Вероятностное тематическое моделирование

Константин Воронцов
ВЦ РАН • МФТИ • МГУ • ВШЭ • Яндекс • FORECSYS

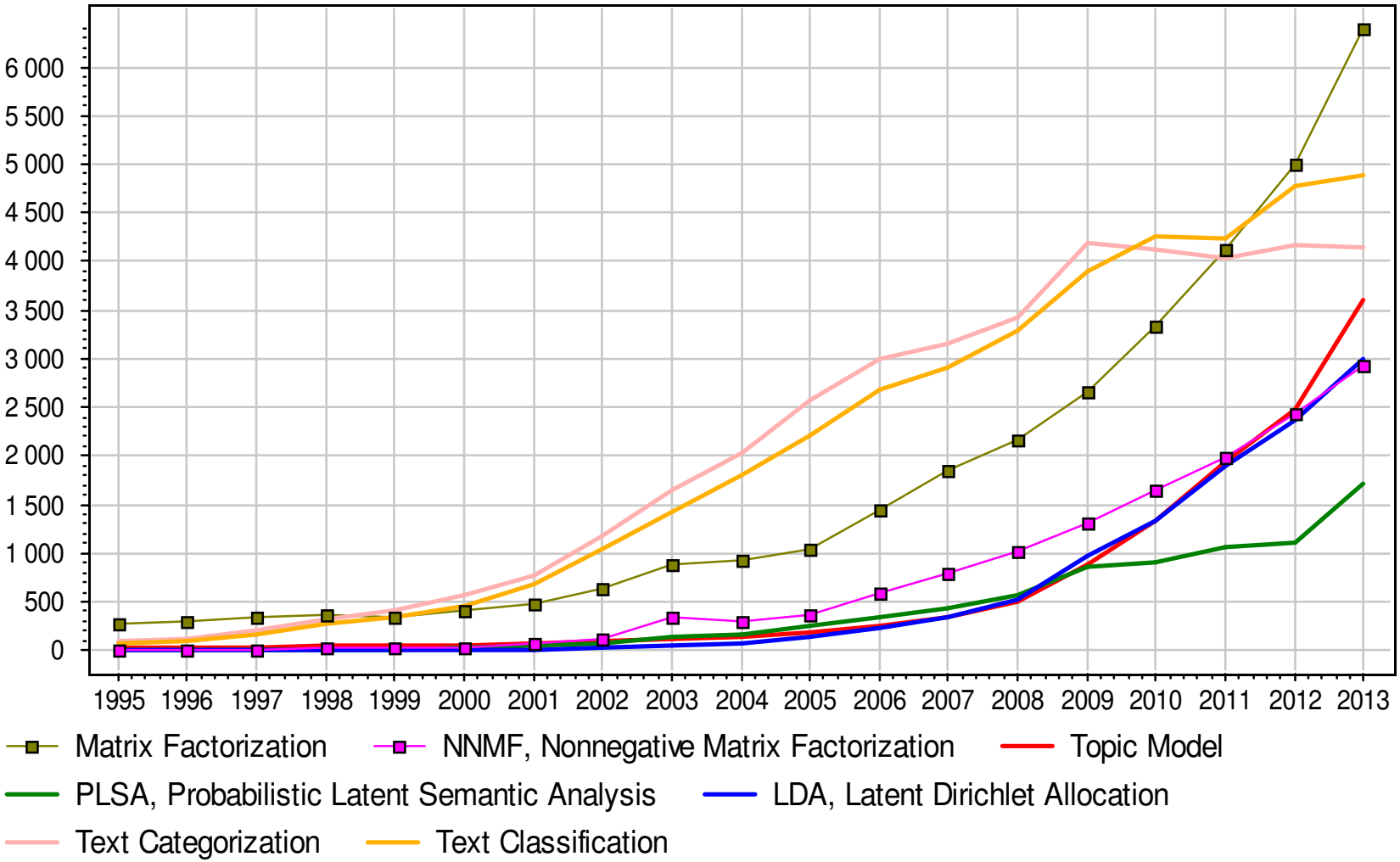
научный семинар • ЛИНИС • 11 сентября 2014

Содержание

1. **Вероятностное тематическое моделирование**
   - Задача тематического моделирования
   - Модель PLSA и EM-алгоритм
   - Модель LDA
2. **Аддитивная регуляризация тематических моделей**
   - Комбинирование регуляризаторов и EM-алгоритм
   - Регуляризаторы для улучшения интерпретируемости
   - Эксперимент
3. **Открытые проблемы, направления исследований**
   - Направления ближайших исследований
   - Лингвистически-ориентированные BTM
   - Мультимодальные, динамические, мультиграммные BTM

Тематическое моделирование и близкие области исследований

Динамика цитирования, по данным Google Scholar:



Понятие «латентной темы»

- *Тема* — специальная терминология предметной области.
- *Тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах.
- *Тема* — вероятностное распределение на терминах: $p(w|t)$ — вероятность встретить термин w в теме t.

Документ d состоит из наблюдаемых терминов $w_1, \ldots, w_{n_d}$, $p(w|d)$ — известная частота термина w в документе d.

Документ имеет ненаблюдаемый *тематический профиль*: $p(t|d)$ — неизвестная частота темы t в документе d.

Когда автор писал термин w в документ d, он думал о теме t.

Тематическая модель пытается выявить латентные темы и оценить распределения $\phi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$.

Постановка задачи

Дана коллекция текстовых документов (мешков слов):
n_{dw} — сколько раз термин w встречается в документе d

Найти модель $p(w|d) = \sum_t \phi_{wt}\theta_{td}$ с параметрами ϕ, θ:
$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Задача максимизации логарифма правдоподобия:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta},$$

при ограничениях нормировки и неотрицательности

$$\phi_{wt} \geqslant 0; \quad \sum_w \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_t \theta_{td} = 1$$

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Теорема

Точка максимума правдоподобия Φ, Θ удовлетворяет системе уравнений со вспомогательными переменными p_{tdw}:

$$\text{E-шаг:} \quad p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{t'} \phi_{wt'}\theta_{t'd}};$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \dfrac{n_{wt}}{\sum_{w'} n_{w't}}; & n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw}; \\ \theta_{td} = \dfrac{n_{td}}{\sum_{t'} n_{t'd}}; & n_{td} = \sum\limits_{w \in d} n_{dw} p_{tdw}; \end{cases}$$

EM-алгоритм — чередование E- и M-шага до сходимости, т. е. решение системы уравнений методом простых итераций.

