Иерархические аддитивно регуляризованные вероятностные тематические модели

Надежда Чиркова [1]
Научный руководитель: Воронцов К. В. [2]

[1]ВМК МГУ [2]ВЦ РАН

Традиционная молодежная школа, 19.06.2015

План выступления

1. Плоская модель
2. Понятие тематической иерархии
3. Построение иерархии
4. Подбор коэффициентов регуляризации

Задача тематического моделирования

Тематическое моделирование = мягкая *кластеризация* документов + построение *распределений слов* для тем

Вход модели:

- Матрица частот слов в документах
- Количество тем

Плоская модель

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:

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

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:

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

с ограничениями $\sum_w \varphi_{wt} = 1 \; \forall \; t$; $\varphi_{wt} \geq 0 \; \forall \; w, t$,

$\sum_t \theta_{td} = 1 \; \forall \; d$; $\theta_{td} \geq 0 \; \forall \; t, d$.

Обучение плоской модели APTM

Решение оптимизационной задачи — метод простой итерации:

EM-алгоритм

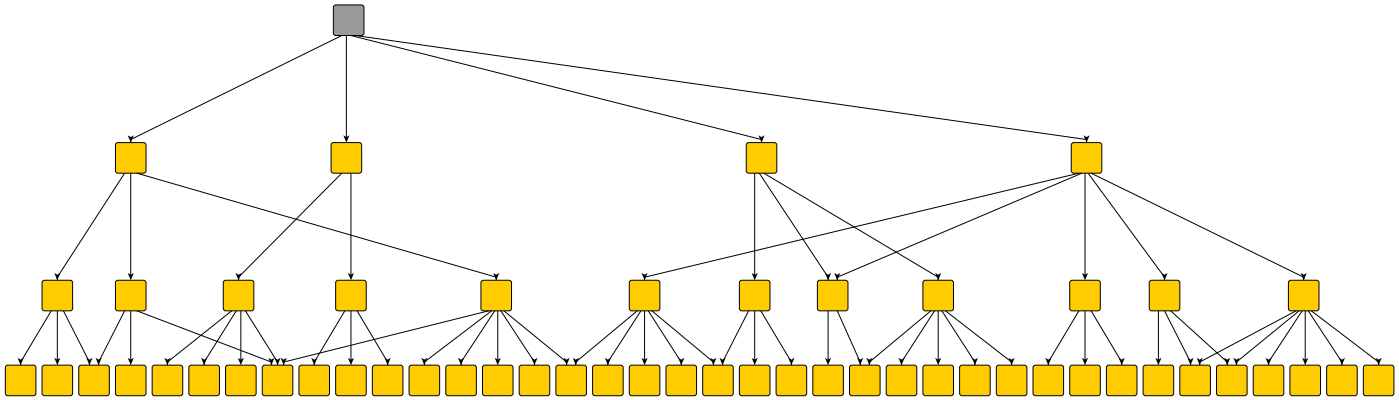
E-шаг: $p(t|d,w) = \dfrac{p(w|t)p(t|d)}{\sum_{t' \in T} p(w|t')p(t'|d)} = \dfrac{\varphi_{wt}\theta_{td}}{\sum_{t' \in T} \varphi_{wt'}\theta_{t'd}}$;

M-шаг: $\varphi_{wt} \propto \left(n_{wt} + \varphi_{wt} \dfrac{\partial R}{\partial \varphi_{wt}} \right)_+$; $n_{wt} = \sum_{d \in D} n_{dw} p(t|d,w)$;

$$\theta_{td} \propto \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+ ; \; n_{td} = \sum_{w \in W} n_{dw} p(t|d,w) \; .$$

Что такое тематическая иерархия

Иерархическая тематическая модель — это ориентированный многодольный граф тем.



✓Иерархическая структура упрощает навигацию по коллекции документов

Способы построения иерархий

Вопрос *полностью автоматического* построения тематических иерархий, а также вопрос их автоматического оценивания остаются открытыми.

Две группы подходов:

- Восходящее построение (объединять мелкие темы в крупные)
- Нисходящее построение (делить крупные темы на мелкие)

Многие [особенно нисходящие] методы строят *дерево*, т.е. не разрешают множественного наследования тем.

Рекурсивное построение 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:

$$n_{dw}^t = n_{dw} p(t|d,w), p(t|d,w) \propto \varphi_{wt}\theta_{td}, t = 1, \ldots, T$$

