МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Магнитогорский государственный технический университет им. Г.И. Носова»

УТВЕРЖДАЮ
Директор ИЭиАС
_____ С.И. Лукьянов

26.02.2020 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

ТЕХНОЛОГИИ ОБРАБОТКИ ПОТОКОВЫХ BIG DATA

Направление подготовки (специальность)
09.04.01 Информатика и вычислительная техника

Направленность (профиль/специализация) программы
Программное обеспечение средств вычислительной техники и автоматизированных систем

Уровень высшего образования - магистратура

Форма обучения
заочная

Институт/ факультет	Институт энергетики и автоматизированных систем
Кафедра	Вычислительной техники и программирования
Курс	2
Семестр	

Магнитогорск

2019

Рабочая программа составлена на основе ФГОС ВО по направлению подготовки 09.04.01 Информатика и вычислительная техника (уровень магистратуры) (приказ Минобрнауки России от 19.09.2017 г. № 918)

Рабочая программа рассмотрена и одобрена на заседании кафедры Вычислительной техники и программирования
19.02.2020 г. протокол № 5

Зав. кафедрой _____ О.С. Логунова

Рабочая программа одобрена методической комиссией ИЭиАС
26.02.2020 г. протокол № 5

Председатель _____ С.И. Лукьянов

Рабочая программа составлена:
доцент кафедры ВТиП, канд. техн. наук _____ Л.Г. Егорова

Рецензент:
Начальник отдела инновационных разработок
ЗАО «КонсОМ-СКС» , канд. техн. наук А.Н. Панов

Лист актуализации рабочей программы

Рабочая программа пересмотрена, обсуждена и одобрена для реализации в 2020 - 2021 учебном году на заседании кафедры Вычислительной техники и программирования

Протокол от 07 октября 2020 г. № 2
Зав. кафедрой _____ О.С. Логунова

Рабочая программа пересмотрена, обсуждена и одобрена для реализации в 2021 - 2022 учебном году на заседании кафедры Вычислительной техники и программирования

Протокол от __ _____ 20__ г. № __
Зав. кафедрой _____ О.С. Логунова

Рабочая программа пересмотрена, обсуждена и одобрена для реализации в 2022 - 2023 учебном году на заседании кафедры Вычислительной техники и программирования

Протокол от __ _____ 20__ г. № __
Зав. кафедрой _____ О.С. Логунова

Рабочая программа пересмотрена, обсуждена и одобрена для реализации в 2023 - 2024 учебном году на заседании кафедры Вычислительной техники и программирования

Протокол от __ _____ 20__ г. № __
Зав. кафедрой _____ О.С. Логунова

1 Цели освоения дисциплины (модуля)

Дисциплина Технологии обработки потоковых Big Data предоставляет практические знания о больших данных, аналитике данных и инструментах по работе с большими данными. Цель дисциплины состоит в обучении эффективному использованию основных методов аналитики больших данных. В результате обучения формируется умение использовать современные технологии и инструментальные средства по работе с большими данными (Hadoop, MapReduce, Spark, NoSQL, язык R и др.)

2 Место дисциплины (модуля) в структуре образовательной программы

Дисциплина Технологии обработки потоковых Big Data входит в часть учебного плана формируемую участниками образоватеьлных отношений образовательной программы.

Для изучения дисциплины необходимы знания (умения, владения), сформированные в результате изучения дисциплин/ практик:

Основы теории машинного обучения

Промышленные информационные системы

Интеллектуальные системы

Методы и средства высокопроизводительного программирования

Знания (умения, владения), полученные при изучении данной дисциплины будут необходимы для изучения дисциплин/практик:

Эволюционные вычисления

Проблемы принятия решений в условиях нечеткой информации

Методы оптимизации

Выполнение и защита выпускной квалификационной работы

3 Компетенции обучающегося, формируемые в результате освоения дисциплины (модуля) и планируемые результаты обучения

В результате освоения дисциплины (модуля) «Технологии обработки потоковых Big Data» обучающийся должен обладать следующими компетенциями:

Структурный элемент компетенции	Планируемые результаты обучения
ПК-4 Обладает способностью к разработке компонентов системы управления базами данных, отладке разрабатываемой системы управления базами данных, документированию разработанной системы управления базами данных в целом и ее компонентов и сопровождению созданной системы управления базами данных	
ПК-4.1 Определяет необходимость разработки компонентов системы управления базами	
ПК-4.2 Оценивает качество разработки компонентов системы управления базами данных	
ПК-8 Обладает способностью к анализу системных проблем обработки информации на уровне инфокоммуникационной системы, подготовке предложений по развитию инфокоммуникационной системы, разработке нормативной и технической документации на аппаратные средства и программное обеспечение	
ПК-8.1 Определяет полноту результатов анализа системных проблем обработки	
ПК-8.2 Оценивает новизну предложений по развитию инфокоммуникационной системы	
ПК-8.3 Оценивает необходимость в разработке нормативной и технической документации	

4. Структура, объём и содержание дисциплины (модуля)

Общая трудоемкость дисциплины составляет 4 зачетных единиц 144 акад. часов, в том числе:

