BigARTM: тематическое моделирование больших текстовых коллекций

Воронцов Константин Вячеславович
ФИЦ ИУ РАН • МФТИ • МГУ • ВШЭ • Яндекс

• First global Big Data Science Meetup •
Москва • 12 сентября 2015

Содержание

Что такое «тема»?

- *Тема* — специальная терминология предметной области.
- *Тема* — набор терминов (слов или словосочетаний), совместно часто встречающихся в документах.

Более формально,

- *тема* — условное распределение на множестве терминов, $p(w|t)$ — вероятность термина w в теме t;
- *тематический профиль* документа — условное распределение $p(t|d)$ — вероятность темы t в документе d.

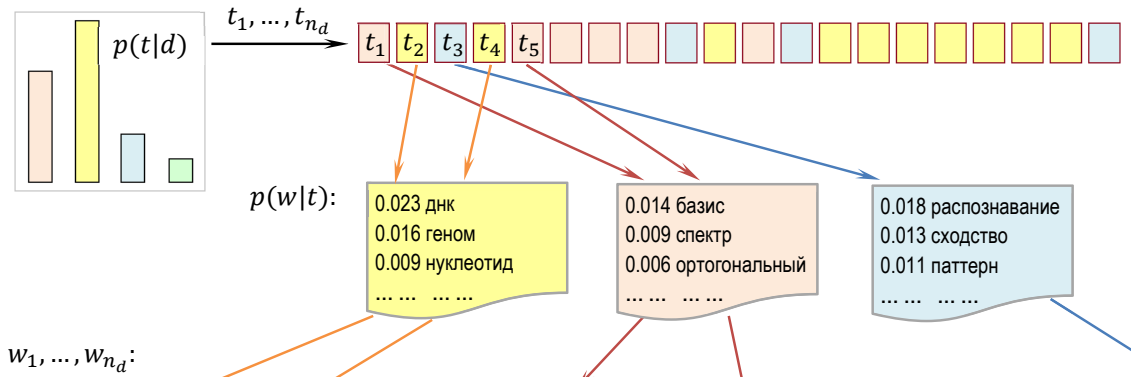
Когда автор писал термин w в документ d, он думал о теме t, и мы хотели бы догадаться, о какой именно.

Тематическая модель выявляет латентные темы по наблюдаемым распределениям слов $p(w|d)$ в документах.

Прямая задача — порождение коллекции по $p(w|t)$ и $p(t|d)$

Вероятностная тематическая модель коллекции документов D описывает появление терминов w в документах d темами t:

$$p(w|d) = \sum_t p(w|t)p(t|d), \quad d \in D$$



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).

Обратная задача — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

Дано: W — словарь терминов
D — коллекция текстовых документов $d = \{w_1 \dots w_{n_d}\}$
n_{dw} = сколько раз термин w встречается в документе d

Найти параметры модели $p(w|d) = \sum_t \phi_{wt}\theta_{td}$:

$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

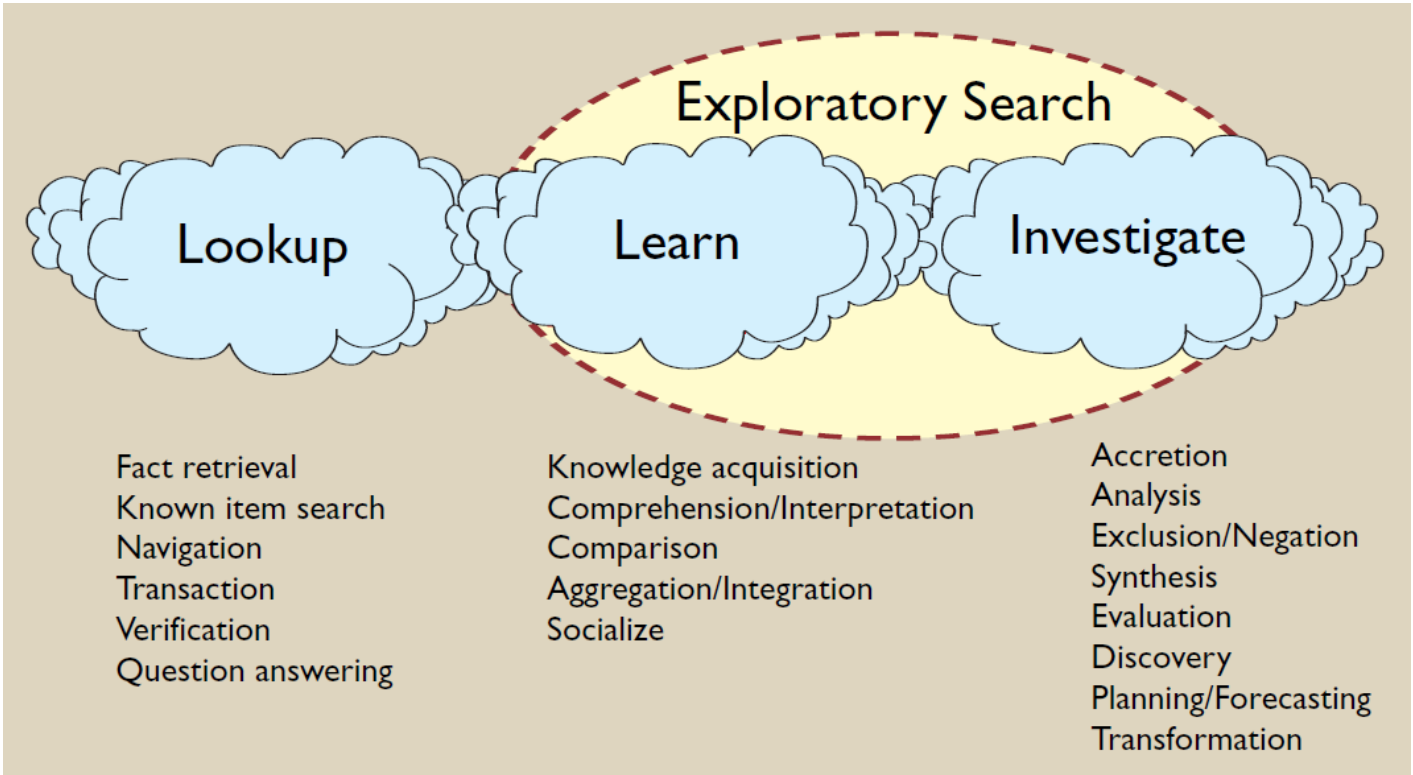
Это задача стохастического матричного разложения,
некорректно поставленная — решение не единственно:

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

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

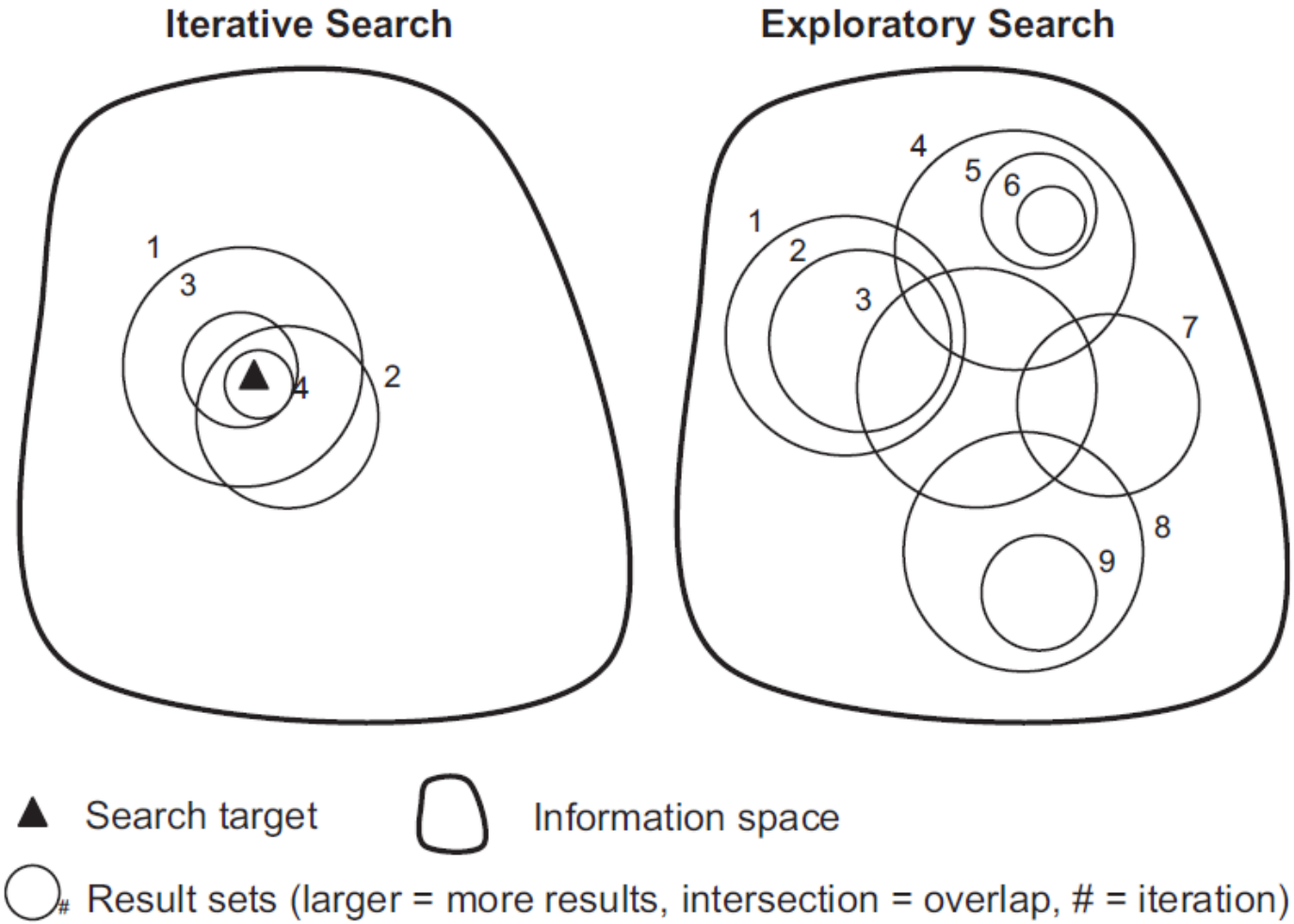
Разведочный поиск — знания «на кончиках пальцев»

