Минобрнауки России

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

(ФГБОУ ВО «ВГУ»)

УТВЕРЖДАЮ

Заведующий кафедрой

Борисов Дмитрий Николаевич

Кафедра информационных систем

10.04.2024

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Б1.В.ДВ.01.03.01 Информационно-поисковые системы

1. Код и наименование направления подготовки/специальности:

09.03.02 Информационные системы и технологии

2. Профиль подготовки/специализация:

Инженерия информационных систем и технологий

3. Квалификация (степень) выпускника:

Бакалавриат

4. Форма обучения:

Очная

5. Кафедра, отвечающая за реализацию дисциплины:

Кафедра информационных систем

6. Составители программы:

Сычев Александр Васильевич, кандидат физико-математических наук, доцент кафедры информационных систем

7. Рекомендована: *НМС ФКН от 05.03.2024 г., протокол № 5*

8. Учебный год:

2026-2027

9.Цели и задачи учебной дисциплины:

Целью является знакомство с основными тенденциями развития мировых информационных ресурсов, моделями систем и методами информационного поиска.

Задачи дисциплины:
- изучение архитектуры информационно-поисковых систем (ИПС), стратегий информационного поиска и методов ранжирования в ИПС, критериев оценки эффективности ИПС;
- практическое знакомство с элементами ИПС путем моделирования их работы.

10. Место учебной дисциплины в структуре ООП:

Учебная дисциплина относится к части блока Б1, формируемой участниками образовательных отношений.

11. Планируемые результаты обучения по дисциплине/модулю (знания, умения, навыки), соотнесенные с планируемыми результатами освоения образовательной программы

(компетенциями выпускников) и индикаторами их достижения:

Код и название компетенции	Код и название индикатора компетенции	Знания, умения, навыки
ПК-3 Выполнение и управление работами по созданию и сопровождению информационных систем	ПК-3.2 Создание программного кода информационной системы в рамках выполнения работ по созданию (модификации) и сопровождению информационной системы	Знает: - основные математические модели и алгоритмы документального поиска; - подходы к измерению эффективности информационного поиска. - подходы к ранжированию документов в ИПС. - основные компоненты архитектуры информационно-поисковых систем (ИПС); - базовые структуры данных, используемые для кодирования поисковых индексов; - принципы работы поисковых роботов в сети Web. Умеет: - применять методы оценки функциональной эффективности к конкретным информационно-поисковым системам. **-** разрабатывать базовые компоненты поискового робота.

12. Объем дисциплины в зачетных единицах/час:

2/72

Форма промежуточной аттестации:

Зачет с оценкой

13. Трудоемкость по видам учебной работы

Вид учебной работы	Семестр 6	Всего
Аудиторные занятия	48	48
Лекционные занятия	32	32
Практические занятия		0
Лабораторные занятия	16	16
Самостоятельная работа	24	24
Курсовая работа		0
Промежуточная аттестация	0	0
Часы на контроль		0

