Additive Regularization of Matrix Factorization for Probabilistic Topic Modeling

(Аддитивная регуляризация тематических моделей)

Konstantin Vorontsov
Yandex • CC RAS • MIPT • HSE • MSU

Analysis of Images, Social Networks and Texts
Ekaterinburg, 10–12 April 2014

Содержание

Задача вероятностного тематического моделирования (BTM)

Тема — это набор терминов, неслучайно часто совместно встречающихся в относительно узком подмножестве документов.

Дано:

W — словарь, множество слов (терминов)

D — множество (коллекция, корпус) текстовых документов

n_{dw} — сколько раз термин $w \in W$ встретился в документе $d \in D$

Найти:

$p(w|t)$ — какими терминами w определяется каждая тема t

$p(t|d)$ — к каким темам t относится каждый документ d

Критерии:

внутренний — точность описания коллекции моделью $p(w|d)$

внешний — качество решения конечной задачи

Цели вероятностного тематического моделирования (BTM)

- Тематический поиск документов и объектов по тексту любой длины или по любому объекту
- Категоризация, классификация, аннотирование, суммаризация текстовых документов

Приложения:

- Поиск научной информации
- Выявление трендов и фронта исследований
- Поиск специалистов (expert search), рецензентов, проектов
- Анализ и агрегирование новостных потоков
- Рубрикация документов, изображений, видео, музыки
- Рекомендующие системы, коллаборативная фильтрация
- Аннотация генома и другие задачи биоинформатики
- Анализ дискретизированных биомедицинских сигналов

Развитие BTM: учёт дополнительной информации

- повышение адекватности тематической модели
- выявление семантики нетекстовых объектов

Виды дополнительной информации:

- последовательность слов документа d: $\{w_1, \ldots, w_{n_d}\}$
- разбиение документа на предложения, разделы
- метаданные: год, авторы, источник, и т.д.
- цитаты и/или гиперссылки: исходящие, входящие
- рубрикатор(ы)
- словари, тезаурусы, онтологии предметных областей
- изображения внутри документов
- именованные сущности в тексте документов
- пользователи документов
- теги, ключевые слова, привязанные к документам

Развитие BTM: расширение функциональных возможностей

- как строить комбинированные и многоцелевые модели?
- как учитывать все дополнительные данные сразу?
- как моделировать изменения тем во времени?
- как обеспечивать интерпретируемость тем?
- как определять правильное число тем?
- как восстанавливать иерархию тем?
- как автоматически именовать темы?
- как учитывать лингвистические знания?
- как строить модели на десятки тысяч тем?
- как строить модели сверхбольших коллекций?
- как делать визуализацию и навигацию по темам?

Цели сегодняшней лекции

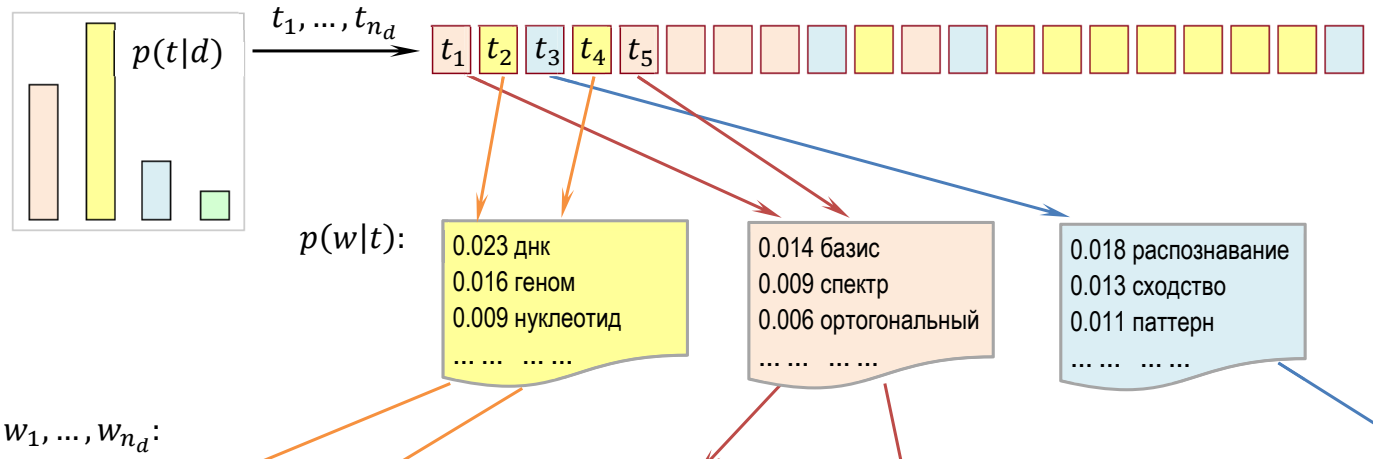
«Представляется важной задача освобождения всюду, где это возможно, от излишних вероятностных допущений»

— А. Н. Колмогоров, создатель современной теории вероятностей (Теория информации и теория алгоритмов, 1987.)

1. Рассказать о разновидностях тематических моделей.
2. Показать простой и мощный математический аппарат для построения сложных тематических моделей.
3. При этом обойтись без распределений Дирихле, байесовского вывода, графических моделей и интегралов.
4. Снизить для исследователей «порог вхождения» в область вероятностного тематического моделирования.

Вероятностная тематическая модель

Порождение слов документа d из $p(w|d) = \sum_{t \in T} p(w|t)p(t|d)$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дупликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Задача тематического моделирования коллекции документов

Базовые предположения:

- коллекция D — выборка троек $(d_i, w_i, t_i)_{i=1}^n \sim p(d, w, t)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- гипотеза условной независимости: $p(w|d, t) = p(w|t)$

Вероятностная модель порождения документов:

$$p(w|d) = \sum_{t \in T} \underbrace{p(w|t)}_{\phi_{wt}} \underbrace{p(t|d)}_{\theta_{td}}$$

- $\phi_{wt} \equiv p(w|t)$ — распределение терминов в темах $t \in T$;
- $\theta_{td} \equiv p(t|d)$ — распределение тем в документах $d \in D$.

Прямая задача: по ϕ_{wt}, θ_{td} сгенерировать документ d.

Обратная задача: по $\hat{p}(w|d) \equiv \frac{n_{dw}}{n_d}$ найти параметры ϕ_{wt}, θ_{td}.

Принцип максимума правдоподобия

Задача: максимизировать логарифм правдоподобия

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

при ограничениях неотрицательности и нормировки

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

Это задача стохастического матричного разложения:

$$\underset{W\times D}{F}\approx\underset{W\times T}{\Phi}\cdot\underset{T\times D}{\Theta}$$

$F=\left\|\hat{p}(w|d)\right\|_{W\times D}$ — известная матрица исходных данных,
$\Phi=\left\|\phi_{wt}\right\|_{W\times T}$ — искомая матрица терминов тем $\phi_{wt}=p(w|t)$,
$\Theta=\left\|\theta_{td}\right\|_{T\times D}$ — искомая матрица тем документов $\theta_{td}=p(t|d)$.

Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Теорема

Если Φ, Θ — решение задачи максимизации правдоподобия, то оно удовлетворяет системе уравнений

$$\begin{cases} \text{E-шаг:} \quad n_{dwt} = n_{dw} \dfrac{\phi_{wt}\theta_{td}}{\sum\limits_{s \in T} \phi_{ws}\theta_{sd}}; & \text{— вспомогательные переменные} \\ \text{M-шаг:} \quad \phi_{wt} = \dfrac{n_{wt}}{n_t}; \quad n_{wt} = \sum\limits_{d \in D} n_{dwt}; \quad n_t = \sum\limits_{w \in W} n_{wt}; \\ \qquad\quad \theta_{td} = \dfrac{n_{td}}{n_d}; \quad n_{td} = \sum\limits_{w \in d} n_{dwt}; \quad n_d = \sum\limits_{t \in T} n_{td}; \end{cases}$$

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

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

Вероятностная интерпретация шагов EM-алгоритма

E-шаг — это формула Байеса:

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

$n_{dwt} = n_{dw}p(t|d,w)$ — 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 $\propto$:

$$p(t|d,w) \propto \phi_{wt}\theta_{td}; \qquad \phi_{wt} \propto n_{wt}; \qquad \theta_{td} \propto n_{td};$$

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 $w \in d$
    $p(t|d,w) = \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$ для всех $t \in T$;
    $n_{wt}, n_{td}, n_t, n_d \mathrel{+}= n_{dw} p(t|d,w)$ 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
      $p(t|d,w) = \dfrac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$ для всех $w \in d,\ t \in T$;
      $\theta_{td} := \frac{1}{n_d} \sum\limits_{w \in d} n_{dw} p(t|d,w)$ для всех $t \in T$;
    пока θ_d не сойдётся;
    $\tilde{n}_{wt}, \tilde{n}_t \mathrel{+}= n_{dw} p(t|d,w)$ 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 $w \in W,\ t \in T$;

Латентное размещение Дирихле [Blei, Ng, Jordan, 2003]

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

- в PLSA — несмещённые оценки максимума правдоподобия:

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

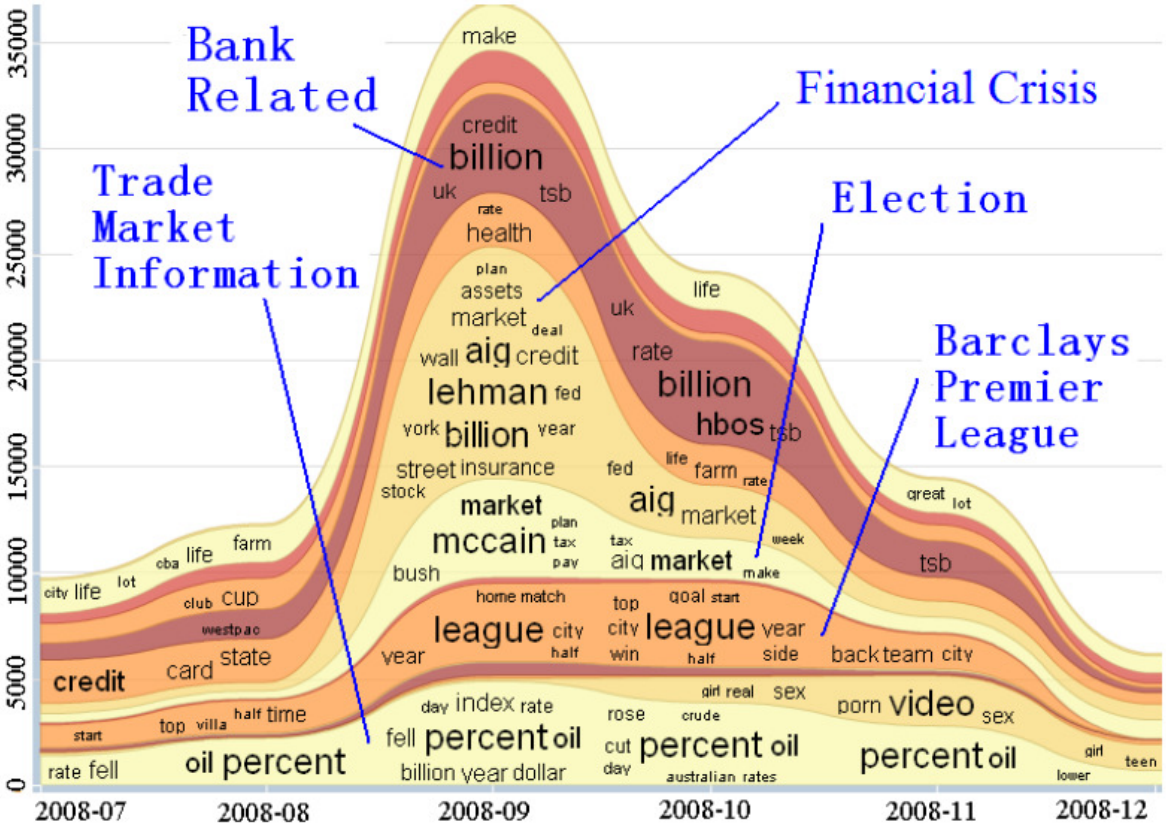
- в LDA — сглаженные байесовские оценки:

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

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.

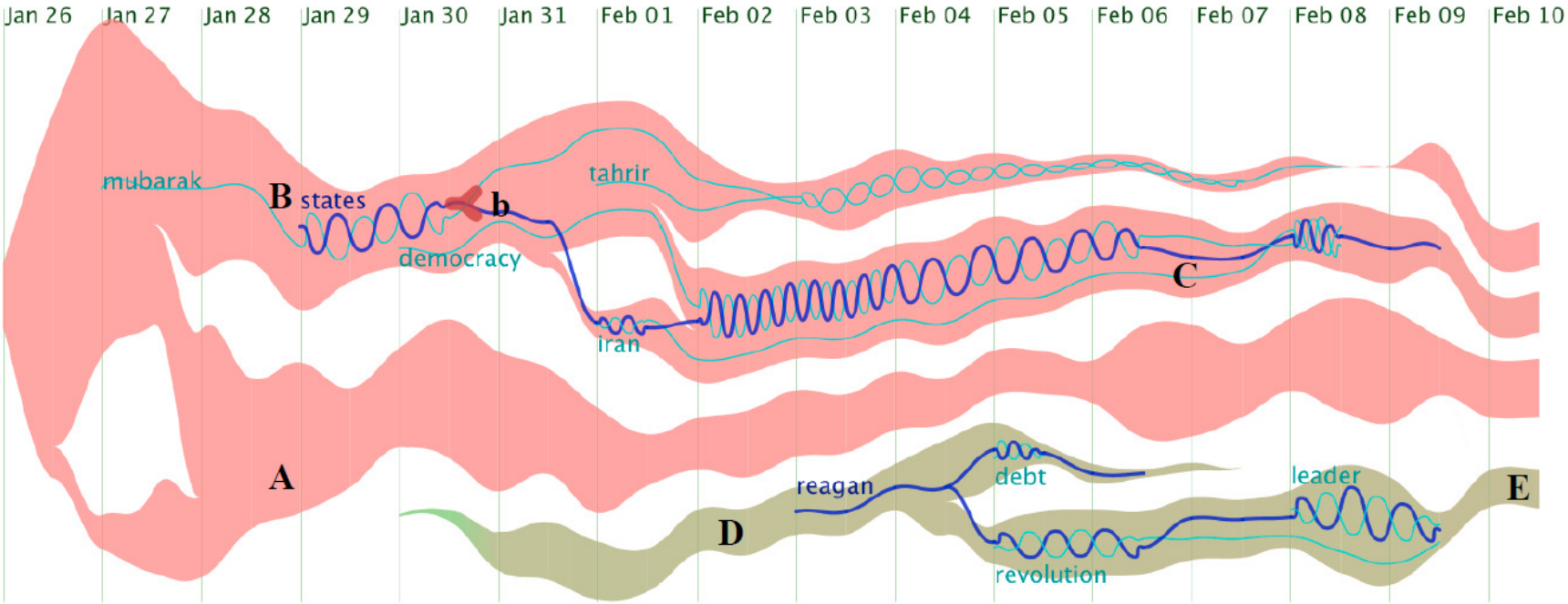
Воронцов К.В., Потапенко А.А. Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных, 2013. — Т. 1, № 6. — С. 657–686.

Динамические модели, учитывающие время



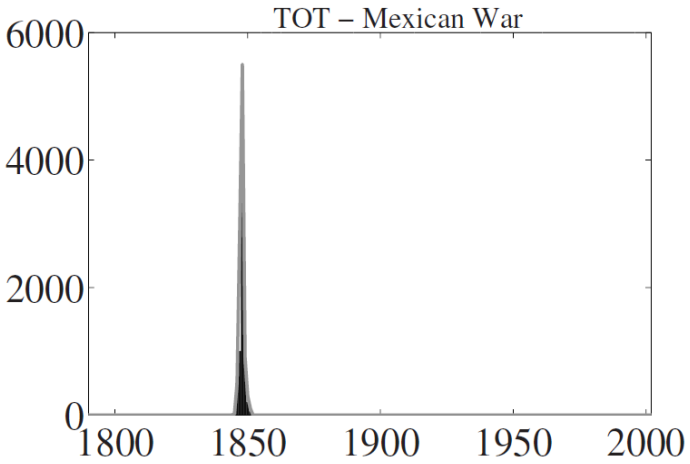
Jianwen Zhang, Yangqiu Song, Changshui Zhang, Shixia Liu Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-varying Corpora // KDD'10, July 25–28, 2010.

