Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д.

Онтологии и тезаурусы

1.1. Определение понятий: онтология, концепт, отношение, аксиомы

Коллекции электронных документов и задачи их автоматической обработки

- **Миллионы текстов в электронной форме**
- **Множество разнообразных насущных задач по автоматической обработке электронных документов**
- **Но: для решения этих задач используются пословные статистические методы ("bag of words" models)**
- **Information retrieval community: текст – это набор features, закономерности которых хорошо учитываются статистическими методами**

Онтологии. Концептуальное индексирование

- Ресурс для автоматического индексирования.
- Индекс: не слова, а понятия.
  - Многозначные слова разведены к разным понятиям
  - Синонимы приводят к одному понятию
  - Отношения могут использоваться для расширения или уточнения запроса

Онтологии. Semantic Web (2001)
Тим Бернес-Ли, Джеймс Хендлер, Ора Лассила

- Страницам сайта приписана некоторое формально описание, которое помогают автоматическим процессам в сети взаимодействовать
- RDF (Resource Description Framework)
- Web Ontology Language (OWL)
- Единицы описаний – из Онтологий
- «Сеть наполнится семантикой»

Онтология: 2 значения

- Философская дисциплина изучает наиболее общие характеристики бытия и сущностей

- Онтология – артефакт, структура, описывающая значения элементов некоторой системы

Онтология (артефакт)

- Неформально, онтология представляет собой некоторое описание взгляда на мир применительно к конкретной области интересов.
- Это описание состоит из терминов и правил использования этих терминов, ограничивающих их значения в рамках конкретной области

Онтология (3)

- На формальном уровне, онтология это система, состоящая из набора понятий и набора утверждений об этих понятиях, на основе которых можно строить классы, объекты, отношения, функции и теории.
- Основные компоненты:
  - Классы или понятия
  - Отношения
  - Функции
  - Аксиомы
  - Примеры

Онтология – спецификация концептуализации (Gruber)

- Концептуализация – структура реальности, независимо от
  - Словаря
  - Конкретной ситуации

  - Кубики на столе: концептуализация: - набор возможных положений, но не конкретное расположение

Онтологией могут быть:

- Глоссарий
- Простая таксономия
- Тезаурус
- Понятийная структура с произвольным набором отношений
- Структура с аксиоматикой

Таксономические отношения

- Варианты названий:
- Is_a – отношение
- Класс - подкласс
- Лингвистика: гипоним – гипероним
- Родовидовое отношение

Свойства таксономических отношений

- Транзитивность: A is_a B, B is_a C,
  - => A is_a C
- Наследование:
  - S= свойство (A)
  - B is_a A
  - => S=свойство (B)

Инициатива (КА)

(KA)2
Knowledge Annotation Initiative of the Knowledge Acquisition Community
(http://www.aifb.uni-karlsruhe.de/Projekte/viewProjektenglish?id_db=4)

- Предметная область разработки – сообщество специалистов по приобретению знаний
- Несколько таксономий: people, publications, events, organizations, research topics

Таксономия публикаций

- Publication
- Article
  - Article in book
  - Conference paper
  - …
- Book
- Journal
  - IEEE expert

Отношения в (КА)

Employee

Head-of-project Project

Works-on-Project Project

Affiliation Organization

Head-of-group Research group

Пример аксиомы

- Работник, являющийся руководителем проекта, работает в проекте
- Переменные E, P
- Forall (E,P) Employee (E) and Head-Of-Project (E,P) => Works-At-Project (E,P)

Языки для описания онтологий

- Ontobroker
- CycL
- Description Logics
- RDF/RDFS
- OWL

Ontobroker

- Подклассы (Subclass): C1::C2 – класс C1 является подклассом C2
- Экземпляры (Instance of): O:C – O является экземпляром C
- Описания атрибутов (Attribute Declaration): C1 [A=>>C2] – для экземпляра класса C1 определен атрибут A, значением которого должен быть экземпляр класса C2

Ontobroker - 2

- Значения атрибутов (Attribute value):
- O [A->>V] – Экземпляр O имеет атрибут A со значением V
- Часть-Целое (Part-of) – O1<:O2 – O1 является частью O2
- Отношения (Relations) предикаты вида p(a1, …an)

Ontobroker - 3

- Запрос
- Forall Obj, FN, EM <-
  - Obj: Researcher [firstName->>FN;
  - Lastname->>»Иванов»; email->>EM].

