

Иерархическая аддитивно регуляризованная тематическая модель конференций ММРО-ИОИ

Надежда Чиркова ¹
Воронцов К. В. ²

¹ВМК МГУ ²ВЦ 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 ARTM

- 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:

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

с ограничениями $\sum_w \varphi_{wt} = 1; \varphi_{wt} \geq 0,$
 $\sum_t \theta_{td} = 1; \theta_{td} \geq 0.$

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 $\Theta = \{\theta_{td}\}_{T \times D}$, $\Phi = \frac{\Phi^1}{\dots}$
 Φ^m

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

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

Оптимизировать регуляризованный логарифм правдоподобия:

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

с ограничениями $\sum_w \varphi_{wt}^m = 1$; $\varphi_{wt}^m \geq 0$, $\sum_t \theta_{td} = 1$; $\theta_{td} \geq 0$,

ξ_m — вес модальности.

Обучение плоской модели АРТМ

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

EM-алгоритм

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

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

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

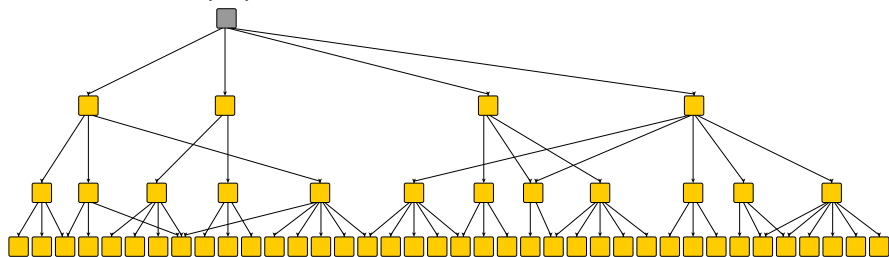
$$\varphi_{wt} = \mathop{\text{norm}}_{w \in W^m} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right),$$

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

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

Понятие тематической иерархии

Иерархическая тематическая модель — это ориентированный многодольный граф тем.



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

Цель работы: обогатить АРТМ возможностью строить тематические иерархии.

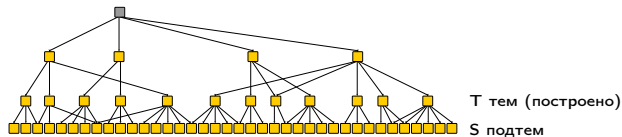
Способы построения иерархий

- 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**
Требуются 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.

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

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

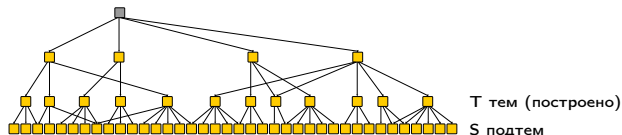
Новая оптимизационная задача:

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

Регуляризатор Θ равносителен 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 $\tilde{\Psi} \in R^{|S| \times |T|}$: $\tilde{\psi}_{st} = p(s|t)$,

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

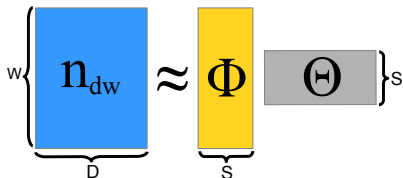
Новая оптимизационная задача:

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

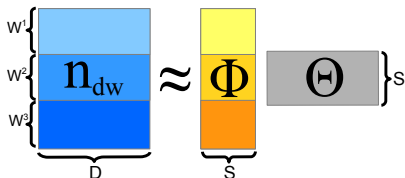
Регуляризатор Φ равносильно добавлению во входную матрицу частот слов $|T|$ псевдодокументов: $n_{dw} = \lambda \varphi_{wt}$, $d \sim t \in T!$

Структурные различия двух подходов

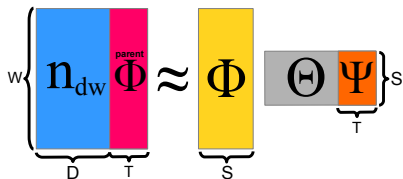
APTM



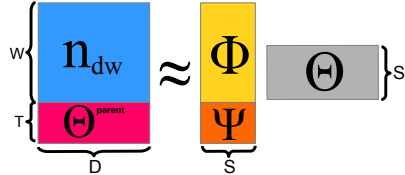
Многомодальный APTM



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».