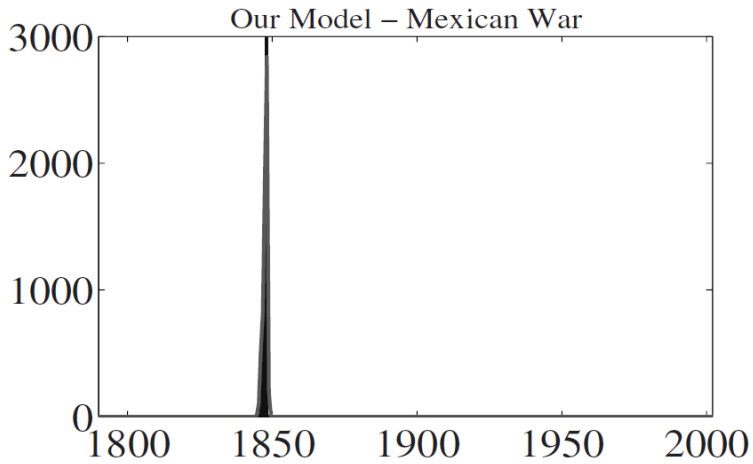
Динамические модели эволюции тем



Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai J. Gao, Xin Tong, Huamin Qu TextFlow: Towards Better Understanding of Evolving Topics in Text // IEEE Transactions On Visualization And Computer Graphics, Vol. 17, No. 12, December 2011.

Совмещение динамической и n-граммной модели



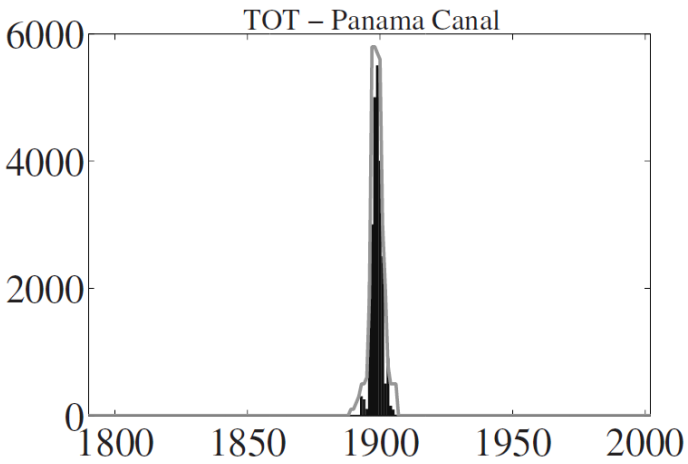


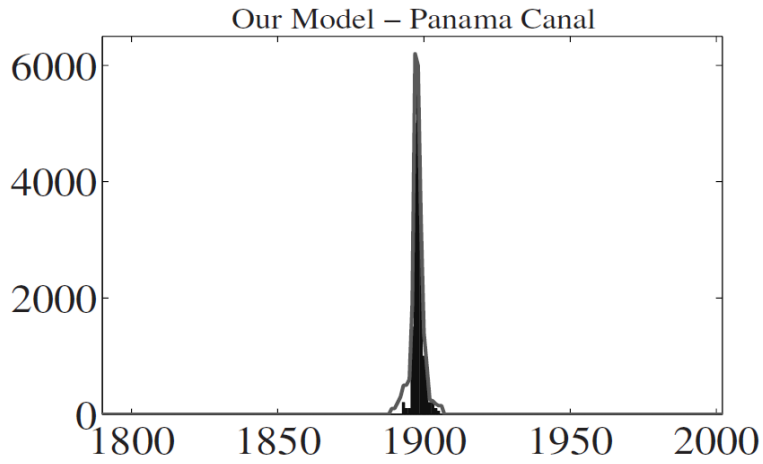
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents // 35'th ECIR 2013, Moscow, March 24–27. — pp. 292–304.

Совмещение динамической и n-граммной модели





1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

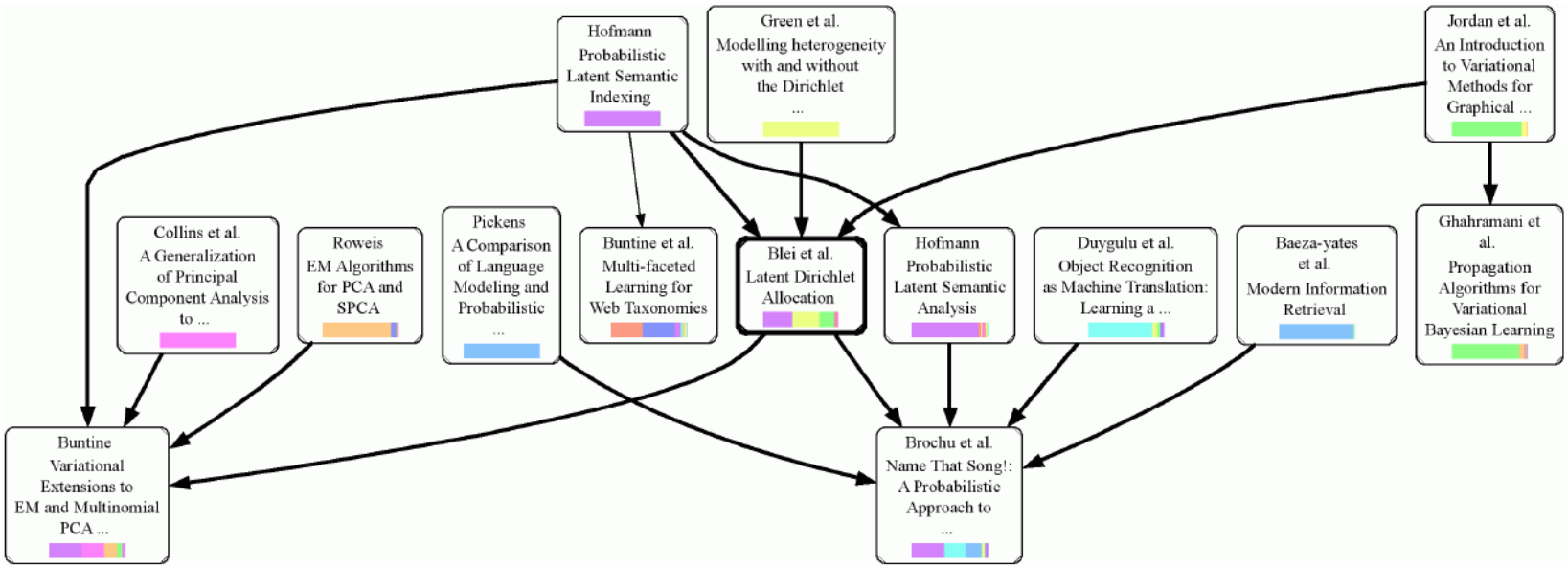
1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents // 35'th ECIR 2013, Moscow, March 24–27. — pp. 292–304.

