УДК. 001.891

БОЛЬШИЕ ДАННЫЕ В ИНФОРМАЦИОННЫХ НАУКАХ

***Роман Геннадьевич Болбаков**, доцент, канд. техн. наук,*
доцент кафедры инструментального
и прикладного программного обеспечения,
e-mail: antaros05@ya.ru,
Институт информационных технологий,
Московский технологический университет (МИРЭА),
https://www.mirea.ru

Дается анализ проблемы «больших данных» в области информационных наук. 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 рост сложности информационных моделей и информационных коллекций. Описан методический и алгоритмический инструментарий, который применим при обработке больших данных в информационных науках.

Ключевые слова: данные; большие данные; информационные объемы; скорость вычислений; разнообразие данных; методы обработки; сложность; вычислительные ресурсы; анализ.

DOI: 10.21777/2312-5500-2017-1-30-35

Введение

В последние годы много говорится о проблеме «больших данных» (Big Data) [1, 2]. Чаще всего эту проблему связывают с необходимостью обработки структурированных и неструктурированных данных больших объемов. Для характеристики «больших данных» используют критерий «три *V*»: объем (*volume*), скорость (*velocity*), многообразие (*variety*).


Р.Г. Болбаков

Большой объем обусловлен тем, что организации собирают данные из различных источников, мультимедийных технологий и информации, полученной в глобальных сетях. Первоначально хранение Big Data было проблемой, но появление новых технологий типа Hadoop немного упростило ситуацию. Скорость обработки обусловлена современными требованиями обработки в реальном времени при больших информационных потоках. Разнообразие данных обусловлено современной тенденцией интеграции [3, 4] данных и технологий. Данные поступают во всех типах форматов – от структурированных числовых данных в традиционных базах данных до неструктурированных текстовых документов, электронной почты, видео- и аудиоданных. Однако три критерия являются существенным упрощением ситуации. Можно согласиться с точкой зрения SAS [5] и включить два дополнительных параметра, когда речь идет о больших данных, – сложность [6] и изменчивость. Изменчивость обусловлена увеличением скорости и разновидности данных, потоки которых могут быть несовместимыми и потоки которых могут иметь различные пики интенсивности.

Развитие проблемы. Появление термина соотносят с 2008 годом [7]. Введение термина «большие данные» связывают с Клиффордом Линчем – редактором журнала Nature [7], подготовившим 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 [8]. Выяснилось, что для обработки фотограмметрической информации, которую получали с первых спутников Земли, потребуются десятилетия с учетом вычислительных ресурсов того времени. Здесь следует отметить, что такую проблему отметили только две страны – СССР и США в довольно узкой области космических исследований Земли. Общественность об этом не подозревала и не

понимала проблемы. Кроме того, необходимо отметить, что проблема больших данных молчаливо соотносится с состоянием вычислительных ресурсов. Уровень проблемы больших данных 1980 года не имел места уже в конце 90-х годов, поскольку вычислительные ресурсы ее преодолели.

Проблему больших данных 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. Однако освещение проблемы дается односторонне. Говорят о больших объемах и сложности, молчаливо считая, что вычислительные ресурсы не развиваются и находятся на одном уровне.

В процессе развития человеческого общества происходит постоянный конфликт между объемами информации и ресурсами для их анализа и обработки. Как результат наблюдения человека за объектами, явлениями и процессами окружающего мира, происходит получение информации в информационном поле [9, 10], накопление опыта и формирование описаний объектов, явлений и процессов. Первичное описание объектов окружающего мира состоит в фиксации количественных и качественных свойств, объектов и отношений между ними [11].

Вторичное описание состоит в формировании моделей данных и моделей объектов, формируемых на основе анализа первичных описаний. 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 актуальными исследования в области методов и алгоритмов анализа больших и сверхбольших наборов данных. В работе [12] 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 [12], а также широкий спектр организаций в самых различных областях. В современной науке возникло новое направление, связанное с анализом больших и сверхбольших наборов данных, Big Data [13]. В информационной науке дополнительным фактором сложности являются информационные модели, которые, в отличие от обычных моделей, представляют собой системный информационный ресурс [14, 15]. С одной стороны, это позволяет решать более широкий круг задач, с другой – возникает проблема организации такого системного ресурса.

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 проблемы. 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, интенсифицированные потоки данных, существенное сокращение допустимого времени анализа данных, предел времени принятия решений при любом количестве данных, возрастание морфологической сложности моделей, возрастание структурной сложности моделей и систем, возрастание вычислительной сложности, относительный рост слабоструктурированной исходной информации, относительный рост нечеткой информации, рост потребностей в параллельных вычислениях и т. д.

Упрощенно проблемы работы с данными большого объема приведены в табл. 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	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 распределение данных, которое само по себе создает проблемы даже при не очень большом объеме. Это мотивирует разработку специальных пространственных моделей данных, которые часто отображают свойства информационного пространства или свойства поля.

