А.В. Замятин

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ



УДК 519.254
ББК 32.81
З26

Замятин А.В.
З26 Интеллектуальный анализ данных : учебное пособие. – Томск : Издательский Дом Томского государственного университета, 2020. – 196 с.

ISBN 978-5-94621-898-6

В учебном пособии рассматриваются вопросы, связанные с популярной сегодня областью машинного обучения и интеллектуального анализа данных. Исследуются основные технологические тренды, наиболее активно использующие алгоритмы интеллектуальной обработки данных, – бизнес, медицина, управление, индустрия. Обсуждаются вопросы терминологии, основные методы анализа и интерпретации данных, методы и инструменты машинного обучения.

Приведены вопросы для самопроверки.

Для студентов университетов и втузов.

УДК 519.254
ББК 32.81

Рецензенты:
доктор технических наук, профессор *Л.Г. Гагарина*
доктор технических наук, профессор *С.П. Сущенко*

ISBN 978-5-94621-898-6

© Замятин А.В., 2020
© Томский государственный университет, 2020

ОГЛАВЛЕНИЕ

Введение 6
1. Актуальность 7
2. Терминология 11
2.1. Data Mining / Data Science 15
2.2. Big Data 18
2.2.1. Основные понятия 20
2.2.2. Свойства Big Data 21
2.3. Data Mining и Big Data 22
2.4. Дедукция и индукция 22
3. Примеры применения 23
3.1. Интеллектуальный анализ данных в бизнесе 23
3.1.1. Розничная торговля 24
3.1.2. Сфера развлечений 25
3.1.3. Маркетинг, страхование, работа с персоналом 26
3.1.4. Примеры применения классификации, кластеризации и прогнозирования 27
3.2. Интеллектуальный анализ данных в решении сложных прикладных задач 29
3.2.1. Медицина 30
3.2.2. Государственное управление 31
3.3. Интеллектуальный анализ данных в ранней диагностике опасных заболеваний 32
3.4. Интеллектуальный анализ данных в индустриальной предиктивной аналитике 33
4. Основные задачи и классификация методов анализа данных 37
4.1. Этапы интеллектуального анализа данных 37
4.2. Общие типы закономерностей при анализе данных 37
4.3. Группы задач анализа данных 38
4.4. Классификация методов 42
4.5. Сравнительные характеристики основных методов 44
5. Принципиальные основы машинного обучения 46
6. Основные методы анализа и интерпретации данных 50
6.1. Предварительная обработка данных 50

6.2. Оптимизация признакового пространства 56
6.2.1. С трансформацией пространства признаков 57
6.2.2. Без трансформации пространства признаков 59
6.3. Классификация 61
6.3.1. Постановка задачи классификации 61
6.3.2. Контролируемая непараметрическая классификация 65
6.3.3. Контролируемая непараметрическая нейросетевая классификация 66
6.3.4. Классификация по методу машины опорных векторов 70
6.3.5. Деревья решений 72
6.4. Неконтролируемая классификация (кластеризация) 84
6.5. Регрессия 89
6.5.1. Понятие регрессии 89
6.5.2. Основные этапы регрессионного анализа 90
6.5.3. Методы восстановления регрессии 91
6.6. Ассоциация 92
6.6.1. Описание алгоритма 95
6.6.2. Пример исполнения алгоритма 96
6.7. Последовательная ассоциация 98
6.7.1. Алгоритмы семейства «Априори» 99
6.7.2. Алгоритм GSP 102
6.8. Многоуровневое машинное обучение 107
6.8.1. Бутстрэппинг 108
6.8.2. Бэггинг 109
6.8.3. Стекинг 111
6.8.4. Бустинг 112
6.9. Обнаружение аномалий 115
7. Визуализация 118
8. Нейросетевые подходы и глубокое обучение 120
8.1. Функции активации 120
8.2. Основные типы искусственных нейронных сетей 123
8.3. Сверточные нейронные сети (Convolutional Neural Networks) 131
8.4. Популярные архитектуры CNN 134
8.5. Среды и фреймворки глубинного обучения 138

