УДК 004.6

Главацкий С.Т., Бурыкин И.Г.

Московский государственный университет имени М.В. Ломоносова, г. Москва, Россия

О ЦИКЛЕ КУРСОВ «АНАТИТИКА БОЛЬШИХ ДАННЫХ ДЛЯ МАТЕМАТИКОВ»*

АННОТАЦИЯ

В статье излагается подход к преподаванию специализации в области науки о данных для математиков на кафедре теоретической информатики механико-математического факультета МГУ имени М.В. Ломоносова. Предложен авторский взгляд на выбор тем и курсов (как основных, так и специальных).

КЛЮЧЕВЫЕ СЛОВА

Высшее образование; математика; большие данные; базы данных; анализ больших наборов данных; глубинный анализ процессов.

Sergey Glavatsky, Ilya Burykin

M.V. Lomonosov Moscow State University, Moscow, Russia

ABOUT COURSES CYCLE "DATA SCIENCE AND DATA MINING FOR MATHEMATICIANS"

ABSTRACT

The article sets out the approach to teaching specialization in the field of Data science for mathematicians at the Department of Theoretical Computer Science at Mechanics and Mathematics Faculty of Moscow State University. The author's view on the selection of topics and courses (both basic and special) is offered.

KEYWORDS

Higher education; mathematics; big data; database; data mining; processes mining.

На кафедре теоретической информатики механико-математического факультета МГУ имени М.В. Ломоносова в течение трех последних лет разрабатывается цикл специальных курсов и практикумов под общим наименованием “Аналитика больших данных для математиков” (“Data Science and Data Mining for Mathematicians”). В рамках этого направления преподавания планируется на базе имеющихся общих курсов, а также за счет введения новых общих дисциплин осуществить преподавание отдельной специализации по методам и алгоритмам представления, моделирования и анализа больших наборов данных.

В последние годы в мировом сообществе о больших наборах данных сложилось представление как о наборах данных, характеризующихся следующими основными особенностями:

- объемом (Volume);
- скоростью обновления (Velocity);
- разнообразием и неоднородностью (Variety);
- достоверностью (Veracity);
- стоимостью (Value).

В настоящее время для больших наборов данных успешно развиваются методы их представления и интеллектуального анализа [1]. Имеется много подходов к классификации основных научно-инженерных направлений в решении различных актуальных задач в этой сфере.

Выделяют, например, следующие 10 направлений исследований [2]:

1. Большие наборы данных. Сбор и обработка больших объемов данных (большие объемы) из различных источников и различных видов (большой выбор), при больших скоростях (с высокой скоростью);
2. Анализ данных. Широкая область бизнеса, которая имеет дело с использованием данных для построения средств поддержки принятия решений, которые помогают

* **Труды XI Международной научно-практической конференции «Современные информационные технологии и ИТ-образование» (SITITO’2016), Москва, Россия, 25-26 ноября, 2016**

бизнес-менеджерам принимать решения на регулярной основе;