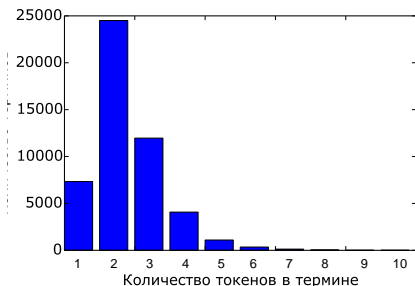
$|D| = 865$; $|W| = 42000$

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:

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

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

✓ 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**

Именованье 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.

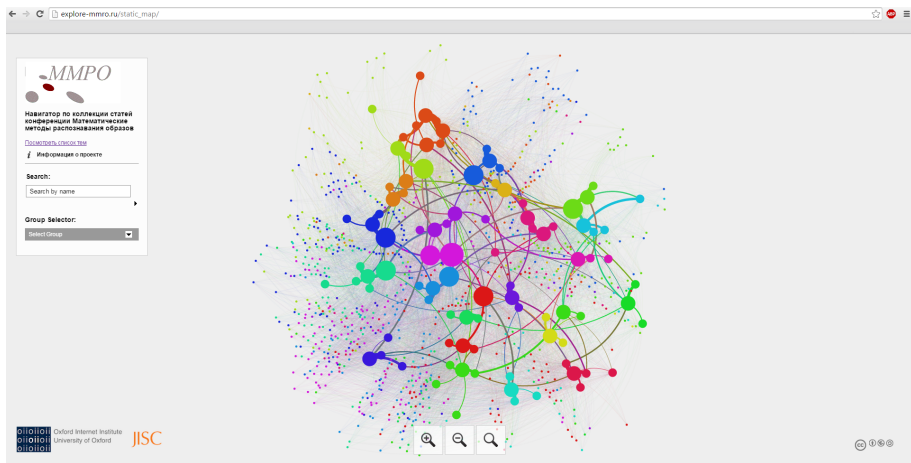
Тематический навигатор (Плавин А.):

- 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

← → C explore-mmro.ru/static_map/#Анализ сигналов и изображений

MMPO
Навигатор по коллекции статей конференции Математические методы распознавания образов
Посетить сайт
i Информация о проекте

Search:
Search by name

Group Selector:
Select Group

Закрыть

Профиль темы "Анализ сигналов и изображений"

Посмотреть тему

- нейрон
- кадр
- нейронный
- этап
- движение
- защита
- сеть
- по крести отрезков
- сканирующей
- нейроны скрытым в слое
- нейронная сеть

Документы:

- Получение оптимальных нейронных сетей в задачах распознавания изображений: оценка метода (0.1.16)
- Методы и средства распознавания образов: обзор с примерами нейронных и ненейронных технологий (0.402)
- Математические методы для распознавания изображений: анализ количественных сетей (0.404)
- Параллельные алгоритмы вычисления скорости обработки изображений в адаптивных нейронных сетях (0.349)
- Распознавание объектов в динамических базах данных: алгоритмы и программное обеспечение (0.334)
- Алгоритмы систем распознавания объектов: анализ в видеосекции (0.203)
- Визуализация процессов обучения нейронных сетей (0.272)
- СРС-методы на методах вейвлет-преобразования: анализ тематических образов объектов распознавания (0.272)
- Исследования в области моделирования процессов распознавания образов: обзор (0.202)
- Системы и анализ алгоритмов распознавания контурных сигналов

OXFORD Internet Institute University of Oxford JISC

⊕ 🔍 🔍

Визуализация и навигатор доступны по адресу
explore-mmro.ru

Навигатор по коллекции статей ММРО-ИОИ

The screenshot shows a web browser window with the URL `explore-mmro.ru/browse/?query=&present_as=topics`. The page title is "Навигатор по коллекции статей ММРО-ИОИ". Below the title, there is a search bar with the placeholder text "enter query here..." and a "Log in" button. A red button labeled "List of topics" is visible. The main content area is titled "Topics" and lists three topics:

- Topic 1: Прикладные задачи анализа данных**
 - Documents:** котельников_пшпго3_исследование котельников_пшпго4_построение кондраков_ип8_анализ чувальни_пшпго3_сигналы чувальни_ип9_адаптивное чичагов_пшпго4_исследование неймарк_пшпго4_возможности неймарк_пшпго4_новое
 - Words:** параметр траектория фазовый выборка ациклических при графов соседства аттрактор за параметров с портретом распознавание область ...
- Topic 2: Динамические системы**
 - Documents:** котельников_пшпго4_построение котельников_пшпго3_исследование неймарк_пшпго4_новое неймарк_пшпго3_применение котельников_пшпго3_синдромальные медведево_пшпго3_эффективности
 - Words:** траектория фазовый аттрактор портрет фазовый портрет рэкс очередь синхрон индикатор ...
- Topic 3: Динамические системы и управление**
 - Documents:** теклюва_пшпго6_сигналы неймарк_пшпго4_возможности неймарк_пшпго3_поставка

Визуализация и навигатор доступны по адресу
explore-mmro.ru

Критерии качества тематических моделей

Интерпретируемость

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

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 $\sum_{d_1, d_2 : \theta_{s, d_1} \neq 0, \theta_{s, d_2} \neq 0} sim(n_{d_1}^s, n_{d_2}^s) p(d_1, d_2 | s),$

внутрикластерное
расстояние

$$n_d^s = \{n_{dw}^s\}_{w \in W}, n_{dw}^s = n_{dw} p(s|d, w),$$
$$sim(d_1, d_2) = \sum_{w \in W} n_{d_1 w} n_{d_2 w}$$

Разреженность иерархической структуры

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

Качество аппроксимации родительских тем

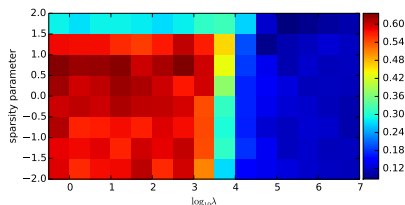
Расстояние Хеллингера между матрицами $A = \{a_{ij}\}_{i,j=1}^{m,n}, B = \{b_{ij}\}_{i,j=1}^{m,n} :$

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

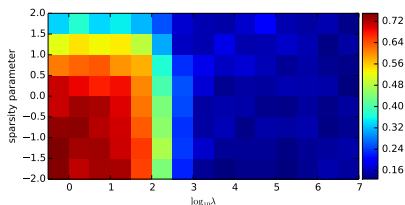
Эксперименты. Исследование взаимодействия разреживающих и сглаживающих регуляризаторов

Разреженность Ψ

Регуляризатор Φ



Регуляризатор Θ



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

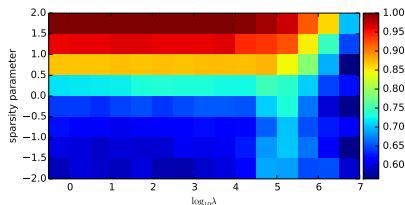
По оси ординат — параметр k разреживания предметных тем ($\tau_i = 10^k \tau_i^0, i = 1, 3, 4, 5$).

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

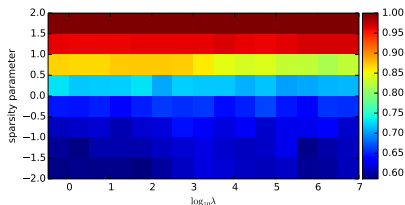
Эксперименты. Исследование взаимодействия разреживающих и сглаживающих регуляризаторов

Чистота

Регуляризатор Φ



Регуляризатор Θ



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

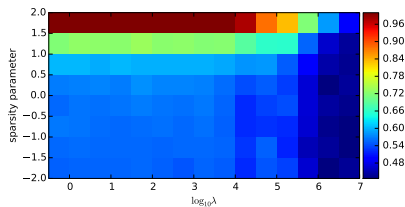
По оси ординат — параметр k разреживания предметных тем ($\tau_i = 10^k \tau_i^0, i = 1, 3, 4, 5$).