9. Обработка естественного языка 140
9.1. Основные задачи обработки текста 140
9.2. Этапы предварительной обработки текста 143
10. Критерии точности 145
10.1. Метрики качества классификации 145
10.2. Гипотеза A/B 150
10.3. Каппа-индекс согласия 150
10.4. ROC-кривая 152
10.5. Метрика качества прогноза временного ряда 153
10.6. Метрики качества кластеризации 155
11. Высокопроизводительная обработка данных 157
11.1. Принципы высокопроизводительных вычислений 157
11.2. Особенности построения вычислительного кластера 161
11.3. Среды и инструменты высокопроизводительных вычислений 171
12. Инструменты Data Mining 176
12.1. Программные инструменты для высокопроизводительной обработки данных 177
12.1.1. Программная среда 177
12.1.2. Базы данных 178
12.1.3. Языки программирования 179
12.2. Примеры программных систем 179
12.2.1. Примеры самостоятельных систем 179
12.2.2. Примеры облачных систем 180
Вопросы и темы для самопроверки 182
Литература 184

1. АКТУАЛЬНОСТЬ

С 1960-х гг. информационно-коммуникационные технологии (ИКТ) последовательно эволюционировали от простых систем обработки файлов до сложных, мощных систем управления базами данных (БД). Исследования в области БД с 1970-х гг. смещались от ранних иерархических и сетевых баз данных к реляционным системам управления базами данных (СУБД), инструментам моделирования данных, а также к вопросам индексирования и организации данных. Пользователи получили гибкий и удобный интерфейс доступа к данным с помощью языков запросов (типа SQL), пользовательские интерфейсы, управление транзакциями и т.п. При этом создаваемые и поддерживаемые БД преимущественно имели ограниченный регистрирующий характер, поддерживая рутинные операции линейного персонала. Основным требованием к таким системам было обеспечение транзакционности и оперативности выполнения всех изменений.

Технология баз данных начиная с середины 1980-х гг. характеризовалась популяризацией, широким внедрением и концентрацией исследовательских усилий на новых, все более мощных СУБД. Появились новые модели данных, такие как объектно-ориентированные, объектно-реляционные, дедуктивные модели. Возникали различные предметно-ориентированные базы данных и СУБД (пространственные, временные, мультимедийные, научные и пр.). Эффективные методы онлайновой обработки транзакций (*On-Line Transaction Processing*; *OLTP*[1]) внесли большой вклад в эволюцию и широкое внедрение реляционной технологии в качестве одного из главных универсальных инструментов эффективного хранения, извлечения и управления большими объемами структурированных данных реляционных СУБД.

С развитием сети Интернет получили развитие и вопросы построения распределенных баз данных, создания распределенных

[1] Способ организации БД, при котором система работает большим потоком с небольшими по размерам транзакциями при минимальном времени отклика системы.

глобальных информационных систем. Многократно возросла интенсивность формирования и архивирования различных данных, следствием чего стало развитие масштабируемых программно-аппаратных комплексов, дорогостоящих мощных и недорогих пользовательских компьютеров и накопителей данных.

Все это способствовало всплеску развития индустрии ИКТ и сделало огромное количество баз данных доступными для хранения разнородной информации в значительных объемах и управления транзакциями в них. При этом все больше актуализировалась потребность анализа имеющихся данных в разновременном аспекте с возможностью построения произвольных запросов при условии обработки сверхбольших объемов данных, полученных в том числе из различных регистрирующих БД. Использование для реализации таких задач традиционных регистрирующих систем и БД крайне затруднительно. Например, в регистрирующей системе информация актуальна исключительно на момент обращения к БД, а в следующий момент времени по тому же запросу можно ожидать другой результат. Интерфейс подобных систем рассчитан на проведение определенных стандартизованных операций, и возможности получения результатов на нерегламентированный произвольный запрос ограничены. Возможности обработки больших массивов данных также могут быть ограничены вследствие ориентации СУБД на нормализованные данные, характерные для стандартных реляционных регистрирующих БД.

Ответом на возникшую потребность стало появление новой технологии организации баз данных – технологии *хранилищ данных* (*Data Warehouse*[1]), предполагающей некоторую предварительную обработку данных и их интеграцию, а также онлайновую аналитическую обработку (*On-Line Analytical Processing*; *OLAP*[2]).

[1]Предметно-ориентированная информационная база данных, предназначенная главным образом для поддержки принятия решений с помощью отчетов.

[2] Технология анализа данных, предполагающая подготовку агрегированной структурированной многомерной информации на основе больших массивов данных (*OLAP*-куба), используемой в реляционной БД при построении сложных многотабличных запросов.