– контактная работа – 8,1 акад. часов:

– аудиторная – 8 акад. часов;

– внеаудиторная – 0,1 акад. часов

– самостоятельная работа – 132 акад. часов;

Форма аттестации - зачет с оценкой

Раздел/ тема дисциплины	Семестр	Аудиторная контактная работа (в акад. часах)			Самостоятельная работа студента	Вид самостоятельной работы	Форма текущего контроля успеваемости и промежуточной аттестации	Код компетенции
		Лек.	лаб. зан.	практ. зан.				
1. Введение в большие данные. Методы многомерного статистического анализа и анализа нечисловой информации								
1.1 Определение больших данных и причины их появления. Примеры возможностей для бизнеса. Различие между Business Intelligence и Big Data. Факторный анализ. Дискриминантный анализ. Кластерный анализ. Многомерное шкалирование. Методы контроля качества.	2		2/2И		22	1. Подготовка к лабораторным занятиям 2. Выполнение лабораторных работ 3. Самостоятельное изучение учебной и научной литературы	1. Беседа - обсуждение 2. Проверка индивидуальных заданий 3. Устный опрос.	
Итого по разделу			2/2И		22			
2. Технологии хранения и обработки больших данных								
2.1 Высокопроизводительные вычисления: Распределенные вычисления на нескольких серверах, вычислительная парадигма MapReduce. Проект Apache Hadoop и его экосистема. Apache Spark и его компоненты. Вычисления в реальном времени, Apache Storm, Flink.	2		2/2И		28	1. Подготовка к лабораторным занятиям 2. Выполнение лабораторных работ 3. Самостоятельное изучение учебной и научной литературы	1. Беседа - обсуждение 2. Проверка индивидуальных заданий 3. Устный опрос.	

2.2 Масштабирование и многоуровневое хранение данных: Теорема CAP. Парадигма NoSQL. Классификация NoSQL баз данных			2/1И		28	1. Подготовка к лабораторным занятиям 2. Выполнение лабораторных работ 3. Самостоятельное изучение учебной и научной литературы	1. Беседа - обсуждение 2. Проверка индивидуальных заданий 3. Устный опрос.	
2.3 Визуализация данных и результатов анализа: Техники визуализации данных, введение в язык R. Визуализация данных в R			2/1И		28	1. Подготовка к лабораторным занятиям 2. Выполнение лабораторных работ 3. Самостоятельное изучение учебной и научной литературы	1. Беседа - обсуждение 2. Проверка индивидуальных заданий 3. Устный опрос.	
Итого по разделу			6/4И		84			
3. Аналитика в больших данных								
3.1 Жизненный цикл аналитики данных. Роли, необходимые для успешного создания проекта по аналитике данных. Сложные методы аналитики: Классификация задач анализа: Text, Data, Web, Social Mining. Применение машинного обучения в аналитике. K-means и C-means кластеризация, классификация. Логистическая регрессия, ассоциации, алгоритм Априори.	2				26	1. Подготовка к лабораторным занятиям 2. Выполнение лабораторных работ 3. Самостоятельное изучение учебной и научной литературы	1. Беседа - обсуждение 2. Проверка индивидуальных заданий 3. Устный опрос.	
Итого по разделу					26			
Итого за семестр			8/6И		132		зао	
Итого по дисциплине			8/6И		132		зачет с оценкой	

5 Образовательные технологии

1. Традиционные образовательные технологии, ориентированные на организацию образовательного процесса и предполагающие прямую трансляцию знаний от преподавателя к студенту.

Формы учебных занятий с использованием традиционных технологий:

Лабораторная работа – организация учебной работы с реальными материальными и информационными объектами, экспериментальная работа с аналоговыми моделями реальных объектов.

2. Технологии проблемного обучения – организация образовательного процесса, которая предполагает постановку проблемных вопросов, создание учебных проблемных ситуаций для стимулирования активной познавательной деятельности студентов.

Формы учебных занятий с использованием технологий проблемного обучения:

Практическое занятие в форме практикума – организация учебной работы, направленная на решение комплексной учебно-познавательной задачи, требующей от студента применения как научно-теоретических знаний, так и практических навыков.

6 Учебно-методическое обеспечение самостоятельной работы обучающихся

Представлено в приложении 1.

7 Оценочные средства для проведения промежуточной аттестации

Представлены в приложении 2.

8 Учебно-методическое и информационное обеспечение дисциплины (модуля)

а) Основная литература:

1. Форман, Д. Много цифр. Анализ больших данных при помощи Excel / Форман Д.; Пер. с англ. Соколовой А. - Москва :Альпина Пабл., 2016. - 461 с. ISBN 978-5-9614-5032-3. - Текст : электронный. - URL: https://znanium.com/catalog/product/551044 (дата обращения: 28.10.2020). – Режим доступа: по подписке

б) Дополнительная литература:

1. Блануца, В. И. Социально-экономическое районирование в эпоху больших данных: Монография / Блануца В.И. - Москва :НИЦ ИНФРА-М, 2019. - 194 с. (Научная мысль) ISBN 978-5-16-013259-4. - Текст : электронный. - URL: https://znanium.com/catalog/product/1014727 (дата обращения: 28.10.2020). – Режим доступа: по подписке.

