

ИСТОРИЧЕСКИЙ ФАКУЛЬТЕТ МГУ ИМЕНИ М. В. ЛОМОНОСОВА
АССОЦИАЦИЯ «ИСТОРИЯ И КОМПЬЮТЕР»

ИСТОРИЧЕСКИЕ ИССЛЕДОВАНИЯ В КОНТЕКСТЕ НАУКИ О ДАННЫХ:

**Информационные ресурсы,
аналитические методы
и цифровые технологии**

Материалы международной конференции

Москва, 4–6 декабря 2020 г.



Москва – 2020

УДК 930.2
ББК 63ф1я431
И90

*Конференция поддержана Российским фондом фундаментальных исследований,
грант 20-09-22024\20*

Ответственный редактор: *И. М. Гарскова*

Редколлегия:

чл.-корр РАН, д.и.н. *Л. И. Бородин*, д.и.н. *В. Н. Владимиров*,
к.и.н. *И. М. Гарскова*

Рецензенты:

чл.-корр. РАН *Н. М. Арсентьев*, д.т.н. *Г. З. Залаев*

Редакционный совет:

д.и.н. *С.А. Баканов* (Челябинск), д.и.н. *Ш. Д. Батырбаева* (Бишкек),
д.и.н. *С. И. Корниенко* (Пермь), д.и.н. *Л. Н. Мазур* (Екатеринбург),
д.и.н. *Ю. Ю. Юмашева* (Москва)

**Исторические исследования в контексте науки о данных:
И90 информационные ресурсы, аналитические методы и цифро-
вые технологии.** Материалы международной конференции. Мо-
сква, 4–6 декабря 2020 г. – Москва : МАКС Пресс, 2020. – 484 с.
ISBN 978-5-317-06529-4

<https://doi.org/10.29003/m1781.978-5-317-06529-4>

Сборник содержит тексты докладов Международной конференции.
В рамках указанной международной конференции была проведена XVII
конференция Межрегиональной Ассоциации «История и компьютер».

Ключевые слова: история, историческая информатика, наука о дан-
ных, методы исследования, информационные ресурсы, цифровые тех-
нологии.

УДК 930.2
ББК 63ф1я431

ISBN 978-5-317-06529-4

© Ассоциация «История и компьютер», 2020
© Исторический факультет МГУ имени
М. В. Ломоносова, 2020
© Оформление. ООО «МАКС Пресс», 2020

<https://doi.org/10.29003/m1847.978-5-317-06529-4/447-453>

Оськин А.Ф.

Полоцкий государственный университет

Оськин Д.А.

Белорусский государственный экономический университет, Минск

Об учебной программе курса «Data Mining – Интеллектуальный анализ данных» для магистрантов специальности «История»

Аннотация. В статье рассматривается учебная программа курса «Data Mining – Интеллектуальный анализ данных» для студентов второй ступени высшего образования, обучающихся по специальности «История». Дается определение термину «Data Mining», перечисляются области применения и подчеркивается важность освоения данных технологий магистрантами этой специальности. Рассматривается содержание лекционной компоненты дисциплины, лабораторного практикума, приводится перечень полезных Интернет-ресурсов.

Ключевые слова: Data Mining, интеллектуальный анализ данных, учебная программа, магистратура.

Oskin A.F.

Polotsk State University, Polotsk

Oskin D.A.

Belarusian State Economic University, Minsk

About the curriculum of the course “Data Mining – Data Mining” for undergraduates of the specialty “History”

Abstract. The article discusses the curriculum of the course “Data Mining – Data Mining” for graduate students studying in the specialty “History”. The definition of the term “Data Mining” is given, the areas of application are listed and the importance of mastering these technologies by undergraduates of this specialty is emphasized. The content of the lecture component of the discipline, laboratory workshop is considered, a list of useful Internet resources is provided.

Keywords: Data Mining, Curriculum, Master Study.

«За последние годы, когда, стремясь к повышению эффективности и прибыльности бизнеса, при создании баз данных все стали пользоваться средствами обработки цифровой информации, появился и побочный продукт этой активности – горы собранных данных: И вот все больше распространяется идея о том, что эти горы полны золота». Эти слова принадлежат известному исследователю в области искусственного интеллекта, автору термина «Data Mining», Григорию Пиатецкому-Шапиро (Gregory Piatetsky-Shapiro).

