МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное учреждение
высшего профессионального образования
"Казанский (Приволжский) федеральный университет"
Институт вычислительной математики и информационных технологий

подписано электронно-цифровой подписью

Программа дисциплины

Разработка тезаурусов и антологий Б3.ДВ.1

Направление подготовки: 010400.62 - Прикладная математика и информатика
Профиль подготовки: Математическое и программное обеспечение вычислительных машин и сетей
Квалификация выпускника: бакалавр
Форма обучения: очное
Язык обучения: русский

Автор(ы):

Соловьев В.Д.

Рецензент(ы):

Иванов В.В.

СОГЛАСОВАНО:

Заведующий(ая) кафедрой: Таланов М. О.
Протокол заседания кафедры No ___ от "____" _____ 201__г
Учебно-методическая комиссия Института вычислительной математики и информационных технологий:
Протокол заседания УМК No ____ от "____" _____ 201__г

Регистрационный No 9128914

Казань
2014

Содержание

Программу дисциплины разработал(а)(и) главный научный сотрудник, д.н. (профессор) Соловьев В.Д. Научно-образовательный центр по лингвистике им.И.А.Бодуэна де Куртене Институт филологии и межкультурной коммуникации , Valery.Solovyev@kpfu.ru

1. Цели освоения дисциплины

Целью освоения дисциплины "Обработка естественного языка" является знакомство с понятиями онтологии и тезаурусы, их месте и роли в современных информационных технологиях и концепции "Семантический Веб". Даются основы языков описания онтологий (RDF, OWL), математического аппарата их обработки;

2. Место дисциплины в структуре основной образовательной программы высшего профессионального образования

Данная учебная дисциплина включена в раздел " Б3.ДВ.1 Профессиональный" основной образовательной программы 010400.62 Прикладная математика и информатика и относится к дисциплинам по выбору. Осваивается на 4 курсе, 8 семестр.

Данная дисциплина относится к профессиональным дисциплинам.

Читается на 4 курсе 7 семестр для студентов, обучающихся по направлению "Прикладная математика и информатика".

3. Компетенции обучающегося, формируемые в результате освоения дисциплины /модуля

В результате освоения дисциплины формируются следующие компетенции:

Шифр компетенции	Расшифровка приобретаемой компетенции
ПК-8 (профессиональные компетенции)	способность профессионально владеть базовыми математическими знаниями и информационными технологиями, эффективно применять из для решения научно-технических задач и прикладных задач, связанных с развитием и использованием информационных технологий
ПК-9 (профессиональные компетенции)	способность осуществлять на практике современные методологии управления жизненным циклом и качеством систем, программных средств и сервисов информационных технологий

В результате освоения дисциплины студент:

1. должен знать:

что такое онтологии и тезаурусы, их место и роль в современных информационных технологиях и концепции "Семантический Веб";

2. должен уметь:

ориентироваться в тех задачах, где применение онтологий и тезаурусов наиболее эффективно;

3. должен владеть:

теоретическими знаниями о языках описания онтологий (RDF, OWL), математическом аппарате их обработки, о структуре с способах использования ресурсов типа WordNet;

4. должен демонстрировать способность и готовность:

приобрести навыки создания онтологий конкретных предметных областей, использовать тезаурусы для организации информационного поиска.

4. Структура и содержание дисциплины/ модуля

Общая трудоемкость дисциплины составляет 3 зачетных(ые) единиц(ы) 108 часа(ов).

Форма промежуточного контроля дисциплины экзамен в 8 семестре.

Суммарно по дисциплине можно получить 100 баллов, из них текущая работа оценивается в 50 баллов, итоговая форма контроля - в 50 баллов. Минимальное количество для допуска к зачету 28 баллов.

86 баллов и более - "отлично" (отл.);

71-85 баллов - "хорошо" (хор.);

55-70 баллов - "удовлетворительно" (удов.);

54 балла и менее - "неудовлетворительно" (неуд.).