в) Методические указания:

Представлены в приложении 1

г) Программное обеспечение и Интернет-ресурсы:

Программное обеспечение

Наименование ПО	№ договора	Срок действия лицензии

MS Office 2007 Professional	№ 135 от 17.09.2007	бессрочно
STATISTICA в.6	К-139-08 от 22.12.2008	бессрочно
Anaconda Python	свободно распространяемое ПО	бессрочно
Oracle SQL Developer	свободно распространяемое ПО	бессрочно
Oracle SQL Developer Data Modeler	свободно распространяемое ПО	бессрочно
MS Windows 7 Professional(для классов)	Д-1227-18 от 08.10.2018	11.10.2021

Профессиональные базы данных и информационные справочные системы

Название курса	Ссылка
Информационная система - Единое окно доступа к информационным ресурсам	URL: http://window.edu.ru/
Национальная информационно-аналитическая система – Российский индекс научного цитирования (РИНЦ)	URL: https://elibrary.ru/project_risc.asp

9 Материально-техническое обеспечение дисциплины (модуля)

Материально-техническое обеспечение дисциплины включает:

1. Лекционная аудитория ауд. 282. Мультимедийные средства хранения, передачи и представления информации.

2. Компьютерные классы Центра информационных технологий ФГБОУ ВО «МГТУ». Персональные компьютеры, объединенные в локальные сети с выходом в Internet, оснащенные современными программно-методическими комплексами для решения задач в области информатики и вычислительной техники.

3. Аудитории для самостоятельной работы: компьютерные классы; читальные залы библиотеки. Все классы УИТ и АСУ с персональными компьютерами, выходом в Интернет и с доступом в электронную информационно-образовательную среду университета.

4. Аудиторий для групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации. Ауд. 282 и классы УИТ и АСУ.

5. Помещения для самостоятельной работы обучающихся, оснащенных компьютерной техникой с возможностью подключения к сети «Интернет» и наличием доступа в электронную информационно-образовательную среду организации. Классы УИТ и АСУ.

6. Помещения для хранения и профилактического обслуживания учебного оборудования. Центр информационных технологий – ауд. 372.

Приложение 1

Оценочные средства для проведения промежуточной аттестации

а) Планируемые результаты обучения и оценочные средства для проведения промежуточной аттестации:

Код индикатора	Индикатор достижения компетенции	Оценочные средства
ПК-4: Обладает способностью к разработке компонентов системы управления базами данных, отладке разрабатываемой системы управления базами данных, документированию разработанной системы управления базами данных в целом и ее компонентов и сопровождению созданной системы управления базами данных		
ПК-4.1	Определяет необходимость разработки компонентов системы управления базами данных	*Перечень теоретических вопросов* 1. Основные направления развития методов обработки и хранения данных. 2. Volume. 3. Закон Мура. 4. Velocity. Variety. 5. Фреймворк Hadoop. 6. Проблема хранения неструктурированных данных. 7. Проблема преобразования данных. 8. Семантические анализаторы. 9. Самообучающиеся автоматы. 10. Oracle Big Data Preparation. 11. Аналитика Big Data — реалии и перспективы в России и мире. 12. Data Mining. 13. Краудсорсинг. 14.Смешение и интеграция данных. 15. Базы данных для Big Data. 16. Машинное обучение.
ПК-4.1	ПК-4.2: Оценивает качество разработки	*Практические задания*

Код индикатора	Индикатор достижения компетенции	Оценочные средства
	компонентов системы управления базами данных	Hadoop имеет высокий уровень использования в IT-компаниях, данная технология начинает все шире внедряться в различных секторах ранка, включая производство, правительственные учреждения, учреждения здравоохранения. Необходимо выполнить следующие задачи: – выявить преимущества и недостатки Hadoop; – проанализировать структуру и основные этапы внедрения Hadoop; – исследовать основные недостатки Hadoop. *Задания на решение задач из профессиональной области, комплексные задания* Провести анализ кластера MongoDB : – проанализировать возможности, предоставляемые MongoDB; – проанализировать этапы развертывания кластера MongoDB; – проанализировать структуру кластера MongoDB. Провести анализ MongoDB с точки зрения замены традиционных хранилищ данных.
ПК-8: Обладает способностью к анализу системных проблем обработки информации на уровне инфокоммуникационной системы, подготовке предложений по развитию инфокоммуникационной системы, разработке нормативной и технической документации на аппаратные средства и программное обеспечение		
ПК-8.1	Определяет полноту результатов анализа системных проблем обработки информации на уровне инфокоммуникационной системы	*Перечень теоретических вопросов* 1. Искусственные нейронные сети. 2. Распознавание образов. 3. Прогнозная аналитика. 4. Имитационное моделирование. 5. Пространственный анализ. 6. Статистический анализ. 7. Визуализация аналитических данных. 8. Языки для Big Data.