В 1989 году он дал следующее определение новому научному направлению [1]: «Data Mining – это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности». В русском языке эквивалентом английского термина «Data Mining» является словосочетание «Интеллектуальный анализ данных». В дальнейшем в данной статье мы будем использовать термин «Data Mining».

Существует ещё несколько определений данного понятия.

- Data Mining – это процесс выделения из данных неявной и неструктурированной информации и представления ее в виде, пригодном для использования.

- Data Mining – это процесс выделения, исследования и моделирования больших объемов данных для обнаружения неизвестных до этого структур (patterns) с целью достижения преимуществ в бизнесе (определение SAS Institute).

- Data Mining – это процесс, цель которого – обнаружить новые значимые корреляции, образцы и тенденции в результате просеивания большого объема хранимых данных с использованием методик распознавания образов плюс применение статистических и математических методов (определение Gartner Group).

Наиболее удачным, по нашему представлению, является определение, данное автором термина, Пиатецким-Шапиро.

В соответствии с этим определением суть и цель технологий Data Mining можно охарактеризовать так: это технологии, которые предназначены для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей.

- Неочевидных – значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем.

- Объективных – значит, что обнаруженные закономерности будут полностью соответствовать действительности, в отличие от экспертного мнения, которое всегда является субъективным.

- Практически полезных – значит, что выводы имеют конкретное значение, которому можно найти практическое применение.

Технологии анализа данных, основанные на применении классических статистических подходов, имеют ряд недостатков. Эти методы основаны на использовании усредненных показателей, анализируя которые трудно выяснить истинное положение дел в исследуемой области (классический пример – средняя температура по больнице).

Прогнозный потенциал методов математической статистики также невысок – исследования специалистов Гарвардского института показали, что на основе имеющейся информации с помощью стандартных статистических методов нельзя было предсказать великую депрессию конца 1920-х годов. Основная область применения этих методов – технологии OLAP (Online Analytical Processing), которые используются для проверки заранее сформулированных гипотез и «грубого» разведочного анализа.

Классические статистические методы имеют ещё один важный недостаток – они пренебрегают аномальными наблюдениями, так называемыми пиками и всплесками. При этом известно, что отдельные нетипичные значения могут составлять самостоятельный интерес для исследования, характеризуя некоторые исключительные, но важные явления. Сама идентификация этих наблюдений, не говоря уже об их последующий анализ, может быть полезной для понимания сущности изучаемых объектов или явлений. Как показывают современные исследования, именно такие события могут стать решающими относительно будущего поведения и развития сложных систем.

Перечисленные недостатки статистических методов послужили толчком к развитию новых методов исследования сложных систем, сочетающих в себе последние достижения искусственного интеллекта, разнообразных математических методов, статистики и эвристических подходов. Эти методы и легли в основу новых технологий анализа данных, получивших название Data Mining.

Нам представляется, что современный уровень развития исторической науки требует от будущих специалистов владения этим важным инструментом научных исследований.

Перейдём к рассмотрению содержания учебной программы дисциплины «Data Mining – Интеллектуальный анализ данных». В соответствии с учебным планом, по данной дисциплине предусмотрено 72 часа аудиторных занятий – 36 часов лекционных и столько же лабораторных занятий.

На рис. 1 приведена интеллект-карта учебной программы с указанием тем лекционных и лабораторных занятий и их объёмом в академических часах.

Как видно из рисунка, курс «Data Mining – Интеллектуальный анализ данных» состоит из 5 модулей.

В первом модуле – «Введение. Что такое Data Mining», дается определение понятию «Data Mining», рассматриваются научные дисциплины, на базе которых возникла эта междисциплинарная область,

проводится сравнение статистики, машинного обучения и Data Mining. На основе проведённого сравнения выявляются отличия Data Mining от других методов анализа данных и обозначаются наиболее перспективные области применения этой технологии.