Модели, учитывающие цитирования или гиперссылки

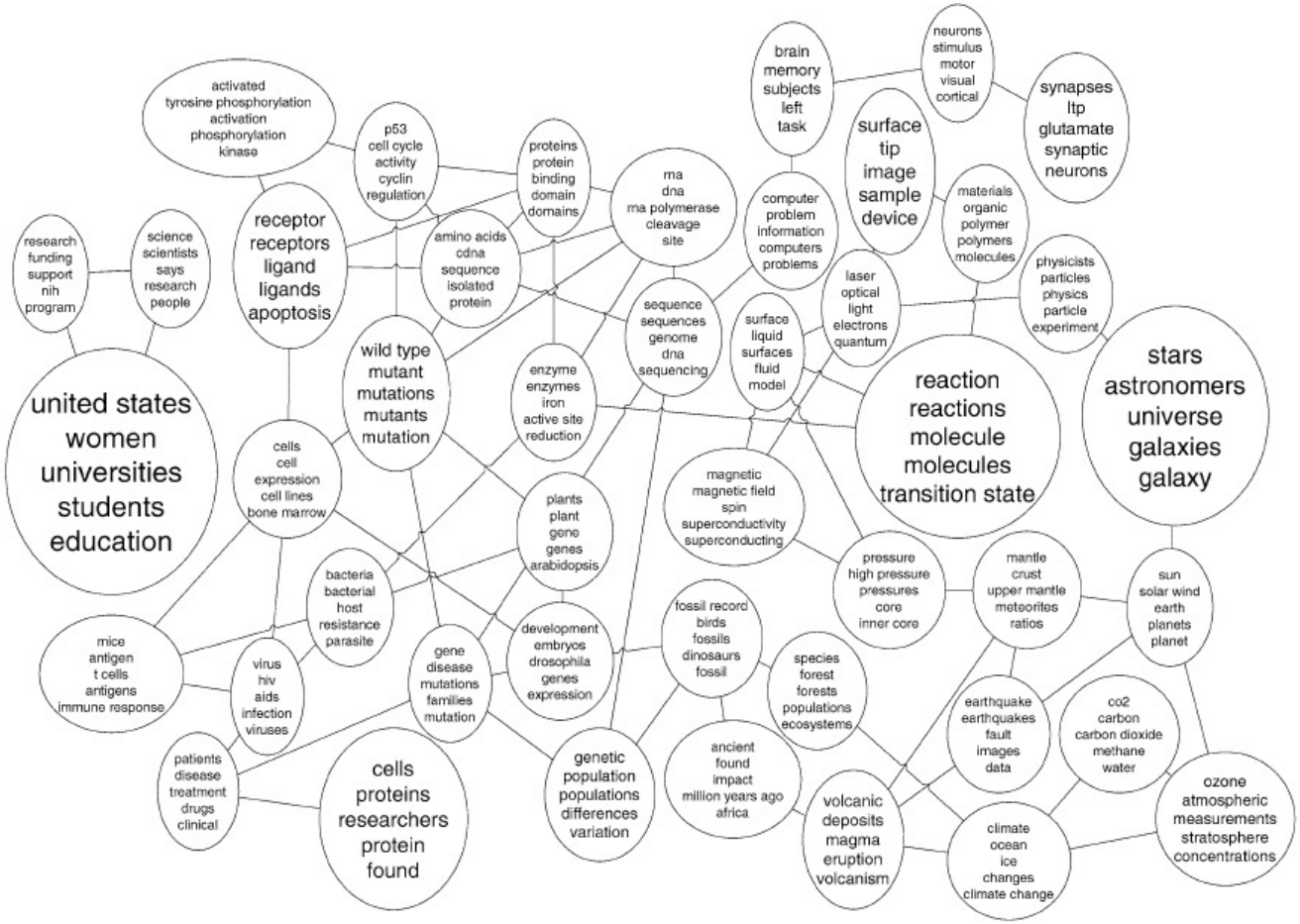
Учёт ссылок уточняет тематическую модель

Тематическая модель выявляет самые влиятельные ссылки



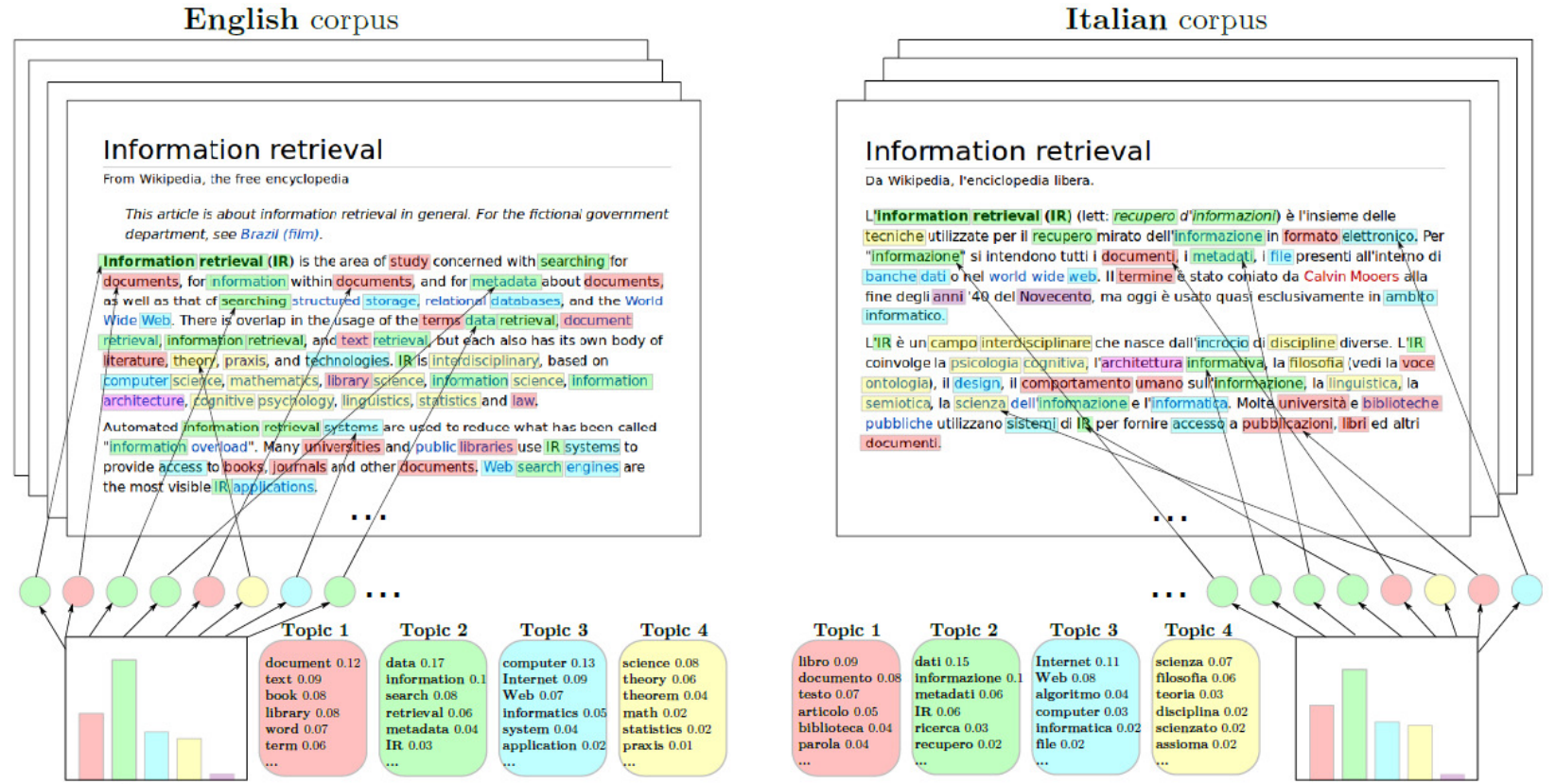
Laura Dietz, Steffen Bickel, Tobias Scheffer. Unsupervised prediction of citation influences // ICML-2007, Pp. 233–240.