Код индикатора	Индикатор достижения компетенции	Оценочные средства
		9. Фреймворки для Big Data 10. Big data: применение и возможности. 11. Решения на основе Big data. 12. Рынок Big data в России.
ПК-8.2	Оценивает новизну предложений по развитию инфокоммуникационной системы	*Практические задания* 1. Ознакомьтесь с доступными способами обработки данных. Для предложенных преподавателем данных выполните консолидацию, трансформацию, визуализацию данных. 2. Выполните построение ассоциативных правил для предложенных преподавателем данных, используя различные параметры построения ассоциативных правил. Сравните полученные результаты. Опишите 4-5 ассоциативных правил, полученных в ходе выполнения работы. 3. Используя механизм кластеризации реализованный на алгоритме k-means, основываясь на данных предложенных преподавателем, решите задачу распределения данных на кластеры и выявления скрытых закономерностей. Проанализируйте получившуюся картину распределения. 4. Постройте прогноз для предложенных преподавателем данных с помощью линейной регрессии. Проанализируйте построенную с помощью линейной регрессии модель прогноза. 5. Постройте карты Кохонена для предложенных преподавателем данных. Проанализируйте результаты. Используя различные отображения карты Кохонена, постройте 3-4 правила. 6. Постройте дерево решения для предложенных преподавателем данных. Попробуйте использовать различные значения параметров обучения дерева решения и сравните полученные деревья. Выведите 5 правил из построенного дерева решений. Приведите 4-5 примеров, для которых можно использовать метод обработки дерево решений.

Код индикатора	Индикатор достижения компетенции	Оценочные средства
ПК-8.3	Оценивает необходимость в разработке нормативной и технической документации на аппаратные средства и программное обеспечение	*Задания на решение задач из профессиональной области, комплексные задания* Анализ применения 1. Big data в банках. 2. Big data в бизнесе. 3. Big data в маркетинге. 4. Big data в промышленности

б) Порядок проведения промежуточной аттестации, показатели и критерии оценивания:

Промежуточная аттестация по дисциплине включает теоретические вопросы, позволяющие оценить уровень усвоения обучающимися знаний, и практические задания, выявляющие степень сформированности умений и владений, проводится в форме зачета с оценкой.

Показатели и критерии оценивания:

– на оценку **«отлично»** (5 баллов) – обучающийся демонстрирует высокий уровень сформированности компетенций, всестороннее, систематическое и глубокое знание учебного материала, свободно выполняет практические задания, свободно оперирует знаниями, умениями, применяет их в ситуациях повышенной сложности.

– на оценку **«хорошо»** (4 балла) – обучающийся демонстрирует средний уровень сформированности компетенций: основные знания, умения освоены, но допускаются незначительные ошибки, неточности, затруднения при аналитических операциях, переносе знаний и умений на новые, нестандартные ситуации.

– на оценку **«удовлетворительно»** (3 балла) – обучающийся демонстрирует пороговый уровень сформированности компетенций: в ходе контрольных мероприятий допускаются ошибки, проявляется отсутствие отдельных знаний, умений, навыков, обучающийся испытывает значительные затруднения при оперировании знаниями и умениями при их переносе на новые ситуации.

– на оценку **«неудовлетворительно»** (2 балла) – обучающийся демонстрирует знания не более 20% теоретического материала, допускает существенные ошибки, не может показать интеллектуальные навыки решения простых задач.

– на оценку **«неудовлетворительно»** (1 балл) – обучающийся не может показать знания на уровне воспроизведения и объяснения информации, не может показать интеллектуальные навыки решения простых задач.

Приложение 2

Учебно-методическое обеспечение самостоятельной работы обучающихся

Лабораторная работа

ВЫБОР ПРЕДМЕТНОЙ ОБЛАСТИ

Поставленная перед обучающимися задача не привязана к какой-либо конкретной предметной области. Предполагается отойти от принципа выполнения заранее поставленных и четко сформулированных задач, чтобы предоставить исполнителю гибкость и возможность творческого подхода выполнения. Таким образом, обучающемуся предоставляется возможность самостоятельного выбора интересующей его прикладной области, над которой в рамках курса будет проводиться работа. Если же обучающейся не имеет своих собственных предпочтений, то ему предлагаются на выбор предметные области, перечисленные ниже:

— «Анализ данных социальных сетей». Например, электронные ресурсы Vkontakte, Twitter, Facebook, LinkedIn и др.;

— «Анализ рынка вакансий». Например, электронный ресурс HeadHunter;

— «Анализ фильмов». Например: интернет-проект «Кинопоиск»;

— «Анализ журнала запросов к сайту Wikipedia»;

— «Технический радар». Анализ информации с ресурса StackOverFlow;