Несмотря на очевидную пользу такого инструмента анализа данных, он ориентирован на хорошо нормализованные табличные данные и не предполагает использования целого ряда дополнительного аналитического инструментария типа классификации, кластеризации, регрессионного анализа, моделирования, прогнозирования и интерпретации многомерных данных и т.п.

Таким образом, сегодня наблюдается высокий уровень развития масштабируемой аппаратно-программной ИКТ-инфраструктуры, позволяющей увеличивать и без того значительные архивы данных. Имеется достаточно существенный задел в области компьютерных наук и информационных технологий, разработаны теория и прикладные аспекты теории вероятностей и математической статистики. Вместе с тем следует признать, что присутствует заметный *избыток данных*[1] при *дефиците информации*[2] *и знаний*[3]. Быстро растущие объемы накопленных и пополняемых (автоматически, а не людьми – как это было когда-то) архивов данных пока существенно превышают способности человека в их практически полезной обработке. Для обострения этого тезиса иногда говорят, что «*…большие базы данных стали могилами, которые редко посещаются…*» Как следствие, важные решения порой принимаются не на основе аналитических выводов из информативных БД, а на основе интуиции человека, не имеющего подходящих инструментов для извлечения полезных знаний из имеющихся огромных объемов данных.

Поэтому в последние годы стремительное развитие получила область *Data Science / Data Mining*[4] (в отечественной литературе наиболее используемая аналогия – *интеллектуальный анализ*

[1] Под *данными* будем понимать представление некоторых фактов в формализованном виде, пригодном для хранения, обработки и передачи.

[2] Под *информацией* будем понимать сведения в любой форме; в отличие от данных информация имеет некоторый *контекст*.

[3] Под знаниями будем понимать совокупность информации о мире, свойствах объектов, закономерностях процессов и явлений, а также правилах их использования *для принятия решений*.

[4] Вопросам терминологии посвящена глава 2.

данных, ИАД), направленная на поиск и разработку методов извлечения из имеющихся данных *знаний*, позволяющих принимать на их основе конкретные, в высокой степени обоснованные, практически полезные управленческие решения.



Рис. 1. Пример обобщенного иерархического представления методологий обработки данных при принятии управленческих решений

На рис. 1 приведен пример обобщенного иерархического представления методологий обработки данных, начиная от интеграции разнородных источников данных и завершая использованием методов *Data Mining* для принятия управленческих решений.

2. ТЕРМИНОЛОГИЯ

Рассматривая вопросы терминологии, описывающей обсуждаемую предметную область интеллектуального анализа данных, логично изучить существующую и наиболее устоявшуюся в мире *англоязычную терминологию*, и уже ориентируясь на нее обсудить удачные *терминологические аналогии*, используемые в русскоязычных публикациях по данной тематике.

Выше, в главе 1, отмечено, что с развитием ИКТ-индустрии стремительно развиваются возможности генерирования значительных массивов данных, при умелом анализе которых могут быть найдены полезные знания, позволяющие повысить эффективность принятия управленческих решений в бизнесе, медицине или государственном управлении. Область, изучающую эти вопросы, принято называть *Data Mining* (сегодня за рубежом даже чаще встречается понятие *Data Science*), а специалиста этой области – *Data Scientist*.

На рис. 2 изображен график роста востребованности данных специалистов за последние несколько лет. На рис. 3 приведена диаграмма, отражающая число вакансий специалистов *Data Scientist* в последнее время на портале для поиска работы. Интересно отметить, что на сегодняшний день около 85% всех вакантных позиций такого типа открыты в США, а 15% – в странах Западной Европы [108].