4.1 Структура и содержание аудиторной работы по дисциплине/ модулю

Тематический план дисциплины/модуля

N	Раздел Дисциплины/ Модуля	Семестр	Неделя семестра	Виды и часы аудиторной работы, их трудоемкость (в часах)			Текущие формы контроля
				Лекции	Практические занятия	Лабораторные работы	
1.	Тема 1. Основные определения.	8		0	4	0	домашнее задание
2.	Тема 2. Классификация онтологий.	8		0	4	0	домашнее задание
3.	Тема 3. Область применения онтологий.	8		0	4	0	домашнее задание
4.	Тема 4. Онтологии верхнего уровня.	8		0	4	0	домашнее задание
5.	Тема 5. Прикладные онтологии.	8		0	4	0	контрольная работа
6.	Тема 6. Языки описания онтологий.	8		0	4	0	домашнее задание
7.	Тема 7. Инструментальные средства проектирования онтологий.	8		0	4	0	домашнее задание
8.	Тема 8. Лингвистическая онтология WordNet.	8		0	4	0	домашнее задание
9.	Тема 9. Информационно-поисковые тезаурусы.	8		0	4	0	домашнее задание

N	Раздел Дисциплины/ Модуля	Семестр	Неделя семестра	Виды и часы аудиторной работы, их трудоемкость (в часах)			Текущие формы контроля
				Лекции	Практические занятия	Лабораторные работы	
10.	Тема 10. Информационно-поисковые тезаурусы и автоматическая обработка текстов.	8		0	4	0	контрольная работа
.	Тема . Итоговая форма контроля	8		0	0	0	экзамен
	Итого			0	40	0	

4.2 Содержание дисциплины

Тема 1. Основные определения.

практическое занятие (4 часа(ов)):

Основные определения. Основные понятия: концепт, отношение, аксиома, онтология. Компоненты онтологии: классы, отношения, функции, аксиомы, примеры. Назначение онтологий.

Тема 2. Классификация онтологий.

практическое занятие (4 часа(ов)):

Классификация онтологий. Классификация по степени формальности. Спектр онтологий. Классификация по цели создания. Онтологии представления. Классификация по содержимому. Лексические онтологии. Онтологии для обработки текстов на естественном языке.

Тема 3. Область применения онтологий.

практическое занятие (4 часа(ов)):

Область применения онтологий. Проект Semantic Web. Использование онтологий в информационном поиске. Способы обработки запросов. Оценка качества информационного поиска. Методы улучшения качества информационного поиска на основе онтологических процессов. Интеграция разнородных источников данных. Спецификация содержимого разнородных источников данных. Концептуальный уровень репрезентаций. Логический уровень.

Тема 4. Онтологии верхнего уровня.

практическое занятие (4 часа(ов)):

Онтологии верхнего уровня. Отличительные черты онтологий верхнего уровня. Характеристика основных онтологий: OpenCyc, DOLCE, SUMO, Онтология Джона Совы, Верхние уровни WordNet.

Тема 5. Прикладные онтологии.

практическое занятие (4 часа(ов)):

Прикладные онтологии. Онтологии предметных областей. Онтология предметной документации в сфере культурного наследия: CIDOC CRM. Онтология товаров и услуг. Рубрикаторы. Онтология доступа к Web OntoSeek.

Тема 6. Языки описания онтологий.

практическое занятие (4 часа(ов)):

Языки описания онтологий. Основные синтаксические структуры: классы, отношения, аксиомы. Архитектура метаданных в WWW: данные, метаданные и связи. Форма метаданных. Пространство имен атрибутов. Связи. Язык RDF. Модель данных RDF. RDF-граф. RDF-литералы. Сравнение литералов. Определение значения типизированного литерала. Языки представления

Тема 7. Инструментальные средства проектирования онтологий.

практическое занятие (4 часа(ов)):

Инструментальные средства проектирования онтологий. Редакторы онтологий. Protégé. Поддерживаемые редактором формализмы и форматы представления. Функциональность. Модель знаний Protégé. Пользовательский интерфейс. Вкладки. Сравнение редакторов.

Тема 8. Лингвистическая онтология WordNet.

практическое занятие (4 часа(ов)):

Лингвистическая онтология WordNet. Описание ресурса. EuroWordNet. Основные принципы. Описание существительных. Синонимы, гипонимы, синсеты. Отношение Часть-целое. Описание прилагательных. Описание глаголов. Отношения между синсетами глаголов. Тропонимия. Отношение причины. Индекс ILI. Онтология верхнего уровня в WordNet. Применение WordNet в информационном поиске. Векторная модель поиска. Расширение запросов на основе WordNet. Проект Meaning. Семантическое индексирование в рамках Meaning.