3. Наука о данных. Разработка и использование статистических и математических моделей, алгоритмов и визуализаций с целью помочь объяснить данные различных видов, будь то структурированные или неструктурированные, используя методы статистический обработки, машинное обучение, искусственный интеллект или иные подходы. В практическом применении науки о данных часто используют большие наборы данных и специализированные алгоритмы для создания и тестирования различных моделей;
4. Интеллектуальный анализ данных. Извлечение новых знаний из различных наборов данных, как правило, с помощью структурированных запросов. В первую очередь интеллектуальный анализ данных ассоциируется с исследованием больших текстовых документов и открытия новых шаблонов при анализе текста. В настоящее время интеллектуальный анализ данных широко применяется при анализе как структурированных (например, представленных в реляционных БД), так и малоструктурированных данных (NoSQL);
5. Бизнес-аналитика. Набор инструментов и подходов, которые позволяют менеджерам управлять процессами на основе данных, собранных в результате каких-то процессов в так называемые хранилища данных. Как правило, – это инструментальные панели, предоставляющие комбинацию запросов, визуализаций и отчетов, предназначенных для достижения конкретных бизнес-целей;
6. Эконометрика. Отрасль прикладной статистики, предназначенной, в частности, для исследования экономических процессов с помощью статистических методов и анализа данных различных видов;
7. Статистика. В практических приложениях здесь выделяют описательную и выведенную статистики. Описательная статистика предоставляет характеризацию данных, используя различные методы измерений, а выведенная статистика предназначается для установления и проверки гипотез (теорий). Статистическая инженерия является смежной областью, где статистические модели построены как на основе данных, так и на основе дедуктивного и индуктивного подходов;
8. Машинное обучение. Процесс построения (чаще – статистических) моделей для нахождения комплексных решений различных задач, например: предсказание значений переменных данных на основе имеющихся данных (регрессия), классификация точек данных или кластеризации точек данных в многомерных пространствах с какой-то метрикой. Машинное обучение включает в себя разработку статистических моделей для преобразования сложных наборов данных в более простые, приближенные представления взаимосвязей между факторами, а также включает в себя кросс-проверки и оптимизации гипер-параметров для оценки качества моделей;
9. Искусственный интеллект. Совокупность методов, охватывающих статистические подходы к обучению и нейронные сети для моделирования поведения экспертных систем с использованием контролируемых (где данные и целевые показатели задаются в явном виде) и неконтролируемых (где шаблоны отыскиваются) подходов к обучению. Методы искусственного интеллекта в большей степени, нежели машинное обучение, базируются на сопоставлении с шаблонами и на распознавании образов;
10. Математическое моделирование. Очень широкое направление, включающее в себя построение и проверку моделей самых различных классов: статистические модели; модели, основанные на алгоритмах линейной алгебры; модели, основанные на системах обыкновенных дифференциальных уравнений и дифференциальных уравнений с частными производными; модели, основанные на методах теории групп, и многие другие. Модели могут описывать реальные или воображаемые сценарии, но, как правило, предназначены для решения и описания реальных проблем.

Приведенная классификация не вполне отражает особенности математического содержания применяемых в рассматриваемой сфере методов и алгоритмов, используя в большей степени функциональные подходы. Так, область “Наука о данных“ (“Data Science”) другими исследователями представляется именно как содержательный раздел науки, а не просто как одно из направлений исследований. В частности, предлагается [3] следующая структура этого раздела (Рис. 1):

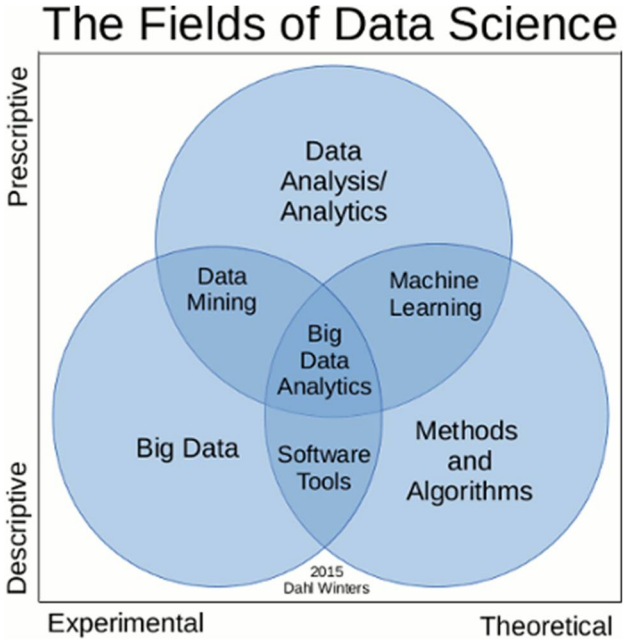

Рис.1. Схема содержания области "Наука о данных"

Здесь:

- по вертикальной оси отложена шкала "Описательные - предписывающие"; а
- по горизонтальной – "Экспериментальные - теоретические";

Термины:

- "Big Data" означает большие наборы данных;
- "Data Analysis/Analytics" – анализ/аналитика данных;
- "Methods and Algorithms" – методы и алгоритмы;
- "Data Mining" – интеллектуальный анализ данных;
- "Machine Learning" – машинное обучение;
- "Software Tools" – программные средства;
- "Big Data Analytics" – аналитика больших наборов данных.