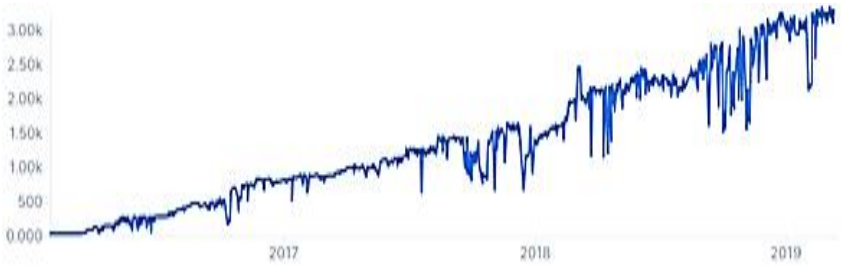


Рис. 2. График уровня востребованности специалистов Data Scientist

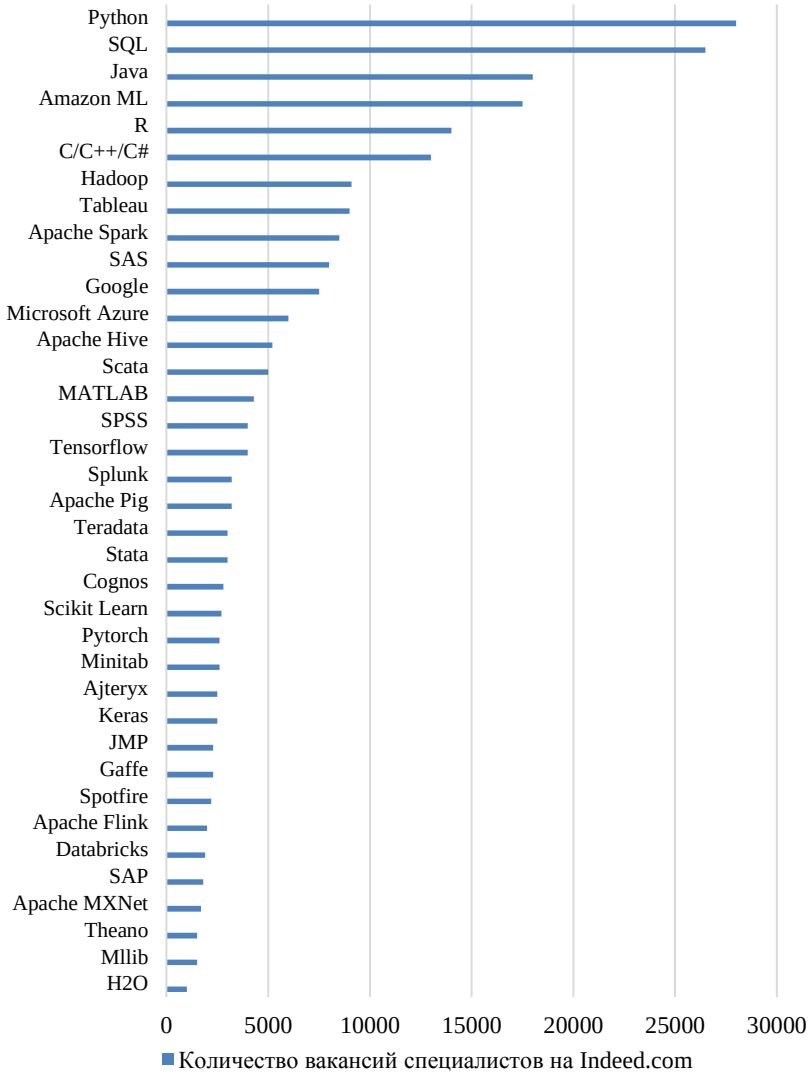


Рис. 3. Иллюстрация востребованности специалистов Data Scientist.
Источник: http//www.indeed.com/jobtrends June 2016

Интересной выглядит визуализация частоты запросов к поисковой системе, наглядно демонстрирующая широту использования основной и вспомогательной терминологии в данной предметной области (рис. 4).



а *б*

Рис. 4. Пример профиля частотности запросов:
а – о предметной области анализа данных; *б* – о специалистах

При этом, если посмотреть на желаемый профиль специалиста *Data Science*, то видно, насколько разносторонней (междисциплинарной) квалификацией, с точки зрения современного работодателя, он должен сегодня обладать (рис. 5):

- SQL – 54%;
- Python – 46%;
- R – 44%;
- SAS – 36%;
- Hadoop – 35%;
- Java – 32%;
- optimization – 23%;
- C++ – 21%4
- visualization – 20%;
- MATLAB – 18%;
- Business Intelligence – 17%;
- distributed – 16%;
- regression – 16%;
- unstructured – 16%;
- Hive – 16%;
- mobile – 15%.

