Иерархическая аддитивно регуляризованная тематическая модель конференций ММРО-ИОИ

Надежда Чиркова [1]
Воронцов К. В. [2]

[1]ВМК МГУ [2]ВЦ РАН

Математические методы распознавания образов
Светлогорск, 21.09.2015

План выступления

1. Аддитивно регуляризованные тематические модели
   - Задача тематического моделирования
   - Многомодальные тематические модели
   - Обучение модели
2. Тематические иерархии в АРТМ
   - Способы построения иерархий
   - Два подхода к построению иерархий в BigARTM
3. Иерархическая модель конференций ММРО-ИОИ
   - Регуляризаторы повышения интерпретируемости
   - Визуализация модели
   - Критерии качества тематических моделей
   - Эксперименты
4. Перспективы

Задача тематического моделирования

D — коллекция текстовых документов, W — множество терминов.

Дана коллекция текстовых документов:

n_{dw} — матрица частот терминов в документах: $F_{dw} = p(w|d) = \frac{n_{dw}}{n_d}$

Построить модель:

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

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

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

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

Оптимизировать регуляризованный логарифм правдоподобия:

$$\mathscr{L}(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

с ограничениями $\sum_w \varphi_{wt} = 1;\ \varphi_{wt} \geqslant 0,$

$\sum_t \theta_{td} = 1;\ \theta_{td} \geqslant 0.$

Многомодальные тематические модели

Модальности — конечные непересекающиеся множества W^m, $m \in M$ (авторы, классы, метки времени ...).

Построить модель:

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

с параметрами $\Phi^m = \{\varphi_{wt}\}_{W^m \times T}$, $m \in M$ и $\Theta = \{\theta_{td}\}_{T \times D}$, $\Phi = \begin{matrix}\Phi^1 \\ \hline \dots \\ \hline \Phi^m\end{matrix}$:

$\varphi^m_{wt} = p(w|t)$ — распределение терминов в теме t;
$\theta_{td} = p(t|d)$ — распределение тем в документе d.

Оптимизировать регуляризованный логарифм правдоподобия:

$$\mathscr{L}(\Phi, \Theta) = \sum_{m \in M} \xi_m \left(\sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_{t \in T} \varphi^m_{wt}\theta_{td} \right) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

с ограничениями $\sum_w \varphi^m_{wt} = 1$; $\varphi^m_{wt} \geqslant 0$, $\quad \sum_t \theta_{td} = 1$; $\theta_{td} \geqslant 0$,

ξ_m — вес модальности.

Обучение плоской модели APTM

Решение оптимизационной задачи — метод простой итерации (*EM-алгоритм*):

EM-алгоритм

E-шаг: $p(t|d,w) = \underset{t\in T}{norm}(\varphi_{wt}\theta_{td})$;

M-шаг: $n_{wt} = \sum\limits_{d\in D} \xi_{m(w)} n_{dw} p(t|d,w)$,

$$n_{td} = \sum_{w\in W} \eta_{m(w)} n_{dw} p(t|d,w),$$

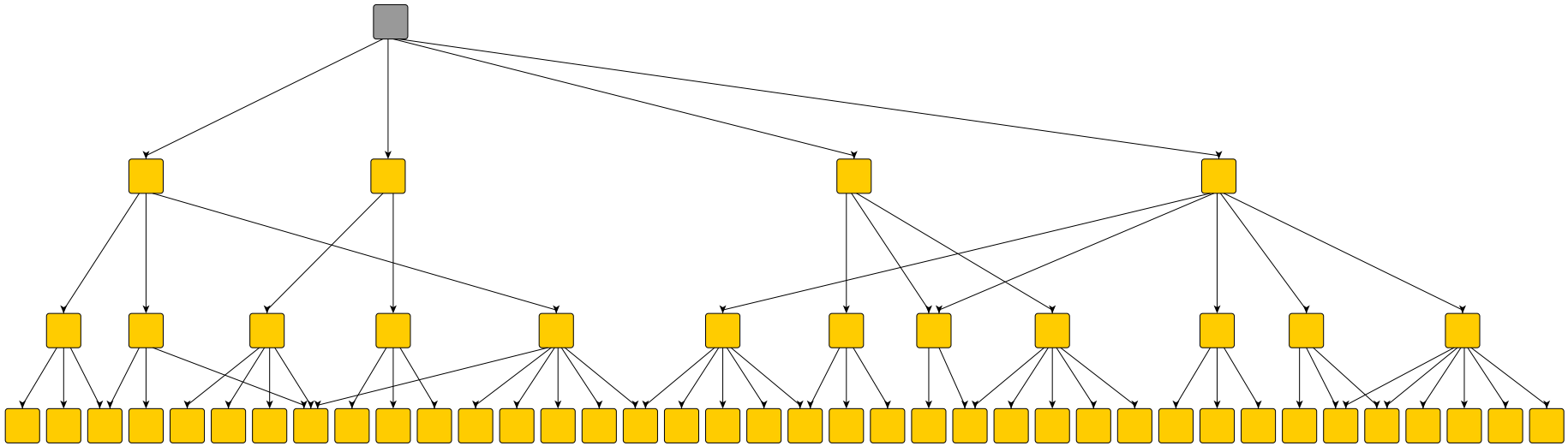
$$\varphi_{wt} = \underset{w\in W^m}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right),$$

$$\theta_{td} = \underset{t\in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right).$$

Оператор $\underset{t\in T}{\text{norm}}(x_t) = \frac{\max\{x_t,0\}}{\sum_{s\in S}\max\{x_s,0\}}$.

Понятие тематической иерархии

Иерархическая тематическая модель — это ориентированный многодольный граф тем.



Иерархическая структура помогает пользователю *постепенно* знакомиться с тематикой коллекции документов.

Цель работы: обогатить APTM возможностью строить тематические иерархии.

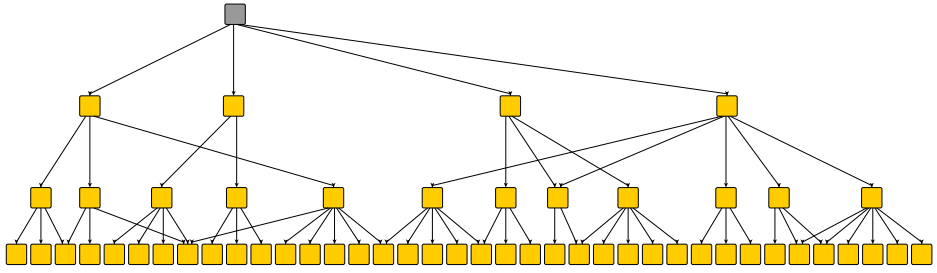
Способы построения иерархий

- различаются по направлению построения графа:
  - ***Нисходящее построение***
    Соответствует человеческой логике разделения тем на подтемы, однако интерпретируемость тем нижних уровней зависит от верхних уровней.
  - ***Восходящее построение***
    Требует строить большое число тем сразу и не позволяет контролировать их интерпретируемость.
- различаются по количеству тематических моделей:
  - можно ***строить отдельную тематическую модель в каждом узле иерархии***
    Требуется задавать параметры модели и количество тем в каждой вершине графа.
  - можно ***строить целый уровень как одну тематическую модель***
    Параметры задаются для целого уровня, разрешается множественное наследование тем, но нужен способ поиска связей между темами.

