

Автоматическое построение ОНТОЛОГИИ ПО КОЛЛЕКЦИИ ТЕКСТОВЫХ ДОКУМЕНТОВ

*Мозжерина Елена Станиславовна
аспир., ассистент каф. Технологии Программирования
ф-та ПМ-ПУ, СПбГУ*

Информационный взрыв

- ~ М. Hilbert & P. Lopez (2011). 'The World's Technological Capacity to Store, Communicate, and Compute Information', *Science Magazine*.
- ~ 1986 год и 2007 год в сравнении
 - ~ Объем хранимой информации:
2,6 эксабайт $\rightarrow$ 295 эксабайт
 - ~ Объем хранимой текстовой информации:
8,7 тыс. терабайт $\rightarrow$ 19,4 тыс. терабайт

Классические подходы к информационному поиску

- ~ Логические модели
- ~ Векторные модели
- ~ Вероятностные модели

Семантический поиск

- ~ Полнотекстовый автоматизированный информационный поиск, проводимый с учетом смыслового содержания запроса пользователя и текстов информационных ресурсов

Онтологический подход

Онтологический подход

?

Онтологический подход

?



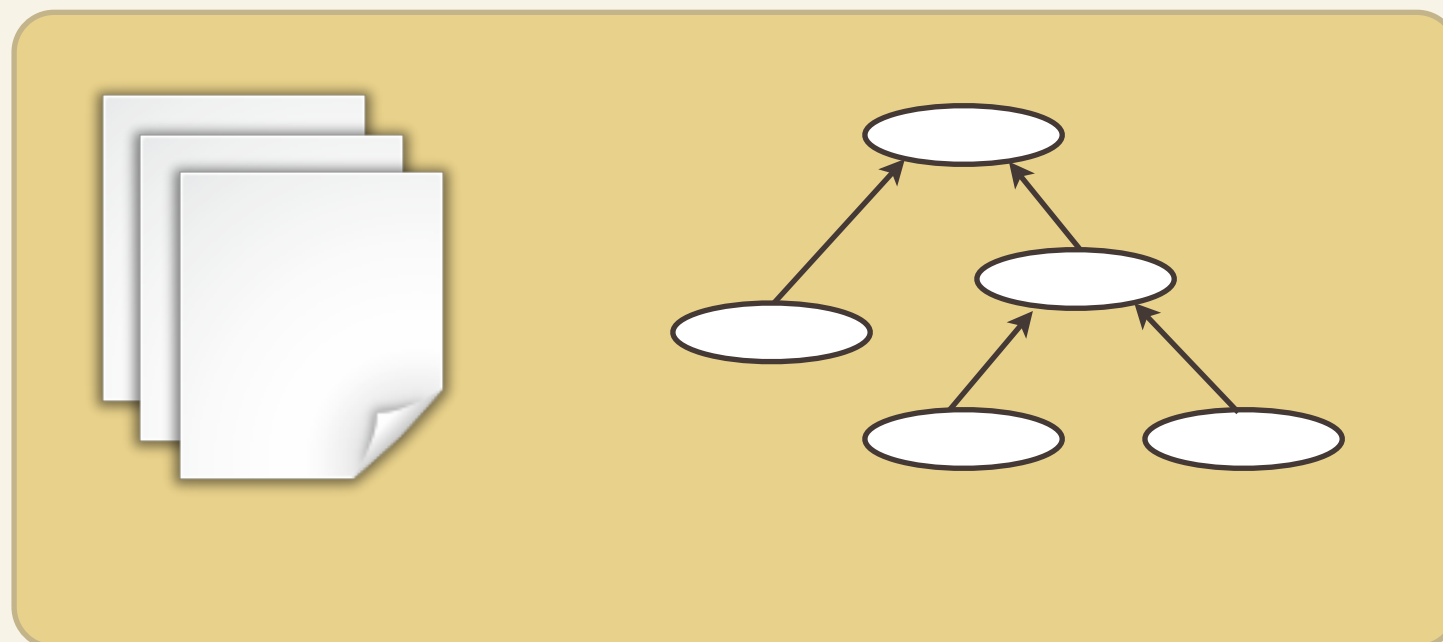
Онтологический подход

?



Онтологический подход

?



