Иерархические вероятностные тематические модели

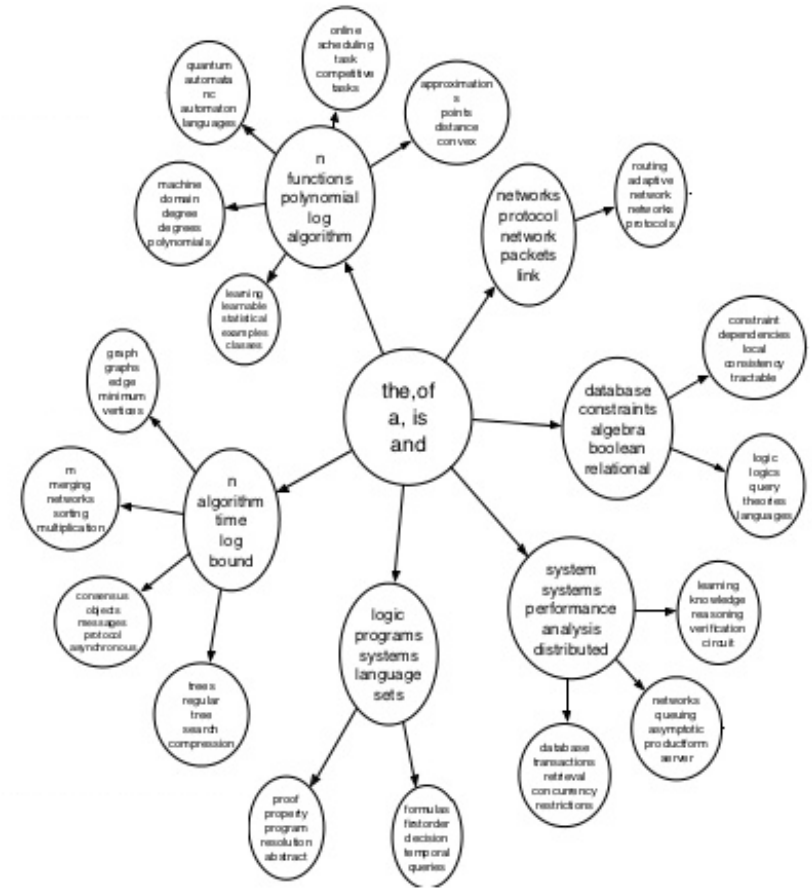
Надежда Чиркова

57 научная конференция МФТИ, 29.11.2014

План выступления

1. Плоские вероятностные тематические модели
   - Постановка задачи
   - Классический подход к решению плоской задачи
2. Иерархические модели
   - Постановка задачи
   - Подход к построению иерархии
   - Требования к темам на разных уровнях
   - Регуляризация тематических моделей
   - Результат

Иерархическая тематическая модель — это дерево тем.
Каждая тема — это набор слов, авторов, документов + множество подтем



✓Тематические иерархии легко интерпретируемы и понятны человеку

David M Blei, Probabilistic Topic Models, 2011

Постановка задачи тематического моделирования

D - коллекция текстовых документов

W- множество терминов

Дана коллекция текстовых документов:

n_{dw} - матрица частот слов в документах (мешок слов):

$$F_{dw} = p(w|d) = \frac{n_{dw}}{n_d}$$

Найти модель

$$p(w|d) = \sum_t p(w|t)p(t|d) = \sum_t \varphi_{wt}\theta_{td} \Leftrightarrow F = \Phi\Theta$$

с параметрами $\Phi = \{\varphi_{wt}\}_{W\times T}$ и $\Theta = \{\theta_{td}\}_{T\times D}$:

$\varphi_{wt} = p(w|t)$ — распределение слов в теме t;

$\theta_{td} = p(t|d)$ — распределение тем в документе d.