$\Rightarrow$ Будем строить иерархию сверху вниз, уровень за уровнем, постоянно увеличивая количество тем.

Первый подход: регуляризатор матрицы Θ

Пусть построен (l-1)-й уровень иерархии с множеством тем T, требуется построить l-й уровень с множеством подтем S, $|S| > |T|$.



Добавим новую матрицу параметров $\Psi \in R^{|T|\times|S|}$: $\psi_{ts} = p(t|s)$,

$$p(t|d) \approx \sum_{s\in S} p(t|s)p(s|d) = \sum_{s\in S} \psi_{ts}\theta_{sd} \Leftrightarrow \Theta^{parent} \approx \Psi\Theta.$$

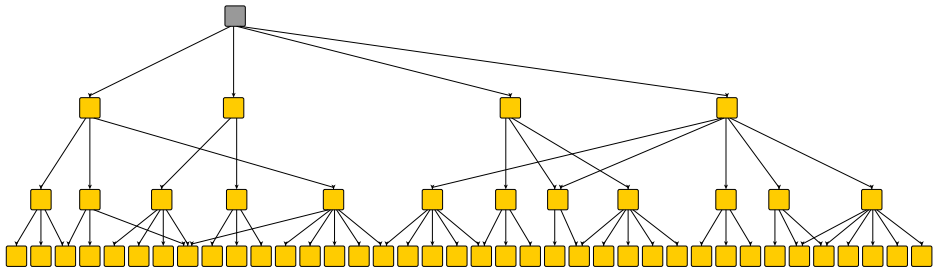
Новая оптимизационная задача:

$$\sum_{d,w} n_{dw} \ln \sum_{s\in S} \varphi_{ws}\theta_{sd} + \lambda \sum_{d,t} \theta_{td}^{parent} \ln \sum_{s\in S} \psi_{ts}\theta_{sd} + R(\Phi,\Theta,\Psi) \to \max_{\Phi,\Theta,\Psi}$$

Регуляризатор Θ равносилен добавлению в модель новой модальности, элементы которой — темы t родительского уровня!

Второй подход: регуляризатор матрицы Φ

Пусть построен (l-1)-й уровень иерархии с множеством тем T, требуется построить l-й уровень с множеством подтем S, $|S| > |T|$.



Добавим новую матрицу параметров $\widetilde{\Psi} \in R^{|S|\times|T|}$: $\tilde{\psi}_{st} = p(s|t)$,

$$p(w|t) \approx \sum_{s\in S} p(w|s)p(s|t) = \sum_{s\in S} \varphi_{ws}\tilde{\psi}_{st} \Leftrightarrow \Phi^{parent} \approx \Phi\widetilde{\Psi}.$$

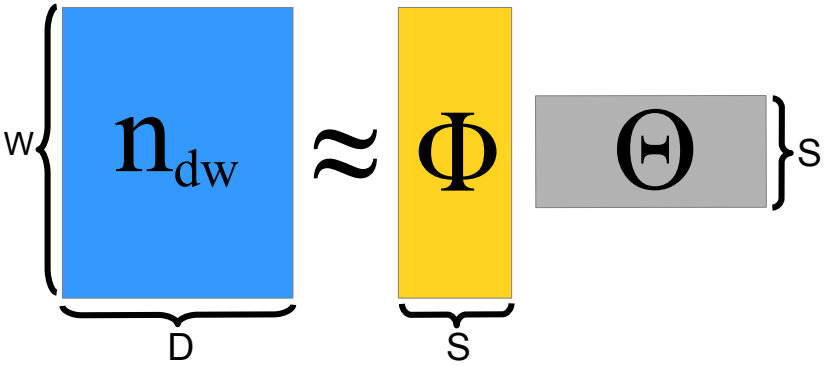
Новая оптимизационная задача:

$$\sum_{d,w} n_{dw} \ln \sum_{s\in S} \varphi_{ws}\theta_{sd} + \lambda \sum_{t,w} \varphi_{wt}^{parent} \ln \sum_{s\in S} \varphi_{ws}\tilde{\psi}_{st} + R(\Phi, \Theta, \widetilde{\Psi}) \to \max_{\Phi,\Theta,\widetilde{\Psi}}.$$

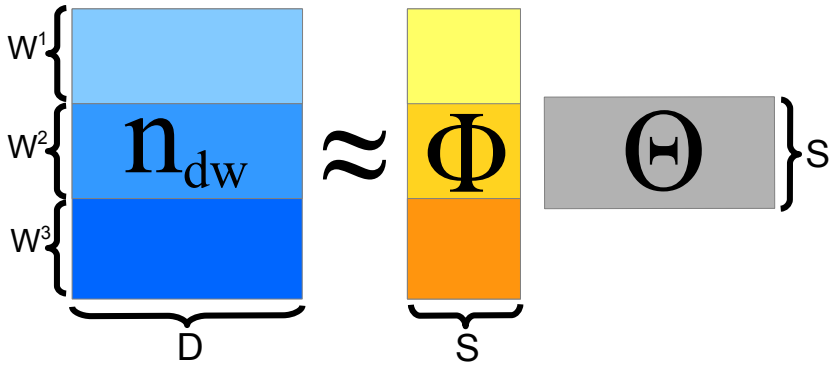
Регуляризатор Φ равносилен добавлению во входную матрицу частот слов $|T|$ псевдодокументов: $n_{dw} = \lambda\varphi_{wt}, d \sim t \in T$!