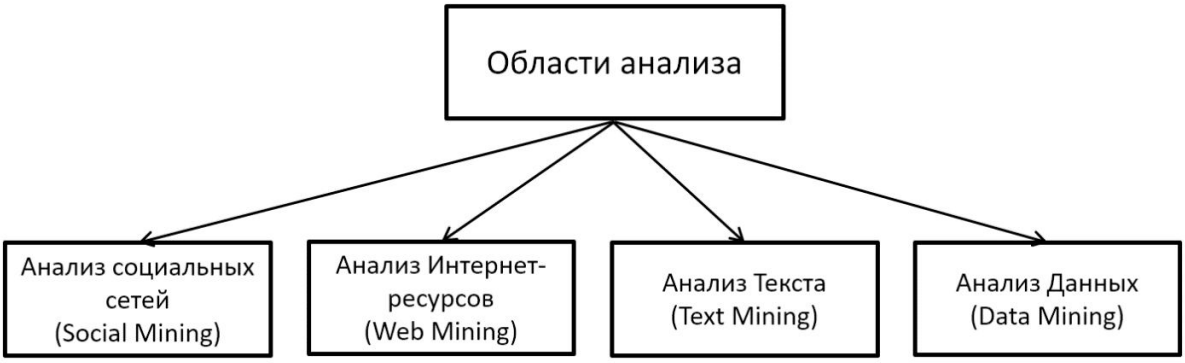
— «Использование существующих решений и наборов данных». Например, информация с ресурса Kaggle (см. условия выставления итоговой оценки). Например, «задача Титаника».

Приветствуются темы из следующих областей: «Образование», «Наука», «Здравоохранение», «Информационные технологии» (ИТ) и др.

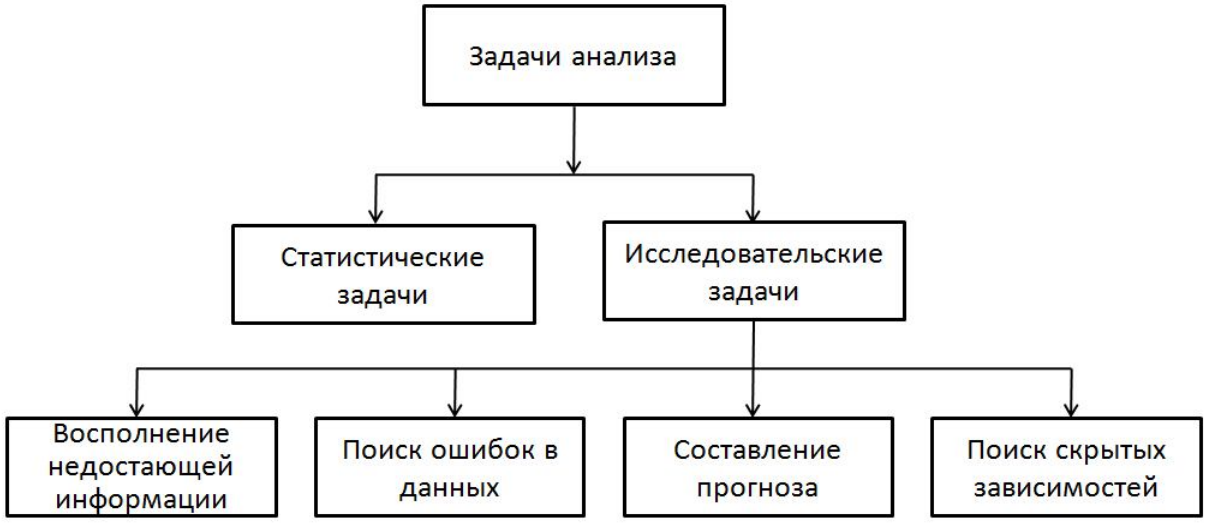
Для выбранной предметной области требуется сформулировать от 5 до 20 задач для проведения анализа. Задачи могут быть отнесены к следующим областям анализа: анализ социальных сетей (Social Mining), анализ Интернет-ресурсов (Web Mining), анализ текста (Text Mining), анализ данных (Data Mining).

1. Vkontakte. [Электронный ресурс]. Режим доступа: // http://www.vk.com.
2. Twitter. [Электронный ресурс]. Режим доступа: // http://www.twitter.com.
3. Facebook. [Электронный ресурс]. Режим доступа: // http://www.facebook.com.
4. LinkedIn. [Электронный ресурс]. Режим доступа: // http://www.linkedin.com.
5. HeadHunter — *качественная база резюме* и *вакансий* и *современные сервисы* для поиска работы и персонала. [Электронный ресурс]. Режим доступа: // http://www.hh.ru.
6. Кинопоиск — русскоязычный интернет-проект, посвящённый кинематографу, [Электронный ресурс]. Режим доступа: // http://www.kinopoisk.ru.
7. Wikipedia — свободная общедоступная мультиязычная универсальная интернет-энциклопедия, [Электронный ресурс]. Режим доступа: // http://www.wikipedia.org.
8. StackOverFlow — популярная система вопросов и ответов о программировании, [Электронный ресурс]. Режим доступа: // http://www.stackoverflow.com.
9. Kaggle - англоязычный ресурс, посвященный задачам анализа и науке о данных, [Электронный ресурс]. Режим доступа: // http://www.kagle.com.
10. «Титаник» (англ. *Titanic*) — британский трансатлантический пароход. «Задача Титаника» - создание модели для предсказания выживших пассажиров парохода в зависимости от характеристик пассажира: его пол, возраст, номер каюты и т. д.