Тема 9. Информационно-поисковые тезаурусы.

практическое занятие (4 часа(ов)):

Информационно-поисковые тезаурусы. Разработка, создание и использование традиционных информационно-поисковых тезаурусов. Назначение тезаурусов. Дескрипторы. Семантические отношения в тезаурусах: иерархические, ассоциативные. Автоматическое индексирование по традиционным тезаурусам. Сочетание свободных запросов и запросов на основе информационно-поисковых тезаурусов.

Тема 10. Информационно-поисковые тезаурусы и автоматическая обработка текстов.

практическое занятие (4 часа(ов)):

Информационно-поисковые тезаурусы и автоматическая обработка текстов. Специфика тезаурусов для автоматической обработки текстов. Примеры тезаурусов. Современные подходы к определению отношений (сравнение). Условия надежности. Автоматическое концептуальное индексирование. Алгоритм АЛОТ. Сеть тематических узлов. Механизмы выделения основных узлов. Тематическая аннотация. Связная аннотация. Автоматическая рубрикация текстов. Критерии оценки качества. Методы автоматического рубрицирования. Применение методов машинного обучения.

4.3 Структура и содержание самостоятельной работы дисциплины (модуля)

N	Раздел Дисциплины	Семестр	Неделя семестра	Виды самостоятельной работы студентов	Трудоемкость (в часах)	Формы контроля самостоятельной работы
1.	Тема 1. Основные определения.	8		подготовка домашнего задания	2	домашнее задание
2.	Тема 2. Классификация онтологий.	8		подготовка домашнего задания	2	домашнее задание
3.	Тема 3. Область применения онтологий.	8		подготовка домашнего задания	4	домашнее задание
4.	Тема 4. Онтологии верхнего уровня.	8		подготовка домашнего задания	4	домашнее задание
5.	Тема 5. Прикладные онтологии.	8		подготовка к контрольной работе	4	контрольная работа
6.	Тема 6. Языки описания онтологий.	8		подготовка домашнего задания	4	домашнее задание

N	Раздел Дисциплины	Семестр	Неделя семестра	Виды самостоятельной работы студентов	Трудоемкость (в часах)	Формы контроля самостоятельной работы
7.	Тема 7. Инструментальные средства проектирования онтологий.	8		подготовка домашнего задания	4	домашнее задание
8.	Тема 8. Лингвистическая онтология WordNet.	8		подготовка домашнего задания	4	домашнее задание
9.	Тема 9. Информационно-поисковые тезаурусы.	8		подготовка домашнего задания	2	домашнее задание
10.	Тема 10. Информационно-поисковые тезаурусы и автоматическая обработка текстов.	8		подготовка к контрольной работе	2	контрольная работа
	Итого				32	

5. Образовательные технологии, включая интерактивные формы обучения

Обучение происходит в форме лекций, лабораторных занятий, а также самостоятельной работы студентов.

6. Оценочные средства для текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины и учебно-методическое обеспечение самостоятельной работы студентов

Тема 1. Основные определения.

домашнее задание , примерные вопросы:

Углубленное изучение литературы. Выполнение заданий по изучаемой теме.

Тема 2. Классификация онтологий.

домашнее задание , примерные вопросы:

Углубленное изучение литературы. Выполнение заданий по изучаемой теме.

Тема 3. Область применения онтологий.

домашнее задание , примерные вопросы:

Углубленное изучение литературы. Выполнение заданий по изучаемой теме.

Тема 4. Онтологии верхнего уровня.

домашнее задание , примерные вопросы:

Углубленное изучение литературы. Выполнение заданий по изучаемой теме.

Тема 5. Прикладные онтологии.