Критерий: максимизация логарифма правдоподобия:

$$\mathscr{L}(\Phi,\Theta) = \sum_{d,w} n_{dw} ln \sum_t \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

Классический подход к решению плоской задачи

PLSA — Probabilistic Latent Semantic Analysis

Основан на итерационном алгоритме:

EM-алгоритм

E-шаг: $p(t|d,w) = \dfrac{p(w|t)p(t|d)}{\sum_{s\in T} p(w|s)p(s|d)} = \dfrac{\varphi_{wt}\theta_{td}}{\sum_{s\in T}\varphi_{ws}\theta_{sd}}$;

$$n_{dw}^t = n_{dw}p(t|d,w);$$

M-шаг: $\varphi_{wt} = \dfrac{n_{wt}}{n_t} = \dfrac{\sum_d n_{dw}^t}{\sum_w\sum_d n_{dw}^t}$;

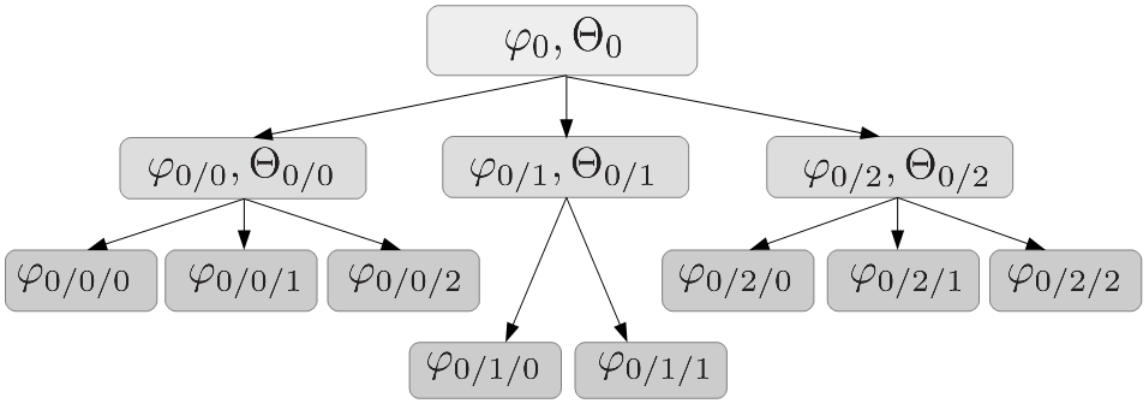
$$\theta_{td} = \frac{n_{td}}{n_d} = \frac{\sum_w n_{dw}^t}{\sum_t\sum_w n_{dw}^t}\ .$$

Постановка задачи для иерархической модели

В иерархической модели ставится дополнительная задача выделения взаимосвязей между темами.

Дана n_{dw} - матрица частот слов в документах.

Построить иерархию - дерево тем:



$$p(s) = p(s|t)p(t),$$
$$(s|d) = p(s|t)p(t|d),$$
$$p(w|t) = \sum_s p(w|s)p(s),$$
$$p(t) = \sum_s p(s|t),$$
t — topic,
s – subtopic

Используя эти формулы, можно привести локальные (относящиеся к отдельным вершинам) столбцы φ_t и матрицы Θ_t к глобальным (описывающим всю иерархию) Φ и Θ.

Критерий: максимизация логарифма правдоподобия:

$$\mathscr{L}(\Phi, \Theta) = \sum_{d,w} n_{dw} ln \sum_t \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

Подход к построению иерархии

Будем строить иерархию рекурсивно, от корня к листьям.

