Федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский университет)»

УТВЕРЖДЕНО
Директор высшей школы программной инженерии
А.В. Малеев

Рабочая программа дисциплины (модуля)

по дисциплине: Информационный поиск
по направлению: Программная инженерия
профиль подготовки: Разработка программно-информационных систем
высшая школа программной инженерии
высшая школа программной инженерии МФТИ - Яндекс
курс: 4
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 8 (весенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:
лекции: 30 час.
семинары: 30 час.
лабораторные занятия: 0 час.

Самостоятельная работа: 48 час.

Всего часов: 108, всего зач. ед.: 3

Количество контрольных работ, заданий: 1

Программу составил: А.В. Созыкин, канд. техн. наук

Программа обсуждена на заседании высшей школы программной инженерии МФТИ - Яндекс 28.04.2023

Аннотация

В курсе рассматриваются общие вопросы построения информационно-поисковых систем: задачи информационного поиска и архитектура поисковых систем, машинное обучение в поиске и компьютерная лингвистика, построение поискового индекса и обнаружение дубликатов, поисковый робот и оценка качества.

Решение предлагаемых практических заданий связано со знакомством с широким спектром технологий и алгоритмов, применяемых на практике при построении компонентов поисковой системы.

1. Цели и задачи

Цель дисциплины

- ознакомление студентов с основными принципами построения информационно-поисковых систем.

Задачи дисциплины

- освоение студентами базовых знаний (понятий, концепций, методов и моделей) в области информационного поиска;
- приобретение теоретических знаний и практических умений и навыков в области построения информационно-поисковых систем;
- оказание консультаций и помощи студентам в построении собственных поисковых архитектур.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-3 Способен осуществлять социальное взаимодействие и реализовывать свою роль в команде	УК-3.2 Взаимодействует с другими членами команды для достижения поставленной задачи
ОПК-4 Способен участвовать в разработке стандартов, норм и правил, а также технической документации, связанной с профессиональной деятельностью	ОПК-4.2 Владеет на практике методологией составления научно-технических отчетов (проектов)
	ОПК-4.1 Знает основные правила оформления технической документации на различных стадиях жизненного цикла информационной системы
ОПК-7 Способен применять в практической деятельности основные концепции, принципы, теории и факты, связанные с информатикой	ОПК-7.2 Понимает принципы работы баз данных и умеет проектировать структуру данных для эффективного хранения информации
ПК-4 Способен разрабатывать тесты, подготавливать тестовые данные, проводить тестирование, разрабатывать документы для тестирования	ПК-4.1 Обладает навыками проведения необходимых видов тестирования в соответствии с планом тестирования
	ПК-4.2 Умеет оценивать важность различных тестов на основе приоритетов пользователя, проектных задач и рисков возникновения ошибки
	ПК-4.4 Умеет выполнять анализ полученных результатов тестирования и оформлять их в соответствии с требуемым форматом
	ПК-4.3 Имеет практический опыт работы с тестовыми средами и системами управления тестированием в своей профессиональной деятельности

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные понятия, законы информационного поиска;
- архитектуру поискового робота;
- современные проблемы соответствующих разделов теории сложных сетей;
- существующие модели случайных сетей и веб-графов.

уметь:

- понять поставленную задачу;
- написать собственный поисковый робот;
- строить индекс по коллекции документов и организовывать поиск по нему;
- использовать свои знания для построения собственной поисковой архитектуры;
- самостоятельно находить способы выполнения поставленных задач, в том числе и нестандартных, и проводить их анализ;
- самостоятельно видеть следствия полученных результатов.

владеть:

- навыками освоения большого объема информации и решения задач (в том числе, сложных);
- навыками самостоятельной работы и освоения новых дисциплин;
- культурой постановки, анализа и решения теоретических задач теории сложных сетей, а также задач исследования сетевых структур;
- методами индексации страниц и обнаружения дубликатов.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Введение.	4	3		5
2	Query - Text Matching.	3	4		6
3	Query - Text Matching, глубинные модели.	3	3		5
4	Поисковый робот.	4	3		5
5	Обнаружение дубликатов.	3	4		6
6	Crawling Order.	3	3		5
7	Построение и использование инвертированного индекса. Сжатие.	4	3		6
8	Обучение ранжированию.	3	4		5
9	Оптимизация индекса. Алгоритмы и эвристики.	3	3		5
Итого часов		30	30		48
Подготовка к экзамену		0 час.			
Общая трудоёмкость		108 час., 3 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 8 (Весенний)

1. Введение.

