного пространства или свойства поля [9].

Методики и методы работы с большими данными. Для приложений, ориентированных на суперобработку больших объемов данных, характерны потребность в обработке сверхбольших наборов данных и возрастающая вычислительная сложность. Требования к вычислениям нелинейно возрастают при росте объемов данных; для обеспечения правильного вида данных требуется применение сложных методов поиска и интеграции. Ключевыми исследовательскими проблемами являются разработка новых алгоритмов, генерация сигнатур данных и создание специализированных вычислительных платформ, включающих аппаратные ускорители. К числу приложений, которым свойственны соответствующие характеристики, относятся следующие.

A/B testing. Методика, в которой контрольная выборка поочередно сравнивается с другими. Тем самым удается выявить оптимальную комбинацию показателей для достижения, например, наилучшей ответной реакции потребителей на маркетинговое предложение. Большие данные позволяют провести огромное количество итераций и таким образом получить статистически достоверный результат.

Ad-hoc GRID - непосредственное формирование сотрудничающих гетерогенных вычислительных узлов в логическое сообщество без предварительно сконфигурированной фиксированной инфраструктуры и с минимальными административными требованиями.

Association rule learning. Набор методик для выявления взаимосвязей, т.е. ассоциативных правил, между переменными величинами в больших массивах данных. Используется в data mining.

BOINC-грид. Как правило, для обработки больших массивов данных используются суперкомпьютеры или вычислительные кластеры. Для достижения большей производительности вычислительные кластеры объединяются высокоскоростными каналами связи в специализированные ГРИД-системы. Однако с развитием сети Интернет появился и другой подход в построении ГРИД-систем, позволяющий объединить значительное число источников сравнительно небольших вычислительных ресурсов для решения задач обработки больших и сверхбольших объемов данных. В большинстве случаев такие системы построены на использовании свободных вычислительных ресурсов частных лиц и организаций, добровольно присоединяющихся к этим системам (volunteer computing). Однако существуют и примеры построения подобных частных (в масштабах организации или группы организаций) распределенных систем [15]. Наиболее эффективно использование таких распределенных систем для проведения серий независимых вычислительных экспериментов [16].

Calculation acceleration - ускорение вычислений - изменение скорости вычислений в одной системе при сравнении со скоростью вычислений в другой системе.

Classification. Набор методик, которые позволяет предсказать поведение потребителей в определенном сегменте рынка (принятие решений о покупке, отток, объем потребления и проч.). Используется в data mining.

Global GRID - глобальные ГРИД - устанавливаются в Интернете, предоставляя отдельным пользователям или организациям мощность ГРИД независимо от того, где в мире эти пользователи находятся. Это также называют Интернет-компьютингом

Cluster analysis. Статистический метод классификации объектов по группам за счет выявления наперед не известных общих признаков. Используется в data mining.

Cluster - кластер - доступная по сети группа рабочих узлов (при необходимости вместе

с головным узлом), размещённая на некотором сайте. Согласно определению в схеме GLUE, кластер это контейнер, который группирует вместе подкластеры или компьютерные узлы.

Cluster and multi-cluster GRIDs model - кластерная и мультикластерная модель ГРИД.

Crowdsourcing. Методика сбора данных из большого количества источников.

Data GRID - проект, финансируемый Европейским Союзом. Цель проекта - создание следующего поколения вычислительной инфраструктуры обеспечения интенсивных вычислений и анализа общих крупномасштабных баз данных (от сотен терабайт до петабайт) для международных научных сообщ

Data fusion and data integration. Набор методик, который позволяет анализировать комментарии пользователей социальных сетей и сопоставлять с результатами продаж в режиме реального времени.

Data mining. Процессы поиска данных большого объема по образцам. Называется также извлечение знания для баз данных (Knowledge-Discovery in Databases – KDD). Нетривиальное извлечение информации из явных первичных данных или использование потенциальной информации из данных. Схема извлечения полезной информации из больших наборов данных или больших баз данных

Ensemble learning. В этом методе задействуется множество предикативных моделей за счет чего повышается качество сделанных прогнозов.