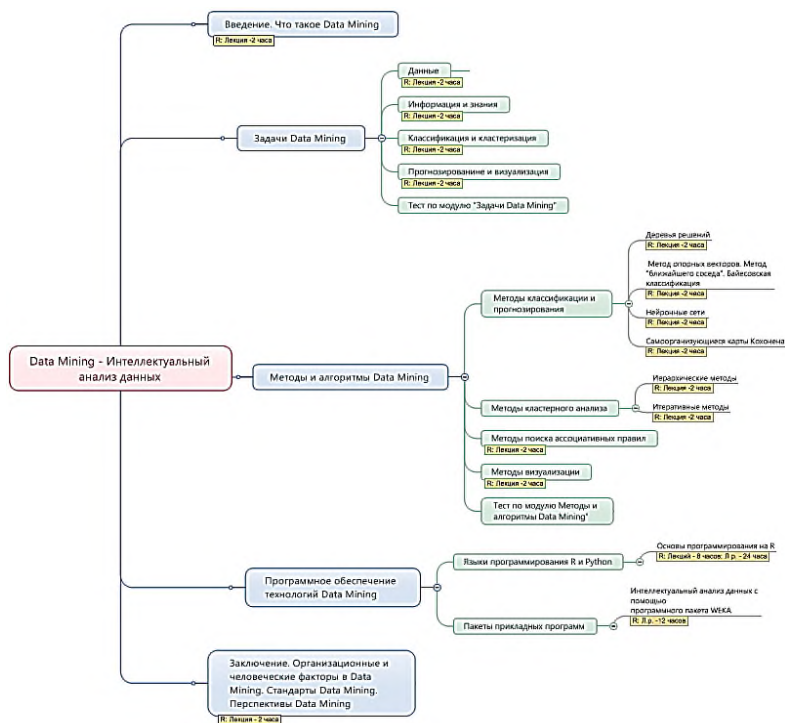


Рис. 1. Интеллект-карта учебной программы дисциплины

В модуле «Задачи Data Mining» описывается суть задач, решаемых с применением технологий Data Mining и приводится классификация этих задач. Подробно рассматриваются такие понятия как «Данные», «Информация», «Знания» и их взаимные связи и отношения.

Модуль «Методы и алгоритмы Data Mining» является главным модулем в теоретической части курса. Он состоит из восьми лекций, описывающих все основные методы Data Mining, а также алгоритмы их реализации. Методы объединены в четыре группы. В группе «Методы классификации и прогнозирования» рассматриваются де-

ревя решений, байесовская классификация, методы опорных векторов и «ближайшего соседа», а также нейронные сети и карты Кохонена. Группа «Методы кластерного анализа» посвящена описанию двух основных подходов к решению задачи кластеризации – иерархическому и итеративному. В этом же модуле кратко рассматриваются методы поиска ассоциативных правил и методы визуализации.

Следующий модуль описывает программное обеспечение, используемое для реализации алгоритмов Data Mining. Рассматриваются языки программирования, предназначенные для решения задач интеллектуального анализа данных, а также специализированные пакеты прикладных программ, используемые для этих целей. Освоение знаний, умений и навыков из этого раздела предполагается путём прохождения лабораторного практикума, общим объёмом 36 академических часов.

И, наконец, в последнем модуле «Заключение. Организационные и человеческие факторы в Data Mining. Стандарты Data Mining. Перспективы Data Mining» подводятся итоги, рассматриваются стандарты и обозначаются перспективы технологий Data Mining.

Для организации практикума по данному курсу необходимо загрузить с соответствующих ресурсов и установить на компьютерах лабораторного класса следующий набор инструментов:

- R – свободно распространяемое программное обеспечение для статистических расчетов и визуализации [2].
- RStudio Desktop – свободно распространяемая среда разработки приложений для языка программирования R [3]
- WEKA – свободно распространяемое программное обеспечение для решения задач Data Mining [4]

Для выполнения лабораторных работ необходимы заранее подготовленные наборы данных (учебные датасеты). На сайте [5] доступны для бесплатной загрузки и использования в исследовательских или учебных целях более пятидесяти тысяч различных наборов данных. Мы использовали датасет «All Space Missions from 1957». Он состоит из 4322 наблюдений и содержит сведения о стране, выполнившей запуск космической ракеты, месте и дате запуска, организации, реализующей проект, стоимости миссии, названии ракеты и её статусе (действующая или списанная), успешности миссии.

Используя данный датасет можно познакомиться со следующими видами анализа данных: разведочный анализ данных, визуализация, кластеризация, прогнозирование.

Рассмотрим теперь подходы к организации учебного процесса по дисциплине.

В последние годы при проведении занятий с магистрантами мы активно используем технологии смешанного обучения [6].

По определению, под смешанным обучением понимается совместное использование классических методик обучения и современных дистанционных технологий.