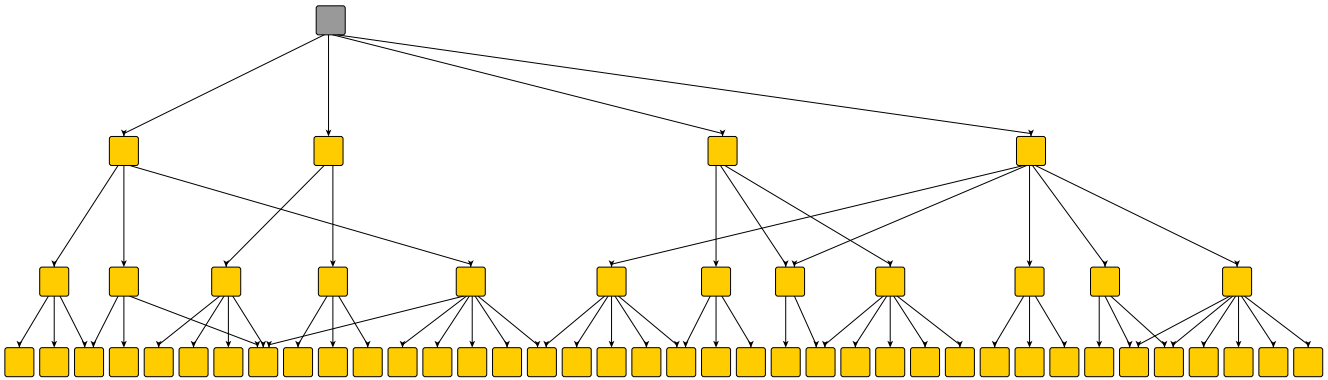
3. для всех $t = 1, \ldots, T$:
   Построить Узел (n_{dw}^t, T_t)

Недостатки:

- Количество тем задается вручную для каждого узла дерева
- Нет множественного наследования
- Высокая чувствительность к ошибкам на верхних уровнях (например, дублирование или смешение тем)

Иерархический регуляризатор

Будем строить иерархию уровень за уровнем, постоянно увеличивая количество тем:



T тем (построено)

S тем

Обычная модель:

параметры Φ, Θ, $\varphi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$, цель: $F \approx \Phi\Theta$

Добавим новую матрицу параметров $\Psi \in R^{T \times S}$:

$$\psi_{ts} = p(t|s), \Theta^{parent} \approx \Psi\Theta$$

Новая оптимизационная задача:

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

Обучение уровня иерархической модели APTM

Метод простой итерации:

EM-алгоритм

E-шаг: $p(s|d,w) \propto \varphi_{ws}\theta_{sd}$

$$p(s|t,d) \propto \psi_{ts}\theta_{sd}$$

M-шаг: $n_{ws} = \sum_{d \in D} n_{dw} p(s|d,w), \quad n^1_{sd} = \sum_{w \in W} n_{dw} p(s|d,w)$

$$n_{ts} = \sum_{d \in D} \theta^{par}_{td} p(s|t,d), \quad n^2_{sd} = \sum_{t \in T} \theta^{par}_{td} p(s|t,d)$$

$$\phi_{ws} \propto \left(n_{ws} + \phi_{ws} \frac{\partial R}{\partial \phi_{ws}} \right)_+$$

$$\psi_{ts} \propto \left(n_{ts} + \psi_{ts} \frac{\partial R}{\partial \psi_{ts}} \right)_+$$

$$\theta_{sd} \propto \left(n^1_{sd} + \lambda n^2_{sd} + \theta_{sd} \frac{\partial R}{\partial \theta_{sd}} \right)_+$$