Типы онтологий

- Общие
- Предметно-ориентированные

- Различаются по способу применения
- Онтологии для автоматического анализа текста

Проблемы построения общих онтологий: верхние уровни

- Верхние уровни в разных онтологиях: CYC, EuroWordNet, WordNet
- Сравнение. Почему они различаются
- Критический анализ Nicola Guarino и предложения, как нужно строить верхний уровень
- Онтология SUMO

Онтология CYC

- Lenat D.
- Самый амбициозный проект
- Начат в 1984
- 1 млн. утверждений “common sense”
- Микротеории: пространство, время, причинность
- Онтология 3 тысяч понятий верхнего уровня – в открытом доступе
- www.cyc.com

Лингвистические онтологии

- The main characteristic of this kind of ontologies is that they are bound to the semantics of grammatical units (words, nominal groups, etc)

- Основной источник понятий в онтологии
  – значения языковых единиц

- Лингвистические онтологии:
  WordNet, Mikrokosmos, Sensus, PyTez

WordNet

- **Реляционное описание лексики английского языка**
- **Иерархическая сеть понятий (synset)**
- **Каждое слово относится к одному или нескольким понятиям**
- **Отдельная иерархическая сеть для различных частей речи – психолингвистическое обоснование**
- **Автор: George Miller (50-е годы статья «Магическое число 7»)**
- **Версия 1.6: 95 тысяч понятий, около 130 тысяч слов и понятий**

EuroWordNet

- Структурные лингвистические ресурсы
- Интерлингва:
  английский WordNet
- Первоначально:
  испанский, итальянский, голландский
- Далее:
  немецкий, французский, чешский, эстонский
- Известны попытки создать свои структурные ресурсы на базе WordNet:
  японский, болгарский, румынский, шведский и др.

Онтология MikroKosmos

New Mexico State University

Nierenburg Sergey

5 тысяч понятий

Автоматический перевод английский – испанский

Узкая предметная область: слияния предприятий

Тезаурус русского языка РуТез

- Ресурс для автоматической обработки текстов
- Содержит общезначимые лексические единицы и терминологию общественно политической области – 115 тысяч слов и выражений
- Иерархическая сеть

МГУ им. М.В.Ломоносова
Научно-исследовательский
вычислительный центр

АНО Центр
информационных
исследований

Университетская информационная система РОССИЯ

УИС РОССИЯ университетская информационная система РОССИЯ

Главная | Поиск | Статистика РФ | Бюджет РФ | Выборы РФ | Парламент РФ | Соционет/RePEc | OECD | Технологии | Справка/Практикум | УИС РОССИЯ v3 | UIS RUSSIA v3 |

USER: BORIS
Доступ: CIR
Имя: Пароль:

Регистрация
Забыли пароль?

WWW.CIR.RU

Поиск по ИСТОЧНИКАМ : все коллекции (CIR)

искать ▶ | очистить | история

добавление условия в запрос

AND | Тезаурус ЦИИ | добавить | список...

спец. атрибуты | свернуть

Выберите коллекции докуме... Переход по ссылке с имени ... коллекции.

- ☑ ***Все коллекции*** свернуть/
  - ☑ ***Издания государст***
    - ☑ НТЦ "Система". Норма
    - ☑ НТЦ "Система". Между
    - ☑ Государственная Дума описание
    - ☑ Государственная Дума
    - ☑ Совет Федерации ФС Р
    - ☑ Госкомстат России. Еж
    - ☑ Госкомстат России. Кр
    - ☑ Госкомстат России. Со описание
    - ☑ Срочная информация
    - ☑ Министерство экономи описание
    - ☑ Банк России. Вестник Б
    - ☑ Счетная палата РФ. Бк
    - ☑ Межгосударственный 1998 года) описание
    - ☑ Нормативно-правовые