✓ *Идея на будущее: можно использовать и другие методы!*

Вероятностная интерпретация шагов EM-алгоритма

E-шаг — это формула Байеса:

$$p_{tdw} = p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T}\phi_{ws}\theta_{sd}}$$

$n_{dwt} = n_{dw}p(t|d,w)$ — оценка числа троек (d,w,t) в коллекции

M-шаг — это частотные оценки условных вероятностей:

$$\phi_{wt} = \frac{n_{wt}}{n_t} \equiv \frac{\sum\limits_{d\in D} n_{dwt}}{\sum\limits_{d\in D}\sum\limits_{w\in d} n_{dwt}}, \qquad \theta_{td} = \frac{n_{td}}{n_d} \equiv \frac{\sum\limits_{w\in d} n_{dwt}}{\sum\limits_{w\in W}\sum\limits_{t\in T} n_{dwt}},$$

Краткая запись через знак пропорциональности $\propto$:

$$\phi_{wt} \propto n_{wt}; \qquad \theta_{td} \propto n_{td};$$

Рациональный EM-алгоритм

Идея: E-шаг встраивается внутрь M-шага

Вход: коллекция D, число тем $|T|$, число итераций $i_{\max}$;
Выход: матрицы терминов тем Θ и тем документов Φ;

инициализация ϕ_{wt}, θ_{td} для всех $d \in D,\ w \in W,\ t \in T$;
для всех итераций $i = 1, \ldots, i_{\max}$
  $n_{wt}, n_{td}, n_t, n_d := 0$ для всех $d \in D,\ w \in W,\ t \in T$;
  для всех документов $d \in D$ и всех слов $w \in d$
    $p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$ для всех $t \in T$;
    $n_{wt}, n_{td}, n_t, n_d \mathrel{+}= n_{dw} p_{tdw}$ для всех $t \in T$;
  $\phi_{wt} := n_{wt}/n_t$ для всех $w \in W,\ t \in T$;
  $\theta_{td} := n_{td}/n_d$ для всех $d \in D,\ t \in T$;

Онлайновый EM-алгоритм (для больших коллекций)

инициализировать ϕ_{wt} для всех $w \in W,\ t \in T$;
$n_{wt} := 0,\ n_t := 0$ для всех $w \in W,\ t \in T$;
для всех пачек документов $D_j,\ j = 1, \ldots, J$
  $\tilde{n}_{wt} := 0,\ \tilde{n}_t := 0$ для всех $w \in W,\ t \in T$;
  для всех документов d из пачки D_j
    инициализировать θ_{td} для всех $t \in T$;
    повторять
      $p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$ для всех $w \in d,\ t \in T$;
      $\theta_{td} := \frac{1}{n_d} \sum\limits_{w \in d} n_{dw} p_{tdw}$ для всех $t \in T$;
    пока θ_d не сойдётся;
    $\tilde{n}_{wt}, \tilde{n}_t \mathrel{+}= n_{dw} p_{tdw}$ для всех $w \in d,\ t \in T$;
  $n_{wt} := \rho_j n_{wt} + \tilde{n}_{wt};\ \ n_t := \rho_j n_t + \tilde{n}_t$ для всех $w \in W,\ t \in T$;
  $\phi_{wt} := n_{wt}/n_t$ для всех $w \in W,\ t \in T$;

LDA — Latent Dirichlet Allocation [Blei 2003]

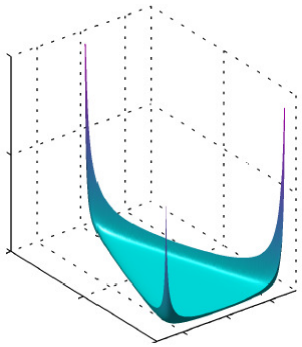
Гипотеза. Вектор-столбцы $\phi_t = (\phi_{wt})_{w \in W}$ и $\theta_d = (\theta_{td})_{t \in T}$ порождаются распределениями Дирихле, $\alpha \in \mathbb{R}^{|T|}$, $\beta \in \mathbb{R}^{|W|}$:

$$\mathrm{Dir}(\phi_t|\beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \beta_0 = \sum_w \beta_w, \ \beta_t \geqslant 0;$$

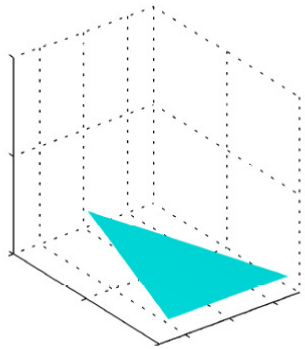
$$\mathrm{Dir}(\theta_d|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \alpha_0 = \sum_t \alpha_t, \ \alpha_t \geqslant 0;$$

Пример:

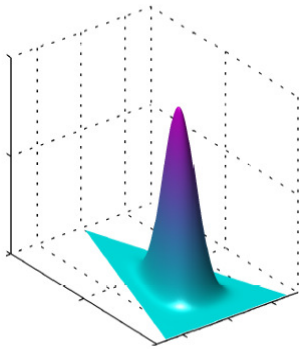
$\mathrm{Dir}(\theta|\alpha)$
$|T| = 3$
$\theta, \alpha \in \mathbb{R}^3$



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$



$\alpha_1 = \alpha_2 = \alpha_3 = 1$



$\alpha_1 = \alpha_2 = \alpha_3 = 10$

Главное отличие LDA от PLSA

Оценки условных вероятностей $\phi_{wt} \equiv p(w|t)$, $\theta_{td} \equiv p(t|d)$:

- в PLSA — несмещённые оценки максимума правдоподобия:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad \theta_{td} = \frac{n_{td}}{n_d}$$

- в LDA — сглаженные байесовские оценки:

$$\phi_{wt} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}, \qquad \theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}.$$

Различие проявляется только при малых n_{wt}, n_{td}.
Робастные LDA и PLSA почти одинаковы по качеству.

Asuncion A., Welling M., Smyth P., Teh Y. W. On smoothing and inference for topic models. Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.

Potapenko A. A., Vorontsov K. V. Robust PLSA Performs Better Than LDA. ECIR-2013, Moscow, Russia, 24-27 March 2013. LNCS, Springer. Pp. 784–787.

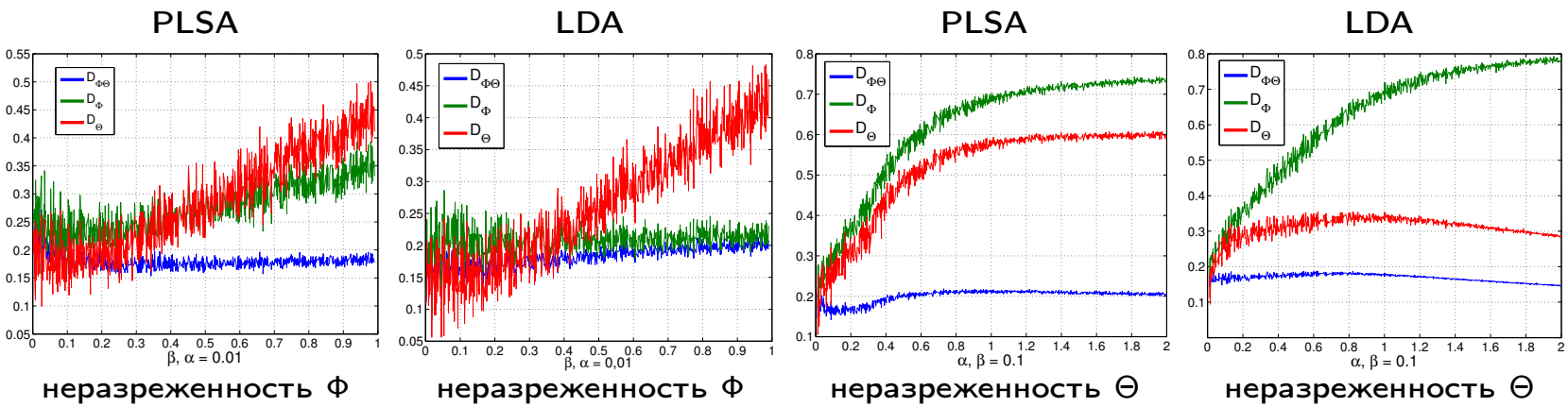
Задача построения BTM — некорректно поставленная

Неединственность стохастического матричного разложения:

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

для любых $S_{T\times T}$ таких, что Φ', Θ' — стохастические.