Постановка задачи. Что должна уметь поисковая система? Базовые компоненты поисковой системы. Векторные модели документа.

2. Query - Text Matching.

Языковые модели. Сглаживание. Как учитывать контекст. Тематическое моделирование.

3. Query - Text Matching, глубинные модели.

Обзор существующих архитектур.

4. Поисковый робот.

Алгоритмы. Взаимодействие с администратором ресурса. Метрики качества обхода. Page Rank.

5. Обнаружение дубликатов.

Зачем это нужно? Виды дублей. Шинглы. Odd Sketch. SimHash.

6. Crawling Order.

OPIC. Обход свежих страниц. Кластеризация свежих страниц. Анализ источников ссылок. Выделение ресурсов. Использование структуры сайта. Политики обходов.

7. Построение и использование инвертированного индекса. Сжатие.

Что такое инвертированный индекс. Подходы к построению. Построение на MapReduce. Использование инвертированного индекса. Подходы к сжатию. Varint.

8. Обучение ранжированию.

Метрики ранжирования. Pointwise, pairwise, listwise подходы.

9. Оптимизация индекса. Алгоритмы и эвристики.

Document Identifier Reordering. Index Pruning. Signature Files.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Учебная аудитория, оснащенная медиапроектором и экраном.

6.Перечень рекомендуемой литературы

Основная литература

1. Автоматическая обработка, хранение и поиск информации [Текст]/Г. Сэлтон , -М., Советское радио, 1973

Дополнительная литература

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

не предусмотрено

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

На лекционных занятиях демонстрируются презентации с помощью мультимедийных технологий.

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Студент, изучающий дисциплину, должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике.

В результате изучения дисциплины студент должен знать основные определения, понятия.

Успешное освоение курса требует напряжённой самостоятельной работы студента. В программе курса приведено необходимое время для работы студента над темой. Самостоятельная работа включает в себя:

- чтение и конспектирование рекомендованной литературы;
- проработку учебного материала (учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств;
- подготовку к дифференцированному зачету.

Руководство и контроль за самостоятельной работой студента осуществляется в форме индивидуальных консультаций.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями к лектору.

ПРИЛОЖЕНИЕ

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Программная инженерия
профиль подготовки: Разработка программно-информационных систем
высшая школа программной инженерии МФТИ - Яндекс
высшая школа программной инженерии
курс: 4
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 8 (весенний) - Дифференцированный зачет

Разработчик: А.В. Созыкин, канд. техн. наук

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-3 Способен осуществлять социальное взаимодействие и реализовывать свою роль в команде	УК-3.2 Взаимодействует с другими членами команды для достижения поставленной задачи
ОПК-4 Способен участвовать в разработке стандартов, норм и правил, а также технической документации, связанной с профессиональной деятельностью	ОПК-4.2 Владеет на практике методологией составления научно-технических отчетов (проектов)
	ОПК-4.1 Знает основные правила оформления технической документации на различных стадиях жизненного цикла информационной системы
ОПК-7 Способен применять в практической деятельности основные концепции, принципы, теории и факты, связанные с информатикой	ОПК-7.2 Понимает принципы работы баз данных и умеет проектировать структуру данных для эффективного хранения информации
ПК-4 Способен разрабатывать тесты, подготавливать тестовые данные, проводить тестирование, разрабатывать документы для тестирования	ПК-4.1 Обладает навыками проведения необходимых видов тестирования в соответствии с планом тестирования
	ПК-4.2 Умеет оценивать важность различных тестов на основе приоритетов пользователя, проектных задач и рисков возникновения ошибки
	ПК-4.4 Умеет выполнять анализ полученных результатов тестирования и оформлять их в соответствии с требуемым форматом
	ПК-4.3 Имеет практический опыт работы с тестовыми средами и системами управления тестированием в своей профессиональной деятельности

2. Показатели оценивания компетенций

В результате изучения дисциплины «Информационный поиск» обучающийся должен:

знать:

- основные понятия, законы информационного поиска;
- архитектуру поискового робота;
- современные проблемы соответствующих разделов теории сложных сетей;
- существующие модели случайных сетей и веб-графов.