Результат

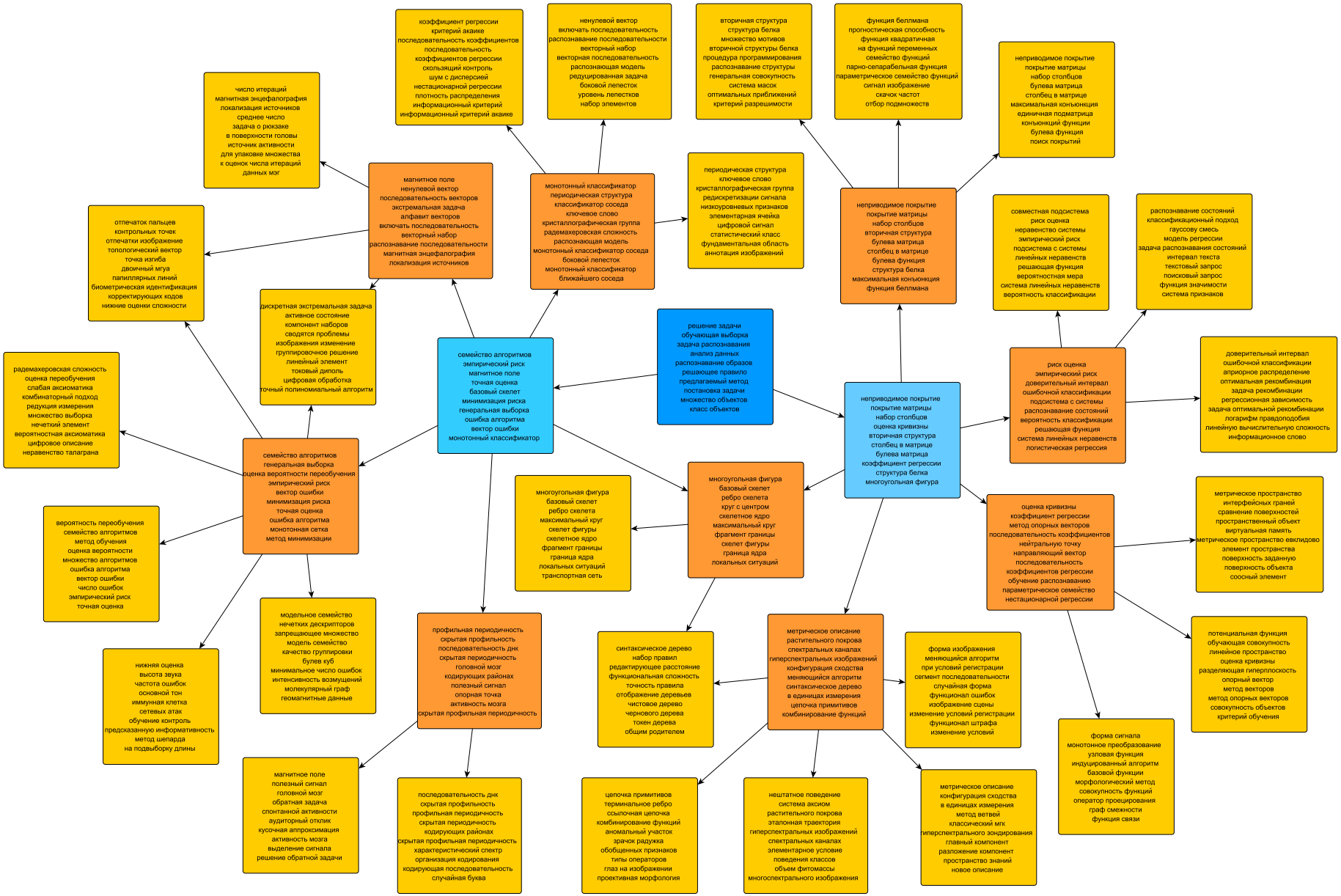
Навигатор по коллекции статей конференций «Интеллектуализация обработки информации» и «Математические методы распознавания образов»:

MMROnavigator.vv.si

$D = 850$; $W = 42000$

Мультиграммы (словосочетания) выделены с использованием внешнего софта

Визуализация иерархии



Критерии качества тематических моделей

Размер ядра темы:
$\text{size} = |W_t|, \quad W_t = \{w : p(t|w) > 0.25\}$

Контрастность темы: $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$

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

Когерентность: $\frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j=1}^{i-1} PMI(w_i, w_j), PMI(w_i, w_j) = \ln \frac{p(w_i, w_j)}{p(w_i)p(w_j)}$

слова в теме t отсортированы по убыванию $p(w|t)$.

Сравнение плоской и двух иерархических моделей

Сравнение качества плоской модели на 60 тем и 3-х уровней иерархических моделей:

Модель	**Чистота**	**Контрастность**
Плоская	0.999	0.961
Рекурсивная	0.878	0.788
Регуляризатор	0.998	0.959

Выбор траектории регуляризации

Эксперименты по сравнению траекторий регуляризации проведены для рекурсивного подхода.

Первый уровень:

Комбинация регуляризаторов	**Ср. чистота**	**Ср. контрть**	**Разреж. Φ**	**Разреж. Θ**
Сглаж. фон. тема	0.981	0.907	0.708	0.225
Разреж. фон. тема	0.971	0.857	0.787	0.214
Разреж. и сглаж. фон. темы	0.994	0.957	0.839	0.106
Разреж. и сглаж. фон. темы + разреж. и сглаж. Θ	0.989	0.883	0.769	0.57

Выбор траектории регуляризации

Эксперименты по сравнению траекторий регуляризации проведены для рекурсивного подхода.
Второй уровень:

Комбинация регуляризаторов	**Ср. чистота**	**Ср. контрть**	**Разреж. Φ**	**Разреж. Θ**
Сглаж. фон. тема	0.846	0.757	0.944	0.754
Разреж. фон. тема	0.866	0.754	0.943	0.753
Разреж. и сглаж. фон. темы	0.870	0.768	0.947	0.766
Разреж. и сглаж. фон. темы + разреж. и сглаж. Θ	0.896	0.796	0.966	0.984

Выбор траектории регуляризации

Эксперименты по сравнению траекторий регуляризации проведены для рекурсивного подхода.

Третий уровень:

Комбинация регуляризаторов	**Ср. чистота**	**Ср. контрть**	**Разреж. Φ**	**Разреж. Θ**
Сглаж. фон. тема	0.819	0.743	0.976	0.780
Разреж. фон. тема	0.820	0.757	0.976	0.778
Разреж. и сглаж. фон. темы	0.838	0.758	0.977	0.785
Разреж. и сглаж. фон. темы + разреж. и сглаж. Θ	0.878	0.788	0.990	0.996

Перспективы

Дальнейшие исследования:

- оценивание и улучшение качества иерархии;
- частичное обучение (использование экспертной разметки);
- автоматическое именование тем.