Эксперимент. Произведение $\Phi\Theta$ восстанавливается устойчиво, матрица Φ и матрица Θ — только когда сильно разрежены:



Вывод 1: нужны дополнительные требования к модели.

Вывод 2: требований сглаживания в LDA не достаточно.

Аддитивная регуляризация тематической модели

Пусть, наряду с правдоподобием, требуется максимизировать ещё n критериев — *регуляризаторов* $R_i(\Phi, \Theta)$, $i = 1, \ldots, n$.

Метод многокритериальной оптимизации — *скаляризация*.

Задача: максимизировать регуляризованное правдоподобие

$$\underbrace{\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\text{log-likelihood } \mathscr{L}(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta},$$

при ограничениях неотрицательности и нормировки

$$\phi_{wt} \geqslant 0; \quad \sum_{w \in W} \phi_{wt} = 1; \qquad \theta_{td} \geqslant 0; \quad \sum_{t \in T} \theta_{td} = 1$$

где $\tau_i > 0$ — *коэффициенты регуляризации*.

EM-алгоритм с регуляризацией M-шага

Теорема

Максимум $\mathscr{L}(\Phi,\Theta)+R(\Phi,\Theta)$ удовлетворяет системе уравнений со вспомогательными переменными p_{tdw}, n_{wt}, n_{td}:

$$\text{E-шаг:}\quad p_{tdw}=\frac{\phi_{wt}\theta_{td}}{\sum_{t'}\phi_{wt'}\theta_{t'd}};$$

$$\text{M-шаг:}\quad \begin{cases} \phi_{wt}\propto\left(n_{wt}+\phi_{wt}\dfrac{\partial R}{\partial\phi_{wt}}\right)_+; \quad n_{wt}=\sum\limits_{d\in D}n_{dw}p_{tdw}; \\ \theta_{td}\propto\left(n_{td}+\theta_{td}\dfrac{\partial R}{\partial\theta_{td}}\right)_+; \quad n_{td}=\sum\limits_{w\in d}n_{dw}p_{tdw}; \end{cases}$$

где $(x)_+=\max(x,0)$ — операция положительной срезки.

PLSA: $R(\Phi,\Theta)=0$

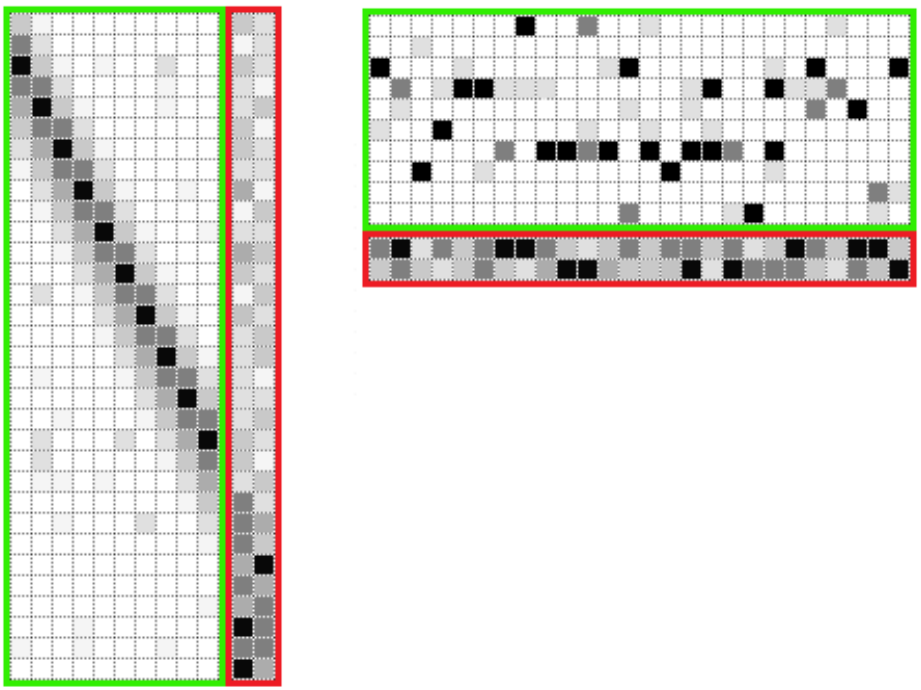
LDA: $R(\Phi,\Theta)=\sum\limits_{t,w}\beta_w\ln\phi_{wt}+\sum\limits_{d,t}\alpha_t\ln\theta_{td}$

Требования интерпретируемости и гипотеза о структуре тем

Предметные темы S содержат термины предметной области, распределения $p(w|t)$ разреженные, существенно различные

Фоновые темы B содержат слова общей лексики, распределения $p(w|t)$ и $p(t|d)$ не разреженные

$\Phi_{W \times T}$ $\Theta_{T \times D}$



Справочные сведения. Дивергенция Кульбака–Лейблера

Функция расстояния между распределениями $P = (p_i)_{i=1}^n$ и $Q = (q_i)_{i=1}^n$:

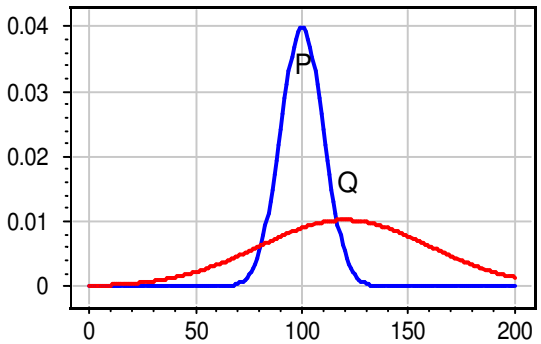
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

1. $\mathrm{KL}(P\|Q) \geqslant 0$; $\quad \mathrm{KL}(P\|Q) = 0 \;\Leftrightarrow\; P = Q$;

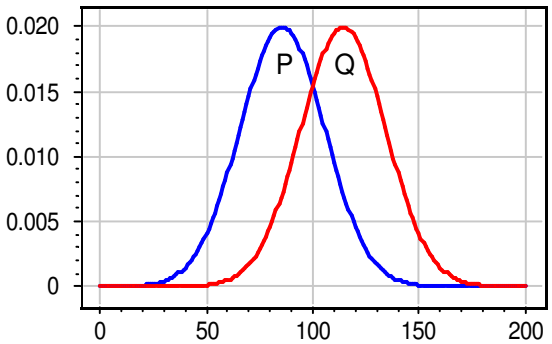
2. Минимизация KL эквивалентна максимизации правдоподобия:

$$\mathrm{KL}(P\|Q(\alpha)) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i(\alpha)} \to \min_\alpha \quad\Longleftrightarrow\quad \sum_{i=1}^n p_i \ln q_i(\alpha) \to \max_\alpha.$$

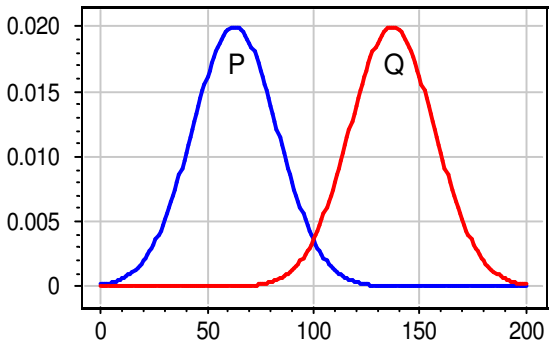
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