уметь:

- понять поставленную задачу;
- написать собственный поисковый робот;
- строить индекс по коллекции документов и организовывать поиск по нему;
- использовать свои знания для построения собственной поисковой архитектуры;
- самостоятельно находить способы выполнения поставленных задач, в том числе и нестандартных, и проводить их анализ;
- самостоятельно видеть следствия полученных результатов.

владеть:

- навыками освоения большого объема информации и решения задач (в том числе, сложных);
- навыками самостоятельной работы и освоения новых дисциплин;
- культурой постановки, анализа и решения теоретических задач теории сложных сетей, а также задач исследования сетевых структур;
- методами индексации страниц и обнаружения дубликатов.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

1. Постановка задачи.
2. Что должна уметь поисковая система?
3. Языковые модели.

4. Обзор существующих архитектур.

5. Метрики качества обхода.

6. Виды дублей.

7. Анализ источников ссылок.

8 Что такое инвертированный индекс?

9. Метрики ранжирования.

10. Оптимизация индекса.

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

1. Сколько существует различных симхешей длины n на расстоянии не более чем m от заданного?

2. На сколько делят трехмерное пространство n различных плоскостей, проходящих через начало координат?

3. Опишите, как с помощью технологии MapReduce можно реализовать Join двух таблиц по ключу.

4. Предложите алгоритм для сегментации запросов. Дан словарь, в котором каждому слову и словосочетанию сопоставлена частота появления в корпусе. Определить наиболее вероятную сегментацию запроса.

5. Имеется словарь запросов Q большого размера N=10 в 8-ой степени. Предложите структуру данных, которая сможет обабатывать запросы вида "дан запрос q, вернуть 10 запросов из словаря Q, похожих на q с точностью до перестановки и добавления слов в порядке увеличения слов в порядке увеличения словности". Приведите оценки по памяти и по времени.

6. Информатика и семиотика.

7. Общие принципы организации информационно-поисковых систем.

8. Метаданные и обработка электронных ресурсов.

9. Модель информационно-поисковой системы.

10. Структура логических компонентов информационно-поисковой системы.

11. Построение тезаурусов и онтологий информационно-поисковых систем.

12. Использование методов машинного обучения для обработки документов.

Критерии оценивания

отлично (10) - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

отлично (9) - выставляется студенту, показавшему свободное оперирование знаниями учебной программы дисциплины, выполнение заданий творческого характера.

отлично (8) - выставляется студенту, показавшему владение программным учебным материалом с наличием несущественных ошибок в действиях, самостоятельно исправляемых учащимся.

хорошо (7) - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускается в ответе или в решении задач некоторые неточности.

хорошо (6) - выставляется студенту если он осознает воспроизведение программного учебного материала, в том числе и различной степени сложности, с несущественными ошибками, затруднения в применении отдельных навыков.

хорошо (5) - выставляется студенту если теоретическое содержание освоено не полностью, некоторые практические навыки сформированы недостаточно, в некоторых случаях были допущены ошибки.

удовлетворительно (4) - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации.

удовлетворительно (3) - выставляется студенту в случае большого количества недочетов и неправильных ответов, а также пассивной работе в ходе занятий, многие учебные задания не выполнены.

неудовлетворительно (2) - выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач.

неудовлетворительно (1) - выставляется студенту, который не освоил теоретическое и практическое содержание курса, все выполненные учебные задания содержат грубые ошибки.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Дифференцированный зачет может проводиться по итогам текущей успеваемости и сдачи заданий и других видов работ, предусмотренных программой дисциплины и (или) путем организации специального опроса, проводимого в устной и (или) письменной форме.

При проведении дифференцированного зачета студенту предоставляется 30 минут на подготовку. Опрос обучающегося по билету на дифференцированном зачете не должен превышать одного астрономического часа.