Как мы видим, здесь в науку о данных включается ряд разделов, другими авторами отнесенных к иным направлениям. По-нашему же мнению, такое представление в большей степени соответствует математическому подходу к сфере анализа больших наборов данных.

Мы не станем углубляться далее в анализ различных концепций, взаимоотношений и взаимосвязей между понятиями аналитики данных (Data Analytics), анализа данных (Data Analysis), интеллектуального анализа данных (Data Mining), науки о данных (Data Science), машинного обучения (Machine Learning) и большими наборами данных (Big Data). Нашей целью является, в определенном смысле, выделение подходов к исследованиям, основанных на:

- использовании математических теорий, понятий и моделей;
- постановке и решении математических задач;
- применении разработанных алгоритмов в науке о данных.

Условно говоря, набор тем, включенных в наш цикл курсов, включает в себя следующее:

1. Типы данных, структуры данных, модели данных;
2. Представление данных, хранение и передача данных;
3. Методы и алгоритмы первичной обработки данных, базы данных, языки манипулирования данными;
4. Проектирование баз данных. Языки определения данных. Нормальные формы в проектировании реляционных баз данных;
5. Структурированные и неструктурированные данные, хранилища данных;
6. Анализ больших (неструктурированных) наборов данных, технологии распараллеливания обработки и сжатия информации;
7. Задачи интеллектуального анализа больших наборов данных. Проблемы больших объемов и размерностей;
8. Вероятностные методы первичного сжатия данных, хеширование и статистические оценки;
9. Задача обнаружения схожих документов, предлагаемые методы и алгоритмы, применение технологий распараллеливания обработки;

10. Метрические пространства. Кластерные методы в снижении размерности задачи;
11. Рекомендательные системы. Матричное представление данных. Алгоритмы линейной алгебры и их использование в снижении размерности задачи;
12. Всемирная паутина, методы сбора данных и первичного анализа;
13. Структура WWW и ее использование в задачах ранжирования информации;
14. Интеллектуальный анализ информационных процессов;
15. Методы обнаружения бизнес-процессов. Альфа-алгоритм. Эвристические алгоритмы обнаружения бизнес-процессов;
16. Продвинутые техники баз данных. In-Memory базы данных как технологическая платформа для обработки больших наборов данных;
17. Базы данных NoSQL как набор технологических платформ для обработки больших наборов данных.

Более детально эти темы распределены по двум годовым (четырем полугодовым) спецкурсам объединенного курса “Анализ больших данных” (“The analysis of big data”):

1. Модели данных и базы данных (Data models and databases) – годовой:
   1.1. Модели данных и основы систем баз данных (Data models and fundamentals of database systems) – полугодовой;
   1.2. Базы данных. Дополнительные главы (Databases: additional chapters) – полугодовой.
2. Аналитика больших данных (Big Data Analytics) – годовой:
   2.1. Аналитика больших данных. Основные алгоритмы (Big Data Analytics: basic algorithms) – полугодовой;
   2.2. Аналитика больших данных. Дополнительные главы (Big Data Analytics: additional chapters) – полугодовой.

Отметим, что предлагаемые спецкурсы:

- включают в себя теоретическую и практическую составляющие;
- являются, с одной стороны, взаимозависимыми, а с другой – не требуют обязательного предварительного изучения содержания остальных спецкурсов из предлагаемого набора;
- отражают как уже ставшие классическими модели и алгоритмы, так и современные взгляды и понятия.

Предполагается, что слушатели спецкурсов уже владеют материалом из основных курсов по:

- линейной алгебре и ее приложениям;
- по теории вероятностей и статистике;
- по теории кодирования (такой курс предлагается сделать основным);
- по программированию.

Ниже приведены более детальные программы для каждого из спецкурсов.