Вид учебной работы	Семестр 6	Всего
Всего	72	72

13.1. Содержание дисциплины

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
1.		**Лекции**	
1.1	Мировые информационные ресурсы	Мировые информационные ресурсы. Оценки их объема и их динамики. Актуальность проблемы информационного поиска. Языки гипертекстовой разметки, их недостатки и перспективы.	Онлайн курс на edu.vsu.ru
1.2	Исторический обзор зарубежных и отечественных информационно-поисковых систем.	Фазы жизненного цикла информационного ресурса: становление, каталогизация, автоматическое индексирование, канонизация поисковых сервисов, угасание. Краткий исторический обзор зарубежных и отечественных информационно-поисковых систем.	Онлайн курс на edu.vsu.ru
1.3	Проблема информационного поиска	Проблема информационного поиска, возможность ее формального математического описания. Основные понятия информационного поиска. Формы релевантности: формальная, содержательная, индивидуально-прагматическая (пертинентность).	Онлайн курс на edu.vsu.ru
1.4	Математические модели информационного поиска	Математические модели информационного поиска. Теоретико-множественная модель. Идеальное качество поиска. Энтропийная модель. Коэффициенты релевантности, выдачи, полноты, специфичности, точности. Матричная модель. Типы сопряженности: “документ-документ”,”термин-термин”,”документ-термин”.	Онлайн курс на edu.vsu.ru
1.5	Методы документального поиска	Методы документального поиска: полнотекстовый поиск, файлы сигнатур (хэширование), инверсия, векторно-кластерные методы. NLP (обработка естественного языка), LSI (индексирование на основе скрытой семантики). SVD – декомпозиция.	Онлайн курс на edu.vsu.ru
1.6	Ранжирование документов: основные подходы и модели	Ранжирование документов в выдаче.	Онлайн курс на edu.vsu.ru
1.7	Анализ гиперссылок и его применение для ранжирования документов. Учет контекста для повышения релевантности поиска.	Использование гиперссылок для ранжирования. Схемы ранжирования, зависящие от запросов и независящие от запросов. Алгоритмы PageRank и HITS	Онлайн курс на edu.vsu.ru
1.8	Архитектура информационно-поисковой системы Web	Архитекура информационно-поисковой системы Web и проблемы ее реализации на примере ИПС Google (1998).	Онлайн курс на edu.vsu.ru

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
1.9	Стратегии обхода веб-графа. Контекстно-сфокусированный поиск. Архитектура системы сфокусированного поиска.	Стратегии обхода веб-графа. Контекстно-сфокусированный поиск. Архитектура системы сфокусированного поиска.	Онлайн курс на edu.vsu.ru
1.10	Самоорганизация в сети WWW. Поиск веб-сообществ.	Самоорганизация в сети WWW. Социальные сети WWW. Модели организации и развития веб-сообществ. Поиск веб-сообществ в сети WWW. Алгоритмы Форда-Фолкерсона, FLG и др. Блогосфера. Современное состояние, динамика развития, исследования.	Онлайн курс на edu.vsu.ru
2.		**Практические занятия**	
3.		**Лабораторные занятия**	
3.1	Модель документа «мешок слов». Закон Ципфа. Принцип Луна.	Разработка приложения для анализа веб-страницы.	Онлайн курс на edu.vsu.ru
3.2	Закон Ципфа. Принцип Луна.	Расчет частот терминов веб-страницы, построение и анализ графика зависимости «частота-ранг».	Онлайн курс на edu.vsu.ru
3.3	Матричная модель документального поиска.	Разработка приложения для анализа коллекции веб-страницы.	Онлайн курс на edu.vsu.ru
3.4	Матричная модель документального поиска.	Анализ коллекции веб-страниц и построение матриц сопряженности типа «термин-документ», «документ-документ», «термин-термин».	Онлайн курс на edu.vsu.ru
3.5	Реализация модуля для скачивания веб-страницы из сети Веб.	Разработка приложения для скачивания веб-страницы из сети Веб по протоколу HTTP на основе сокета.	Онлайн курс на edu.vsu.ru
3.6	Моделирование поискового робота	Разработка модуля для скачивания и анализа веб-страниц.	Онлайн курс на edu.vsu.ru
3.7	Моделирование поискового робота	Разработка модуля для работы с очередью гиперссылок и преобразования относительных ссылок в каноническую форму.	Онлайн курс на edu.vsu.ru
3.8	Моделирование поискового робота	Изучение работы программной реализации поискового робота и характеристик построенного графа гиперссылок на примере выбранного веб-сайта.	Онлайн курс на edu.vsu.ru

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела)	Лекционные занятия	Практические занятия	Лабораторные занятия	Самостоятельная работа	Всего
1	Мировые информационные ресурсы	2			1	3

№ п/п	Наименование темы (раздела)	Лекционные занятия	Практические занятия	Лабораторные занятия	Самостоятельная работа	Всего
2	Исторический обзор зарубежных и отечественных информационно-поисковых систем.	2			2	4
3	Проблема информационного поиска	2			1	3
4	Математические модели информационного поиска	6		4	5	15
5	Алгоритмы документального поиска	6		2	5	13
6	Ранжирование документов: основные подходы и модели	2		4		6
7	Анализ гиперссылок и его применение для ранжирования документов. Учет контекста для повышения релевантности поиска.	4		4	4	12
8	Архитектура информационно-поисковой системы Web	3		2	4	9
9	Стратегии обхода веб-графа. Контекстно-сфокусированный поиск. Архитектура системы сфокусированного поиска.	3			1	4
10	Самоорганизация в сети WWW. Поиск веб-сообществ.	2			1	3
		32	0	16	24	72