Структурные различия двух подходов

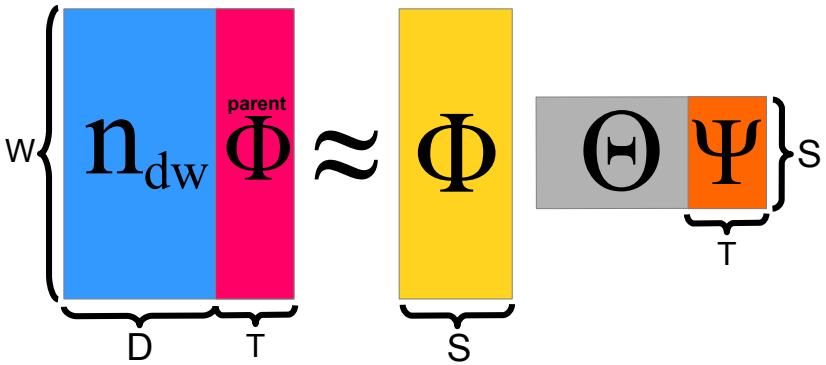
APTM



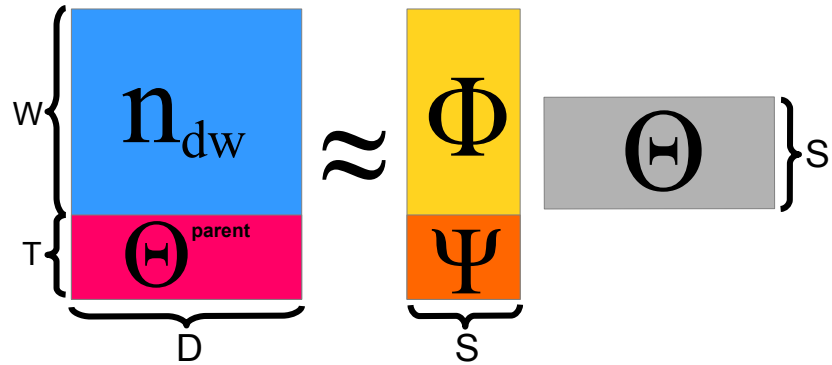
Многомодальный APTM



APTM с регуляризацией Φ



APTM с регуляризацией Θ



Иерархическая модель конференций ММРО-ИОИ

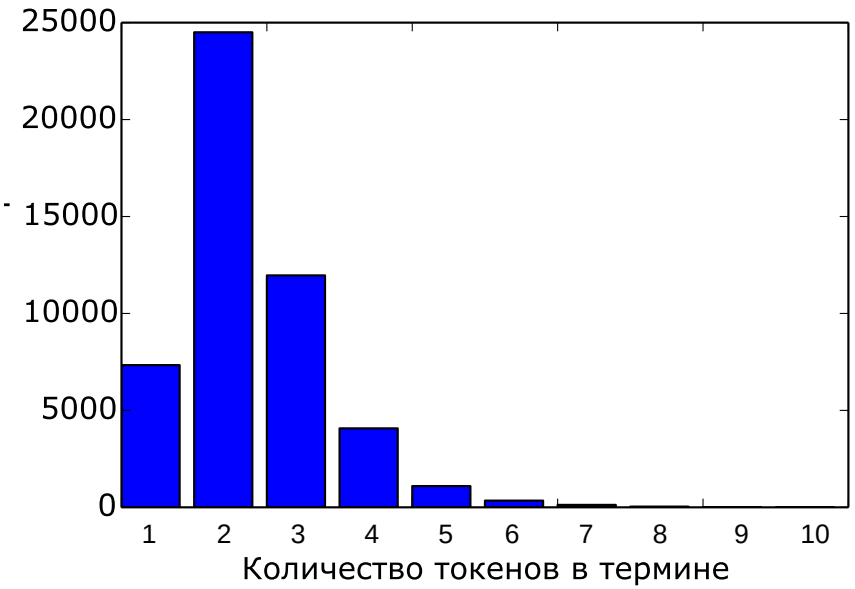
Коллекция составлена из статей конференций *«Интеллектуализация обработки информации»* и *«Математические методы распознавания образов»*.
$|D| = 865; |W| = 42000$

Предварительная обработка:

- лемматизация,
- выделение мультиграмм,
- составление матрицы частот слов.

Множество терминов разделено на 10 модальностей
(по количеству токенов в термине).

Соотношение модальностей:



Регуляризаторы повышения интерпретируемости

Множество тем S одного уровня иерархии разбивается 2 подмножества:

- *фоновые* B (общая лексика уровня, сглаженные строки Θ)

  В модели ММРО: b_1 — разреженная фоновая тема (для общей лексики языка и коллекции),

  b_2 — сглаженная фоновая тема (для общей лексики данного уровня)
- *предметные* S_0 (специализированная лексика каждой темы, разреженные и декоррелированные строки Φ, разреженные строки Θ)

Итоговая комбинация регуляризаторов на М-шаге:

$$\varphi_{wt} = \underset{w \in W^m}{norm}\Big(n_{wt} - \tau_1 \underbrace{\beta_w^1[s=b_1]}_{\substack{\text{разреживание}\\ \text{фоновой}\\ \text{темы 1}}} + \tau_2 \underbrace{\beta_w^2[s=b_2]}_{\substack{\text{сглаживание}\\ \text{фоновой}\\ \text{темы 2}}} - \tau_3 \underbrace{\beta_w^3[s\in S_0]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} - \tau_4 \underbrace{\phi_{wt} \sum_{s\in S\setminus t} \varphi_{ws}}_{\text{декорреляция}} \Big)$$

$$\theta_{td} = \underset{s\in S}{norm}\Big(n_{td} + \tau_4 \underbrace{\alpha_s[s\in B]}_{\substack{\text{сглаживание}\\ \text{фоновых}\\ \text{тем}}} - \tau_5 \underbrace{\alpha_t[t\in S_0]}_{\substack{\text{разреживание}\\ \text{предметных}\\ \text{тем}}} \Big)$$

✓ Эффективнее включать регуляризаторы разреживания после 20 итерации EM-алгоритма.

Визуализация модели и создание тематического навигатора

Итоговая модель:

3 уровня (исключая корневую вершину), 10/30/60 тем, 2 фоновые темы на каждом уровне.

Модель построена с помощью библиотеки BigARTM

Именование тем выполнялось экспертом (Воронцов К. В).

Визуализация (Айсина Р., Чиркова Н.):

- в виде графа, *вершины:* темы и документы, *ребра* соответствуют значениям матриц Θ и Ψ, превышающим порог ϵ (значения ниже считаются нулем).
- *Раскладка вершин* силовым алгоритмом (в реализации Gephi).
- *Поддержка интерактивного интерфейса:* javascript-библиотека Sigma.