$\mathrm{KL}(P\|Q) = 0.442$
$\mathrm{KL}(Q\|P) = 2.966$



$\mathrm{KL}(P\|Q) = 0.444$
$\mathrm{KL}(Q\|P) = 0.444$



$\mathrm{KL}(P\|Q) = 2.969$
$\mathrm{KL}(Q\|P) = 2.969$

Регуляризатор сглаживания (переосмысление LDA)

Гипотеза сглаженности фоновых тем $t \in B$:
распределения ϕ_{wt} близки к заданному распределению β_w
распределения θ_{td} близки к заданному распределению α_t

$$\sum_{t\in B} \mathrm{KL}_w(\beta_w \| \phi_{wt}) \to \min_{\Phi}; \qquad \sum_{d\in D} \mathrm{KL}_t(\alpha_t \| \theta_{td}) \to \min_{\Theta}.$$

Максимизируем сумму этих регуляризаторов:

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in B}\sum_{w\in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d\in D}\sum_{t\in B} \alpha_t \ln \theta_{td} \to \max.$$

Подставляем, получаем формулы M-шага LDA, для всех $t \in B$:

$$\phi_{wt} \propto n_{wt} + \beta_0\beta_w, \qquad \theta_{td} \propto n_{td} + \alpha_0\alpha_t.$$

Это новая, не-байесовская интерпретация LDA [Blei 2003].

Регуляризатор для разреживания предметных тем

Гипотеза разреженности предметных тем $t \in S$:
среди ϕ_{wt}, θ_{td} много нулевых значений.

Максимизируем дивергенцию между заданными распределениями β_w, α_t и искомыми ϕ_{wt}, θ_{td}:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in S} \alpha_t \ln \theta_{td} \to \max.$$

Подставляем, получаем «анти-LDA» для всех $t \in S$:

$$\phi_{wt} \propto \left(n_{wt} - \beta_0 \beta_w\right)_+, \qquad \theta_{td} \propto \left(n_{td} - \alpha_0 \alpha_t\right)_+.$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions.

Регуляризатор для декоррелирования предметных тем

Гипотеза некоррелированности предметных тем $t \in S$:
чем различнее темы, тем лучше они интерпретируются.

Минимизируем ковариации между вектор-столбцами ϕ_t:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in S} \sum_{s \in S \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max .$$

Подставляем, получаем ещё один вариант разреживания —
постепенное контрастирование строк матрицы Φ:

$$\phi_{wt} \propto \left(n_{wt} - \tau \phi_{wt} \sum_{s \in S \setminus t} \phi_{ws} \right)_+ .$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

Регуляризатор для сокращения числа тем

Гипотеза: если в теме слишком мало слов, то она не нужна.

Разреживаем распределение $p(t) = \sum_d p(d)\theta_{td}$, максимизируя KL-дивергенцию между $p(t)$ и равномерным распределением:

$$R(\Theta) = -\tau \sum_{t \in S} \ln \sum_{d \in D} p(d)\theta_{td} \to \max.$$

Подставляем, получаем:

$$\theta_{td} \propto \left(n_{td} - \tau \frac{n_d}{n_t} \theta_{td} \right)_+.$$

Эффект:
строки матрицы Θ могут целиком обнуляться для тем t, собравших мало слов по коллекции, $n_t = \sum_d \sum_w n_{dwt}$.

Регуляризатор для максимизации когерентности тем

Гипотеза: тема лучше интерпретируется, если она содержит *когерентные* (часто встречающиеся рядом) слова $u, w \in W$.

Пусть C_{uw} — оценка когерентности, например $\hat{p}(w|u) = \frac{N_{uw}}{N_u}$.
Согласуем ϕ_{wt} с оценками $\hat{p}(w|t)$ по когерентным словам,

$$\hat{p}(w|t) = \sum_u p(w|u)p(u|t) = \frac{1}{n_t}\sum_u C_{uw} n_{ut};$$

$$R(\Phi,\Theta) = \tau \sum_{t\in T} n_t \sum_{w\in W} \hat{p}(w|t) \ln \phi_{wt} \to \max.$$

Подставляем, получаем ещё один вариант сглаживания:

$$\phi_{wt} \propto n_{wt} + \tau \sum_{u \in W\setminus w} C_{uw} n_{ut}.$$

Mimno D., Wallach H. M., Talley E., Leenders M., McCallum A. Optimizing semantic coherence in topic models // Empirical Methods in Natural Language Processing, EMNLP-2011. — Pp. 262–272.

Эксперимент по комбинированию регуляризаторов

Задача: улучшить интерпретируемость, не ухудшив перплексию

Набор регуляризаторов:

1. сглаживание фоновых тем — столбцов Φ, строк Θ
2. разреживание предметных тем — столбцов Φ, строк Θ
3. декоррелирование предметных тем — столбцов Φ
4. удаление незначимых тем — строк Θ

Данные: NIPS (Neural Information Processing System)

- $|D| = 1566$ статей конференции NIPS на английском языке;
- суммарной длины $n \approx 2.3 \cdot 10^6$,
- словарь $|W| \approx 1.3 \cdot 10^4$.
- контрольная коллекция: $|D'| = 174$.