Функция обработки узла (темы)

Построить Узел (n_{dw}, T):

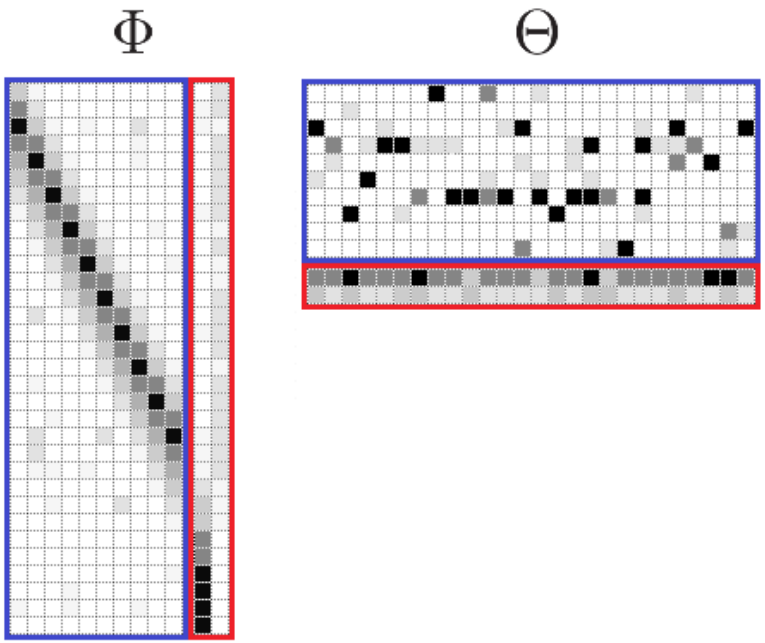
1. Построить плоскую модель — получить матрицы Φ и Θ;
2. Разделить входную коллекцию на T коллекций:

$$n_{dw} \Rightarrow n_{dw}^t, t = 1, \ldots, T$$

3. для всех $t = 1, \ldots, T$:
   Построить Узел (n_{dw}^t, T_t)

Количество тем (пока) задается вручную для каждого узла дерева.

Требования к темам на разных уровнях



- В каждой вершине выделяем небольшое число предметных тем-детей и несколько фоновых тем;
- В корневой вершине фоновая тема - это слова общей лексики языка, в некорневых вершинах - слова общей лексики данного уровня;
- Предметные темы декоррелированы между собой и с фоновыми темами;
- Предметные темы разрежены, т.е. содержат небольшое количество слов.

Регуляризация тематических моделей

Задача тематического моделирования некорректно поставлена: она имеет бесконечно много решений

$$\Phi\Theta = (\Phi S)(S^{-1}\Theta)$$

$\Rightarrow$ регуляризация — введение некоторого количества n дополнительных критериев $R_i(\Phi, \Theta), i = 1, \ldots, n$. Новые формулы M-шага:

$$\varphi_{wt} \propto \left(n_{wt} + \varphi_{wt} \frac{\partial R(\Phi, \Theta)}{\partial \varphi_{wt}} \right)_+, \theta_{td} \propto \left(n_{td} + \theta_{td} \frac{\partial R(\Phi, \Theta)}{\partial \theta_{td}} \right)_+,$$

где $R(\Phi, \Theta)$ - линейная комбинация регуляризаторов:
$R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)$.
Регуляризация позволяет удовлетворить требованиям к модели, предъявленным на предыдущем слайде.

Тематическая модель с регуляризаторами сглаживания и разреживания

Формулы регуляризованного M-шага
применительно к иерархической модели:

$$\varphi_{wt} \propto \Big(n_{wt} + \tau_1 \underbrace{\beta_w[t \in B]}_{\substack{\text{сглаживание} \\ \text{фоновых} \\ \text{тем}}} - \tau_2 \underbrace{\beta_w[t \in S]}_{\substack{\text{разреживание} \\ \text{предметных} \\ \text{тем}}} - \tau_3 \phi_{wt} \underbrace{\sum_{s \in S \setminus t} \varphi_{ws}}_{\text{декорреляция}} \Big)_+$$

$$\theta_{td} \propto \Big(n_{td} + \tau_4 \underbrace{\alpha_t[t \in B]}_{\substack{\text{сглаживание} \\ \text{фоновых} \\ \text{тем}}} - \tau_5 \underbrace{\alpha_t[t \in S]}_{\substack{\text{разреживание} \\ \text{предметных} \\ \text{тем}}} - \tau_6 \underbrace{\frac{n_d}{n_t}\theta_{td}}_{\substack{\text{удаление} \\ \text{малых тем}}} \Big)_+$$

,
B — множество фоновых тем, S — множество предметных тем,
$\tau_i, i = 1, \ldots, 6$ - коэффициенты регуляризации.