Genetic algorithms. В этой методике возможные решения представляют в виде `хромосом`, которые могут комбинироваться и мутировать. Как и в процессе естественной эволюции, выживает наиболее приспособленная особь.

GRID (грид, сеть) - географически распределенная информационная система - технология распределённых вычислений, в которой вычислительная система («суперкомпьютер») представлена в виде соединенных сетью вычислительных узлов, слабосвязанных, гомогенных или гетерогенных компьютеров, работающих вместе для выполнения большого количества заданий. ГРИД-технология применяется для решения разного рода научных задач, требующих значительных вычислительных ресурсов.

GRID infrastructure - инфраструктура ГРИД - географически распределённая инфраструктура, объединяющая множество ресурсов разных типов (процессоры, долговременная и оперативная память, хранилища и базы данных, сети), доступ к которым пользователь может получить из любой точки, независимо от места их расположения.

Machine learning. Направление в информатике (исторически за ним закрепилось название `искусственный интеллект`), которое преследует цель создания алгоритмов самообучения на основе анализа эмпирических данных.

MIMD, Multiple Instruction Multiple Data - Вычислительная система со множественным потоком команд и множественным потоком данных

Natural language processing (NLP). Набор заимствованных из информатики и лингвистики методик распознавания естественного языка человека.

Network analysis. Набор методик анализа связей между узлами в сетях. Применительно к социальным сетям позволяет анализировать взаимосвязи между отдельными пользователями, компаниями, сообществами и т.п.

Optimization. Набор численных методов для редизайна сложных систем и процессов для улучшения одного или нескольких показателей. Помогает в принятии стратегических решений, например, состава выводимой на рынок продуктовой линейки, проведении инвестиционного анализа и проч.

Pattern recognition. Набор методик с элементами самообучения для предсказания поведенческой модели потребителей.

Predictive modeling. Набор методик, которые позволяют создать математическую модель наперед заданного вероятного сценария развития событий. Например, анализ базы данных CRM-системы на предмет возможных условий, которые подтолкнут абоненты сменить провайдера.

Regression. Набор статистических методов для выявления закономерности между изменением зависимой переменной и одной или несколькими независимыми. Часто применяется для прогнозирования и предсказаний. Используется в data mining.

Sentiment analysis. В основе методик оценки настроений потребителей лежат технологии распознавания естественного языка человека. Они позволяют вычленить из общего информационного потока сообщения, связанные с интересующим предметом (например, потребительским продуктом). Далее оценить полярность суждения (позитивное или негативное), степень эмоциональности и проч.

Signal processing. Заимствованный из радиотехники набор методик, который преследует цель распознавания сигнала на фоне шума и его дальнейшего анализа.

Spatial analysis. Набор отчасти заимствованных из статистики методик анализа пространственных данных – топологии местности, географических координат, геометрии объектов. Источником больших данных в этом случае часто выступают геоинформационные системы (ГИС).

Statistics. Наука о сборе, организации и интерпретации данных, включая разработку опросников и проведение экспериментов. Статистические методы часто применяются для оценочных суждений о взаимосвязях между теми или иными событиями.

Supervised learning. Набор основанных на технологиях машинного обучения методик,

которые позволяют выявить функциональные взаимосвязи в анализируемых массивах данных.

Simulation. Моделирование поведения сложных систем часто используется для прогнозирования, предсказания и проработки различных сценариев при планировании.

Time series analysis. Набор заимствованных из статистики и цифровой обработки сигналов методов анализа повторяющихся с течением времени последовательностей данных. Одни из очевидных применений – отслеживание рынка ценных бумаг или заболеваемости пациентов.

Unsupervised learning. Набор основанных на технологиях машинного обучения методик, которые позволяют выявить скрытые функциональные взаимосвязи в анализируемых массивах данных. Имеет общие черты с Cluster Analysis.