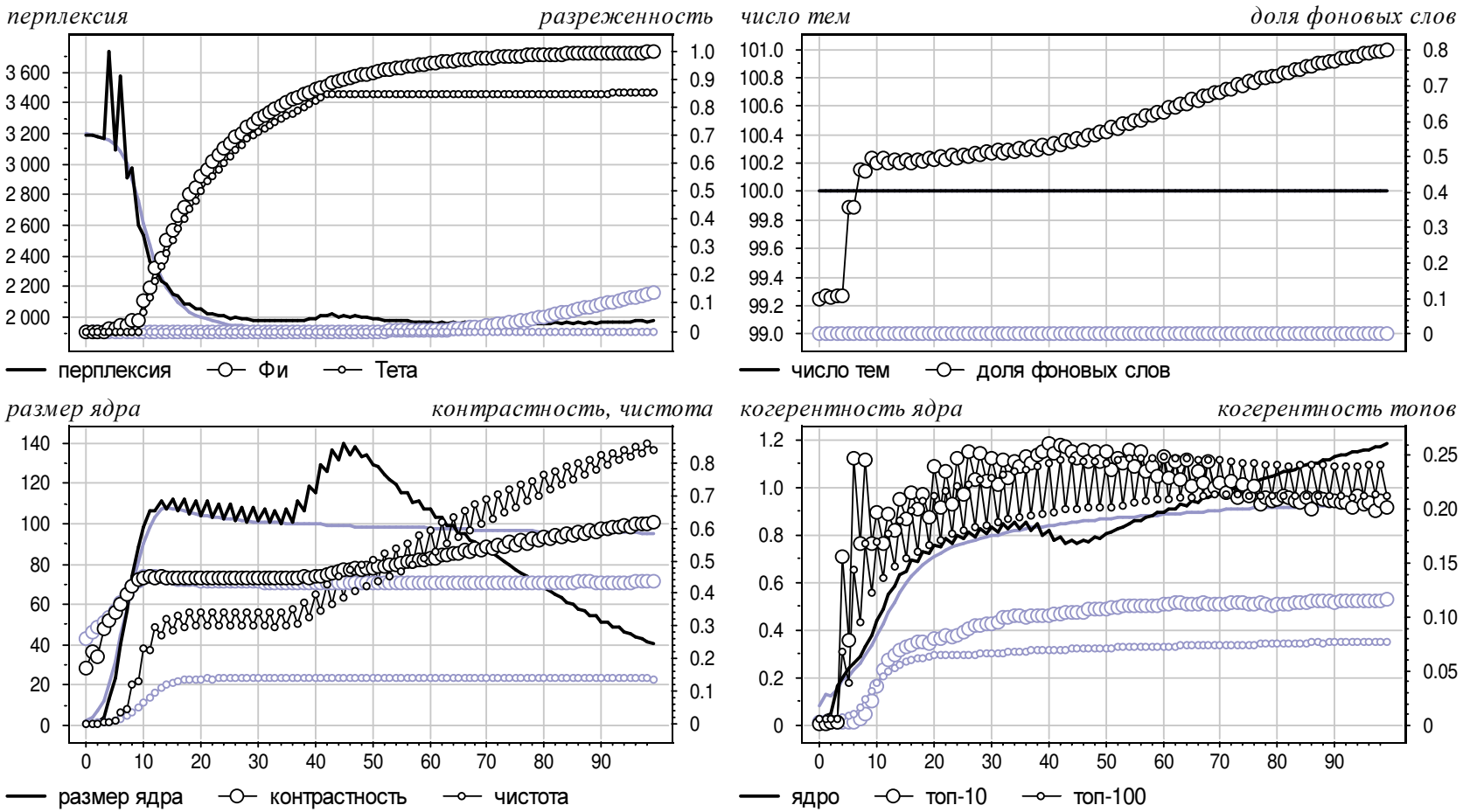
Критерии качества модели

Построение BTM — многокритериальная оптимизация.
Поэтому критериев для контроля качества модели тоже много.

- Перплексия контрольной коллекции: $\mathcal{P} = \exp\left(-\frac{1}{n'}\mathscr{L}(D')\right)$
- Разреженность — доля нулевых элементов в Φ и Θ
- Характеристики интерпретируемости тем:
  - когерентность темы [Newman, 2010]
  - размер ядра темы: $|W_t|$, ядро $W_t = \{w\colon p(t|w) > 0.25\}$
  - чистота темы: $\sum\limits_{w \in W_t} p(w|t)$
  - контрастность темы: $\frac{1}{|W_t|}\sum\limits_{w \in W_t} p(t|w)$
- Вырожденность тематической модели:
  - число тем: $|T|$
  - доля фоновых слов: $\frac{1}{n}\sum\limits_{d \in D}\sum\limits_{w \in d}\sum\limits_{t \in B} p(t|d,w)$

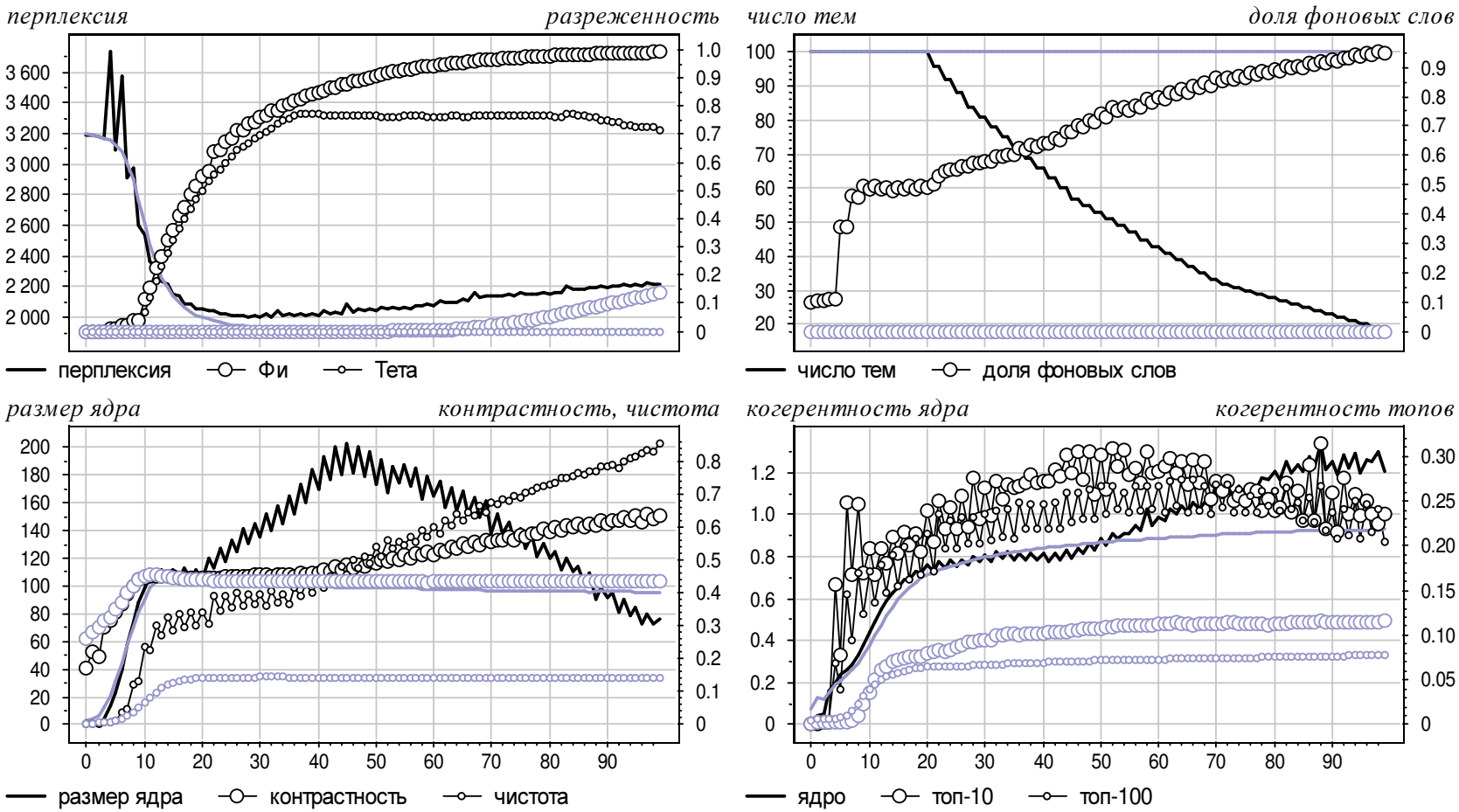
Разреживание, сглаживание, декорреляция, сокращение тем

Зависимости критериев качества от итераций EM-алгоритма (серый — PLSA, чёрный — ARTM)



Все те же, с удалением незначимых тем

Зависимости критериев качества от итераций EM-алгоритма (серый — PLSA, чёрный — ARTM)



Выводы по результатам экспериментов

Одновременное улучшение многих показателей:

- *разреженность* выросла от 0 до 95%–98%
- *когерентность тем* выросла от 0.1 до 0.3
- *чистота тем* выросла от 0.15 до 0.8
- *контрастность тем* выросла от 0.4 до 0.6
- *размер ядер тем* вырос от 0 до 150 терминов
- почти без потери *перплексии* (правдоподобия) модели

