АННОТАЦИИ РАБОЧИХ ПРОГРАММ ДИСЦИПЛИН (МОДУЛЕЙ)
ООП ВЫСШЕГО ОБРАЗОВАНИЯ – ПРОГРАММЫ МАГИСТРАТУРЫ
Направление подготовки 01.04.02 «Прикладная математика и информатика»

Направленность программы (магистерская программа)
«Большие данные: инфраструктуры и методы решения задач»

Английский язык

Курс английского языка направлен как на дальнейшее развитие таких видов речевой деятельности, как чтение, аудирование, письменная речь в профессионально значимых ситуациях, так и на формирование умений переводческой деятельности, которые также являются частью профессиональной коммуникативной компетенции выпускника. Наибольшее внимание при обучении уделяется продуктивным видам речевой деятельности (письменная речь и говорение), интегративным умениям чтения, аудирования и письменной речи (аннотирование, резюмирование), а также различным видам перевода.

История и методология прикладной математики и информатики

Целью курса является краткое изложение основных фактов, событий и идей в ходе многовековой истории развития математики в целом и одного из её важнейших направлений – «прикладной» (вычислительной) математики, зарождения и развития вычислительной техники и программирования. В курсе делается попытка представить математику как единое целое, где тесно перемежаются проблемы так называемой «чистой» и «прикладной» математики, граница между которыми зачастую весьма условная. Показывается роль математики и информатики в истории развития цивилизации, Дается характеристика научного творчества наиболее выдающихся учёных - генераторов научных идей. Особое внимание уделяется развитию математики и информатики в России.

Современная философия и методология науки

Цель дисциплины – формирование у слушателя целостного видения науки, понимания им специфики научной деятельности, характера исторического развития науки, ее взаимодействия с другими сферами человеческой деятельности.

Задачи дисциплины – заложить теоретические предпосылки для выработки умения анализировать реальную научную деятельность на основе теоретической концепции науки, выявлять специфический характер различных областей науки (специфику понимания строгости, обоснованности, доказательности научного знания, методов его получения, функций научного знания и др.), дифференцировать знание на научное и вненаучное на основе критериев научности.- ориентировать слушателя на понимание исторически изменяющегося характера науки (и ее параметров), восприятия ее за пределами науки в других областях культуры, а также связей науки и общества.- ознакомить с существующими концепциями науки, которые позволяют глубже понимать природу, сущность науки, перспективы развития самой науки, общества, активно использующего науку, и культуру, их породившую.

Модуль Математическое моделирование

Непрерывные математические модели

Излагаются и обсуждаются методы математического моделирования физических, биологических и экономических процессов. Выводятся уравнения, составляющие основу рассматриваемых моделей. Обсуждаются постановки задач. Подробно изучаются методы решения задач, которые возникают в процессе моделирования этих процессов. Приводиться также обзор некоторых результатов в области суперкомпьютерного моделирования.

Дискретные и вероятностные модели

В части посвященной Дискретным моделям рассматриваются способы представления таких функций и их основные свойства, а также вопросы полноты и выразимости; разбираются основные свойства графов, деревья и остовные деревья, раскраски графов, экстремальные графы и теория Рамсея. Приводятся примеры применения свойств дискретных функций и графов в различных областях.

В части посвященной Вероятностным моделям изучаются принципы выбора математических моделей реальных явлений и процессов, протекающих в условиях стохастической неопределенности. Основной упор делается на описание асимптотических аппроксимаций и на энтропийный подход. Значительное внимание уделяется обсуждению условий применимости вероятностных моделей и, в частности, предельных теорем теории вероятностей. Обсуждаются обобщения классических предельных теорем на выборки случайного объема.

Оптимизация и численные методы

Излагаются и обсуждаются методы исследования и решения задач оптимизации и операторных уравнений в гильбертовых пространствах. Рассматриваются вопросы существования решений, условия оптимальности и основные итерационные вычислительные процедуры градиентного типа и метода Ньютона. В конечномерных пространствах для задач линейного и квадратичного программирования описываются конечно-шаговые алгоритмы симплекс-метода и метода сопряженных градиентов.