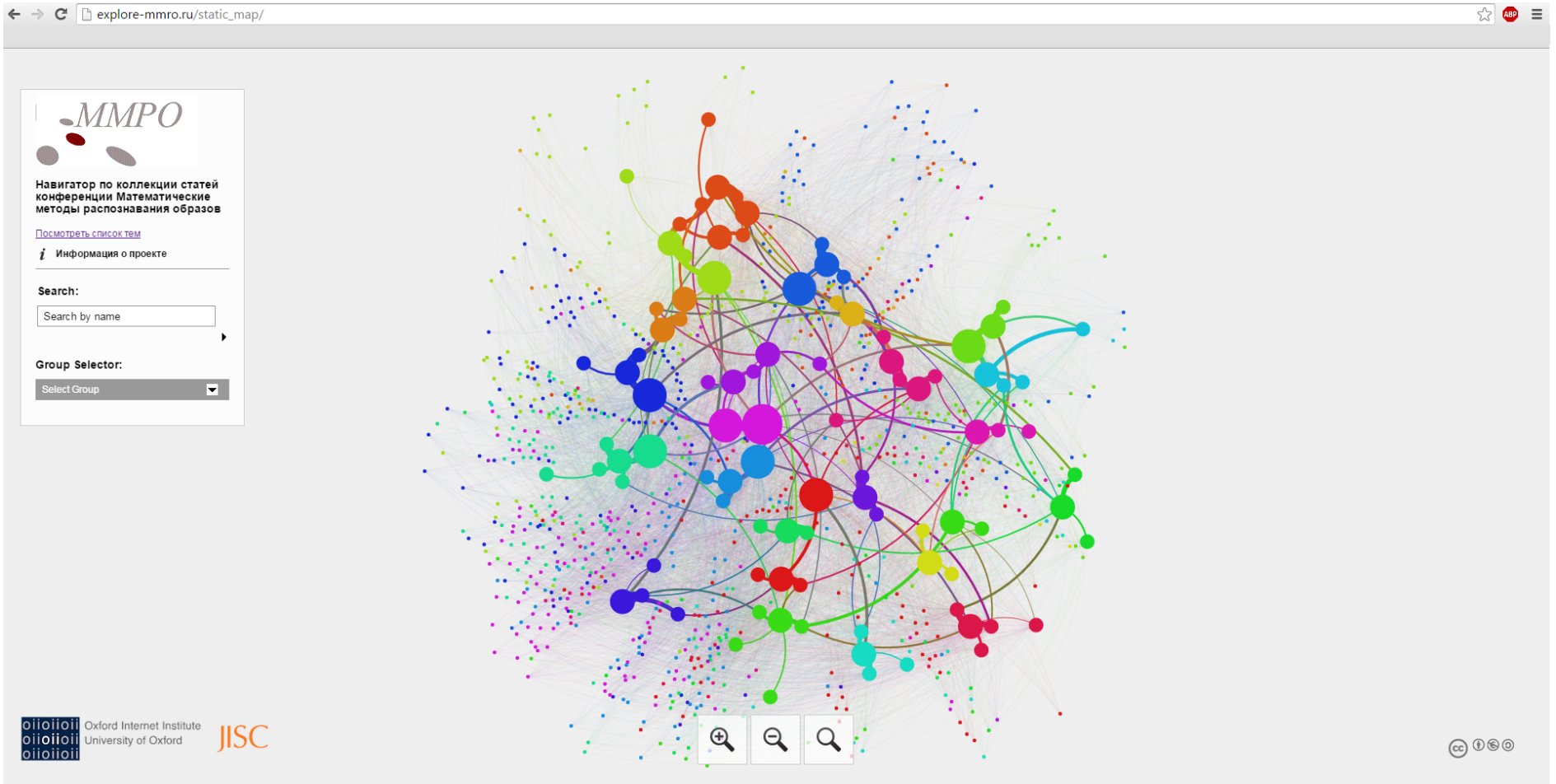
Тематический навигатор (Плавин А.):

- в виде веб-сайта,
- организует удобный доступ к темам, документам, терминам и переходы между ними,
- поддерживает обратную связь от пользователя (Оценивайте темы!)