контрольная работа , примерные вопросы:

Проведение контрольной работы в виде теста с ответами на следующие вопросы: 1. Что такое онтология? Составные части онтологий 2. Чем отличаются онтологии верхнего уровня от онтологий предметной области? 3. Чем отличаются онтологии предметной области от прикладных онтологий? 4. Перечислите основные характеристики лексических онтологий. 5. Перечислите известные Вам проекты онтологий верхнего уровня. 6. Что такое универсалии? 7. Чем существенно отличается отношение НАДКЛАСС-ПОДКЛАСС от ЧАСТЬ-ЦЕЛОЕ? 8. Перечислите традиционные подходы к обработке запроса. В чем их недостатки? 9. Чем критерий полноты отличается от критерия точности? 10. Назовите способы улучшения поиска при помощи тезаурусов и онтологий. 11. Перечислите основные элементы ER-модели. 12. В чем проявляется ?интеллектуальность? агентов Semantic Web? 13. Что такое рубрикатор? Использование рубрикаторов в интернет-системах по товарам и услугам 14. Система Ontoseek: какие проблемы пословного поиска, и какими средствами предполагалось решать? 15. Почему AND-список высказываний о любом ресурсе может быть представлен неупорядоченным множеством?

Тема 6. Языки описания онтологий.

домашнее задание , примерные вопросы:

Углубленное изучение литературы. Выполнение заданий по изучаемой теме.

Тема 7. Инструментальные средства проектирования онтологий.

домашнее задание , примерные вопросы:

Углубленное изучение литературы. Выполнение заданий по изучаемой теме.

Тема 8. Лингвистическая онтология WordNet.

домашнее задание , примерные вопросы:

Углубленное изучение литературы. Выполнение заданий по изучаемой теме.

Тема 9. Информационно-поисковые тезаурусы.

домашнее задание , примерные вопросы:

Углубленное изучение литературы. Выполнение заданий по изучаемой теме.

Тема 10. Информационно-поисковые тезаурусы и автоматическая обработка текстов.

контрольная работа , примерные вопросы:

Углубленное изучение литературы. Выполнение заданий по изучаемой теме.

Тема . Итоговая форма контроля

Примерные вопросы к экзамену:

По данной дисциплине предусмотрено проведение зачета. Примерные вопросы для зачета - Приложение1.

ВОПРОСЫ НА ЗАЧЕТ:

1. Что такое онтология? Составные части онтологий
2. Чем отличаются онтологии верхнего уровня от онтологий предметной области?
3. Чем отличаются онтологии предметной области от прикладных онтологий?
4. Перечислите основные характеристики лексических онтологий.
5. Перечислите известные Вам проекты онтологий верхнего уровня.
6. Что такое универсалии?
7. Чем существенно отличается отношение НАДКЛАСС-ПОДКЛАСС от ЧАСТЬ-ЦЕЛОЕ?
8. Перечислите традиционные подходы к обработке запроса. В чем их недостатки?
9. Чем критерий полноты отличается от критерия точности?
10. Назовите способы улучшения поиска при помощи тезаурусов и онтологий.
11. Перечислите основные элементы ER-модели.
12. В чем проявляется "интеллектуальность" агентов Semantic Web?
13. Что такое рубрикатор? Использование рубрикаторов в интернет-системах по товарам и услугам

14. Система Ontoseek: какие проблемы пословного поиска, и какими средствами предполагалось решать?
15. Почему AND-список высказываний о любом ресурсе может быть представлен неупорядоченным множеством?
16. Чем отличаются понятия ресурс, объект и документ в контексте Web?
17. Что такое RDF? Что представляет собой модель данных RDF и на чем она основана?
18. Для чего нужен RDFS?
19. Что такое реификация?
20. Чем отличается класс RDFS от класса OWL?
21. Перечислите известные Вам редакторы онтологий. Какой формализм является основным для редактора Protege?
22. Как называются элементарные структурные единицы WordNet? Перечислите основные отношения в WordNet.
23. Что такое концептуальное индексирование и концептуальный поиск?
24. Какие проблемы использования онтологии в информационном поиске?
25. Основные этапы работы вопросно-ответной системы. Как можно использовать онтологию в вопросно-ответной системе?
26. Обработка булевского запроса в вопросно-ответной системе
27. Какие проблемы возникают при использовании WordNet для автоматической обработки текста?
28. Опишите проблему лексической многозначности.
29. Как в WordNet происходит разрешение многозначности?
30. Перечислите основные виды отношений в ИПТ.
31. Почему традиционные ИПТ мало используются для автоматического индексирования текстов.
32. Методы использования традиционных ИПТ в автоматических технологиях обработки текстов (запросов).
33. В чем состоят отличительные особенности Тезауруса для автоматического концептуального индексирования?
34. Каковы возможные способы установление отношений в тезаурусах?
35. Что такое отношения онтологической зависимости?
36. Перечислите этапы автоматической обработки текстов на основе Тезауруса.
37. Как моделируется связность текста?
38. Каков принцип построения связной аннотации текста?
39. Перечислите методы автоматической рубрикации
40. По каким причинам возникают сложности в задачах автоматической рубрикации текстов?