Модуль Программное обеспечение современных вычислительных комплексов

Современные операционные системы

В курсе «Современные операционные системы» рассматриваются базовые концепции функционирования операционных систем, утилиты, обеспечивающие подсистемы, процессы и управление процессами, управление файловыми системами и устройствами хранения данных, элементы обеспечения безопасности и защиты от несанкционированного доступа.
Курс «Современные операционные системы» направлен на формирование у студентов компетенций, необходимых для решения задач системного администрирования, включающих в себя:
- самостоятельное администрирование операционных систем;
- управление учетными записями и правами пользователей;
- решение проблем функционирования операционных систем.

Сетевые технологии

Задачи курса «Сетевые технологии» - формирование у слушателей структурированного представления о современных сетевых технологиях, включая принципы передачи данных в современных сетях, технологии локальных и глобальных сетей, проблемы информационной безопасности в современных сетях и основные подходов к их решению, овладение слушателями терминологией, необходимой для описания современных сетевых технологий. Курс является вводным к другим курсам магистратуры: технологии сети Интернет, телекоммуникационные технологии, математические основы безопасности ИТ.

Архитектура и программное обеспечение высокопроизводительных вычислительных систем

Количество ядер в современных процессорах уже измеряется десятками, в графических ускорителях – тысячами, а в суперкомпьютерах– миллионами. Многоядерные вычислительные системы широко применяются в машинном обучении, науках о материалах, биоинформатике, автоматизации проектирования, вычислительной химии и физике. Эффективно использовать эту значительную вычислительную мощность – непростая задача, требующая применения современных под-ходов, составляющих основное содержание предлагаемого спецкурса.

Целью освоения дисциплины «Архитектура и программное обеспечение высокопроизводительных вычислительных систем» является получение студентами знаний в области параллельных и распределенных вычислений, выработка у студентов навыков разработки, отладки и исследования производительности параллельных программ. Задачи дисциплины состоят в изучении и практическом освоении современных суперкомпьютерных технологий..

Управление разно-структурированными большими данными

В курсе рассматривается специальный вид стека для параллельных архитектур оперирования данными в аналитических приложениях Big Data. Эти архитектуры полностью отличаются от архитектур суперкомпьютеров. Параллельная архитектура оперирования данными основана на кластере процессоров, обычно соединяемых быстрой сетью (например, гигабитной Ethernet). Центральной в таком архитектурном стеке является парадигма программирования, называемая

MapReduce. Свободно распространяемая реализация такого стека включат HDFS, Hadoop Distributed File System, и поддержку MapReduce (в Hadoop). Такие архитектуры поддерживают разно-структурированные данные, которые могут быть представлены в разнообразных моделях данных (структурированных, слабоструктурированных, неструктурированных).

В курсе рассматриваются основные идеи и подходы параллельных архитектур оперирования разно-структурированными данными. Рассматриваются вопросы реализации различных алгоритмов в среде map-reduce (таких как матрично-векторное умножение, поддержка SQL-подобных операций и операций реляционной алгебры), сравнения реализации таких операций с традиционными. Map-reduce программирование в курсе изучается применяя собственно язык map-reduce Hadoop'а наряду с декларативными языками над Hadoop'ом (такими как PigLatin, Hive, Jaql (IBM)).

Также в курсе рассматривается перспективные методы анализа данных (в дополнении к MapReduce) в середе Hadoop 2.0, основанные на парадигме распределения ресурсов YARN (Yet Another Resource Negotiator). Yarn поддерживает выполнение любых программ, которые могут выполняться параллельно, и позволяет уйти от традиционной парадигмы программирования в Hadoop (map-shuffle-reduce). Это позволяет эффективно программировать сложные задачи, такие как ETL, обработку графов (Giraph), массивно параллельные алгоритмы машинного обучения и моделирования в среде Hadoop. Данная область является широко перспективной и открыта для множества исследований.

В комбинации с Hadoop'ом в курсе рассматриваются базы данных NoSQL (такие как HBase). Их использование совместно с Hadoop'ом изучается на примерах приложений. Подходы к интеграции Hadoop'а в хранилище данных также рассматриваются. В курсе рассматриваются методы применения аналитики данных над Hadoop'ом на примере методов извлечения информации из текстовых документов.