Классификация задач анализа по областям приведена на рис.1.



В тоже время задачи анализа можно классифицировать по типу: задачи статистического типа и задачи исследовательского типа. Классификация приведена на рис.2.



Статистические задачи относятся к традиционной обработке известного набора данных, объектов и их атрибутов для получения численных характеристик. Традиционно принято считать, что статистические задачи относятся к категории бизнес-аналитики (Business Intelligence). Они призваны помочь ответить на вопросы: «Какие численные показатели получила отрасль за прошлое время?», «Как правильно настроить рабочие процессы на основе прошлых, исторических данных?». Иными словами, результаты решения статистических задач помогают понять, что же произошло в прошлом и как на основе этих данных оптимизировать бизнес или производственные процессы и получить выгоду, зачастую экономическую. Особенностью реализации этого типа задачи являются: большое количество записей, большой объем информации и реализация алгоритмов обработки средствами и фреймворками для высокопроизводительных и распределенных вычислений.

Исследовательские задачи (Data Science), в отличие от статистических, подразумевают поиск скрытых зависимостей и паттернов в данных, восстановление недостающей информации, поиск ошибок в данных, а также составление некоторых прогнозов на будущее. Особенностью этого типа задач является использование инновационных, современных и прогрессивных методов анализа, которые в том числе позволяют построить своего рода экспертную систему.

При формулировании задач анализа необходимо, чтобы были представлены на утверждение задачи из каждой категории. Проработка каждой задачи анализа требует

проявления фантазии и собственной заинтересованности в получении ответа на поставленный вопрос, потому что именно личностная заинтересованность может привести к высокому качеству выполнения проекта.

Стоит принять во внимание, что данные, подвергаемые анализу, могут обладать рядом неприятных свойств: неполнота, противоречивость, некорректность и разнородность. Если не учитывать возможность наличия таких свойств в данных, то результаты решения задач анализа могут находиться в другой плоскости относительно истинного решения. Для того, чтобы результаты решения задач были корректными, необходимо осуществлять валидацию и верификацию подвергаемой анализу информации. Зачастую применяют следующие подходы для проверки данных на корректность: методы машинного обучения, поиск нечетких связей и соответствий, и выявление обратной связи между атрибутами объектов, результатами решения задачи и входных данных.

Если рассматривать предметную область «Вакансии» с web-ресурса «HeadHunter», то в роли задач анализа могут выступать следующие приведенные статистические и исследовательские задачи.

Статистические задачи:

— анализ наиболее востребованных на рынке информационных технологий языков программирования в заданные интервалы времени (начиная с 2002 по 2016 гг.);

— определение распределения вакансий в области информационных технологий по регионам в зависимости от года;

— поиск наиболее популярных профессий в Российской Федерации;

— нахождение зависимости зарплаты от специализации;

Исследовательские задачи:

— поиск скрытых зависимостей между характеристиками работодателя и представленных вакансий;

— прогнозирование заработной платы в области IT на 2030 год.

Для предметной области «Социальные сети» в роли статистических задач анализа могут выступать:

— определение перечня городов, из которых в вузы Москвы приезжают для поступления абитуриенты, в том числе и зарубежные;

— нахождения перечня стран и городов, в которых работают выпускники вузов Москвы;

— установление параметров корреляции популярных тем обсуждений в социальных сетях с событиями в новостях.

Исследовательскими задачами для социальных сетей могут быть:

— прогнозирование количества приезжих абитуриентов в вузы Москвы;

— поиск скрытых зависимостей между родным городом абитуриента и Москвы.

Лабораторная работа

ФОРМИРОВАНИЕ НАБОРА ДАННЫХ

Во время выполнения проекта может потребоваться работать с информацией разного типа. Традиционно принято выделять четыре типа данных: структурированные данные, полуструктурированные данные, квазиструктурированные данные и не структурированные данные. На рис. 1 приведена классификация информации по типам. При этом стоит отметить, что объем занимаемых данных растет с уменьшением их структурированности.