Выявление взаимосвязей между темами



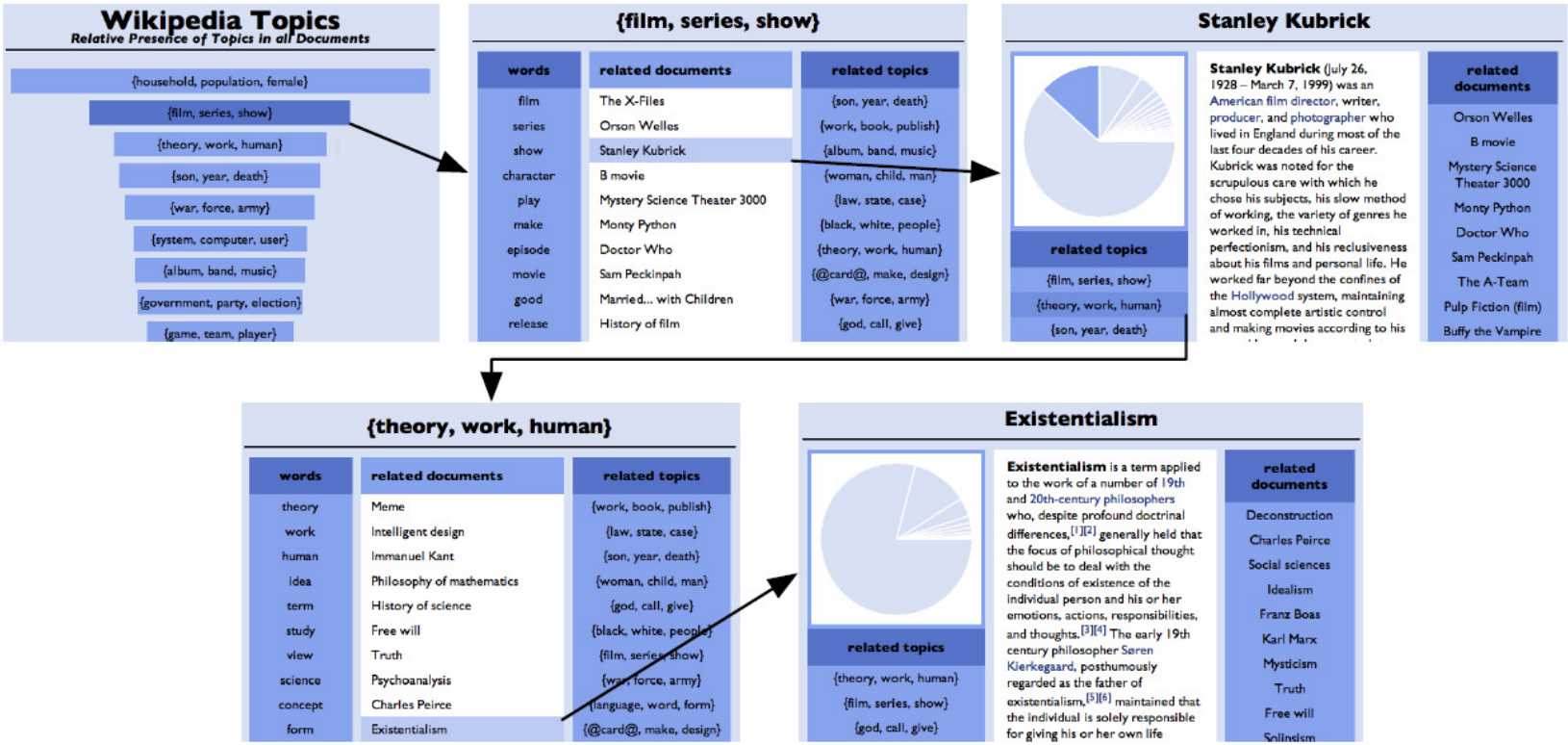
D. Blei, J. Lafferty. A correlated topic model of Science // Annals of Applied Statistics, 2007. Vol. 1, Pp. 17-35.

Многоязычные модели



I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: a short overview of its methodology with applications // NIPS, 7–8 December 2012. — Pp. 1–11.

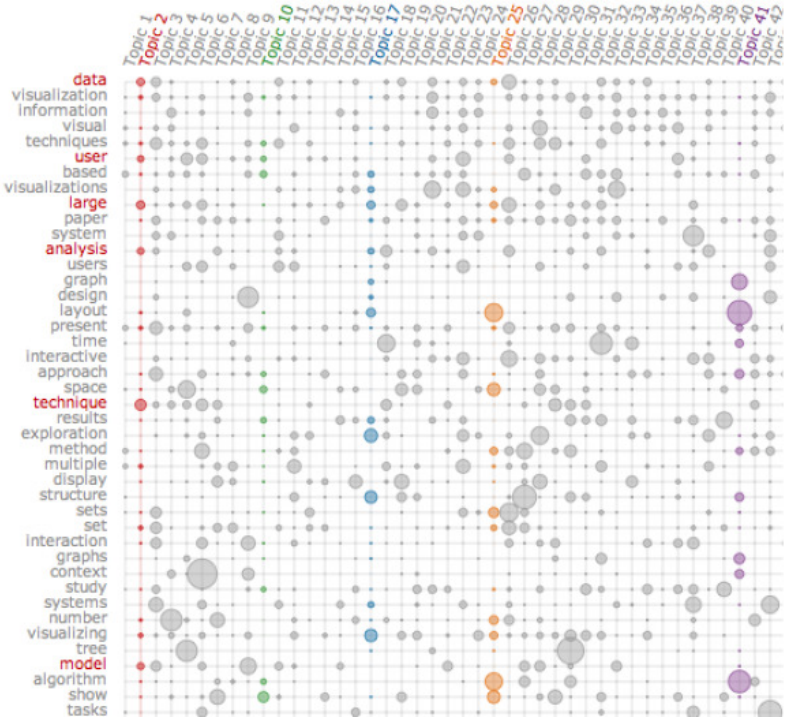
Визуализация тематической модели

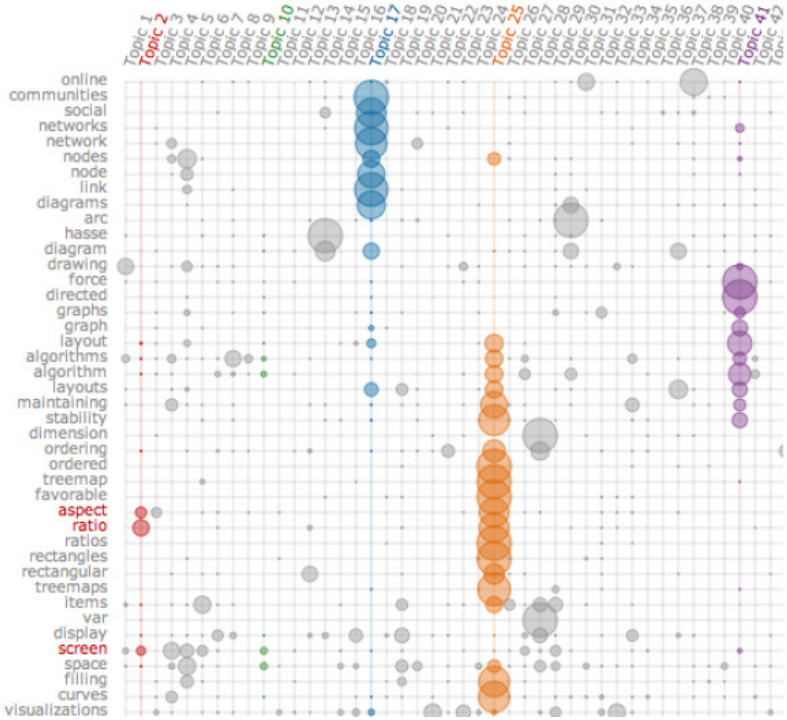


A. Chaney, D. Blei. Visualizing topic models // International AAAI Conference on Social Media and Weblogs, 2012.

Визуализация тематической модели

Группирование ядер из слов в каждой теме

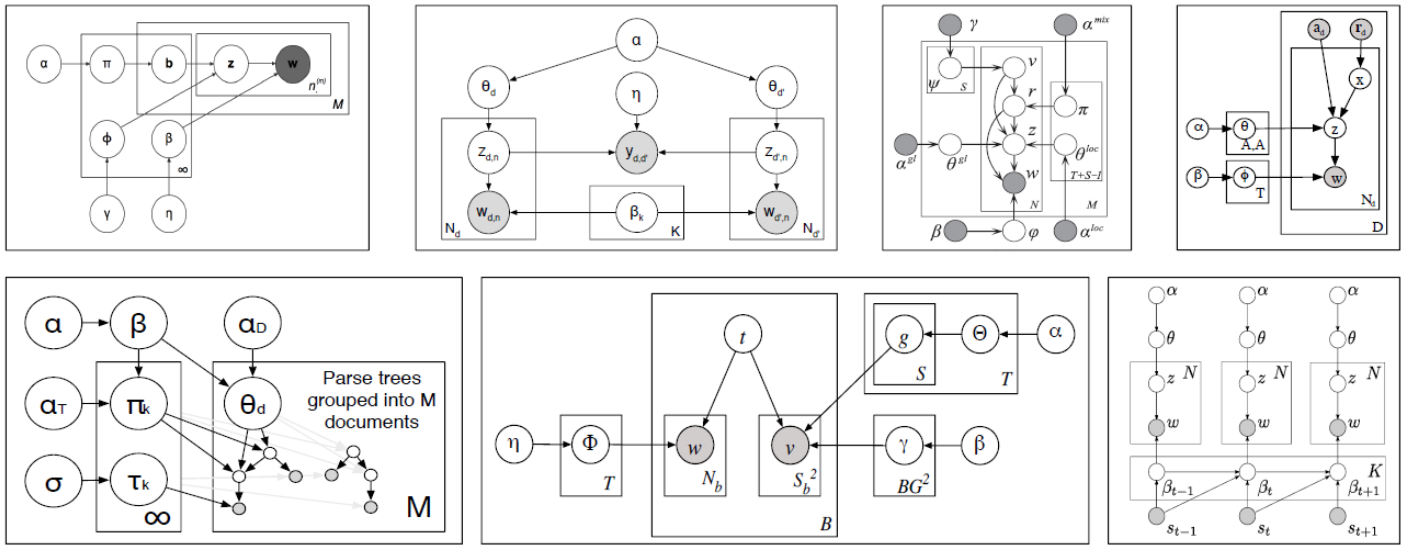




Jason Chuang, Christopher D. Manning, Jeffrey Heer.
Termite: Visualization Techniques for Assessing Textual Topic Models // Advanced Visual Interfaces, 2012