Результат

Навигатор по коллекции статей конференций «Интеллектуализация обработки данных» и «Математические методы распознавания образов»:

MMROnavigator.vv.si

тематический навигатор

Главная | Корневая тема | О проекте

Тематический навигатор

Тематический навигатор - это визуализация иерархической тематической модели, построенной по коллекции статей двух научных конференций:

- Интеллектуализация обработки информации
- Математические методы распознавания образов

Быстрая навигация по дереву тем:

- Корневая тема
  - Теоретические исследования
    - Логические методы анализа данных
      - Решающие деревья
      - Линейные локальные признаки
      - Алгоритмы
    - Классификация
      - Fris-функции
      - Теория переобучения
      - Классификаторы, метод ближайшего соседа
    - Теория риска, физика, химия
      - Теория риска
      - Задача структура-свойство
      - Классификация, кластеризация
  - Прикладные исследования
    - Финансы, тексты, оптимизация
      - Фондовые рынки
      - NP-трудные задачи
      - Тематические модели
    - Изображения, тексты
      - Практические применения
        - Анализ последовательностей
        - Гистологические изображения
        - Энергетическая безопасность
      - Анализ текстов, оптимизация
        - Поиск текстов
        - Размещение преприятий
        - Контроль знаний
      - Изображения
        - Аннотация изображений
        - Распознавание изображений
        - Симметрия изображений
    - Последовательности

ММРО

тематический навигатор

Главная | Корневая тема | О проекте

Тема 0/1: Прикладные исследования

Надтема

Подтема 0: Финансы, тексты, оптимизация

следовать задача анализ кластер mssc

Подтема 1: Изображения, тексты

изображение признак объект задача дать

Подтема 2: Последовательности

последовательность длина число задача

Фоновая тема 3

определённый сеть здравый comboost

Слова(2813)

последовательность
задача
изображение
анализ
число
объект
следовать
длина
структура
слово
дать
набор
являться
признак
фрагмент
информация
строка
элемент
преобразование
группа
иметь
поиск
проблема
распознавание
текст
алфавит
документ
состояние
результат
таблица
уровень
основа
метод
повтор
символ
исходный
дискретный

Авторы(389)

Кельманов А. В.
Чалей М. Б.
Михайлова Л. В.
Хамидуллин С. А.
Кутыркин В. А.
Дорофеюк А. А.
Леухин А. Н.
Емельянов Г. М.
Назипова Н. Н.
Тетуев Р. К.
Федотов Н. Г.
Панкратов А. Н.
Михайлов Д. В.
Дедус Ф. Ф.
Сулимова В. В.
Моттль В. В.
Пятков М. И.
Торшин И. Ю.
Дорофеюк Ю. А.
Кий К. И.
Покровская И. В.
Чернов В. М.
Тюкаев А. Ю.
Воронцов К. В.
Мучник И. Б.
Романов С. В.
Мокшанина Д. А.
Лебедев Л. И.
Парсаев Н. В.
Глумов Н. И.
Романченко С. М.
Ольшевец М. М.
Кудинов П. Ю.
Спиро А. Г.
Разин Н. А.
Кузнецов А. В.
Пяткин А. В.