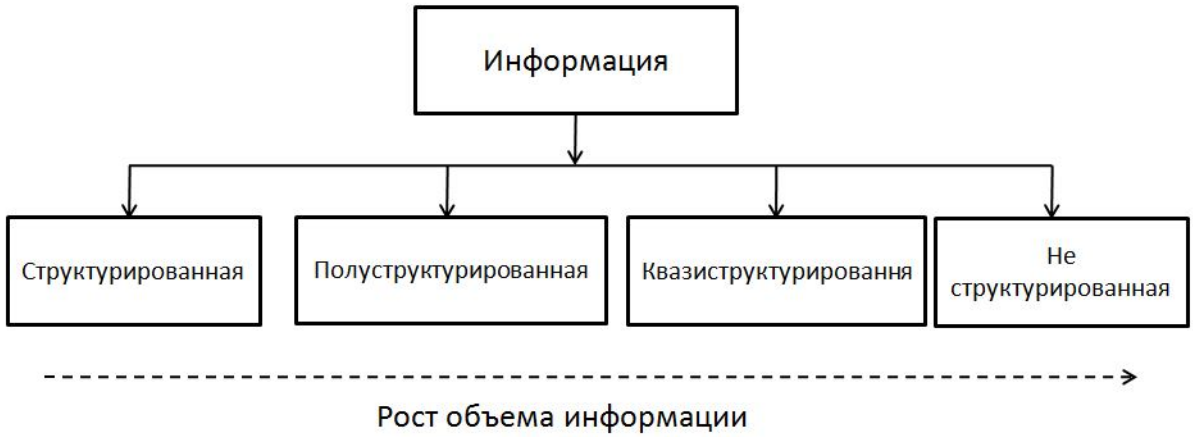


Рис. 1. Классификация информации по типу

Исполнитель самостоятельно выбирает тип данных, с которым в дальнейшем будет работать, но требуется принимать во внимание, что поскольку в курсе рассматриваются подходы и технологии обработки именно большого объема данных, то для выбранной прикладной области рекомендуется иметь для проведения анализа не менее 2 Гб структурированных или полуструктурированных данных, если не используются методы анализа неструктурированного контента. В случае, если используются методы анализа неструктурированного контента, такого как изображения, аудио- и видеозаписи, то рекомендуемый минимальный объем информации - 5Гб.

Особо стоит отметить, что в некоторых случаях обеспечить удовлетворение критерия по объему данных невозможно в связи с природными и техническими ограничениями на количество предоставляемых данных источником.

Важным аспектом используемых для анализа данных является полнота анализируемой выборки. Исполнителю рекомендуется убедиться в том, что исходная выборка достаточна для проведения исследований и сбора статистических данных.

В проекте не предусмотрено предоставление готового набора данных (англ., Data Set) для решения сформулированных задач анализа. Слушателю требуется самостоятельно сформировать собственный набор данных. При этом этап сбора данных является одним из самых трудоемких этапов проекта.

Помочь в формировании набора данных могут web-ресурсы, предоставляющие программный API для получения информации в виде форматов xml, json или любом другом формате.

Одним из таких ресурсов является Интернет-ресурс Head Hunter, предоставляющий информацию по работодателям, вакансиям и соискателям. На рассматриваемом ресурсе в разделе для разработчиков представлено качественное и детальное описание того, как использовать программные интерфейсы (API) сервисов для получения данных, какие типы HTTP-запросов необходимо отправлять на какие адреса (URL). С использованием программного интерфейса ресурса Head Hunter можно сформировать набор данных до 100 Гб информации, который содержит приблизительно 17 000 000 вакансий за временной промежуток от 2002 года до 2017 года, а также информацию приблизительно о 800 000 компаниях. Этого набора данных вполне достаточно для выполнения проекта.

Для поиска других ресурсов, предоставляющих программные интерфейсы получения информации, рекомендуется использовать сервисы-агрегаторы. Такие сервисы обычно хранят перечень сайтов, для которых доступны программные интерфейсы. Предпочтение стоит отдавать национальным ресурсам, нежели зарубежным.

Одним из таких сервисов-агрегаторов является ресурс, http://www.programmableweb.com и Web-Scraping - компьютерная технология, позволяющая извлекать информацию с Интернет ресурсов.

Еще одним способом получения информации с Интернет ресурсов является технологии web-scraping, позволяющие из HTML-разметки страниц

сайтов получать информацию в виде csv, xml или других форматов. К технологиям web-scrapping можно отнести: import.io, Jaunt, JSoup и другие.

- Import.io. [Электронный ресурс]. Режим доступа: *https://www.import.io/*
- Jaunt. [Электронный ресурс]. Режим доступа: http://jaunt-api.com/
- Jsoup. [Электронный ресурс]. Режим доступа: *https://jsoup.org/*

Лабораторная работа

АРХИТЕКТУРА ПРОЕКТИРУЕМОЙ СИСТЕМЫ

Разрабатываемый проект подразумевает создание программного решения, позволяющего автоматически или полуавтоматически решать сформулированные задачи анализа. В основе решения может быть заложена относительно простая, но функциональная и расширяемая модульная схема.

Стоит отметить, что программное решение должно обладать хорошей производительностью, гибким масштабированием, быть распределенным и гарантировать надежность передачи данных между узлами системы. Еще одной важной особенностью

рассматриваемой архитектуры является возможность гибкой настройки проводимых в системе аналитик. На рис.1 представлена рекомендуемая модульная схема разрабатываемой системы.