14. Методические указания для обучающихся по освоению дисциплины

1) При изучении дисциплины рекомендуется использовать следующие средства:

- рекомендуемую основную и дополнительную литературу;
- методические указания и пособия;
- контрольные задания для закрепления теоретического материала;
- электронные версии учебников и методических указаний для выполнения лабораторно-практических работ.

2) Для лучшего усвоения дисциплины рекомендуется проведение письменного опроса (тестирование, решение задач) студентов по материалам лекций. Подборка вопросов для тестирования осуществляется на основе изученного теоретического материала.

3) При проведении лабораторных занятий обеспечивается практическая демонстрация материалов лекционных занятий и осуществляется экспериментальная проверка методов, алгоритмов и

технологий информационного поиска, излагаемых в рамках лекций.

4) При переходе на дистанционный режим обучения для создания электронных курсов, чтения лекций онлайн и проведения лабораторно-практических занятий используются информационные ресурсы образовательного портала "Электронный университет ВГУ (https://edu.vsu.ru), базирующегося на системе дистанционного обучения Moodle, развернутой в университете.

Электронный курс, размещенный на портале Электронный университет ВГУ (https://edu.vsu.ru/course/view.php?id=4154).

15. Перечень основной и дополнительной литературы, ресурсов интернет, необходимых для освоения дисциплины

№ п/п	Источник
1	Маннинг К.Д. Введение в информационный поиск / К.Д. Маннинг, П. Рагхаван, Х.М. Шютце. – М. : Вильямс, 2011. - 528 с.
2	Юре, Л. . Анализ больших наборов данных [Электронный ресурс] / Юре Л. , Ананд Р. , Джеффри Д. У. — Москва : ДМК Пресс, 2016 .— 498 с. — <URL:https://e.lanbook.com/book/93571>
3	Даг, Т. Релевантный поиск с использованием Elasticsearch и Solr / Т. Даг, Б. Джон ; перевод с английского А. Н. Киселев. — Москва : ДМК Пресс, 2018. — 408 с. — ISBN 978-5-97060-592-9. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/111439

б) дополнительная литература:

№ п/п	Источник
1	Симанков, В. С. Методы и алгоритмы поиска информации в Интернете : монография / В. С. Симанков, Д. М. Толкачев. — Москва : Креативная экономика, 2017. — 332 с. — ISBN 978-5-9500501-8-3. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/116049
2	Щербаков, А. Интернет-аналитика. Поиск и оценка информации в web-ресурсах / А. Щербаков.— Москва : Книжный мир, 2012 .— 78 с. (http://biblioclub.ru/index.php?page=book&id=89693)
3	Белов, В. В. Повышение пертинентности поиска в современных информационных средах : учебное пособие / В. В. Белов, А. А. Терехов, В. И. Чистякова. — Москва : Горячая линия-Телеком, 2012. — 158 с. — ISBN 978-5-9912-0223-7. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/5118

в) информационные электронно-образовательные ресурсы:

№ п/п	Источник
1	Курс лекций “Анализ гиперссылок в сети Веб: модели, подходы и алгоритмы” (адрес http://romip.ru/russir2007/slides/haw.zip) для для слушателей летней школы-семинара по информационному поиску “RUSSIR’07”, проводившейся в сентябре в г. Екатеринбурге (http://romip.ru/russir2007/schedule.html).
2	Материалы летней школы-семинара по информационному поиску “RUSSIR’09” (http://romip.ru/russir2009/program.html)
3	Материалы летней школы-семинара по информационному поиску “RUSSIR’10” (http://romip.ru/russir2010/program.html)
4	Материалы летней школы-семинара по информационному поиску “RUSSIR’11” (http://romip.ru/edbt-russir2011/section.php?id=93l)
5	Материалы летней школы-семинара по информационному поиску “RUSSIR’12” (http://romip.ru/russir2012/section.php?id=122l)
6	Материалы летней школы-семинара по информационному поиску “RUSSIR’13” (http://romip.ru/russir2013/section.php?id=152)
7	Материалы летней школы-семинара по информационному поиску “RUSSIR’14” (http://romip.ru/russir2014/section.php?id=187)
8	Материалы летней школы-семинара по информационному поиску “RUSSIR’16” (http://romip.ru/russir2016/lecture-materials/)