Выработаны рекомендации по подбору траектории:

- разреживание включать постепенно после 10-20 итераций
- сглаживание включать сразу
- декорреляцию включать сразу и как можно сильнее
- сокращение числа тем включать постепенно,
- никогда не совмещая с декорреляцией на одной итерации

Направления ближайших исследований

- Лингвистическая регуляризация, отказ от «мешка слов»
  - учёт линейной структуры текста, регуляризация $p(t|d,w)$
  - полностью автоматическое выделение терминов
  - учёт лингвистических ресурсов (тезаурусов, онтологий)
  - гибрид с методом лексических цепочек
  - учёт лингвистических моделей тематики и дискурса
- BigARTM — библиотека с открытым кодом
  - параллельные вычисления
  - распределённое хранение коллекции
  - любые сочетания регуляризаторов и метрик качества
  - расширение библиотеки регуляризаторов и метрик качества

Текущие исследования

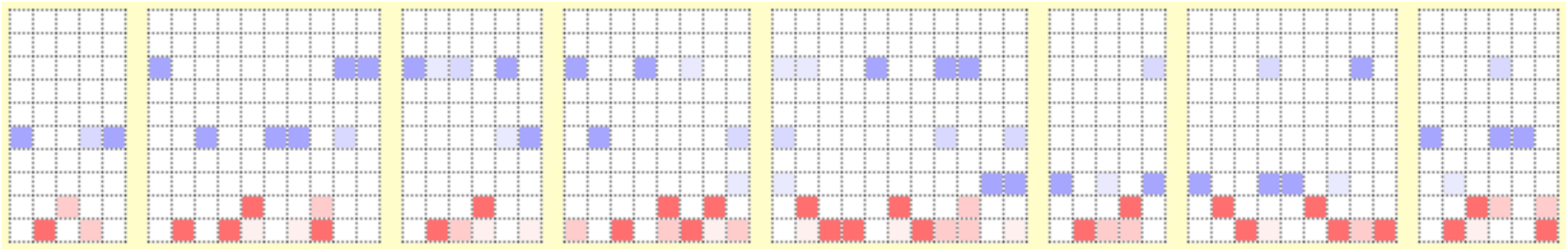
- Оптимизация числа тем
- Мультимодальные модели
- Мультиязычные модели
- Динамические модели
- Нисходящие иерархические модели
- Визуализация модели и результатов тематического поиска
- Автоматическое именование тем

Прикладные задачи

- Диагностика многих заболеваний по одной ЭКГ
- Иерархия тематики конференций ММРО, ИОИ, EURO
- Анализ потока правительственных пресс-релизов
- Персонализация показа рекламных баннеров

Тематическое моделирование предложений (Sentence PTM)

Тематические профили слов $p(t|d,w)$ документа d образуют матрицу размера $T \times n_d$, где n_d — длина документа:



Предположения разреженности и непрерывности тематики:

- каждое предложение относится к 1–2 предметным темам
- слова общей лексики не влияют на тематику предложений
- соседние предложения часто относятся к одним темам
- между абзацами вероятность смены темы выше
- между секциями она ещё выше

Эти предположения легко формализовать регуляризаторами

EM-алгоритм с регуляризацией E-шага

Теорема

Если регуляризатор зависит от Φ, Θ через $p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$,

$$R(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} r_{dw}(p_{1dw}, \ldots, p_{Tdw}),$$

то решение задачи максимизации регуляризованного правдоподобия удовлетворяет системе уравнений

$$\begin{cases} \text{E-шаг:} \quad \tilde{p}_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}} \left(1 + \dfrac{\partial r_{dw}}{\partial p_{tdw}} - \sum_{s \in T} p_{sdw} \dfrac{\partial r_{dw}}{\partial p_{sdw}}\right); \\ \text{M-шаг:} \quad \phi_{wt} \propto \left(\sum_{d \in D} n_{dw} \tilde{p}_{tdw}\right)_+; \quad \theta_{td} \propto \left(\sum_{w \in d} n_{dw} \tilde{p}_{tdw}\right)_+; \end{cases}$$

Регуляризатор разреживания распределений $p(t|d,w)$

Гипотеза разреженности распределений $p(t|d,w)$:
в документе слово может относиться только к одной теме.

Максимизируем KL-дивергенции между $\hat{p}(t) = \frac{1}{|T|}$ и $p(t|d,w)$:

$$R(\Phi,\Theta) = -\tau \sum_{d\in D}\sum_{w\in d} n_{dw} \frac{1}{|T|} \sum_{t\in T} \ln p_{tdw}.$$

Подставляем, получаем формулу модифицированного E-шага:

$$\tilde{p}_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum\limits_{s\in T}\phi_{ws}\theta_{sd}}(1+\tau) - \frac{\tau}{|T|}.$$

Эффект одновременного разреживания:
если $p(t|w) < \frac{1}{|T|}$, то ϕ_{wt} уменьшается;
если $p(t|d) < \frac{1}{|T|}$, то θ_{td} уменьшается.

Лингвистический анализ электрокардиосигналов

Дано:
20 тысяч кодограмм ЭКГ (строки в 6-буквенном алфавите), каждая отнесена к некоторым из 40 заболеваний,
важно учесть случаи сочетания заболеваний.

Найти:

- темы классов (диагностические эталоны заболеваний)
- алгоритм классификации (диагностики заболеваний)

Регуляризаторы:

- разреживание, сглаживание, антикоррелирование
- привязка документов к классам (категоризация)
- учёт различий в степени доверия диагнозам
- учёт несбалансированности классов

Регуляризатор для классификации документов

Пусть C — множество классов (для ЭКГ — заболевания, для текстов — категории, авторы, ссылки, годы, читатели)

Гипотеза:
классификация документа d объясняется его темами:

$$p(c|d) = \sum_{t \in T} p(c|t)p(t|d) = \sum_{t \in T} \psi_{ct}\theta_{td}.$$

Минимизируем дивергенцию между моделью $p(c|d)$ и «эмпирической частотой» классов в документах m_{dc}:

$$R(\Psi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C} m_{dc} \ln \sum_{t \in T} \psi_{ct}\theta_{td} \to \max.$$

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 1–2.

Регуляризатор для классификация документов

EM-алгоритм дополняется оцениванием параметров ψ_{ct}.

E-шаг. По формуле Байеса:

$$p(t|d,w) = \frac{\phi_{wt}\theta_{td}}{\sum_{s\in T}\phi_{ws}\theta_{sd}} \qquad p(t|d,c) = \frac{\psi_{ct}\theta_{td}}{\sum_{s\in T}\psi_{cs}\theta_{sd}}$$

M-шаг. Максимизация регуляризованного правдоподобия:

$$\phi_{wt} \propto n_{wt} \qquad n_{wt} = \sum_{d\in D} n_{dw} p(t|d,w)$$

$$\psi_{ct} \propto m_{ct} \qquad m_{ct} = \sum_{d\in D} m_{dc} p(t|d,c)$$

$$\theta_{td} \propto n_{td} + \tau m_{td} \qquad n_{td} = \sum_{w\in W} n_{dw} p(t|d,w) \qquad m_{td} = \sum_{c\in C} m_{dc} p(t|d,c)$$

Регуляризатор для категоризации документов

Снова регуляризатор для классификации:

$$R(\Psi,\Theta) = \tau \sum_{d\in D}\sum_{c\in C} m_{dc} \ln \sum_{t\in T} \psi_{ct}\theta_{td} \to \max$$

Недостаток: за «эмпирическую частоту классов» не вполне обоснованно принимается равномерное распределение:

$$m_{dc} = n_d \tfrac{1}{|C_d|}[c \in C_d]$$

Ковариационный регуляризатор:

$$R(\Psi,\Theta) = \tau \sum_{d\in D}\sum_{c\in C} m_{dc} \sum_{t\in T} \psi_{ct}\theta_{td} \to \max$$

приводит к естественному аналитическому решению

$$\psi_{ct} = \big[c = c^*(t)\big], \quad c^*(t) = \arg\max_{c\in C} \sum_{d\in D} m_{dc}\theta_{td}$$

Эффект: Каждая категория c распадается на свои темы.