- ☑ ***Средства массовой информации*** свернуть список | описание
  - ☑ Аргументы и Факты *(21845 статей, с 1997 года)* описание
  - ☑ Известия *(53937 статей, с 2000 года)* описание
  - ☑ Финансовые Известия *(1683 документа, с 2001 года)* описание
  - ☑ Ведомости *(54022 статьи, с 1999 года)* описание
  - ☑ Комсомольская правда *(43462 статьи, с 1999 года)* описание
  - ☑ Независимая газета *(90848 статей, с 1998 года)* описание
  - ☑ Слово *(2502 статьи, с 1999 года)* описание
  - ☑ Сегодня *(17072 статьи, с 2000 года по апрель 2001 года. Издание прекращено)* описание
  - ☑ Региональный пресс-бюллетень агентства ВПС *(11883 статьи, с 1990 года по январь 2001 года. Издание прекращено)* описание
  - ☑ Журнал "Эксперт" *(13189 статей, с 2001 года)* описание
  - ☑ Газета "Поиск" *(3142 статьи, с 2002 года)* описание
  - ☑ Материалы агентства Reuters *(21578 документов)*
- ☑ ***Издания исследовательских центров*** свернуть список | описание
  - ☑ РЕЦЭП. Обзор экономики России *(2265 документов, с 1996 года)* описание
  - ☑ Экономическая экспертная группа. Обзор экономических показателей *(1163 документа, с 1998 года)* описание
  - ☑ Фонд "ИПН". Индекс потребительских настроений *(355 документов, с 1998 года)* описание
  - ☑ Бюро экономического анализа. Доклады *(493 документа, с 2001 года)* описание
  - ☑ НИСП. Доклады по программе "Социальная политика: Накануне 21 века. Реалии 21 века" *(208 документов, с 2000 года)* описание
  - ☑ РЕЦЭП. Обзор экономики России *(2131 документ, с 1996 года)*
  - ☑ Аналитические материалы. (по бюджетным вопросам) *(893 документа)* описание
  - ☑ Демоскоп Weekly *(6264 статьи, с 2001 года)* описание
- ☑ ***Научные издания*** свернуть список | описание
  - ☑ Социологический журнал *(103 статьи, с 1999 года)* описание
  - ☑ Журнал "Полис" *(887 статей, с 1991 года)* описание
  - ☑ Вестник МГУ. Серия 6. Экономика *(194 статьи, с 1998 года)* описание
  - ☑ Вестник МГУ. Серия 7. Философия *(123 статьи, с 1999 года)* описание
  - ☑ Вестник МГУ. Серия 8. История. *(118 статей, с 1999 года)* описание
  - ☑ Вестник МГУ. Серия 9. Филология *(104 статьи, с 1998 года)* описание
  - ☑ Вестник МГУ. Серия 10. Журналистика *(244 статьи, с 1999 года)* описание
  - ☑ Вестник МГУ. Серия 12. Политические Науки *(180 статей, с 1999 года)* описание

Лингвистические онтологии и информационный поиск

- Электронные коллекции разнообразных текстов
- Возможности систем автоматической обработки текста для анализа релевантности контекста ограниченны
- Нет возможности подробно проанализировать контекст упоминания понятия в тексте.
- Онтологии специального типа?

Многоязычные онтологии

- EuroWordNet
- MikroCosmos
- RuThes содержит двуязычный ресурс Общественно-политический тезаурус (75 тысяч русский терминов – 70 тысяч англоязычных)
- Чем установления языковых соответствий отличается в традиционных словарях и онтологиях

Онтологии и вопросно-ответные системы

- Система ищет в сверхбольшой текстовой коллекции
- Сравнение систем в соревновании TREC и CLEF
- Конкретные системы
- Практическая актуальность: поиск в Интернет не по краткому запросу, а по развернутому вопросу

Онтологии и вопросно-ответные системы

Постановка задачи:

- 60-е годы: поиск в специальных базах знаний
- Сейчас: поиск в громадных текстовых массивах

Примеры вопросов:

- *What does the Peugeot company manufacture?*
- *How long did the Charles Manson Murder trial last?*
- *Who is the first American in space?*

Как создать онтологию для конкретной области

- Тексты
- Набор словосочетаний: автоматическое извлечение терминов
- Выделение понятия
- Отношения между понятиями:
  - Извлечение из текстов по шаблонам
  - Статистические методы
  - Методы на основе синтаксической структуры

Как использовать созданные онтологии

- Слияние онтологий
- Использование общих онтологий для эффективного создания онтологий в конкретных предметных областях
- Semantic web: одна (или несколько) онтология верхнего уровня, к которой достраиваются специализированные онтологии

Вопросы к лекции

- Что такое онтология?
- Составные части онтологий
- Классификация онтологий