Визуализация и навигатор доступны по адресу

explore-mmro.ru

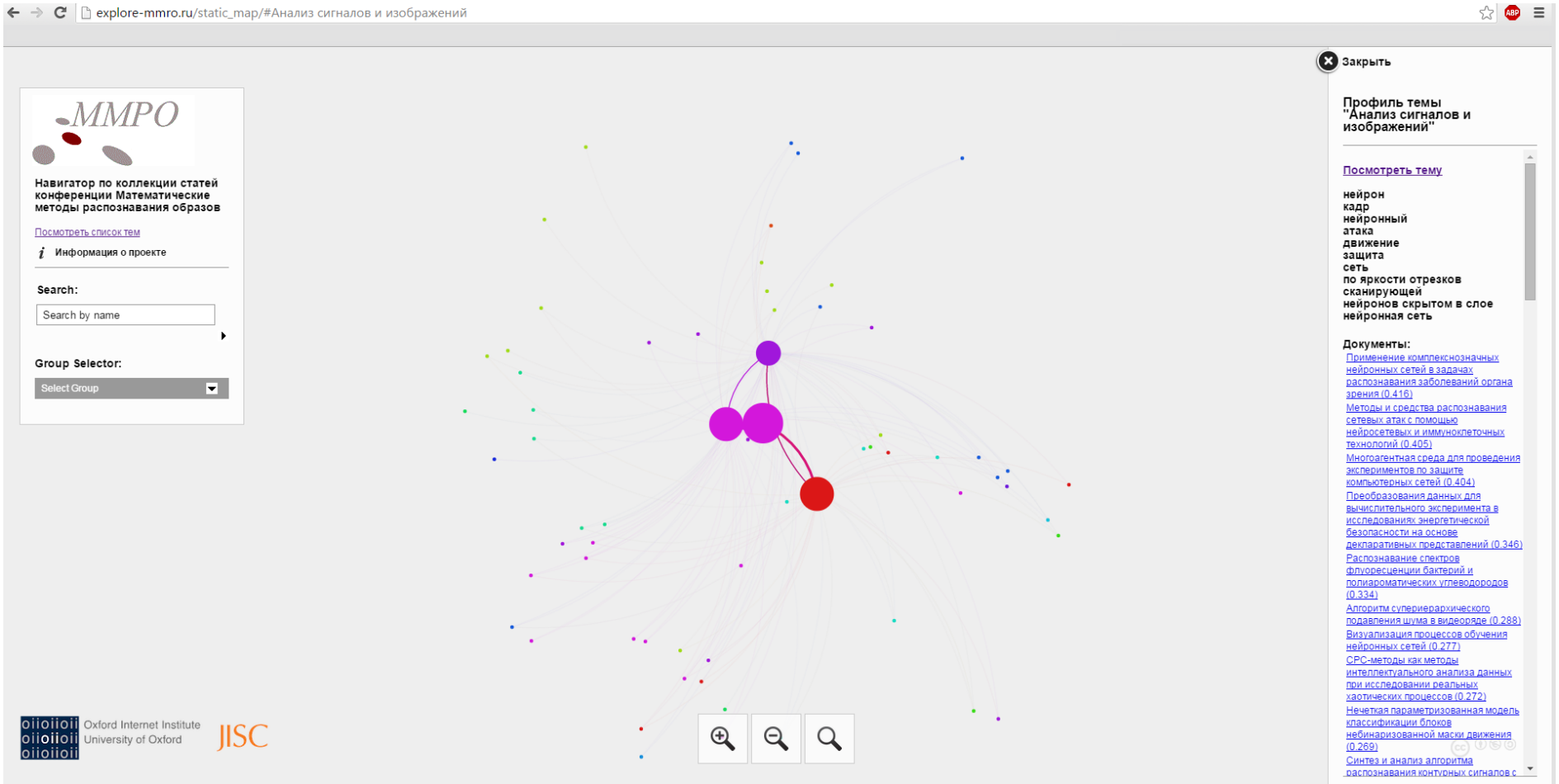
Визуализация модели



Визуализация и навигатор доступны по адресу

explore-mmro.ru

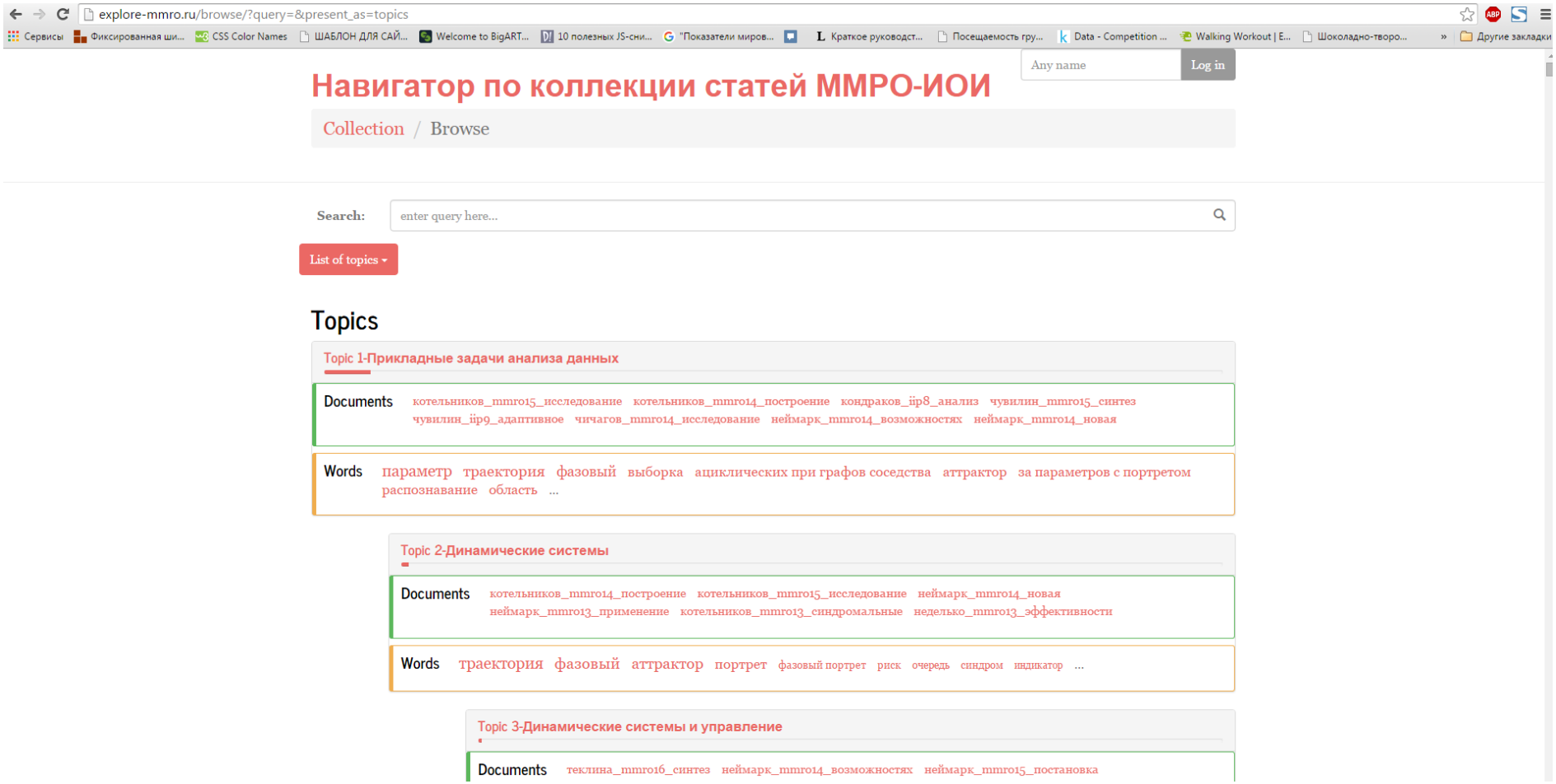
Визуализация модели



Визуализация и навигатор доступны по адресу

explore-mmro.ru

Навигатор по коллекции статей ММРО-ИОИ



Визуализация и навигатор доступны по адресу

explore-mmro.ru

Критерии качества тематических моделей

Интерпретируемость

Размер ядра темы $|W_s|, \quad W_s = \{w : p(s|w) > 0.25\}$

Контрастность темы $\frac{1}{|W_s|} \sum_{s \in W_s} p(s|w)$

Чистота темы $\sum_{w \in W_t} p(w|s)$

Качество кластеризации

Средневзвешенное внутрикластерное расстояние

$$\sum_{\substack{d_1, d_2 \,:\, \theta_{s,d1} \neq 0, \\ \theta_{s,d2} \neq 0}} sim(n^s_{d_1}, n^s_{d_2}) p(d_1, d_2|s),$$

$$n^s_d = \{n^s_{dw}\}_{w \in W}, n^s_{dw} = n_{dw} p(s|d, w),$$

$$sim(d_1, d_2) = \sum_{w \in W} n_{d_1 w} n_{d_2 w}$$

Разреженность иерархической структуры

Разреженность Ψ $\sum_t \sum_s [\psi_{ts} > \epsilon]$

Качество аппроксимации родительских тем

Расстояние Хеллингера между матрицами

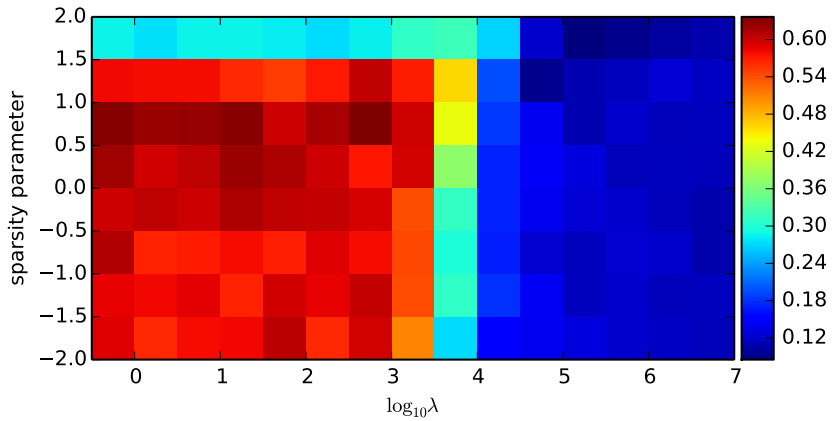
$$A = \{a_{ij}\}_{i,j=1}^{m,n}, \; B = \{b_{ij}\}_{i,j=1}^{m,n}:$$

$$\rho(A, B) = \frac{1}{m} \sum_{j=1}^{m} \sqrt{\frac{1}{2} \sum_{i=1}^{n} (\sqrt{a_{ij}} - \sqrt{b_{ij}})^2}$$

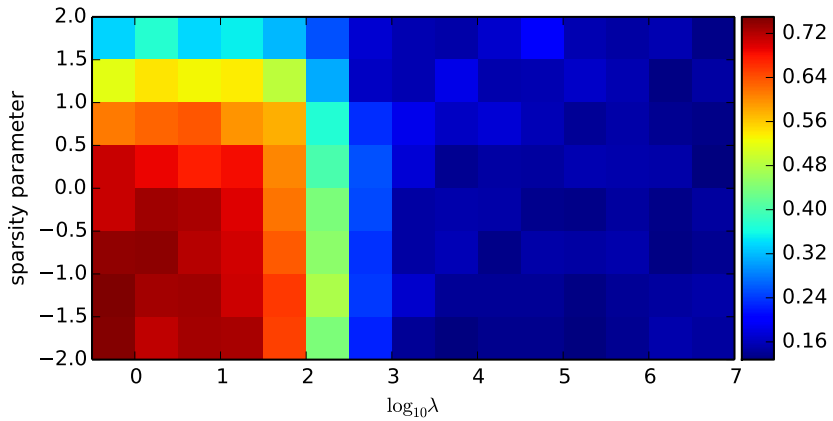
Эксперименты. Исследование взаимодействия разреживаюих и сглаживающих регуляризаторов