Онтологии для семантического поиска

- ~ Использование существующих онтологий верхнего уровня
- ~ Использование онтологий, специально разработанных экспертами для данной предметной области

Цель доклада

- ~ Предложить метод автоматизации процесса построения онтологии по коллекции текстовых документов, относящихся к одной предметной области, с использованием статистических методов анализа естественно-языковых текстов

Понятие онтологии

- ~ $O = \langle T, R, F \rangle$
- ~ T – конечное множество терминов
(концептов, понятий, классов)
предметной области
- ~ R – конечное множество отношений
между понятиями заданной области
- ~ F – конечное множество функций
интерпретации

Состав лингвистической онтологии

- ~ Термины предметной области (классы) и их свойства (атрибуты)
- ~ Экземпляры классов (с конкретными значениями свойств)
- ~ Отношения между классами и отдельными элементами

Этапы разработки онтологии

- ~ Определение классов
- ~ Определение отношений и построение иерархии классов
- ~ Определение слотов и описание допустимых значений для слотов
- ~ Заполнение значений слотов для экземпляров

Построение онтологии на примере

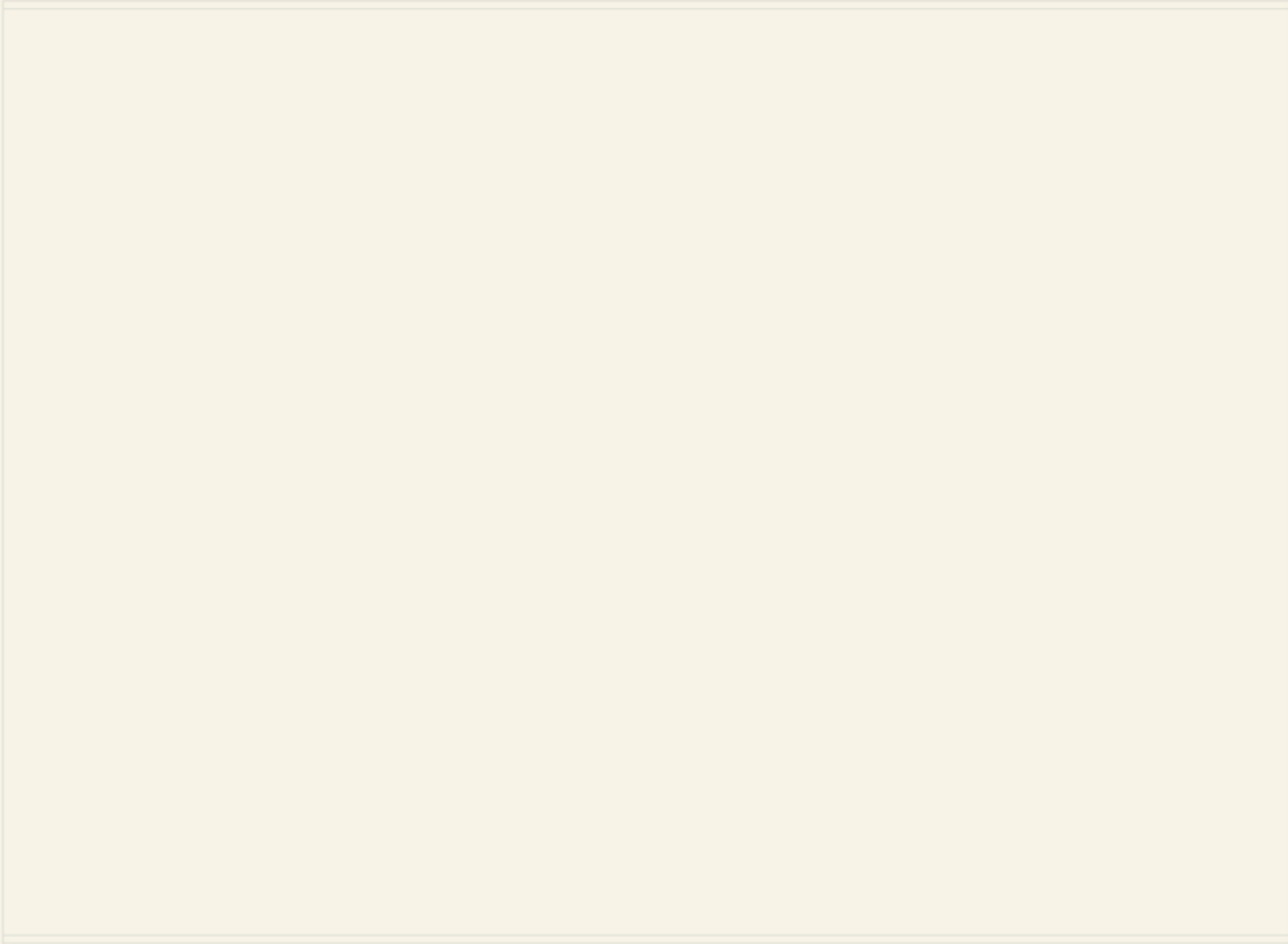
- ~ Музыкальный инструмент флейта относится к ряду деревянных духовых инструментов. Флейты представляют собой целое семейство инструментов (блок флейта, пикколо). Современные флейты изготавливают из дерева, металла, пластика. Профессиональные музыканты предпочитают инструменты производителя Yamaha.