Модели данных и основы систем баз данных:

1. Основы баз данных. История развития. Понятие модели данных. СУБД, устройства хранения данных, языки манипулирования данными;
2. СУБД: основные функции, запросы, транзакции. Компоненты СУБД, архитектура современных СУБД;
3. Модели данных. Иерархическая, сетевая, реляционная и слабоструктурированная модели. Основы реляционной модели данных;
4. Реляционная алгебра, отношения, кортежи, основные операции;
5. Реляционная алгебра: представление сложных запросов. Мультимножества;
6. Язык определения данных и манипулирования данными SQL;
7. SQL: основные стандарты и особенности реализации;
8. Функциональные зависимости и их роль в устранении аномалий манипулирования в реляционных базах данных;
9. Исчисление функциональных зависимостей, алгоритмы вычисления замыканий;
10. Декомпозиция схем отношений. Свойства соединения без потерь и сохранения зависимостей;
11. Нормальные формы схем отношений. Основные теоремы о декомпозиции схем;
12. Алгоритмы приведения к 3-й нормальной форме и нормальной форме Бойса-Кодда;
13. Многозначные зависимости. Исчисление многозначных зависимостей. Замыкание множества функциональных и многозначных зависимостей;
14. Декомпозиции схем отношений, 4-я нормальная форма;
15. Проектирование схем баз данных. Концептуальные модели, модель сущностей-

связей: основные понятия и представления;
16. Модель сущностей-связей: правила описания связей и ограничений, принципы проектирования.

Базы данных. Дополнительные главы:

1. Прошлое, настоящее и будущее корпоративных приложений. Новые требования к корпоративным приложениям. Изменения в аппаратном обеспечении. Характеристики современных корпоративных приложений;
2. Методы хранения баз данных. Словарное кодирование;
3. Сжатие: Prefix Encoding / Run-Length Encoding / Cluster Encoding / Indirect Encoding / Delta Encoding;
4. Размещение данных в оперативной памяти. Секционирование;
5. Структуры и операции. Манипулирование данными (insert / update / delete / insert only);
6. Реконструкция кортежей. Поиск данных. Стратегии материализации;
7. Продвинутые техники баз данных. Архитектура базы данных: дифференциальный буфер, операция слияния;
8. Параллельная обработка данных;
9. Материализованные агрегаты, их кеширование. Управление рабочей нагрузкой и планирование;
10. Механизм старения данных. Актуальное и историческое хранение;
11. Внутренние механизмы базы данных. Индексы;
12. Журналирование. Восстановление. Репликация. Резервирование;
13. Введение в СУБД SAP HANA. Структура данных: таблицы, представления и материализованные агрегаты;
14. СУБД HANA: введение в язык манипулирования данными SQLScript;
15. NoSQL. Идея NoSQL. ACID vs BASE. Теорема CAP: 3 класса распределенных систем;
16. NoSQL модели данных: Document, Graph, Key-value store, Google‘s BigTable.

Аналитика больших данных. Основные алгоритмы:

1. Введение в Data Mining;
2. Технология MapReduce распараллеливания вычислений;
3. Алгоритмы, использующие MapReduce;
4. Операции реляционной алгебры, матричные вычисления с использованием MapReduce;
5. Алгоритмы обнаружения схожих элементов. Сходство по Жаккару;
6. Методы сжатия больших файлов. Хеширование, подписи больших файлов;
7. Локально-чувствительное хеширование документов;
8. Техника группировок. Построение семейств функций;
9. Метрики на пространствах данных;
10. Локально-чувствительные семейства функций хеширования;
11. Методы высоких степеней сходства. Индексация;
12. Методы высоких степеней сходства. Использование позиции и длины;
13. Методы кластеризации в обработке больших данных;
14. Кластеризация в различных метрических пространствах;
15. In-Memory базы данных как технологическая платформа для обработки больших данных;
16. Алгоритмы In-Memory.

Аналитика больших данных. Дополнительные главы:

1. Анализ ссылок в Интернет;
2. Вычисление PageRank;
3. Модифицированные алгоритмы вычисления PageRank;
4. Модель корзины покупок. Задача поиска частых наборов элементов. Ассоциативные правила.
5. Представление данных: triangular matrix method, triples method. Алгоритм A-priori;
6. Глубинный анализ процессов. Введение. Основные понятия и цели;
7. Методы и типы глубинного анализа процессов;
8. Моделирование бизнес-процессов. Control-Flow notations: BPMN, UML, Petri nets, Causal nets, Transition systems;
9. BPMN нотация и модель бизнес-процессов: объекты потока управления, соединяющие объекты, роли, артефакты. Модели и экземпляры;
10. Исследование поведения процесса: deadlock, livelock, remaining activity, dead activity.