Разреженность Ψ

Регуляризатор Φ



Регуляризатор Θ



По оси абсцисс — коэффициент λ [сглаживающего] иерархического регуляризатора (лог. шкала);

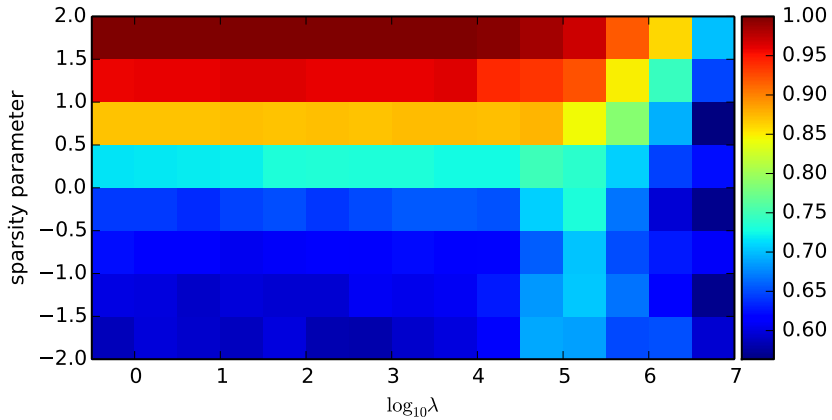
По оси ординат — параметр k разреживания предметных тем ($\tau_i = 10^k \tau_i^0, i = 1, 3, 4, 5$).

Существует λ, при котором тематический граф достаточно разрежен и близок к дереву.

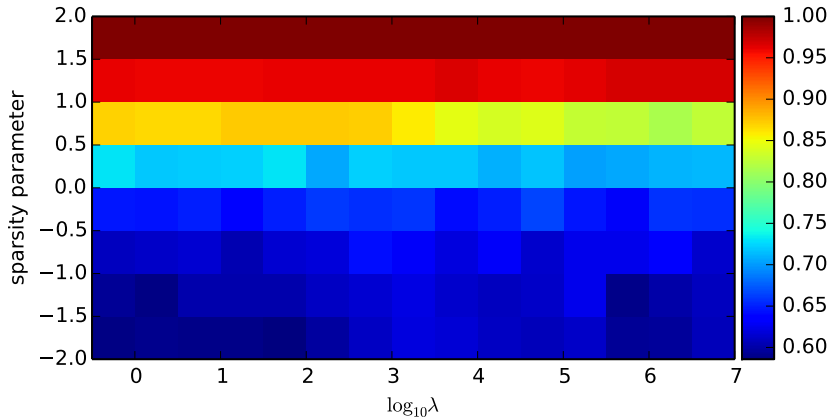
Эксперименты. Исследование взаимодействия разреживаюих и сглаживающих регуляризаторов

Чистота

Регуляризатор Φ



Регуляризатор Θ



По оси абсцисс — коэффициент λ [сглаживающего] иерархического регуляризатора (лог. шкала);

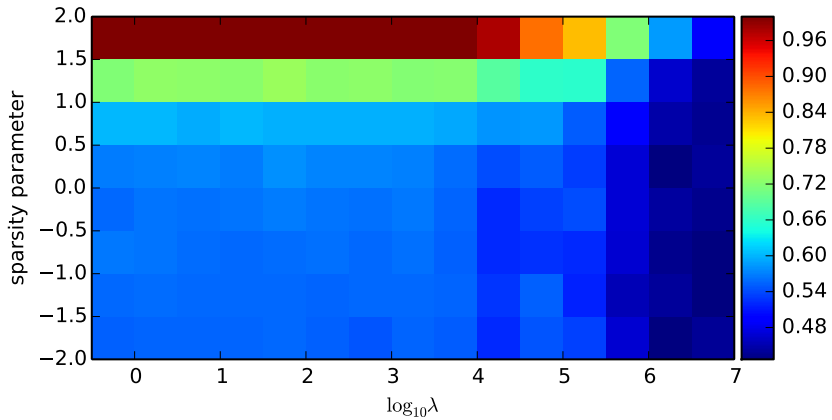
По оси ординат — параметр k разреживания предметных тем ($\tau_i = 10^k \tau_i^0, i = 1, 3, 4, 5$).

Нельзя задавать слишком большие λ — ухудшается интерпретируемость.

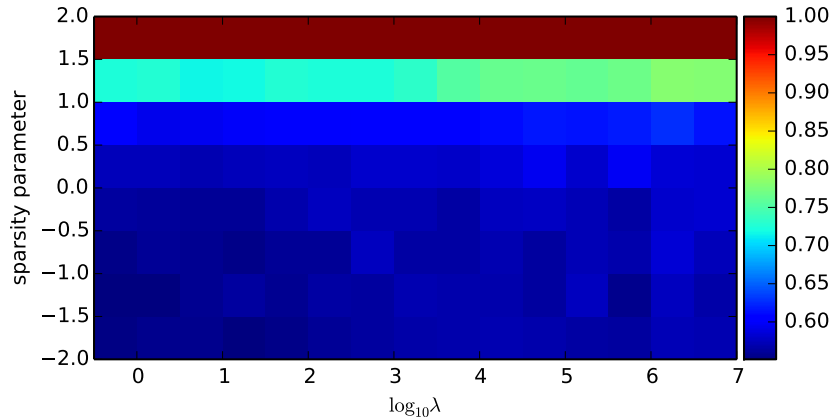
Эксперименты. Исследование взаимодействия разреживаюих и сглаживающих регуляризаторов

Контрастность

Регуляризатор Φ



Регуляризатор Θ



По оси абсцисс — коэффициент λ [сглаживающего] иерархического регуляризатора (лог. шкала);

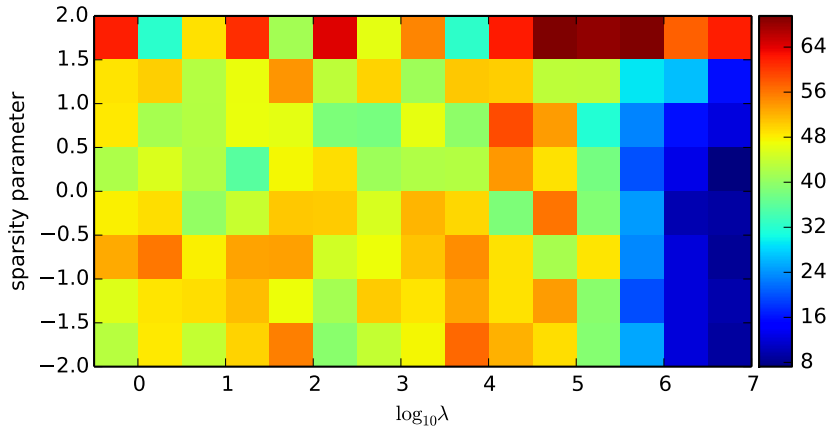
По оси ординат — параметр k разреживания предметных тем ($\tau_i = 10^k \tau_i^0, i = 1, 3, 4, 5$).