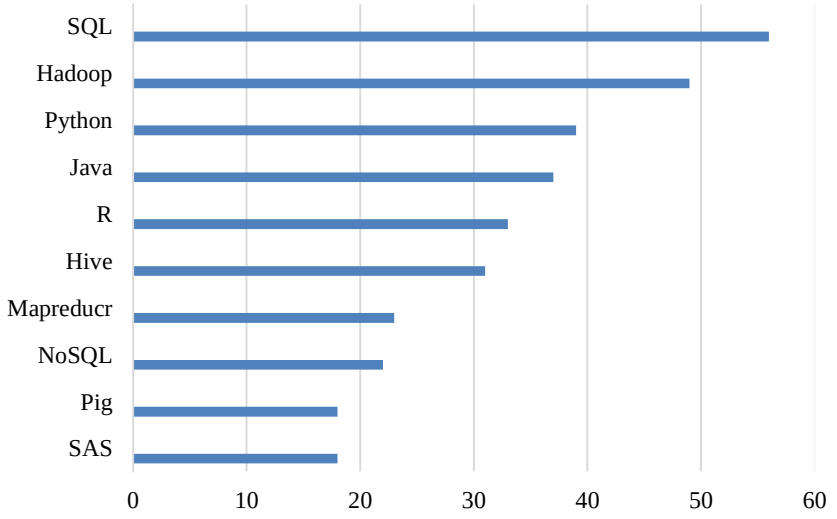


Рис. 5. Ожидаемая квалификация специалиста *Data Science*:
наиболее важные умения специалистов Data Science на Linkedln

Учитывая продолжающееся интенсивное развитие области анализа данных, встречается отличающаяся терминология, описывающая одно и тоже явление или сферу, или один термин, который может быть трактуем по-разному. Например, в англоязычной литературе можно встретить различные термины и их сочетания, описывающие область интеллектуального анализа данных и являющиеся достаточно близкими по значению:

- Data Science;
- Data Mining;
- Big Data;
- Machine Learning;
- Deep Learning;
- Statistical Analysis and Data Mining;
- Predictive Analytics and Data Mining;
- Data Science and Data Mining;
- Discovery Driven Data Mining;
- Knowledge Discovery in Databases и др.

Обзору большинства этих терминов посвящено данное пособие, однако одно из ключевых мест занимает сегодня понятие «глубинного / глубокого обучения» (англ. *Deep Learning*). Такое обучение предполагает применение преимущественно нейросетевых моделей, причем отличающихся именно сложной архитектурой и высокой обучающейся емкостью (например, к таким нейросетевым моделям относят сверточные нейронные сети, рассматриваемые в разд. 8.3).

2.1. Data Mining / Data Science

Существующие массивы данных не только характеризуются значительным объемом и регулярной пополняемостью, но и содержат порой преимущественно тривиальные (неактуальные, ошибочные и т.п.) элементы. Процесс поиска в этих данных чего-то ценного стал сравним с работой на горнорудных предприятиях, где в многотонных завалах руды осуществляется поиск (добыча) драгоценных металлов или камней, полезный выход которых может исчисляться граммами. Учитывая сходную трудоемкость процесса «добычи» (англ. *mining*) знаний из «завалов» данных термин закрепился и для области *Data Mining.*

Для *Data Mining* могут быть даны различные определения, не претендующие на исключительную полноту. *Data Mining* – это*:*

1) процесс обнаружения в базах данных нетривиальных и практически полезных закономерностей [5];

2) процесс выделения, исследования и моделирования больших объемов данных для обнаружения неизвестных до этого структур (паттернов) с целью достижения преимуществ в бизнесе [34];

3) процесс, цель которого – обнаружить новые значимые корреляции, образцы и тенденции в результате просеивания большого объема хранимых данных с использованием методик распознавания образцов и других статистических и математических методов [19];

4) исследование и обнаружение «машиной» (алгоритмами, средствами искусственного интеллекта) в «сырых» данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком [53];

5) процесс обнаружения полезных знаний о бизнесе [45].

Поэтому для более глубокого понимания сути явления *Data Mining* только какого-либо одного определения не вполне достаточно. Под *Data Mining* (наиболее устоявшаяся аналогия в русском языке, и все-таки не прямой перевод, – *интеллектуальный анализ данных*; по-видимому, это более широкий термин) понимают совокупность методов обнаружения в данных таких знаний, которые обязательно обладают следующими свойствами:

– ранее неизвестные, неожиданные;
– практически полезные;
– доступные для интерпретации;
– необходимые для принятия решений в различных сферах человеческой деятельности.