Построение онтологии на примере

~ Музыкальный инструмент флейта
относится к ряду деревянных духовых
инструментов. Флейты представляют
собой целое семейство инструментов
(блок флейта, пикколо). Современные
флейты изготавливают из дерева,
металла, пластика. Профессиональные
музыканты предпочитают инструменты
производителя Yamaha.

Построение онтологии на примере

~ Музыкальный инструмент флейта
относится к ряду деревянных духовых
инструментов. Флейты представляют
собой целое семейство инструментов
(блок флейта, пикколо). Современные
флейты изготавливают из дерева,
металла, пластика. Профессиональные
музыканты предпочитают инструменты
производителя Yamaha.



Музыкальный инструмент

Музыкальный инструмент

is-a ↑

Флейта

:материал

:производитель





Методы построения онтологии

- ~ Лингвистические методы: лексико-синтаксические шаблоны
(Е.А. Рабчевский)
- ~ С использованием подходов из области искусственного интеллекта: система
продукций (Л.В. Найханова)
- ~ Статистические методы

Этапы построения онтологии

- ~ Предварительная подготовка коллекции
- ~ Определение классов онтологии
- ~ Определение отношений «is-a» и «synonym-of», построение иерархии классов

Предварительная подготовка коллекции

- ~ Приведение документов к единому формату
- ~ Подготовка документов
- ~ Кластеризация коллекции

Определение классов онтологии

- ~ Имя класса содержит хотя бы одно существительное (I)
- ~ Общеупотребительные слова обладают большей частотой встречаемости, и приблизительно равной в документах из различных кластеров (II)

Определение классов онтологии

- ~ Определение терминов из одного слова
- ~ Определение терминов из нескольких слов

Построение иерархии классов

- ~ Количество информации термина на более низком уровне иерархии больше, чем количество информации более общего термина (III)

Построение иерархии классов

- ~ $C(t)$ – контекст термина t – множество слов, появляющихся совместно с данным термином
- ~ $C(t) = P(Y|t)$,
где Y принимает значения терминов
данного кластера

Построение иерархии классов

- ~ Пусть A, B – отдельные слова. Если t_1 имеет вид « A » и t_2 имеет вид « AB » или « CA », то « t_2 is-a t_1 »
- ~ Если $I(t_2) < I(t_1)$ и $C(t_1) \approx C(t_2)$, то « t_2 is-a t_1 »
- ~ Если $I(t_2) \approx I(t_1)$ и $C(t_1) \approx C(t_2)$, то « t_2 synonym-of t_1 »

Отличительные особенности предлагаемого подхода

- ~ Работа с предварительно кластеризованной коллекцией
- ~ Расширение статистического метода эвристиками

Оценка качества полученной ОНТОЛОГИИ

- ~ Экспертная оценка
- ~ Опосредованно, через другие приложения
- ~ Сравнение с существующими

Планы развития подхода

- ~ Расширение подхода для выделения других отношений
- ~ Расширение подхода на оставшиеся этапы определения слотов и экземпляров
- ~ Экспорт разработанной онтологии в различные инструментальные системы

Спасибо за внимание!

Елена Станиславовна Мозжерина
mozzherina@gmail.com

~ Контекст
работы

~ Цель доклада

~ Понятие
онтологии

~ Пример

~ Методы и этапы

~ Описание
подхода

~ Особенности

~ Планы по
развитию

~ Дополнения:
коллекция
эвристики

Ohsumed–20000

- ~ Медицинские аннотации из MeSH за 1991 год
- ~ 23 категории сердечно-сосудистых заболеваний
- ~ Число документов 13 929
 - ~ Тренировочная коллекция 6 286
 - ~ Коллекция для тестирования 7 643

Эвристики

- ~ Имя класса содержит хотя бы одно существительное (I)
- ~ Общеупотребительные слова обладают большей частотой встречаемости, приблизительно равной в документах из различных кластеров (II)
- ~ Количество информации термина из нескольких слов больше, чем количество информации отдельных слов, входящих в его состав (III)