Visualization. Методы графического представления результатов анализа больших данных в виде диаграмм или анимированных изображений для упрощения интерпретации облегчения понимания полученных результатов.

Выводы. Анализ данных больших объемов требует привлечения технологий и средств реализации высокопроизводительных вычислений. Основными факторами проблемы являются в первую очередь сложность и во вторую физический объем информационной коллекции. Большие объемы данных порождают проблемы при формировании информационных ресурсов из таких данных [17]. По существу большие данные являются новой формой информационного барьера [4].

Большие данные с одной стороны обуславливают постановку и решение новых задач [18]. С другой стороны они обуславливают развитие интегрированных и комплексных систем и технологий. Преувеличенное внимание к «большим данным» со стороны журналистов и бизнесменов обусловлено отсутствием практики преодоления информационных барьеров и рассмотрением этого явления как совершенно нового, в то время как оно периодически появляется в развитии человечества и «новым» является не само явление, а «новое качество» известного явления. С познавательной точки зрения преодоление информационного барьера «большие данные» способствует развитию познания окружающего мира и построению его целостной картины.

Список литературы

1. Майер-Шенбергер В., Кукьер К. Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим. – Манн, Иванов и Фербер, 2014 -240с.

2. Черняк Л. Большие данные – новая теория и практика //Открытые системы. СУБД – 2011. – №10. – с.18-25.

3. Jacobs , A. The pathologies of big data //Communications of the ACM. – 2009. – Т. 52. – №. 8. – p.36-44.

4. V. Ya. Tsvetkov, A. A. Lobanov. Big Data as Information Barrier // European Researcher, 2014, Vol.(78), № 7-1, p. 1237-1242

5. Lynch C. Big data: How do your data grow? //Nature. – 2008. – Т. 455. – №. 7209. – p.28-29.

6. Космические исследования земных ресурсов. Методы и средства измерений и обработки информации. М.: Наука, 1976. - 386с.

7. Цветков В.Я. Методы и системы обработки и представления видеонформации. - М.:ГКНТ, ВНТИЦентр, 1991. - 113с.

8. Цветков В. Я. Естественное и искусственное информационное поле// Международный журнал прикладных и фундаментальных исследований. -2014. - №5, ч.2. – с.178 -180.

9. Tsvetkov V.Y. Information field**. //** Life Science Journal 2014- 11(5). –pp.551-554.

10. The Fourth Paradigm: Data-Intensive Scientific Discovery, 2009, URL: http://research.microsoft.com/enus/ collaboration/fourthparadigm

11. Loek Essers: CERN pushes storage limits as it probes secrets of universe, URL: http://news.idg.no/cw/art.cfm?id=FF726AD5-1A64-6A71-CE987454D9028BDF.

12. Казенников А.О., Соловьев И.В. Извлечение структурированного новостного сообщения из веб-страниц при использовании дополнительной информации RSS. // Вестник МГТУ МИРЭА «MSTU MIREA HERALD» 2014 - № 2 (3) - с.276-288.

13. V. Ya. Tsvetkov. Complexity Index // European Journal of Technology and Design, 2013, Vol.(1), № 1, p.64-69.

14. V. Ya. Tsvetkov. Spatial Information Models // European Researcher, 2013, Vol.(60), № 10-1, p.2386- 2392.

15. Прорывная технология машинного перевода и вокруг нее. PC WEEK, №9, 12 апреля 2011.

16. Е. Е. Ивашко, Н. Н. Никитина. Организация квантовохимических расчетов с использованием пакета Firefly в гетерогенной грид-системе на базе BOINC // Вычислительные методы и программирование, Том 13, 2012 , с. 8 — 12.

16. Матчин В.Т. Информационные ресурсы как инструмент научного исследования и развития // Вестник МГТУ МИРЭА. - 2014 - № 2 (3) - с.235-256.

17. Herodotou H. et al. Starfish: A Self-tuning System for Big Data Analytics //CIDR. – 2011. – Т. 11. – p.261-272.