Методики и методы работы с большими данными. Потенциальные трудности, с которыми могут столкнуться организации при анализе больших данных, включают в себя отсутствие внутренних средств аналитики и высокую стоимость найма внешних опытных специалистов-экспертов. Количество информации, которая применяется при анализе, и ее разнообразие могут вызывать дополнительную проблему качества данных и вопросы согласованности данных [16]. Обычные базы данных не подходят для хранения Big Data, поэтому чаще всего применяют Hadoop. Однако интеграция системы Hadoop и хранилищ данных может быть проблемой, хотя различные производители в настоящее время предлагают разъемы программного обеспечения между Hadoop и реляционными базами данных, а также инструменты интеграции других данных с большими возможностями данных.

Основным методом преодоления проблемы являются различные платформы, которые разрабатывают известные фирмы [17]. Для приложений, ориентированных на обработку больших объемов данных, характерна 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 данных и создание специализированных 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. Этим удается выявить оптимальную комбинацию показателей для достижения наилучшей ответной реакции потребителей на предложение. Большие данные поз-

воляют провести огромное число итераций и таким образом получить статистически достоверный результат.

Ad-hoc GRID – методика, основанная на формировании сотрудничающих гетерогенных вычислительных узлов в логическое сообщество без предварительно сконфигурированной фиксированной инфраструктуры и с минимальными административными требованиями.

BOINC-грид. В этой методике для обработки больших массивов данных используются 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.

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

Classification. Набор методик, которые позволяют предсказать поведение потребителей в определенном сегменте рынка (принятие решений о покупке, отток, объем потребления и проч.). Используется в data mining. Эта технология включает: обучение ассоциативным правилам (association rule learning), классификацию (методы категоризации новых данных на основе принципов, ранее примененных к уже наличествующим данным), кластерный анализ, регрессионный анализ.

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

Crowd sourcing. 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 сообществ.

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

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.

MIMD, Multiple In struction 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 абонента к смене 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.

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

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. По существу, большие данные являются новой формой информационного барьера [2]. Большие данные, с одной стороны, обусловливают постановку и решение новых задач [18]. С другой стороны, они обусловливают 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.

Литература

1. *Jacobs A.* The pathologies of big data // Communications of the ACM. 2009. Vol. 52. Iss. 8. P. 36–44.

2. *Tsvetkov V. Ya., Lobanov A. A.* Big Data as Information Barrier // European researcher. Series A. 2014. Vol. 78. Iss. 7-1. P. 1237–1242.

3. *Cohen W. W., Richman J.* Learning to match and cluster large high-dimensional data sets for data integration // Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. – ACM, 2002. P. 475–480.

4. *Цветков В. Я.* Ресурсность и интегративность сложной организационно технической системы // Международный журнал прикладных и фундаментальных исследований. 2016. № 5 (часть 4). С. 676–676.

5. http://www.sas.com/en_us/insights/big-data/what-is-big-data.html.

6. *Tsvetkov V. Ya.* Complexity Index // European Journal of Technology and Design. 2013. Vol. 1. Iss. 1. P. 64–69.

7. *Lynch C.* Bigdata: Howdoyourdatagrow? // Nature. 2008. Vol. 455. Iss. 7209. P. 28–29.

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

9. *Tsvetkov V. Ya.* Information field // Life Science Journal. 2014. Vol. 11. Iss. 5. P. 551–554.

10. *Бондур В. Г.* Информационные поля в космических исследованиях // Образовательные ресурсы и технологии. 2015. № 2 (10). С. 107–113.

11. *Цветков В. Я.* Фактофиксирующие и интерпретирующие модели // Международный журнал прикладных и фундаментальных исследований. 2016. № 9-3. С. 487–487.

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

13. http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics.

14. *Павлов А. И.* Информационные модели и информационные единицы // Перспективы науки и образования. 2015. № 6. С. 12–17.

15. *Савиных В. П., Цветков В. Я.* Геоданные как системный информационный ресурс // Вестник Российской академии наук. 2014. Т. 84. № 9. С. 826–829.

16. *Дулин С. К., Розенберг И. Н.* Об одном подходе к структурной согласованности геоданных // Мир транспорта. 2005. Т. 11. № 3. С. 16–29.

17. IBM big data platform – Bringing big data to the Enterprise. https://www.ibm.com/software/data/bigdata.

18. *Herodotou H. et al.* Starfish: A Self-tuning System for Big Data Analytics // CIDR. 2011. Vol. 11. P. 261–272.

Big data in information sciences

***Roman Genad'evich Bolbakov**, Associate Professor, Ph.D., Assistant professor, Institute of Information Technology. Moscow Technologies University (MIREA)*

The article analyzes the problem of «big data» in the field of information sciences. The article describes the causes of the problem and the factors that led to its emergence. The article describes the comparison of large data and normal data. The article shows that the problem of large data is not only large amounts of data collections. Important factors are large data: time constraints on data processing and analysis, as well as increase the complexity of the information models and information collection. This article describes the methodological and algorithmic tools that apply when processing large data in information sciences.

Keywords: data, big data, information volume, computing speed, a variety of data processing methods, complexity, computational resources, analysis.