Резюме по краткому обзору тематическим моделей

- Их много, они разные, их трудно комбинировать
- Математический аппарат местами выносит мозг...



Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad.
Knowledge discovery through directed probabilistic topic models: a survey.
Frontiers of Computer Science in China, Vol. 4, No. 2., 2010, Pp. 280–301.
(русский перевод на www.MachineLearning.ru)

Topic Modeling Bibliography: http://mimno.infosci.cornell.edu/topics.html

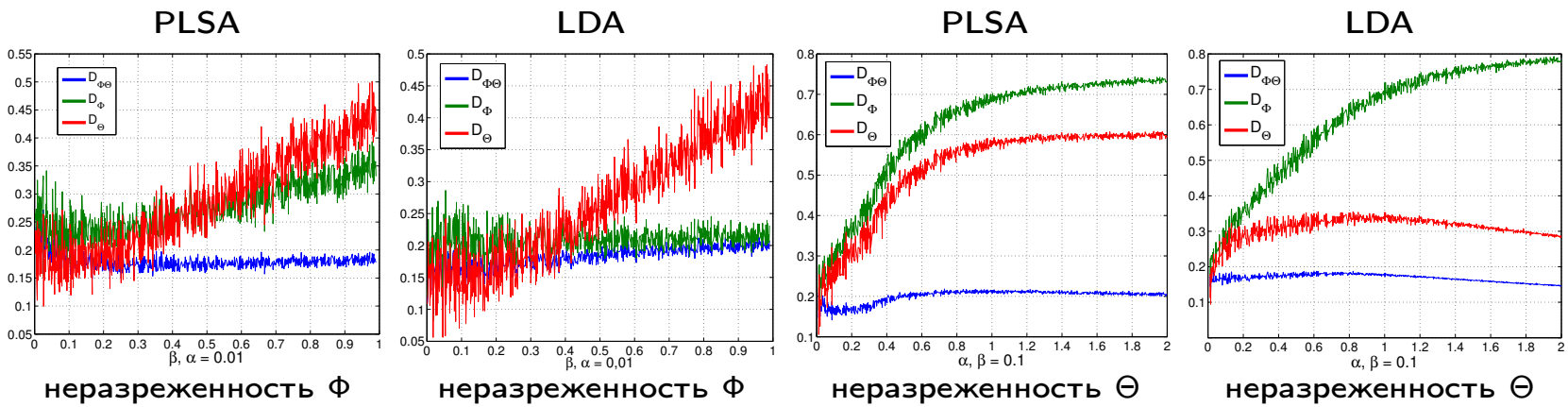
Задача построения BTM — некорректно поставленная

Неединственность 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: вводить дополнительные ограничения можно и нужно!

Аддитивная регуляризация тематической модели

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

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*.

Обоснование регуляризованного EM-алгоритма PLSA

Теорема

Если Φ, Θ — решение задачи максимизации регуляризованного правдоподобия, то оно удовлетворяет системе уравнений

$$\begin{cases} \text{E-шаг:} \quad n_{dwt} = n_{dw} \dfrac{\phi_{wt}\theta_{td}}{\sum\limits_{s \in T} \phi_{ws}\theta_{sd}}; \\ \text{M-шаг:} \\ \phi_{wt} = \dfrac{n_{wt}}{n_t}; \quad n_{wt} = \left(\sum\limits_{d \in D} n_{dwt} + \phi_{wt} \dfrac{\partial R}{\partial \phi_{wt}} \right)_+; \quad n_t = \sum\limits_{w \in W} n_{wt}; \\ \theta_{td} = \dfrac{n_{td}}{n_d}; \quad n_{td} = \left(\sum\limits_{w \in d} n_{dwt} + \theta_{td} \dfrac{\partial R}{\partial \theta_{td}} \right)_+; \quad n_d = \sum\limits_{t \in T} n_{td} \end{cases}$$

При $R(\Phi, \Theta) = 0$ это формулы EM-алгоритма для PLSA.

Математический ликбез. Дивергенция Кульбака–Лейблера

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

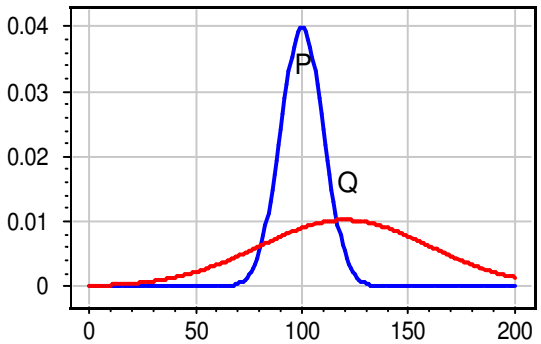
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

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

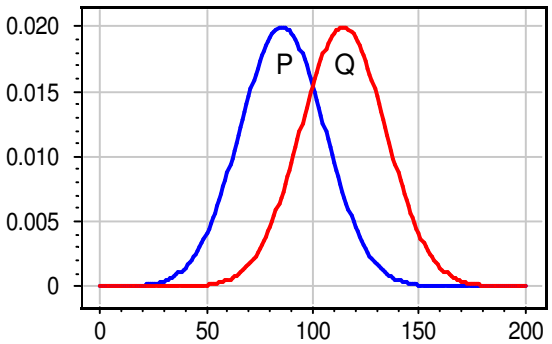
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

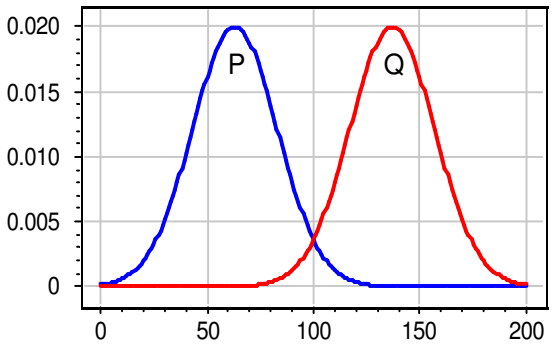
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



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



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



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

Регуляризатор №1: Сглаживание (совпадает с LDA)

Гипотеза сглаженности:

распределения ϕ_{wt} близки к заданным распределениям β_w

распределения θ_{td} близки к заданным распределениям α_t

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

Максимизируем сумму этих регуляризаторов:

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

Подставляем, получаем формулы M-шага LDA:

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

Этого вы не найдёте в *D.Blei, A.Ng, M.Jordan.* Latent Dirichlet allocation // Journal of Machine Learning Research, 2003. — Vol. 3. — Pp. 993–1022.

Регуляризатор №2: Частичное обучение (обобщение LDA)

Пусть имеется дополнительная информация от экспертов:
1) списки тем $T_d \subset T$ для некоторых документов $d \in D_0$,
2) списки терминов $W_t \subset W$ для некоторых тем $t \in T_0$.