Статьи(318)

- Распознавание скрытой периодичности в кодирующих последовательностях ДНК
- Методика исследования функционирования открытых индексных паевых инвестиционных фондов
- Структурные различия кодирующих и некодирующих районов последовательностей ДНК генома человека
- Скрытая профильная периодичность как новый тип периодичности генома
- Задача распознавания статистических таблиц
- Анализ фондовых и валютных рынков с помощью обобщенного непараметрического метода
- Выявление дубликатов объектов в прикладных онтологиях с помощью методов анализа формальных понятий
- О некоторых задачах анализа и распознавания последовательностей, включающих повторяющиеся упорядоченные наборы вектор-фрагментов
- Пространство формализации изображений
- Использование сложностных разложений в задачах анализа символьных последовательностей
- Морфология и синтаксис в задаче семантической кластеризации
- Профильно-статистическая основа локальных сигналов в ДНК
- Задачи анализа и распознавания последовательностей, включающих серии повторяющихся вектор-фрагментов
- Семантическая схожесть текстов в задаче автоматизированного контроля знаний
- О сложности некоторых задач кластеризации векторных последовательностей
- Задачи и методы автоматического построения графа цитирования по коллекции научных документов
- Автоматическая аннотация изображений
- Классификация изображений периодических структур на основе непрерывного преобразования симметрии
- Тернарные системы счисления в кольце целых чисел Эйзенштейна и их приложения

Главная | Корневая тема | О проекте

Тема 0/1/1: Изображения, тексты

Надтема

Подтема 0: Практические применения

внимание информационный

Подтема 1: Анализ текстов

текст предприятие таблица ячейка

Подтема 2: Изображения

изображение преобразование

Фоновая тема 3

признак задача объект дать анализ

Слова(844)

изображение
признак
объект
задача
дать
набор
преобразование
анализ
группа
текст
поиск
ячейка
симметрия
распознавание
сгусток
число
структура
решение
гистограмма
элемент
блок
основа
предприятие
таблица
классификатор
форма
множество
результат
являться
соответствовать
информационный
обработка
характеристика
включать
внимание
активный
петроглиф

Авторы(300)

Кельманов А. В.
Федотов Н. Г.
Емельянов Г. М.
Хамидуллин С. А.
Михайлов Д. В.
Михайлова Л. В.
Кий К. И.
Романов С. В.
Мокшанина Д. А.
Колоколов А. А.
Кудинов П. Ю.
Левашкина А. О.
Поршнев С. В.
Мельниченко А. С.
Мнухин В. Б.
Каркищенко А. Н.
Глумов Н. И.
Кузнецов А. В.
Визильтер Ю. В.
Бекетова И. В.
Каратеев С. Л.
Рогова К. А.
Дюкова Е. В.
Романченко С. М.
Пяткин А. В.
Полежаев В. А.
Березин А. В.
Иванова Е. Ю.
Воронцов К. В.
Дорофеюк А. А.
Васин Ю. Г.
Лебедев Л. И.
Моттль В. В.
Рогов А. А.
Сулимова В. В.
Фрей А. И.
Петровский М. И.

Статьи(222)

- Пространство формализации изображений
- Задача распознавания статистических таблиц
- Классификация изображений периодических структур на основе непрерывного преобразования симметрии
- Сравнительный анализ особенностей CBIR-систем
- Морфология и синтаксис в задаче семантической кластеризации
- Автоматическая аннотация изображений
- Семантическая схожесть текстов в задаче автоматизированного контроля знаний
- Преобразования данных для вычислительного эксперимента в исследованиях энергетической безопасности на основе декларативных представлений
- Анализ формальных понятий и сжатие текстовой информации в задаче автоматизированного контроля знаний
- Исследование некоторых постановок задачи о рюкзаке и алгоритмов их решения с использованием унимодулярных преобразований и L-разбиения
- Исследование математических моделей и декомпозиционных алгоритмов для решения двухстадийной задачи размещения
- Метод геометризованных гистограмм, дуальное описание сцен и его применение
- Применение методов распознавания образов в системе управления коллекциями графических документов
- Линейная комбинация случайных лесов в задаче предсказания релевантности документов
- Применение триплетных признаков распознавания к цветным изображениям
- Инкрементное обучение деревьев решений в задаче распознавания структуры статистических таблиц
- Анализ фондовых и валютных рынков с помощью обобщенного непараметрического метода
- Вычислительный алгоритм поиска на изображении прото-объекта