Главным достоинством классической методики является прямой, непосредственный контакт обучаемого и обучающегося, оказывающий большое эмоциональное воздействие на обучаемого и способствующий лучшему усвоению учебного материала. Применение технологий дистанционного обучения позволяет задать для каждого обучающегося собственную образовательную траекторию, индивидуализировать процесс обучения, сделать его асинхронным. Каждый обучающийся может осваивать учебный материал в любое удобное для него время, в привычной обстановке, с оптимальной скоростью. Сочетание этих двух подходов и образует систему технологий, имеющих общее название «Смешанное обучение» (Blended Learning).

Как показал зарубежный опыт, из всех многочисленных моделей смешанного обучения, наибольшее распространение получила модель, называемая «Перевернутый класс» (Flipped Classroom). Термин «Перевернутый класс» можно отнести к широкому спектру технологий смешанного обучения, при использовании которых студенты, до очной встречи с преподавателем, получают удаленный доступ к образовательному контенту по теме, а во время очного занятия закрепляют изученный контент и выполняют практические задания по теме.

При этом наибольшее распространение получил подход, при котором студенты просматривают дома серию коротких видеолекций по теме предстоящего занятия, а в аудитории совершенствуют свои знания, выполняя практические задания.

Технология «Перевернутый класс» органично вписывается в любую систему информационной поддержки обучения. При этом, как уже отмечалось выше, появляется возможность индивидуализации обучения, реализуется подход, при котором каждый обучающийся усваивает новые знания по изучаемой дисциплине, двигаясь по своей образовательной траектории.

Для организации дистанционного обучения в Полоцком государственном университете используется платформа G Suite for Education, разработанная корпорацией Google LLC. Платформа находится в открытом доступе и может быть использована любым государственным учебным заведением после регистрации. Главным сервисом,

обеспечивающим эффективную организацию учебного процесса на платформе G Suite for Education, является приложение Google Classroom. Это приложение позволяет создавать учебные курсы, назначать и проверять задания, обмениваться со студентами файлами, проводить онлайн-тестирование знаний, обсуждать в режиме живого общения отдельные темы учебного курса. Летом текущего года, внося свой вклад в борьбу с пандемией коронавируса, корпорация Google сделала бесплатным ещё один свой веб-сервис, Google Meet, интегрировав его с Google Classroom. С помощью Google Meet можно организовывать вебинары, онлайн-встречи, совещания, конференции. Таким образом, в настоящее время Google Classroom имеет функционал, позволяющий удовлетворить все потребности пользователей, использующих платформу G Suite for Education для организации дистанционного обучения.

С учетом всего сказанного выше, занятия по дисциплине «Data Mining – Интеллектуальный анализ данных» в текущем семестре будут организованы следующим образом. Перед началом семестра во вновь созданный Google-класс дисциплины будет выложен электронный учебно-методический комплекс. В течение семестра, во время очных занятий, магистрантам будут предлагаться наборы учебных задач, соответствующие различным уровням усвоения дисциплины. Результаты этих тестов для каждого студента будут сохраняться и анализироваться. Это позволит выделить, по крайней мере три группы, состоящие из магистрантов с примерно одинаковыми способностями. Дальнейшая работа будет организовываться с учетом специфики каждой группы, что должно повысить эффективность и, в конечном счете, качество обучения.

Библиография

1. KDNuggets®, сайт Г. Пиатецкого-Шапиро – URL: <https://www.kdnuggets.com/> (дата обращения: 29.08.2020).
2. The R Project for Statistical Computing – URL: <https://www.r-project.org/> (дата обращения: 29.08.2020).
3. RStudio Desktop – URL: <https://rstudio.com/products/rstudio/download/#download> (дата обращения: 29.08.2020).
4. WEKA. Machine learning software to solve data mining problems – URL: <https://sourceforge.net/projects/weka/> (дата обращения: 29.08.2020).
5. Kaggle: Your Machine Learning and Data Science Community – URL: <https://www.kaggle.com/> (дата обращения: 29.08.2020).
6. Оськин А.Ф., Оськин Д.А. Применение технологий смешанного обучения в учебном процессе высшей школы: Сборник материалов IV Международной научно-практической конференции «Педагогический опыт: теория, методика, практика». Том 1. – Чебоксары, 2015. С. 207–209.