ϕ^0_{wt} — распределение, равномерное на W_t
θ^0_{td} — распределение, равномерное на T_d

Максимизируем сумму этих регуляризаторов:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W_t} \phi^0_{wt} \ln \phi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T_d} \theta^0_{td} \ln \theta_{td} \to \max$$

Подставляем, получаем обобщение LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0 \phi^0_{wt} \qquad \theta_{td} \propto n_{td} + \alpha_0 \theta^0_{td}$$

Nigam K., McCallum A., Thrun S., Mitchell T. Text classification from labeled and unlabeled documents using EM // Machine Learning, 2000, no. 2–3.

Регуляризатор №2: Частичное обучение (второе обобщение LDA)

Идея: вместо логарифма можно взять любую другую монотонно возрастающую функцию μ

$$R(\Phi,\Theta) = \beta_0 \sum_{t\in T_0}\sum_{w\in W_t} \phi^0_{wt}\mu(\phi_{wt}) + \alpha_0 \sum_{d\in D_0}\sum_{t\in T_d} \theta^0_{td}\mu(\theta_{td}) \to \max.$$

Подставляем, получаем ещё одно обобщение LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0\phi^0_{wt}\phi_{wt}\mu'(\phi_{wt}) \qquad \theta_{td} \propto n_{td} + \alpha_0\theta^0_{td}\theta_{td}\mu'(\theta_{td}).$$

При $\mu(z) = z$ максимизируется сумма ковариаций $\text{cov}(\theta^0_d, \theta_d)$.

Преимущество ковариационного регуляризатора:
Если θ^0_{td} равномерно на T_d, то ковариация не накладывает ограничений на распределение θ_{td} между темами из T_d.

Регуляризатор №3: Разреживание (третье обобщение LDA)

Гипотеза разреженности: среди ϕ_{wt}, θ_{td} много нулей.

Чем сильнее разрежено распределение, тем ниже его энтропия.
Максимальной энтропией обладает равномерное распределение.

Максимизируем дивергенцию между распределениями β_w, α_t (равномерными?) и искомыми распределениями ϕ_{wt}, θ_{td}:

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

Подставляем, получаем «анти-LDA»:

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

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 №4: Удаление незначимых тем

Гипотеза: если тема собрала 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:

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

Подставляем, получаем:

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

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

Регуляризатор №5: Декорреляция

Гипотеза некоррелированности тем:
чем различнее темы, тем лучше они интерпретируются.

Минимизируем ковариации между вектор-столбцами ϕ_t:

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

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

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

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

Регуляризатор №6: Максимизация 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.

Регуляризатор №7: Связи между документами

Гипотеза: чем больше n_{dc} — число ссылок из d на c, тем более близки тематики документов d и c.

Минимизируем ковариации между вектор-столбцами связанных документов θ_d, θ_c:

$$R(\Phi, \Theta) = \tau \sum_{d,c \in D} n_{dc} \text{cov}(\theta_d, \theta_c) \to \max,$$

Подставляем, получаем ещё один вариант сглаживания:

$$\theta_{td} \propto n_{td} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc}.$$

Dietz L., Bickel S., Scheffer T. Unsupervised prediction of citation influences // ICML 2007. — Pp. 233–240.

Регуляризатор №8: Классификация документов

Пусть C — множество классов документов (категории, авторы, ссылки, годы, пользователи,...)

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 №8: 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:

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

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

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

Регуляризатор №9: Категоризация 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: для «эмпирической частоты классов» приходится необоснованно брать 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.

Регуляризатор №10: Динамическая тематическая модель

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:

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

Преимущества и ограничения

Преимущества APTM:

- легко учитывать различную дополнительную информацию
- больше свободы на этапе формализации требований
- легко комбинировать регуляризаторы в любых сочетаниях
- предельно упростился переход от модели к алгоритму:

Ограничения APTM:

- ещё не все модели успели переформулировать в APTM ;)
- надо подбирать много коэффициентов регуляризации

Подбор траектории регуляризации

Пусть задана линейная комбинация регуляризаторов:

$$R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)$$

Задача: выбрать вектор коэффициентов $\tau = (\tau_i)_{i=1}^{n}$

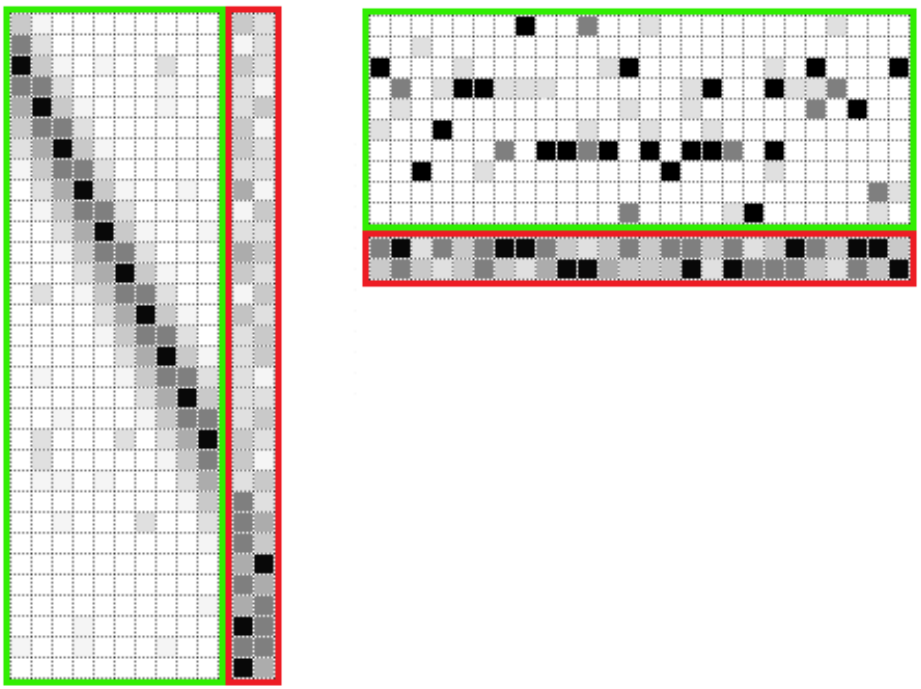
Ближайший аналог: «Regularization Path» в задачах регрессии с L_1- и L_2-регуляризацией (Elastic Net)

Общие соображения о выборе траектории регуляризации:
1) мониторить много критериев качества в ходе итераций,
2) усиливать регуляризаторы постепенно,
3) сначала достичь сходимости нерегуляризованного PLSA,
4) чередовать регуляризаторы, если они мешают друг другу
5) ослаблять регуляризаторы, когда их цель уже достигнута

Гипотеза о структуре интерпретируемых тем

1. Предметные темы разреженные, существенно различные, имеют **ядро**, состоящее из терминов предметной области.
2. Фоновые темы плотные, содержат слова общей лексики.

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



Комбинирование разреживания, сглаживания и декорреляции

Задача: улучшить интерпретируемость, не ухудшив перплексию

Набор регуляризаторов:

№1 сглаживание фоновых тем — столбцов Φ, строк Θ

№3 разреживание предметных тем — столбцов Φ, строк Θ

№4 удаление незначимых тем — строк Θ

№5 декоррелирование предметных тем — столбцов Φ

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: $|D'| = 174$.