ММРО

тематический навигатор

Главная | Корневая тема | О проекте

Тема 0/1/1/1: Анализ текстов

Надтема

Подтема 0: Поиск текстов

текст релевантность атрибут клик

Подтема 1: Размещение предприятий

предприятие размещение

Подтема 2: Контроль знаний

ячейка таблица синтаксический сяу

Фоновая тема 3

лес ответ смысловой разбор задача

Слова(198)

текст
предприятие
таблица
ячейка
петроглиф
синтаксический
лес
сяу
релевантность
ответ
атрибут
подструктура
клик
название
единица
смысловой
ситуация
тезаурус
дерево
размещение
схожесть
флексия
документ
потребитель
сессия
симплексный
сеть
связь
двухстадийный
тестирование
порождение
знание
предметный
разбор
random
языковой
скал

Авторы(74)

Михайлов Д. В.
Емельянов Г. М.
Кудинов П. Ю.
Полежаев В. А.
Леухин А. Н.
Чувилин К. В.
Чучупал В. Я.
Алябушев А. А.
Рогова К. А.
Прокофьев П. А.
Леванова Т. В.
Стрижов В. В.
Царёв Д. В.
Сидоров Ю. В.
Спиридонов К. Н.
Рахманов Х. Э.
Кириллов А. Н.
Воронцов К. В.
Крымова Е. А.
Колоколов А. А.
Куликов А. И.
Покровская И. В.
Царьков С. В.
Местецкий Л. М.
Ботов П. В.
Лисица А. В.
Бекетова И. В.
Сулейманова Е. А.
Кочетова Н. А.
Вьюшнов Д. В.
Давыдов И. А.
Машечкин И. В.
Майсурадзе А. И.
Петровский М. И.
Загоруйко Н. Г.
Фрей А. И.
Виноградов А. П.

Статьи(40)

- Задача распознавания статистических таблиц
- Семантическая схожесть текстов в задаче автоматизированного контроля знаний
- Морфология и синтаксис в задаче семантической кластеризации
- Анализ формальных понятий и сжатие текстовой информации в задаче автоматизированного контроля знаний
- Инкрементное обучение деревьев решений в задаче распознавания структуры статистических таблиц
- Линейная комбинация случайных лесов в задаче предсказания релевантности документов
- Исследование математических моделей и декомпозиционных алгоритмов для решения двухстадийной задачи размещения
- Оптимальные байесовские стратегии анализа релевантности для объектов с заданной структурой
- К вопросу об инварианте графического изображения
- Алгоритм автоматического определения контекста информационного потокового видео
- О подходах к синтезу случайных и решающих лесов
- Формализация и автоматический анализ понятий при обработке неструктурированной информации
- Алгоритм локального поиска для задачи конкурентного размещения предприятий
- Сравнение эвристических алгоритмов выбора линейных регрессионных моделей
- Использование FRiS-функции при решении задачи распознования состояний объектов в функционально-топической диагностике
- Задачи анализа изображений в информационно-поисковой системе PIRS
- Математические методы атрибуции литературных текстов небольшого объема
- Задачи и методы автоматического построения графа цитирования по коллекции научных документов

ММРО

тематический навигатор

Главная | Корневая тема | О проекте

Тема 0/1/1/2: Изображения

Надтема

Подтема 0: Аннотация изображений

сгусток гистограмма контрастный форма

Подтема 1: Распознавание изображений

изображение преобразование блок

Подтема 2: Симметрия изображений

группа симметрия задача анализ

Фоновая тема 3

задача анализ объект структура

Слова(371)