Прикладной многомерный статистический анализ

В рамках данного курса будут рассмотрены основные задачи многомерного статистического анализа. А именно, будет дано описание математических моделей и методов таких разделов математической статистики как корреляционный анализ, регрессионный анализ, дисперсионный анализ, дискриминантный анализ, кластерный анализ. Предложенные методы и алгоритмы иллюстрируются с помощью более-менее реальных примеров.

Виртуальная интеграция неоднородных данных и унификация моделей данных

В курсе рассматриваются проблемы виртуальной интеграции и интероперабельности различных информационных ресурсов (ИР) при создании информационных систем (ИС). Рассматриваются различные технологии интеграции информационных ресурсов, приводятся примеры систем интеграции ресурсов, их сравнительный анализ.

Излагаются теоретические основы виртуальной интеграции, понятие поглощения запросов, алгоритмы переписывания запросов с использованием взглядов. Далее виртуальная интеграция ИР рассматривается применительно к конкретной инфраструктуре предметных посредников, разработанной в ИПИ РАН. Описывается ядро канонической информационной модели предметных посредников (объектно-фреймовый язык с Datalog-подобными логическими правилами); исчисление спецификаций, необходимое для создания посредников; технология создания посредников иллюстрируется на примерах из научных и прикладных предметных областей. Промышленные методы и средства виртуальной интеграции рассматриваются на примере системы IBM Federation Server.

Отдельный раздел курса посвящен методам конструирования канонической информационной модели как унифицирующего расширяемого языка, позволяющего представлять в нем раз-личные языки ИР с сохранением их семантики. Сохраняющие семантику представления в канонической модели разнообразных моделей данных ИР рассматриваются на примерах со-временных моделей данных (многомерных массивов, графовых, онтологических).

Курс основан на классических мировых результатах в области интеграции данных; документации и статьях, опубликованных компанией IBM, на опубликованных работах авторов курса. В процессе чтения лекций студентам предлагается ряд учебных задач, охватывающих основные понятия курса,

некоторые задачи требуют изучения дополнительной литературы. Предлагаются также варианты нерешенных научно-учебных задач в качестве курсовых или магистерских работ.

Материализованная интеграция данных и организация хранилищ больших данных

Целью дисциплины является изучение методов, алгоритмов, архитектур материализованной интеграции данных, а также организации хранилищ больших данных.

В курсе рассматриваются проблемы материализованной интеграции в связи с созданием хранилищ данных (warehouse). Излагаются теоретические основы материализованной интеграции как часть теории баз данных, именуемая в литературе как обмен данными (data exchange). Рассматривается декларативный язык спецификаций отображений схем данных, алгоритмы обмена данными и их сложность, алгоритмы вычисления ответов на запросы при обмене данными.

Рассматриваются архитектуры хранилищ данных, методы проектирования хранилищ данных, методы и средства наполнения хранилищ данных (Extract-Transformation-Load). Рассматриваются языки и методы аналитической обработки многомерных кубов данных (OLAP). Построение хранилищ данных на основе параллельных машин баз данных рассматривается на примере современной платформы IBM Pure Data System for Analytics (Netezza).

Отдельный раздел курса посвящен расширению технологии предметных посредников средствами материализованной интеграции в инфраструктурах параллельных распределенных вычислений (Hadoop).

Дополнительно к лекциям студентам предлагается ряд лабораторных работ, охватывающих основные понятия курса. Лабораторные работы выполняются на виртуальной машине, содержащей установку хранилища данных IBM InfoSphere Warehouse 10.

Курс основан на известных в мире работах по интеграции данных, учебных материалах компании IBM, на опубликованных работах авторов курса. Определен набор учебных задач по материалам курса, решение которых является необходимым условием получения удовлетворительной итоговой оценки. Варианты более сложных научно-учебных задач предлагаются студентам в качестве тем курсовых или магистерских работ.