16. Перечень учебно-методического обеспечения для самостоятельной работы

№ п/п	Источник
1	Электронный курс на образовательном портале «Электронный университет ВГУ».– (https://edu.vsu.ru/course/view.php?id=4154)

17. Образовательные технологии, используемые при реализации учебной дисциплины, включая дистанционные образовательные технологии (ДОТ), электронное обучение (ЭО), смешанное обучение):

Учебный Web-сервер *Apache*, редактор *notepad++,* интерпретаторы *Perl* и *PHP,* клиент для протоколов удалённого доступа *(*включая *SSH).*

18. Материально-техническое обеспечение дисциплины:

Компьютерная лаборатория с локальной сетью из 15 персональных компьютеров с установленным системным и прикладным программным обеспечением и выходом в Интернет.

19. Оценочные средства для проведения текущей и промежуточной аттестаций

Порядок оценки освоения обучающимися учебного материала определяется содержанием следующих разделов дисциплины:

№ п/п	Разделы дисциплины (модули)	Код компетенции	Код индикатора	Оценочные средства для текущей аттестации
1	1-10	ПК-3	ПК-3.2	Письменный опрос, практические задания
2				

Промежуточная аттестация

Форма контроля - Зачет с оценкой

Оценочные средства для промежуточной аттестации

Оценка знаний, умений и навыков, характеризующая этапы формирования компетенций в рамках изучения дисциплины осуществляется в ходе текущей и промежуточной аттестаций.

Текущая аттестация проводится в соответствии с Положением о текущей аттестации обучающихся по программам высшего образования Воронежского государственного университета и Положения о балльно-рейтинговой системы на факультете компьютерных наук Воронежского государственного университета. Текущая аттестация проводится в форме(ах)*: письменного опроса и выполнения практических заданий на лабораторных занятиях.*

Промежуточная аттестация проводится в соответствии с Положением о промежуточной аттестации обучающихся по программам высшего образования и Положением о балльно-рейтинговой системе на факультете компьютерных наук Воронежского государственного университета.

Контрольно-измерительные материалы промежуточной аттестации включают в себя теоретические вопросы, позволяющие оценить уровень полученных знаний и практические задания, позволяющие оценить степень сформированности умений и навыков.

20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1 Текущий контроль успеваемости

Тестовые задания - 1 балл за каждый правильный тест (максимум).

Компетенция ПК-3

1. Выберите правильное соответствие между типовыми компонентами и задачами ИПС в Веб:
a) Сбор документов из сети Веб.
b) Размещение документов для последующего их индексирования.
c) Обработка документов и формирование структур данных, используемых для поиска документов.
d) Работа с документами и запросами пользователей.

a) Индексатор.
b) Обработчик запросов с поддержкой ранжирования.
c) Сетевой робот-”паук”.
d) Хранилище документов.

2. Выберите правильное соответствие для компонент структуры ИПС Google:
a) загрузка веб-страниц из сети WWW;
b) формирование задания на загрузку документов;
c) размещение скачанных из сети документов в хранилище;
d) извлечение документов из хранилища, анализ их структуры, преобразование документов в множество вхождений слов (hits);

e) извлечение гиперссылок из документов и размещение информации о них в файле анкеров;
f) преобразование URL из файла анкеров из относительной в абсолютную форму, отображение их в идентификаторы docID;
g) помещение текста гиперссылки в прямой индекс, связанный с docID;
h) генерация базы данных гиперссылок в виде пары docID;
i) на основе данных из “емкостей”, отсортированных по docID, генерирует обратный индекс;
j) создает список wordID и их смещений в обратном индексе;
k) использует список wordID и словарь, построенный сортировщиком, и строит новый словарь для поискового агента;
l) использует словарь, построенный DumpLexicon, обратный индекс и PageRank для ответа на запросы.