Примерные вопросы к текущим формам контроля

1. Перечислить составные части онтологий
2. Перечислить отличия онтологий верхнего уровня от онтологий предметной области?
3. Отличия онтологии предметной области от прикладных онтологий?
4. Лексических онтологии и их характеристики.
5. Дать обзор по проектам онтологий верхнего уровня.
6. Основные этапы автоматической обработки текстов на основе Тезауруса.
7. Принципы моделирования связности текста.
8. Принцип построения связной аннотации текста.
9. Методы автоматической рубрикации
10. Причины возникновения сложности в задачах автоматической рубрикации текстов?

7.1. Основная литература:

1. Саттон, Р. С. Обучение с подкреплением [Электронный ресурс] / Р. С. Саттон, Э. Г. Барто ; пер. с англ. - Эл. изд. - М.: БИНОМ. Лаборатория знаний, 2012. - 399 с.

http://e.lanbook.com/books/element.php?pl1_id=4405

2. Леонтьева, Нина Николаевна. Автоматическое понимание текстов: системы, модели, ресурсы : учеб. пособие для студ. лингв. фак. вузов / Н. Н. Леонтьева .? М. : Академия, 2006 .? 304 с. ? Рекомендовано УМО

3. Введение в теорию алгоритмических языков и компиляторов: учеб. пособие / Л.Г. Гагарина, Е.В. Кокорева. - М.: ИД ФОРУМ, 2011. - 176 с.: ил.; 60x90 1/16. - (Высшее образование). (переплет) ISBN 978-5-8199-0404-6, 1000

http://znanium.com/bookread.php?book=265617

4.Интеллектуализация сетевых систем поиска экономической информации: Монография / А.Н. Романов, Б.Е. Одинцов. - М.: Вузовский учебник: ИНФРА-М, 2010. - 144 с.: 60x90 1/16. - (Научная книга). (переплет) ISBN 978-5-9558-0156-8, 1000

http://znanium.com/bookread.php?book=189601

7.2. Дополнительная литература:

1.Теория синтаксического анализа перевода и компиляции / А. Ахо, Дж. Ульман .? М. : Мир, 1978. Т.1: Синтаксический анализ .? М. : Мир, 1978 .? 612с.

2.Теория синтаксического анализа перевода и компиляции / А. Ахо, Дж. Ульман .? М. : Мир, 1978. Т.2: Компиляция .? М. : Мир, 1978 .? 487с.

7.3. Интернет-ресурсы:

Semantic Web - http://www.w3.org/2001/sw/

swoogle.umbc.edu - http://swoogle.umbc.edu/

Диалог 21 - http://www.dialog-21.ru/

Справочник по онтологиям - http://www.xml.com/pub/a/2002/11/06/ontologies.html.

Форум по Semantic Web - http://wonderweb.semanticweb.org/.

8. Материально-техническое обеспечение дисциплины(модуля)

Освоение дисциплины "Разработка тезаурусов и антологий" предполагает использование следующего материально-технического обеспечения:

Лекционные занятия по дисциплине проводятся в аудитории, оснащенной доской и мелом(маркером), а так же в специализированных компьютерных кабинетах.

Программа составлена в соответствии с требованиями ФГОС ВПО и учебным планом по направлению 010400.62 "Прикладная математика и информатика" и профилю подготовки Математическое и программное обеспечение вычислительных машин и сетей .

Автор(ы):
Соловьев В.Д. _____
"__" _____ 201 __ г.

Рецензент(ы):
Иванов В.В. _____
"__" _____ 201 __ г.