Интеллектуальный анализ данных

В курсе рассматриваются современные алгоритмы и методы интеллектуального анализа данных для решения поиска ассоциативных правил, тематического моделирования, кластеризации, классификации и прогнозирования В первой части курса, посвященной изучению методов обучения без учителя, рассматриваются: задача поиска ассоциативных правил и основные применяемые для этого алгоритмы - apriori и fp-tree; задача выявления скрытых структур в данных на основе тематического моделирования, в частности метод главных компонент, кластеризация переменных, самоорганизующиеся отображения, неотрицательная матричная факторизация; задача кластеризации данных на основе иерархических, метрических и вероятностных методов. Также обсуждаются методы предобработки данных для эффективного решения данных задач. Вторая часть курса посвящена изучению методов прогнозирования, используемых в системах интеллектуального анализа данных, связанные с этим проблемы, алгоритмы и терминология. Рассматриваются следующие вопросы: понятие проклятия размерности и проблема переобучения; вопросы и критерии для оценки и выбора моделей с использованием валидации и кросс-валидации; алгоритмы и методы необходимой предобработки данных для решения задачи прогнозирования. Далее рассматриваются наиболее популярные и современные алгоритмы и модели машинного обучения и прикладной статистики для решения задач прогнозирования в системах интеллектуального анализа данных, в частности: линейные регрессионные модели; пошаговые методы отбора переменных, регуляризация, преобразование пространства признаков для решения задач прогнозирования; нелинейные регрессионные модели, сплайны, локальная взвешенная регрессия; нейронные сети, их типовые архитектуры RBF и MLP, алгоритмы ранней остановки обучения, методы оптимизации для обучения нейронных сетей; а метод опорных векторов для бинарной классификации, виды ядерных функций, алгоритмы оптимизации для обучения модели на основе опорных векторов; деревья решений, алгоритмы и критерии поиска разбиения при их построении, вопросы управление процессом роста и обрубания ветвей деревьев для борьбы с переобучением; ансамбли моделей на основе бустинга и бэгинга, случайный лес и градиентный бустинг.

Идентификация и слияние сущностей в больших данных

В настоящем курсе изучаются методы и инструменты интеграции информации из различных источников больших данных (в масштабе Веба, социальных сред (Twitter, Linkedin, …), блогов, публикаций в средствах массовой информации, машинных логов, сенсорных данных, и пр.

Большие данные обычно являются неструктурированными (чаще всего текстовыми), слабоструктурированными (например, в виде XML, JSON, баз данных NoSQL). Вместе с тем, образуются также и структурированные большие данные как, например, результат наблюдений (измерений) современными инструментами, накопления многочисленных таблиц в Вебе.

Развитые методы интеграции (ETL) не ориентированы на большие данные. Современные ИТ платформы включают распределенные инфраструктуры типа Hadoop, обеспечивающие параллельную обработку и анализ таких разноструктурированных больших данных на основе парадигмы MapReduce, при этом методы интеграции (ETL) над Hadoop практически отсутствуют.

В курсе рассматриваются современные методы извлечения сущностей (Entity Resolution) и методы слияния данных (Data Fusion) в разрезе интеграции больших данных. В практической части курса рассматриваются масштабируемые методы, реализуемые на языках Jaql и HIL, и выполняющиеся в распределенных инфраструктурах типа Hadoop.

Анализ больших данных в социальных средах

В настоящем курсе изучаются методы, модели и инструменты анализа больших данных в социальных средах. Веб, блоги, социальные сети порождают множество разнообразных связей между сущностями и сетевое взаимодействие между ними. В таком представлении социальные данные взаимосвязаны при помощи явных или неявных ассоциаций. В курсе планируется рассмотреть ряд специальных методов анализа социальных данных как сетей, включая методы, созданные в различных областях, таких как статистика, теория графов, математические модели сетей, социология и пр.

В курсе рассматриваются методы, относящиеся к анализу существующих сетей, в том числе, к оценке различных свойств сетей и характери-стик сущностей в них, распознаванию групп тесно связанных сущностей, выявлению сообществ, важных для планирования рекомендатель-ных систем, анализа рынков, социального поведения, и пр. Рассматриваются различные подходы к моделированию сетевых взаимодействий в статических и эволюционирующих сетях и модели влияния на социальные сети. На основе подобных методов в курсе изучаются вопросы извлечения информации из социальных сетей. Примерами таких вопросов являются аналитика больших данных с эмоциональной окраской, оценка которой влияет на принятие решений; анализ трендов на основе социальных сетей в различные периоды времени, анализ связей и влияния сущностей друг на друга и методы распространения информации в Вебе, устойчивость сетей, а также вопросы использования результатов анализа социальных данных в социальных и экономических науках.