Нельзя задавать слишком большие λ — ухудшается интерпретируемость. Можно комбинировать разреживание и сглаживание — улучшается интерпретируемость.

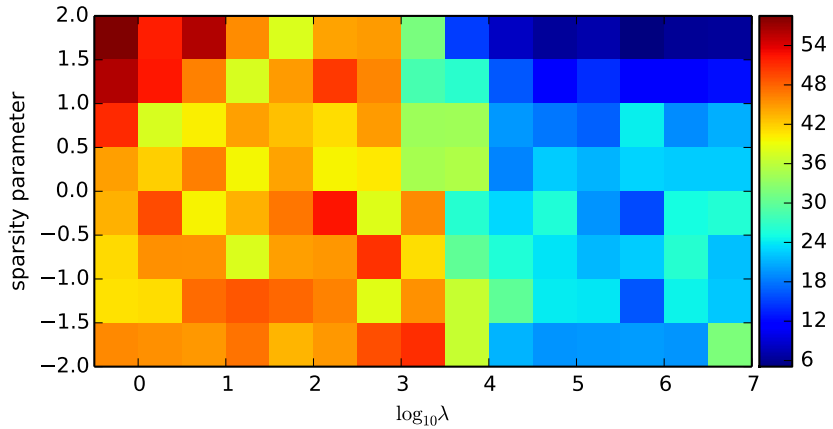
Эксперименты. Исследование взаимодействия разреживаюих и сглаживающих регуляризаторов

Средневзвешенное внутрикластерное расстояние

Регуляризатор Φ



Регуляризатор Θ



По оси абсцисс — коэффициент λ [сглаживающего] иерархического регуляризатора (лог. шкала);

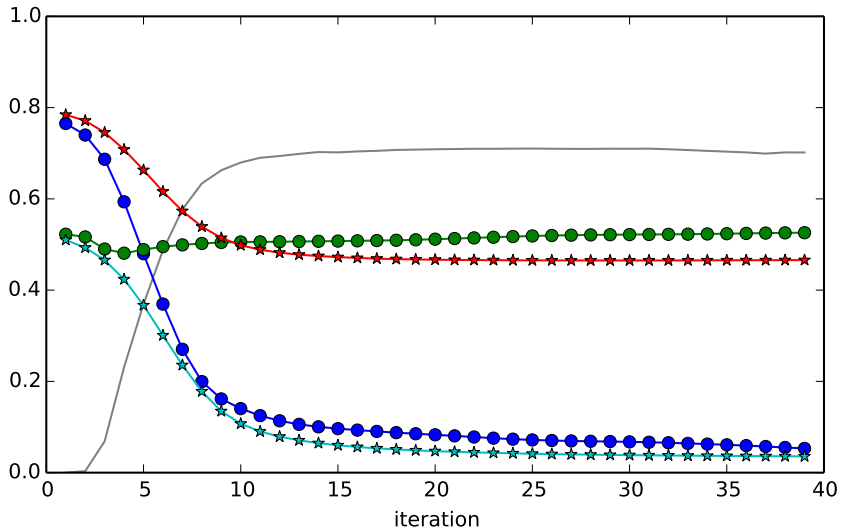
По оси ординат — параметр k разреживания предметных тем ($\tau_i = 10^k \tau_i^0, i = 1, 3, 4, 5$).

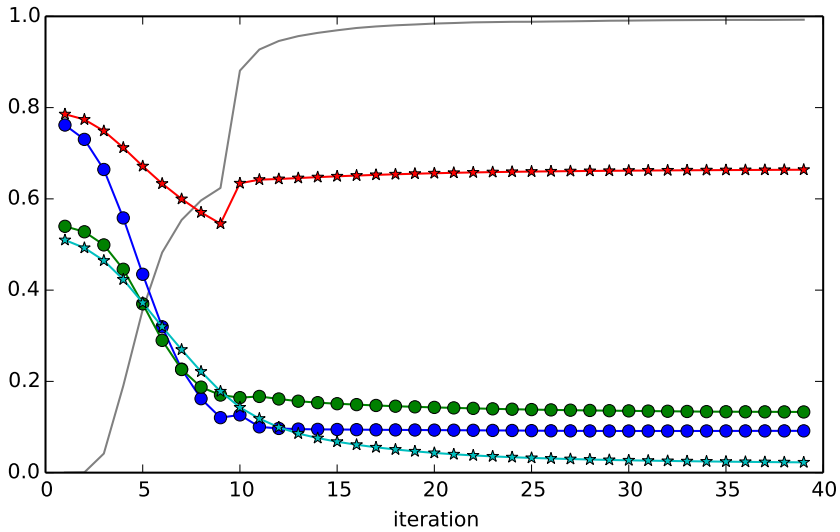
При некоторой комбинации сглаживания и разреживания документы кластеризуются стабильно лучше, чем при других комбинациях.

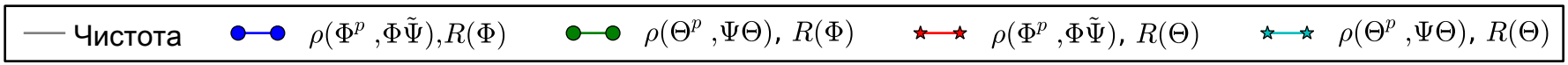
Эксперименты. Исследование сходимости модели

Разреживание с 1-й итерации **Разреживание с 10-й итерации**

Переход между 1-м и 2-м уровнями:







По оси абсцисс — номер итерации EM-алгоритма;

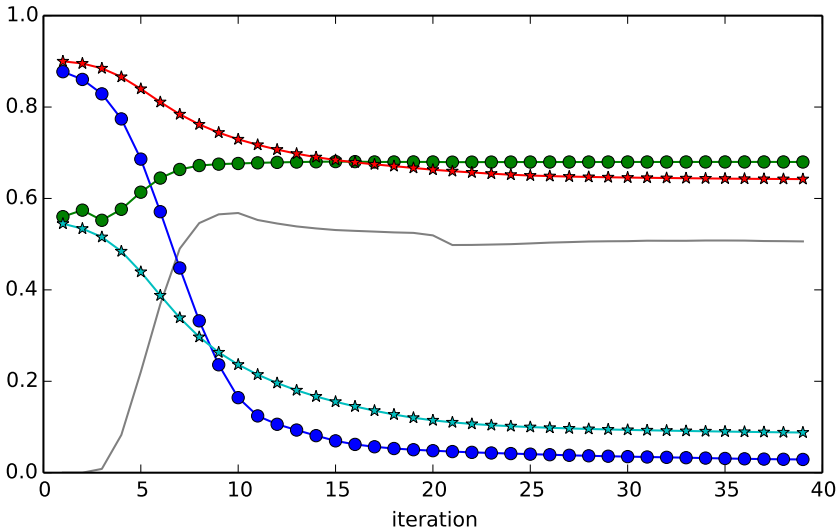
По оси ординат — расстояние между матрицами родительского уровня и их высокоранговыми разложениями на дочернем уровне.