a) DumpLexicon
b) URL-резолвер
c) URL-резолвер
d) URL-резолвер
e) URL-сервер
f) Индексатор
g) Индексатор
h) Поисковый агент
i) Сервер хранилища
j) Сетевой робот
k) Сортировщик
l) Сортировщик

3. Выберите правильные утверждения, относящиеся к кластерной архитектуре Google. Надежность в работе системы может быть обеспечена:
a) в первую очередь программным путем;
b) в первую очередь аппаратным путем;
c) путем использования недорогих ПК;
d) путем использования серверного оборудования.

4. Для матрицы “релевантность-выдача”

	X	^X
Y	A	B
^Y	C	D

укажите, что представляет собой подмножество A?

a) Документы, не попавшие в выдачу поисковой системы.
b) Документы, попавшие в выдачу поисковой системы.
c) Нерелевантные документы, не попавшие в выдачу поисковой системой.
d) Нерелевантные документы, попавшие в выдачу поисковой системой.
e) Нерелевантные документы.
f) Релевантные документы, не попавшие в выдачу поисковой системой.
g) Релевантные документы, попавшие в выдачу поисковой системой.
h) Релевантные документы.

5. Какую интерпретацию имеет элемент td[i][j] матрицы сопряженности типа термин-документ?

a) количество документов, содержащих общие термины.
b) количество общих терминов, содержащихся одновременно в документах d[i] и d[j].
c) количество терминов, содержащихся одновременно в общих документах.
d) наличие термина t[i] в документе d[j].

e) частота термина t[i] в документе d[j].

6. Какие компоненты включает в себя модель документального поиска?

a) Множество пользователей.
b) Множество представлений документа.
c) Множество представлений информационной потребности пользователя.
d) Поисковый робот.
e) Средства анализа гиперссылок в документах.
f) Средства защиты документов от несанкционированного доступа.
g) Средства моделирования представлений документа, запросов и их отношений.
h) Функция ранжирования.

7. Установите правильное соответствие:
a) соответствие содержания документа информационной потребности пользователя;
b) степень соответствия содержания документа, найденного в результате информационного поиска, информационной потребности пользователя, сформулированной в виде информационного запроса;
c) наличие в документе контекстных ситуаций, затребованных пользовательским запросом;

a) Пертинентность.
b) Содержательная релевантность.
c) Формальная релевантность.

8. Выберите правильные утверждения. Контекстный граф показывает:
a) темы, прямо или косвенно связанные с целевой темой.
b) пути связывающие документы в графе с целевыми документами.
c) степень подобия запроса и тематики документов.
d) факторы контекста, привлекаемые для поиска документов.

Вопросы с кратким (вычисляемым) ответом - 1 балл за каждый правильный тест (максимум)

Компетенция ПК-4

1. Для коллекции документов размером 100 в матрице “релевантность-выдача”:

	X	^X
Y	A	B
^Y	C	D

относительно поискового запроса q1 были получены следующие числовые значения для подмножеств A,B,C,D:
A=10; B=2; C=3; D=85.
Рассчитайте значение коэффициента точности поиска. Числовой ответ должен быть приведен с точностью до 2 цифр после запятой.

Вопросы с развернутым ответом

Критерии оценивания развернутого ответа:
В зависимости от степени полноты ответа на вопрос максимально можно получить за развернутый ответ на вопрос 3 балла.

Компетенция ПК-3

1) Что такое спамдексинг? Приведите краткое описание основных приемов спамдексинга.

20.2 Промежуточная аттестация

Промежуточная аттестация проводится в соответствии с Положением о промежуточной аттестации обучающихся по программам высшего образования и Положением о балльно-рейтинговой системе

на факультете компьютерных наук Воронежского государственного университета.