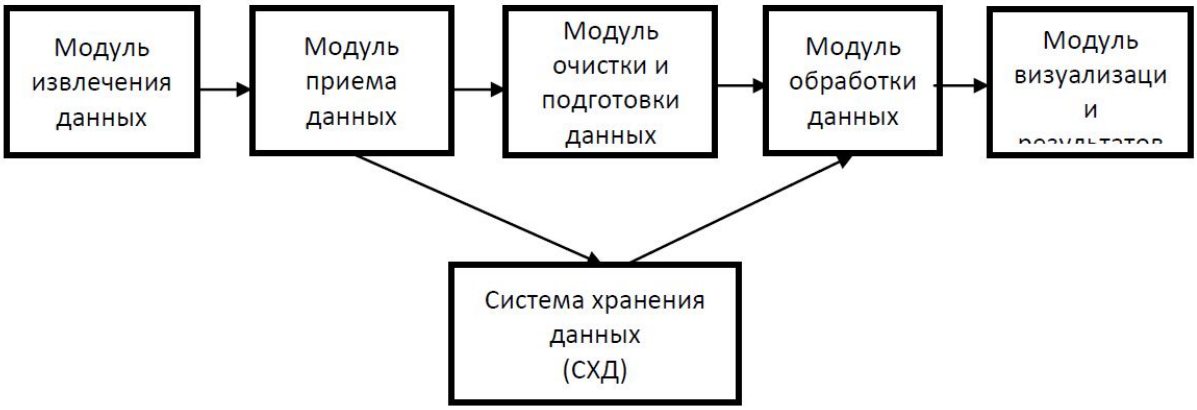


Рис. 1. Модульная схема программной системы анализа данных

Представленный шаблон архитектуры не является обязательным и строго зафиксированным и может видоизменяться в зависимости от выбранной задачи анализа и используемых технологий и методов. Таким образом, за исполнителем проекта остается право видоизменять схему, варьировать количеством модулей и связей между модулями в тех границах, которые позволяют решить поставленную задачу и удовлетворить требованиям на систему, перечисленным выше.

Использование «Модуля извлечения данных» осуществляет получение данных из локального, внешнего или удаленного источника и передачу их в систему. Например, если сформированный набор данных находится на локальном диске, то рассматриваемый модуль просто считывает информацию с локального диска и передает ее в модуль приема.

Если набор данных формируется из удаленного источника, то модуль извлечения осуществляет соединение с удаленным источником и извлекает информацию из него по средствам предоставляемого API удаленного источника или любого поддерживаемого протокола обмена данными. Стоит отметить, что модуль извлечения данных обладает рядом характеристик, таких как скорость извлечения и скорость передачи. В разрабатываемой системе для проведения нагрузочного тестирования итоговой системы следует предусмотреть возможность варьирования перечисленных параметров.

Поскольку предполагается, что проектируемая система должна работать с большим объемом данных и система, в теории, должна обеспечивать высокоскоростной и надежный входной поток данных, то «модуль приема данных» должен гарантировать получение информационных пакетов от «модуля извлечения», их доставку в систему хранения данных (СХД) и отсутствие потерянных пакетов. Для удовлетворения этому требованию возможно использование брокеров сообщений в модуле приема данных. К таким брокерам сообщений относятся RabbitMQ, Kafka и другие системы, удовлетворяющие протоколу AMQP.

RabbitMQ. [Электронный ресурс]. Режим доступа: *https://www.rabbitmq.com/*

Apache Kafka. [Электронный ресурс]. Режим доступа: http://*kafka.apache.org/*

AMQP - Advanced Message Queuing Protocol. [Электронный ресурс]. Режим доступа: http://www.amqp.org/

«Модуль очистки и подготовки данных» отвечает за очистку сырых данных, поступивших от модуля приема, их помещение в хранилище при необходимости, или, в случае обработки в режиме реального времени, их передачу непосредственно обработчику.

Зачастую три модуля в совокупности: извлечение, прием и очистка, формируют процесс, известный как ETL-процесс. Существует огромное число готовых инструментальных средств, которые поддерживают данный процесс. Например: Talend Open Studio, Data Science Studio, GeoKettle и др. При выполнении проекта целесообразно

использовать готовые ETL-инструменты для реализации поставленных задач. Но обязательным является условие сохранения сформулированных свойств системы по производительности, масштабируемости и надежности.

«Модуль обработки данных» в зависимости от сценария использования системы получает данные из «Хранилища» или непосредственно с модуля очистки и подготовки. Он содержит основную логику обработки данных с использованием, например (http://spark.apache.org/) и методов машинного обучения.

После обработки данные передаются на «модуль визуализации результатов», который осуществляет отображение результатов обработки в виде диаграмм, графиков или может даже наносить особые отметки на интерактивную карту.

Проект по аналитике данных предполагает не линейное выполнение, а итерационную работу, то есть на каждом этапе могу быть выявлены недочеты, исправление которых требует возвращения к предыдущему этапу. Например, на этапе реализации алгоритма анализа в модуле обработки может стать очевидным, что используемая модель данных не достаточно полна (а именно, требуется добавление одного или нескольких новых полей). В связи с этим требуется внести изменения в «модуль очистки и подготовки данных» или даже в «модуль извлечения данных». Исходя из этого, особо важным является продумывание архитектуры модулей с заделом на будущее расширение, сопровождение и поддержку. Еще одной причиной обратить внимание на возможность расширения системы является необходимость проведения нескольких итераций и разработки различных алгоритмов вычислений для поиска скрытых зависимостей и прогнозирования.

Совокупное использование представленных модулей позволяет изучить все этапы жизненного цикла проекта по аналитике, начиная с этапа сбора информации, заканчивая этапом представления результатов и внедрения готового продукта.