изображение
преобразование
задача
объект
группа
симметрия
анализ
сгусток
дать
структура
гистограмма
число
являться
характеристика
обработка
блок
форма
признак
контрастный
множество
поиск
результат
применение
активный
цвет
алгоритм
инвариант
виртуальный
распознавание
дасп
решение
периодический
помощь
включать
цветной
случайный
выделение

Авторы(135)

Кий К. И.
Федотов Н. Г.
Кельманов А. В.
Мнухин В. Б.
Каркищенко А. Н.
Мокшанина Д. А.
Романов С. В.
Кузнецов А. В.
Глумов Н. И.
Поршнев С. В.
Левашкина А. О.
Михайлова Л. В.
Хамидуллин С. А.
Рогова К. А.
Визильтер Ю. В.
Бекетова И. В.
Каратеев С. Л.
Фрей А. И.
Мельниченко А. С.
Чуличков А. И.
Зубюк А. В.
Березин А. В.
Иванова Е. Ю.
Пытьев Ю. П.
Цыбульская Н. Д.
Чернов В. М.
Броневич А. Г.
Пяткин А. В.
Дорофеюк А. А.
Дмитриев Е. В.
Козодеров В. В.
Кондранин Т. В.
Генрихов И. Е.
Апарин Г. П.
Абламейко С. В.
Крючков А. Н.
Жарких А. А.

Статьи(95)

- Пространство формализации изображений
- Классификация изображений периодических структур на основе непрерывного преобразования симметрии
- Метод геометризованных гистограмм, дуальное описание сцен и его применение
- Преобразование симметрии периодических структур в частотной области
- Обнаружение локальных искусственных изменений на крупноразмерных изображениях
- Виртуальные граничные кривые: подход к анализу движения
- Выявление следов применения алгоритмов цифровой обработки на изображениях
- Применение триплетных признаков распознавания к цветным изображениям
- Случайная морфология: алгоритмы обучения и классификации
- Сравнительный анализ особенностей CBIR-систем
- Об условиях устойчивости нахождении осей симметрии зашумленного изображения
- Модифицированный метод геометризованных гистограмм и его применение
- Геометризованные гистограммы и понимание изображений
- NP-полнота некоторых задач анализа данных
- Вычислительный алгоритм поиска на изображении прото-объекта
- Сегментация гистологических изображений. Выделение фолликулов и ядер
- Автоматическая аннотация изображений
- Анализ текстур гистологических изображений
- Создание системы распределенного отказоустойчивого хранения цветных крупноформатных изображений
- Комбинированный подход к локализации записей на изображениях произведений живописи
- Трейс-преобразование как источник признаков распознавания

Результаты и перспективы

Результаты:

- предложен способ построения тематических иерархий с помощью рекурсивного применения ARTM с регуляризаторами разреживания, сглаживания и декоррелирования.
- разработана пилотная версия тематического навигатора ММРО-ИОИ

Дальнейшие исследования:

- оценивание и улучшение качества иерархии
- автоматическое определения числа подтем в каждой теме
- графическая визуализация иерархии
- частичное обучение (использование экспертной разметки)
- автоматическое именование тем

Литература

К. В. Воронцов, А. А. Потапенко Регуляризация вероятностных тематических моделей для повышения интерпретируемости и определения числа тем // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.) Вып.13 (20). М: Изд-во РГГУ, 2014. С.676–687. (2012)

E. Zavitsanos, G. Paliouras, G. A. Vouros Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes. – Journal of Machine Learning Research. – 2011. – № 12. – с. 2749-2775

Воронцов К. В., Потапенко А. А., Фрей А. И., Апишев М. А., Дойков Н. В., Шапулин А. В., Чиркова Н. А. Многокритериальные и многомодальные вероятностные тематические модели коллекций текстовых документов // Интеллектуализация обработки информации (ИОИ-2014): Тезисы докл. — Москва: Торус Пресс, 2014. С. 198–199.