Обобщение: мультимодальные модели

Произвольное число модальностей X^j, $j = 1, \ldots, m$.
Вероятностное пространство $D \times T \times X$, $X = X^1 \times \cdots \times X^m$.
Каждый документ d состоит из токенов $x_1, \ldots, x_{n_d} \in X$.
Тематическая модель j-й модальности:

$$p(x|d) = \sum_{t \in T} p(x|t)\, p(t|d) = \sum_{t \in T} \phi^j_{xt} \theta_{td}, \qquad x \in X^j, \quad d \in D.$$

параметры модели $\Phi^j = \left(\phi^j_{xt}\right)_{|X^j| \times |T|}$, $\Theta = \left(\theta_{td}\right)_{|T| \times |D|}$.

Задача максимизации правдоподобия:

$$Q(\Phi, \Theta) = \sum_{j=1}^{m} \tau^j \sum_{d \in D} \sum_{x \in X_j} n^j_{dx} \ln \sum_{t \in T} \phi^j_{xt} \theta_{td} \to \max.$$

где n^j_{dx} — эмпирическая частота элемента $x \in X^j$.

Модифицированный EM-алгоритм

Теорема

Если функция $R(\Phi, \Theta)$ стохастических матриц Φ^j, $j = 1, \ldots, m$, Θ непрерывно дифференцируема и (Φ, Θ) является точкой локального максимума функции $Q(\Phi, \Theta) + R(\Phi, \Theta)$, то для любой темы t и документа d, удовлетворяющих условиям регулярности $\phi_t^j \not\equiv 0$, $\theta_d \not\equiv 0$, выполняется система уравнений:

$$p_{tdx}^j = \frac{\phi_{xt}^j \theta_{td}}{\sum_{s \in T} \phi_{xs}^j \theta_{sd}}; \quad n_{xt}^j = \sum_{d \in D} n_{dx}^j p_{tdx}^j; \quad n_{td}^j = \sum_{x \in X^j} n_{dx}^j p_{tdx}^j;$$

$$\phi_{xt}^j \propto \left(n_{xt}^j + \phi_{xt}^j \frac{\partial R}{\partial \phi_{xt}^j} \right)_+;$$

$$\theta_{td} \propto \left(\sum_{j=1}^m \tau^j n_{td}^j + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+.$$

Задача анализа потока пресс-релизов

Дано: коллекция пресс-релизов МИДов ряда стран.
Более 40 тыс. сообщений, 180Мб текста.

Найти:

- какие темы общие, какие специфичны для источников?
- какие темы «вечные», а какие привязаны к событиям?
- какие темы, и когда коррелируют с заданной темой?

Регуляризаторы:

- разреживание, сглаживание, антикоррелирование
- привязка документов к источникам и моментам времени
- сглаживание тематик во времени
- частичное обучение: привязка ключевых терминов к темам

Регуляризаторы для динамической тематической модели

Y — моменты времени (например, годы публикаций),
$y(d)$ — метка времени документа d,
$D_y \subset D$ — все документы, относящиеся к моменту $y \in Y$.

Гипотеза 1: распределение $p(t|y) = \sum_{d \in D_y} \theta_{td} p(d)$ разрежено:

$$R_1(\Theta) = -\tau_1 \sum_{y \in Y} \sum_{t \in T} \ln p(t|y) \to \max.$$

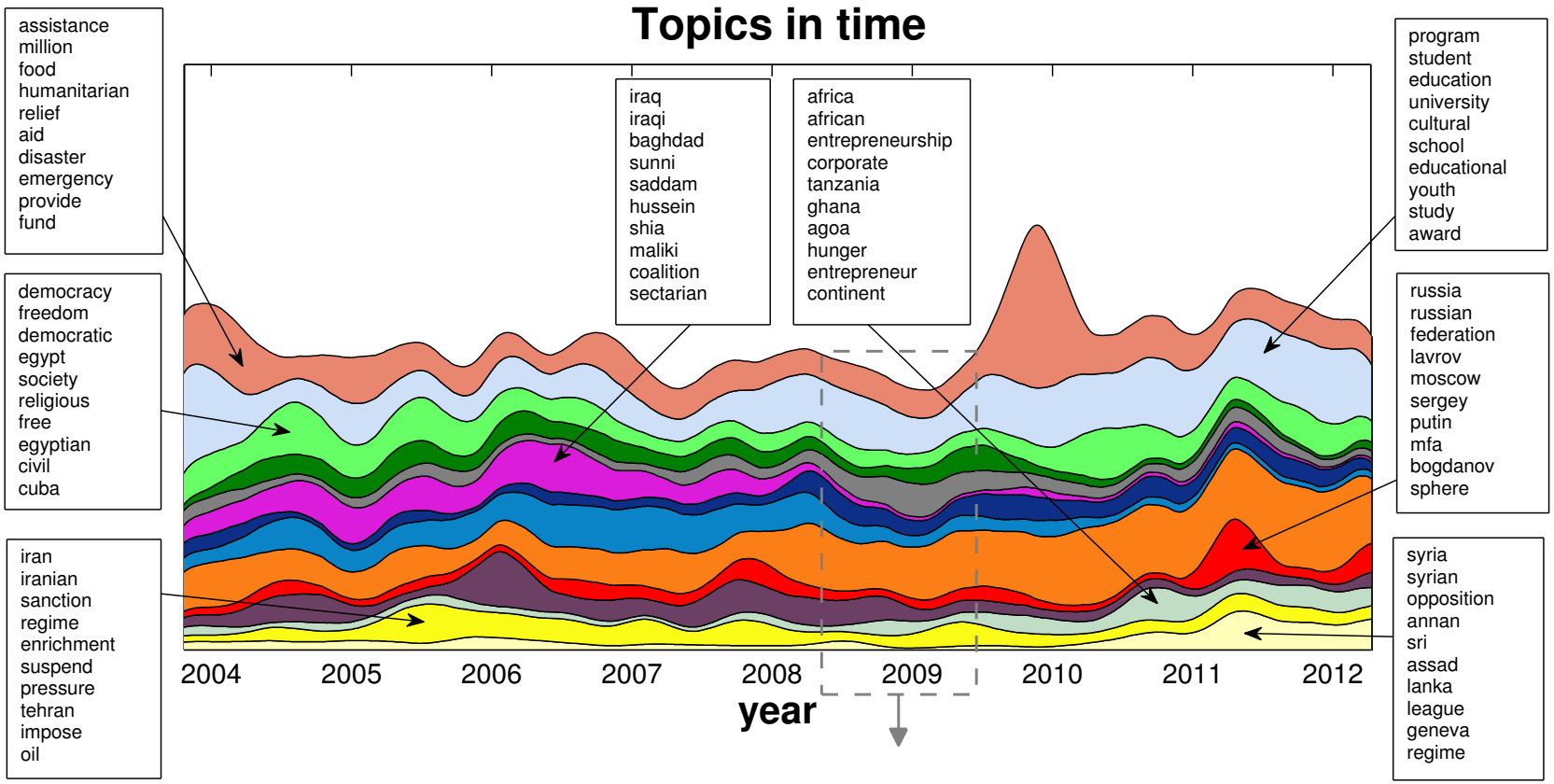
Эффект — разреживание тем t с малым $p(t|y(d))$:

$$\theta_{td} \propto \left(n_{td} - \tau_1 \frac{\theta_{td} p(d)}{p(t|y(d))} \right)_+.$$