В более широком смысле под *Data Mining* понимают концепцию анализа данных, предполагающую, что:

– данные могут быть неточными, неполными (содержать пропуски), противоречивыми, разнородными, косвенными и при этом иметь гигантские объемы; поэтому понимание данных в конкретных приложениях требует значительных интеллектуальных усилий;
– сами алгоритмы анализа данных могут обладать «элементами интеллекта», в частности способностью обучаться по прецедентам, т.е. делать общие выводы на основе частных наблюдений; разработка таких алгоритмов также требует значительных интеллектуальных усилий;
– процессы переработки сырых данных в информацию, а информации в знания уже не могут быть выполнены по старинке «вручную» и требуют порой нетривиальной автоматизации.

Data Mining – мультидисциплинарная область, возникшая и развивающаяся на базе достижений прикладной математической статистики, распознавания образов, методов искусственного интеллекта, теории баз данных и др. (рис. 6). Иногда отмечают иной характер мультидисциплинарности *Data Mining* – это объединение *компьютерных наук* (англ. *Computer Science*), *математики* (англ. *Mathematics*) и представлений о *предметной области* (англ. *Domain Expertise*) (рис. 7). В этом случае компьютерные науки описывают среду

создания информационных продуктов (англ. *data products*), математика выстраивает теоретическую основу для решения поставленных проблем, а представление о предметной области позволяет понять реальность, в которой существует проблемная ситуация.

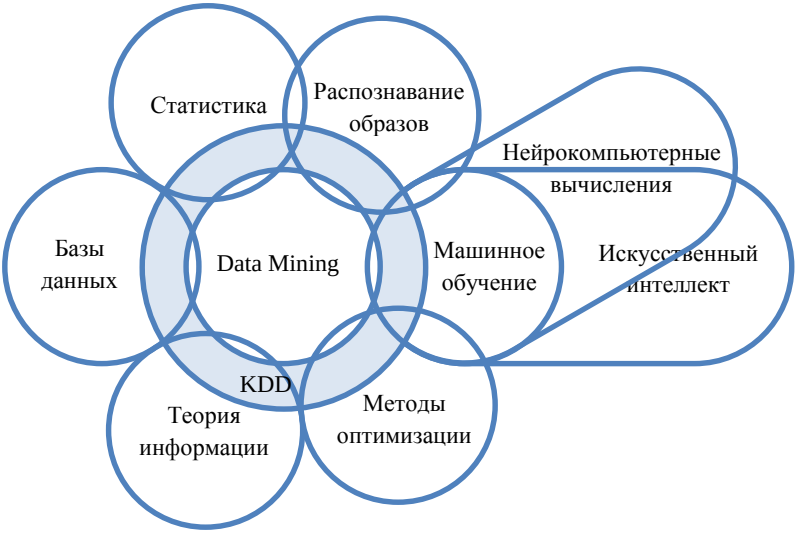


Рис. 6. Иллюстрация Data Mining как междисциплинарной области



Рис. 7. Иллюстрация Data Mining как междисциплинарной области

Именно включение *предметной области* как междисциплинарной компоненты *Data Mining* существенно осложняет практическую интеллектуальную работу в этой сфере, требуя от специалиста (*Data Scientist*) при решении каждой конкретной задачи анализа данных достаточно глубокого погружения в новую, незнакомую ему область человеческой деятельности. Однако очевидно, что без такого погружения сложно найти эффективные решения существующих проблем.

Следует отметить еще одну важную отличительную особенность *Data Mining*. Вычислительная сложность многих традиционных методов математической статистики или технологий БД определяется как $O(n^2)$, или (n^3), где n – объем исходных данных. Однако при сверхбольших объемах данных (превышающих порой несколько миллионов записей), характерных для современного этапа развития отрасли ИКТ, непосредственное применение таких традиционных методов обработки данных крайне затруднительно даже на самой современной мощной вычислительной технике. Учитывая эту особенность, *Data Mining* предполагает создание и развитие *специализированных алгоритмов*, характеризующихся значительно более высокой вычислительной эффективностью ($O(n)$, $O(\log n)$ и т.п.), достигаемой, например, за счет поиска приближенного результата (эвристики) без существенной потери точности.