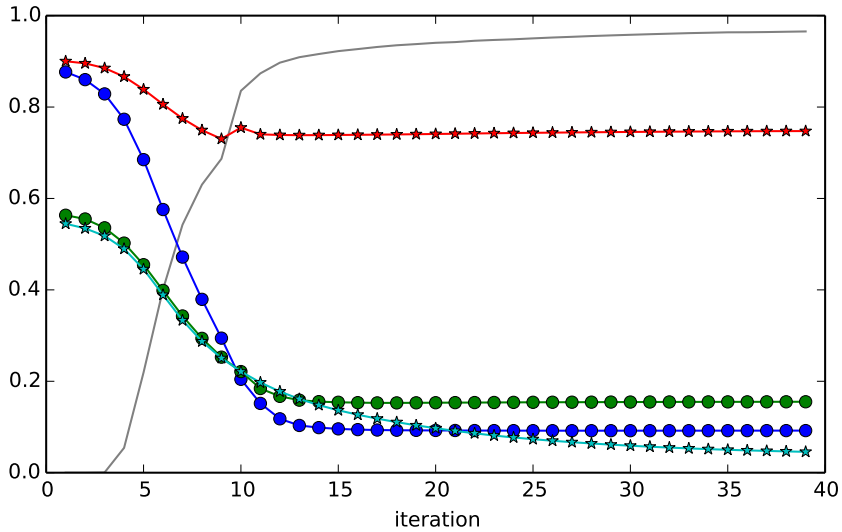
Регуляризатор Θ аппроксимирует только «свою» матрицу, а регуляризатор Φ аппроксимирует еще и родительскую Θ^{parent}.

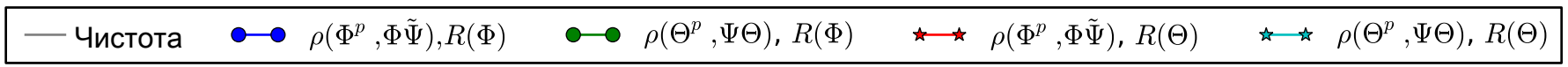
Эксперименты. Исследование сходимости модели

Разреживание с 1-й итерации **Разреживание с 10-й итерации**

Переход между 2-м и 3-м уровнями:







По оси абсцисс — номер итерации EM-алгоритма;

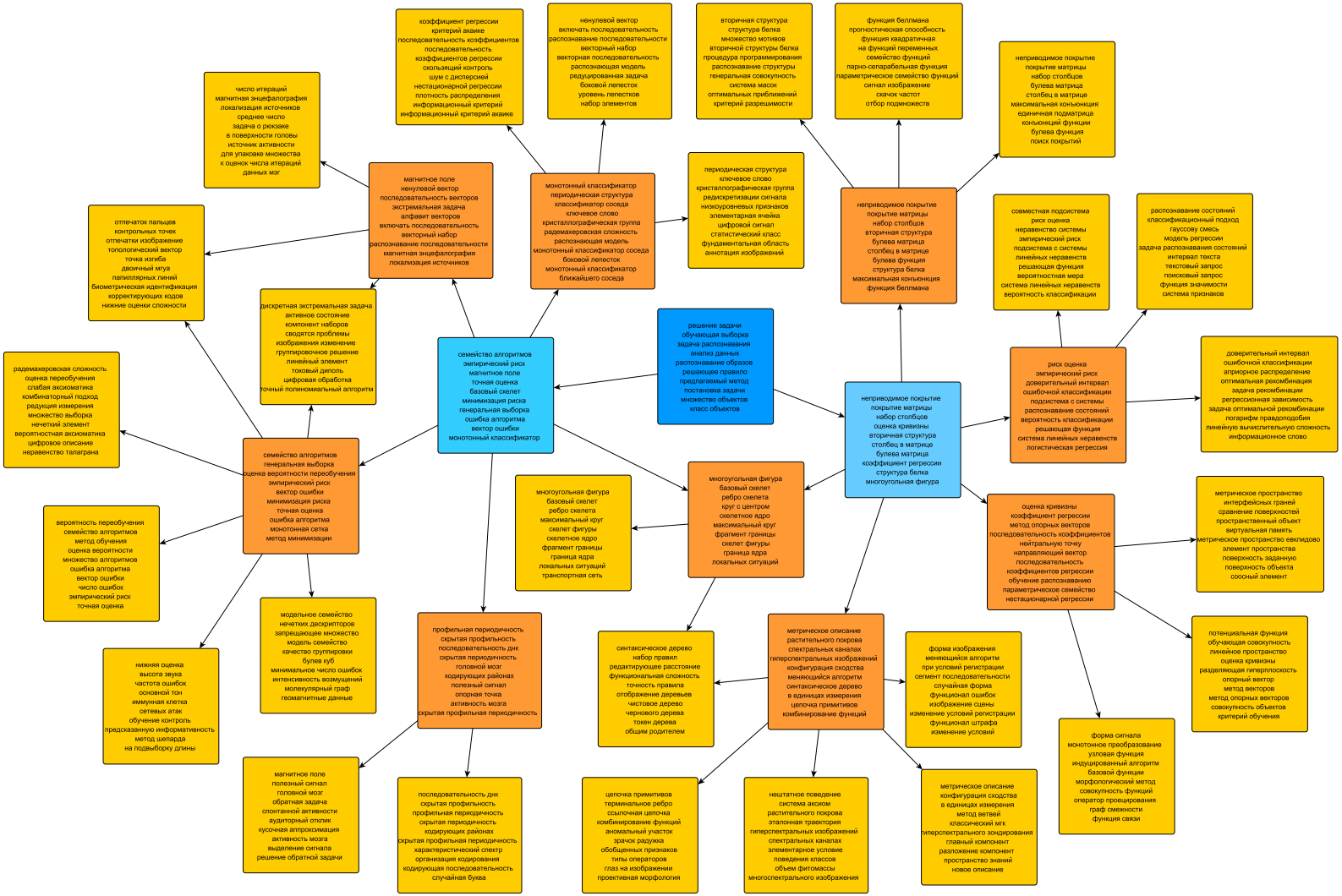
По оси ординат — расстояние между матрицами родительского уровня и их высокоранговыми разложениями на дочернем уровне.

Перспективы

Дальнейшие исследования:

- объединение двух регуляризаторов в один (влияющий на обе матрицы сразу);
- разработка новых критериев качества тематических иерархий (Оценивайте темы!);
- внедрение возможности модификации иерархии «на лету» + создание соответствующего интерфейса для эксперта;
- автоматическое именование тем;
- тестирование подхода на новых датасетах и сравнение с другими моделями.

Фрагмент иерархической тематической модели ММРО-ИОИ



explore-mmro.ru