Нельзя задавать слишком большие λ — ухудшается интерпретируемость.

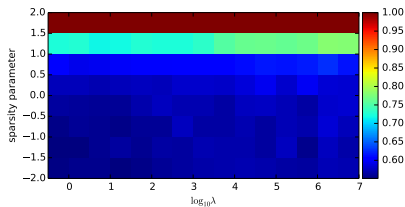
Эксперименты. Исследование взаимодействия разреживающих и сглаживающих регуляризаторов

Контрастность

Регуляризатор Φ



Регуляризатор Θ



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

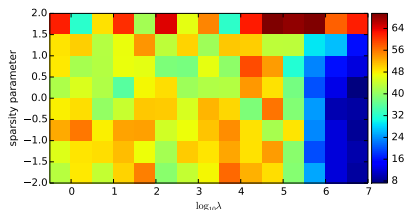
По оси ординат — параметр k разреживания предметных тем ($\tau_i = 10^k \tau_i^0, i = 1, 3, 4, 5$).

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

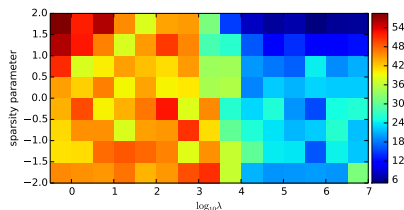
Эксперименты. Исследование взаимодействия разреживающих и сглаживающих регуляризаторов

Средневзвешенное внутрикластерное расстояние

Регуляризатор Φ



Регуляризатор Θ



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

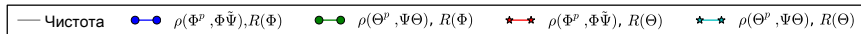
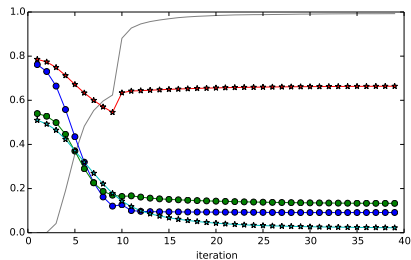
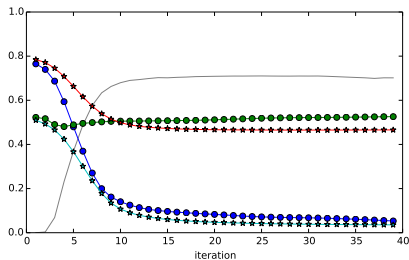
По оси ординат — параметр k разреживания предметных тем ($\tau_i = 10^k \tau_i^0, i = 1, 3, 4, 5$).

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:



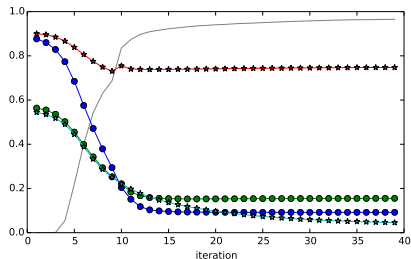
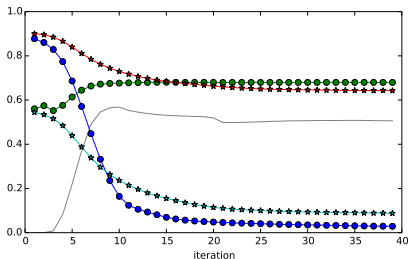
По оси абсцисс — номер итерации EM-алгоритма;

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:



— Чистота ●—● $\rho(\Phi^p, \Phi\tilde{\Psi}), R(\Phi)$ ●—● $\rho(\Theta^p, \Psi\Theta), R(\Phi)$ ★—★ $\rho(\Phi^p, \Phi\tilde{\Psi}), R(\Theta)$ ★—★ $\rho(\Theta^p, \Psi\Theta), R(\Theta)$

По оси абсцисс — номер итерации EM-алгоритма;
По оси ординат — расстояние между матрицами родительского уровня и их высокоранговыми разложениями на дочернем уровне.

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