2.2. Big Data

Анализ терминологии, достаточно полно описывающей область *Data Mining*, позволяет отметить, что в ней часто в схожем контексте применяется и такой термин, как *Big Data* (рус. *большие данные*). При этом широта его употребления специалистами и неспециалистами порой затрудняет однозначное толкование этого термина, в особенности учитывая обсуждение деталей понятия *Data Mining* выше. Как же следует понимать *Big Data?*

Ежедневно в мире создается более 5 эксабайтов[1] информации. В 2012 г. в мире было сгенерировано около 2,43 Зеттабайт (1 ЗБ –

[1] Единица измерения, равная 10^{18}, или 2^{60}, байт.

около 1 млрд Гб), что более чем в 2 раза превосходит объем информации в цифровом виде в 2010 г. (1,2 ЗБ). К 2020 г. информационные системы имеют дело с количеством данных, равным 40 ЗБ (примерно в 57 раз большим, чем количество песчинок на пляжах всей поверхности Земли). Очевидно, появление и активное тиражирование термина *Big Data* во многом вызвано сопровождением этого объективного процесса накопления сверхбольших объемов данных.

Действительно, обеспечивать соответствие возможностей ИКТ-инфраструктуры, предназначенной лишь для надежного хранения стремительно растущих объемов накапливаемых данных, не говоря уже о возможностях их интеллектуальной обработки, – непростая и дорогостоящая задача. Это означает перспективы формирования и развития рынка *Big Data* со значительной финансовой емкостью. Аналитики отмечают, что мировой рынок *Big Data* (технологий и сервисов для обработки данных) в ближайшие годы будет расти в среднем на 13,2% в год и к 2022 г. вырастет до 274,3 млрд долл.

Целесообразно отметить, что основным драйвером продвижения термина *Big Data* во многом являются *рыночные законы маркетинга*[1]. Они позволяют привлечь к проблеме внимание не только ученых, но и государства и бизнеса, и предусмотреть в бюджетах компаний средства на развитие этих технологий (в первую очередь – именно аппаратной ИКТ-инфраструктуры, что объясняет не столь высокую популярность термина *Data Mining*, связанного больше со специализированным алгоритмическим и программным обеспечением интеллектуальной обработки данных). Именно поэтому порой дискуссия о границах того, что является «действительно» «большими данными», а что уже не является, сводится к тому, насколько дорогостоящая инфраструктура требуется для их поддержки[2].

[1] Именно они, главным образом, «реанимировали» традиционные технологии web-хостинга в виде популярных сегодня *облачных сервисов* и способствуют развитию суперкомпьютерной тематики, регулярно продвигая различные рейтинги мощности суперкомпьютеров в мире.

[2] Например, по этой логике компания *Google* или исследовательский коллайдер *CERN*, конечно, работают в концепции *Big Data*, а вот все менее масштабные предприятия – сомнительно…

2.2.1. Основные понятия

Рассмотрим некоторые наиболее типичные определения и толкования термина *Big Data* [56]:

– данные очень большого объема;

– область управления и анализа больших объемов данных;

– область управления и анализа больших объемов данных, представленных (в отличие от реляционных БД) в слабоструктурированных форматах (веб-журналы, видеозаписи, текстовые документы, машинный код или, например, геопространственные данные и т.п.);

– область работы с информацией огромного объема и разнообразного состава, весьма часто обновляемой и находящейся в разных источниках, в целях увеличения эффективности деятельности, создания новых продуктов и повышения конкурентоспособности;

– область, объединяющая техники и технологии, которые извлекают смысл из данных на экстремальном пределе практичности;

– постоянно растущий объем информации, поступающей в оперативном режиме из социальных медиа, от сетей датчиков и других источников, а также растущий диапазон инструментов, используемых для обработки данных и выявления на их основе важных бизнес-тенденций;

– наборы данных, превосходящие возможности традиционных программно-аппаратных инструментов оперирования данными (например, в случае, когда параметры набора данных превосходят возможности обработки стандартными средствами *MS Excel*).

Очевидно, и данный набор понятий не позволит совершенно четко определить, где построить границу между «действительно» «большими данными» и «просто данными». Сравнительно большой объем данных не всегда можно по умолчанию отнести к сфере *Big Data,* так как возможности их анализа зависят не только от объема данных, но и от вычислительной сложности задачи (очевидно,