Soundness criterion;

11. Анализ бизнес-процессов с помощью сетей Петри: boundedness, safeness, deadlock, liveness, WF-net sound;
12. Alpha алгоритм;
13. Alpha алгоритм: ограничения. Alpha Plus алгоритм;
14. Process Mining Tools: ProM, Celonis Process Mining for SAP systems;
15. Базы данных NoSQL как технологическая платформа для обработки больших данных: Google Software Stack, Hadoop 1.0 / 2.0 (YARN) architecture;
16. NoSQL базы данных: XML, JSON Document, Resource Description Framework (triplestores), key-value store, Cassandra Query Language.

Для успешного восприятия материала курсов студентам предлагается не только теоретический материал, но и его практическая поддержка в виде выполнения конкретных проектов с использованием предустановленных программных сред, в частности, SAP SQL Anywhere 16 Developer Edition и SAP HANA express edition (Virtual Machine Method).

В перспективе предполагается развертывание ряда специальных практикумов по отдельным направлениям исследований в науке о данных, а также пополнение методическими материалами разрабатываемого открытого электронного ресурса [4].

Литература

1. Leskovec J., Rajaraman A., Ullman J.D. Mining of Massive Datasets. - 2nd Ed. – Cambridge University Press, 2014. – 511p.
2. Sampathkumar R. What is the difference between big data, analytics, data science, data analysis, data mining, business intelligence, econometrics, statistics, machine learning (artificial intelligence) and mathematical modeling? [Электронный ресурс]. – URL: https://www.quora.com/What-is-the-difference-between-big-data-analytics-data-science-data-analysis-data-mining-business-intelligence-econometrics-statistics-machine-learning-artificial-intelligence-and-mathematical-modelling (дата обращения: 10.10.2016).
3. Winters D. What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, and Big Data? [Электронный ресурс]. – URL: https://www.quora.com/What-is-the-difference-between-Data-Analytics-Data-Analysis-Data-Mining-Data-Science-Machine-Learning-and-Big-Data-1?redirected_qid=1084449 (дата обращения: 10.10.2016).
4. Главацкий С.Т., Бурыкин И.Г. Базы данных как технологическая платформа для разработки учебных материалов в системе дистанционного обучения // Международная научно-практическая конференция "Информационные технологии в образовании XXI века". Сборник научных трудов. – М.: НИЯУ МИФИ, 2015. – С.222-227.

References

1. Leskovec J., Rajaraman A., Ullman J.D. Mining of Massive Datasets. - 2nd Ed. – Cambridge University Press, 2014. – 511p.
2. Sampathkumar R. What is the difference between big data, analytics, data science, data analysis, data mining, business intelligence, econometrics, statistics, machine learning (artificial intelligence) and mathematical modeling? [Электронный ресурс]. – URL: https://www.quora.com/What-is-the-difference-between-big-data-analytics-data-science-data-analysis-data-mining-business-intelligence-econometrics-statistics-machine-learning-artificial-intelligence-and-mathematical-modelling (accessed: 10.10.2016).
3. Winters D. What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, and Big Data? [Электронный ресурс]. – URL: https://www.quora.com/What-is-the-difference-between-Data-Analytics-Data-Analysis-Data-Mining-Data-Science-Machine-Learning-and-Big-Data-1?redirected_qid=1084449 (accessed: 10.10.2016).
4. Glavatsky S.T., Burykin I.G. Bazy dannykh kak tekhnologicheskaya platforma dlya razrabotki uchebnykh materialov v sisteme distantsionnogo obucheniya // Mezhdunarodnaya nauchno-prakticheskaya konferentsiya "Informatsionnye tekhnologii v obrazovanii XXI veka". Sbornik nauchnykh trudov. – M.: NIYaU MIFI, 2015. – S.222-227.

Поступила 11.10.2016

Об авторах:

Главацкий Сергей Тимофеевич, доцент кафедры теоретической информатики механико-математического факультета МГУ имени М.В. Ломоносова, к.ф-м.н., serge@rector.msu.ru;

Бурыкин Илья Геннадиевич, научный сотрудник кафедры теоретической информатики механико-математического факультета МГУ имени М.В. Ломоносова, Ilia.Burykin@sdo.msu.ru.