Гипотеза 2: $p(t|y)$ меняются плавно, с редкими скачками:

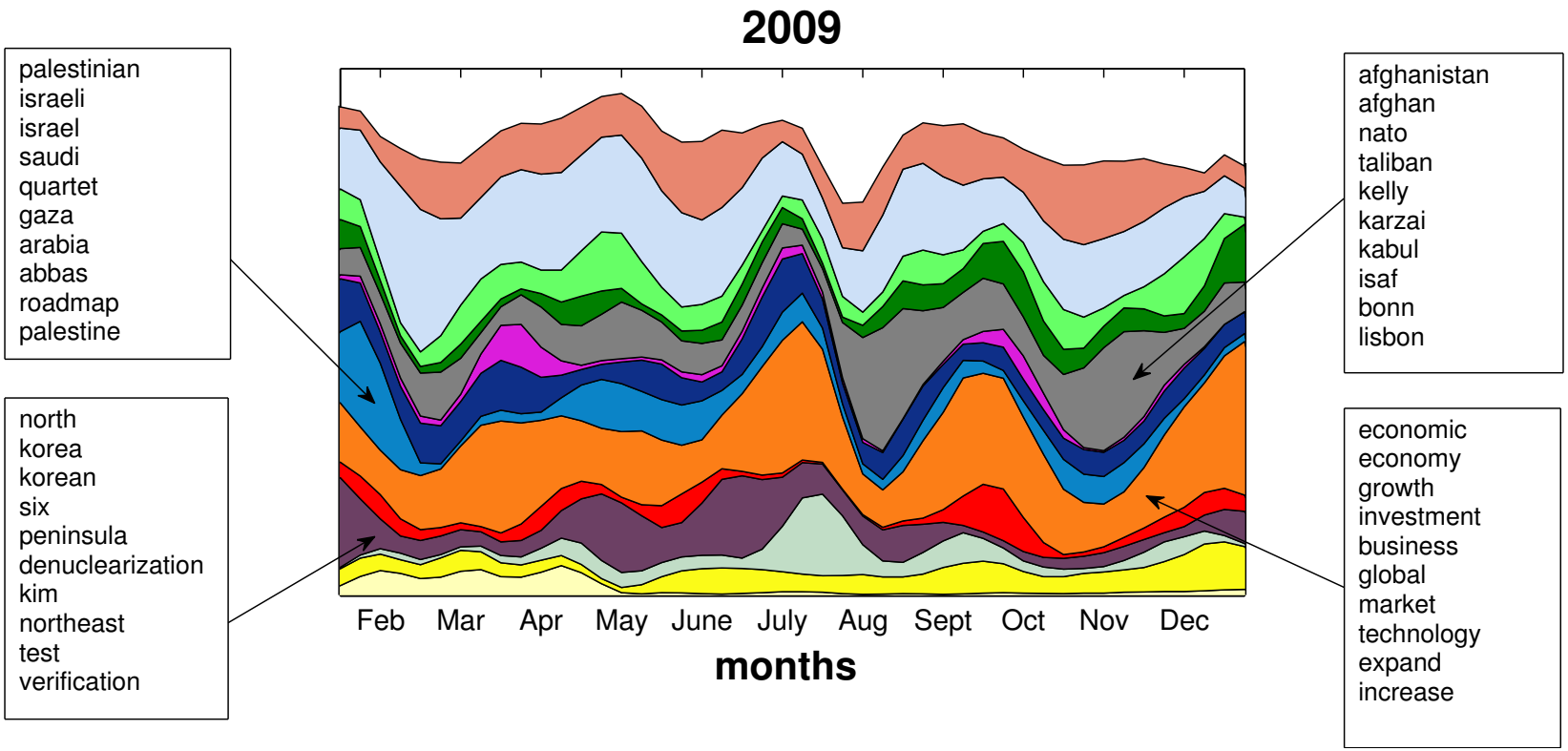
$$R_2(\Theta) = -\tau_2 \sum_{y \in Y} \sum_{t \in T} \left| p(t|y) - p(t|y-1) \right| \to \max.$$

Эксперименты с динамической тематической моделью



Никита Дойков. Курсовая работа, ВМК МГУ, 2014

Эксперименты с динамической тематической моделью



Никита Дойков. Курсовая работа, ВМК МГУ, 2014

Полностью автоматическое выделение терминов

Трёхэтапный отсев словосочетаний для сокращения словаря:

1. морфологический и синтаксический анализ
   (отсев грамматически некорректных словосочетаний)
2. статистический анализ
   (отсев не-коллокаций, CValue + OkapiBM25)
3. тематическое моделирование
   (отсев нетематических словосочетаний)

Воронцов Константин Вячеславович
voron@yandex-team.ru

Страницы на www.MachineLearning.ru:

- Участник:Vokov
- Вероятностные тематические модели (курс лекций, К. В. Воронцов)
- Тематическое моделирование

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов // Доклады РАН. 2014. — Т. 455., № 3. 268–271.

Vorontsov K. V., Potapenko A. A., Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // Analysis of Images, Social Networks and Texts. Ekaterinburg, 10–12 April 2014.

Vorontsov K. V., Potapenko A. A., Additive Regularization of Topic Models // Machine Learning Journal (to appear).

Литература

- *Hofmann T.* Probabilistic Latent Semantic Indexing. SIGIR, 1999.
- *Blei D., Ng A., Jordan M.* Latent Dirichlet Allocation // Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.
- *Teh Y. W., Newman D., Welling M.* A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. NIPS, 2006, Pp. 1353–1360.
- *Porteous I., Newman D., Ihler A., Asuncion A., Smyth P., Welling M.* Fast Collapsed Gibbs Sampling For Latent Dirichlet Allocation. KDD 2008.
- *Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models. Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.
- *Newman D., Lau J. H., Grieser K., Baldwin T.* Automatic Evaluation of Topic Coherence // Human Language Technologies, HLT-2010, Pp. 100–108.
- *Yi Wang.* Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details. 2011.
- *Sato I., Nakagawa H.* Rethinking Collapsed Variational Bayes Inference for LDA. Int'l Conf. on Machine Learning ICML, 2012.
- *Vorontsov K. V.* Additive Regularization for Topic Models of Text Collections // Doklady Mathematics. Pleiades Publisher, 2014. Vol. 88, No. 3.
- *Vorontsov K. V., Potapenko A. A.*, Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // AIST'14. Springer. 2014. (to appear)