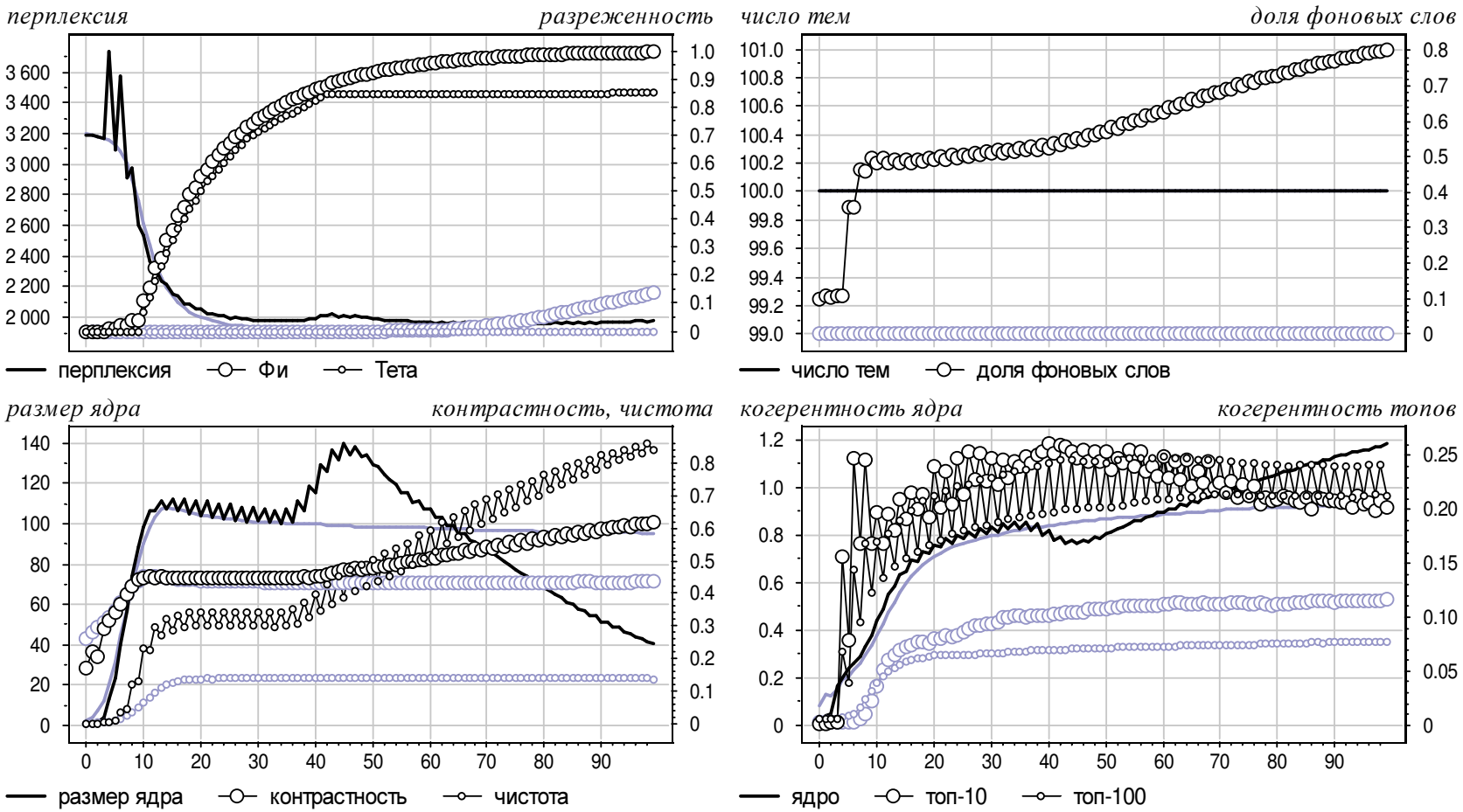
Критерии качества модели

- Перплексия контрольной коллекции: $\mathcal{P} = \exp(-\mathscr{L}/N)$
- Разреженность — доля нулевых элементов в Φ и Θ
- Характеристики интерпретируемости тем:
  - когерентность темы: [Newman, 2010]
  - размер ядра темы: $|W_t| = \#\{w\colon p(t|w) > 0.25\}$
  - чистота темы: $\sum_{w \in W_t} p(w|t)$
  - контрастность темы: $\frac{1}{|W_t|} \sum_{w \in W_t} p(t|w)$
- Вырожденность тематической модели:
  - число тем
  - доля фоновых слов в документах коллекции

Newman D., Lau J.H., Grieser K., Baldwin T. Automatic evaluation of topic coherence // Human Language Technologies, HLT-2010, Pp. 100–108.

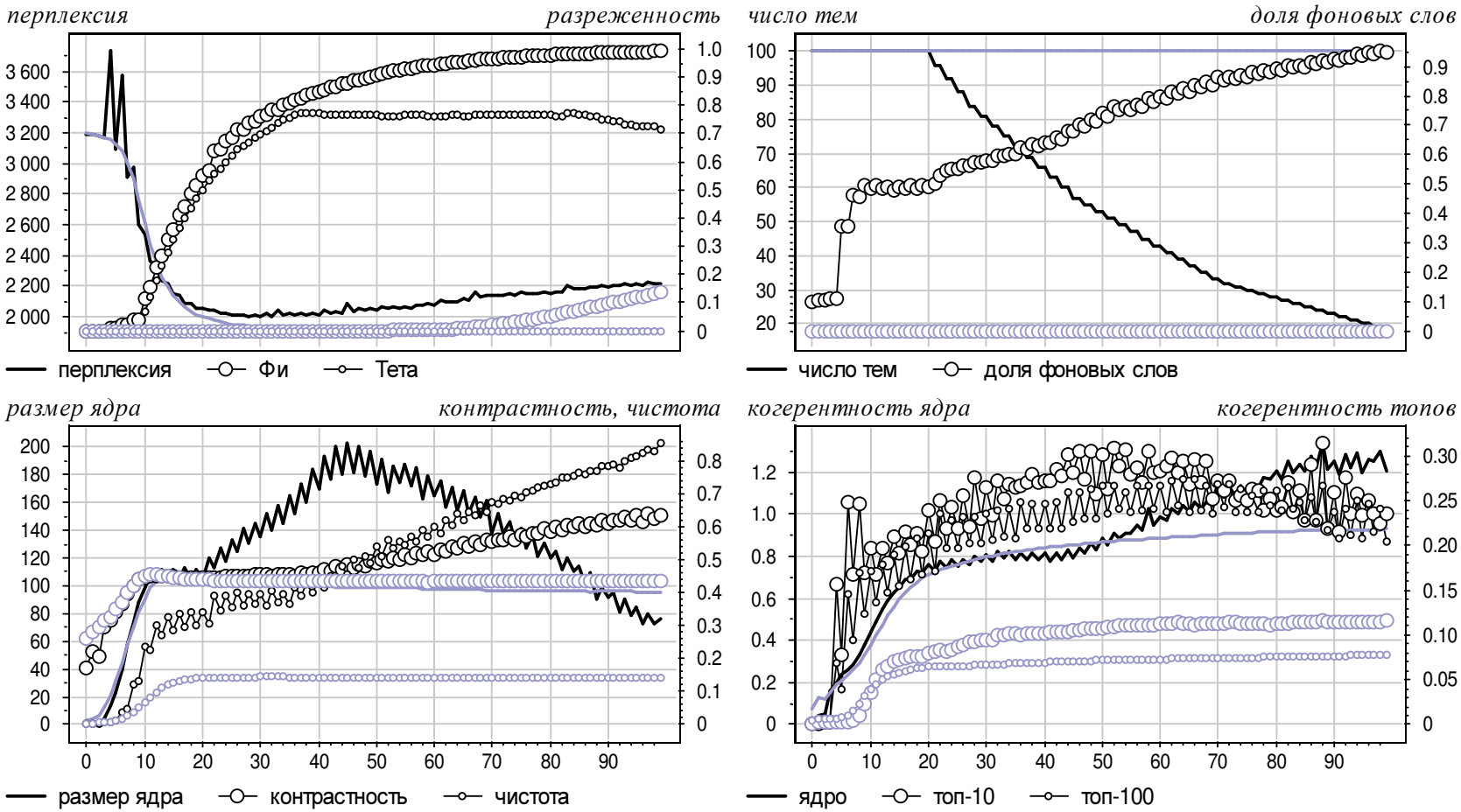
Комбинирование разреживания, сглаживания и декорреляции

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

Показана возможность одновременного:

- усиления разреженности (до 98%)
- улучшения интерпретируемости (когерентности) тем
- повышения различности (чистоты и контрастности) тем
- при размере ядер тем 50–150 слов
- почти без потери перплексии (правдоподобия) модели

Подобраны траектории регуляризации:

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

Направления ближайших исследований

- Лингвистическая регуляризация, отказ от «мешка слов»
  - учёт линейной структуры текста
  - учёт лингвистических ресурсов (тезаурусов, онтологий)
  - выделение терминов-словосочетаний
- Разработка BigARTM — библиотеки с открытым кодом
  - параллельные вычисления
  - распределённое хранение коллекции
  - любые сочетания регуляризаторов
- Открытые проблемы
  - глобальная оптимизация
  - иерархические модели
  - визуализация результатов поиска научной информации

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, №3, 2014.

Vorontsov K. V., Potapenko A. A., Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // AIST'14. Springer. 2014.