- пользователь может не знать ключевых терминов
- пользователя может интересовать множество ответов



Gary Marchionini. Exploratory Search: from finding to understanding. Communications of the ACM. 2006, 49(4), p. 41–46.

От поиска “query-browse-refine” к разведочному поиску



R.W.White, R.A.Roth. Exploratory Search: beyond the Query-Response paradigm. San Rafael, CA: Morgan and Claypool, 2009.

Возможный сценарий разведочного поиска

Поисковый запрос:

- документ любой длины или даже коллекция документов

Цели поиска:

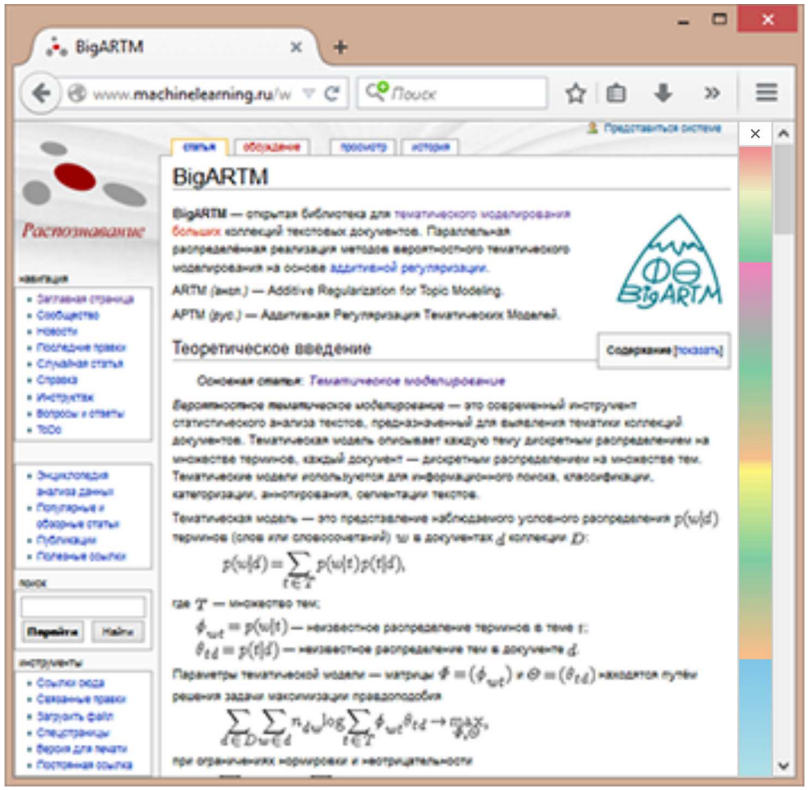
- к каким темам относится мой запрос?
- что ещё известно по этим темам?
- какова тематическая структура этой предметной области?
- что ещё есть понятного, обзорного, важного, свежего?

Сценарий поиска:

1. имея любой текст под рукой, в любом приложении,
2. хотим получить картину содержащихся в нём тем-подтем,
3. и «дорожную карту» предметной области в целом

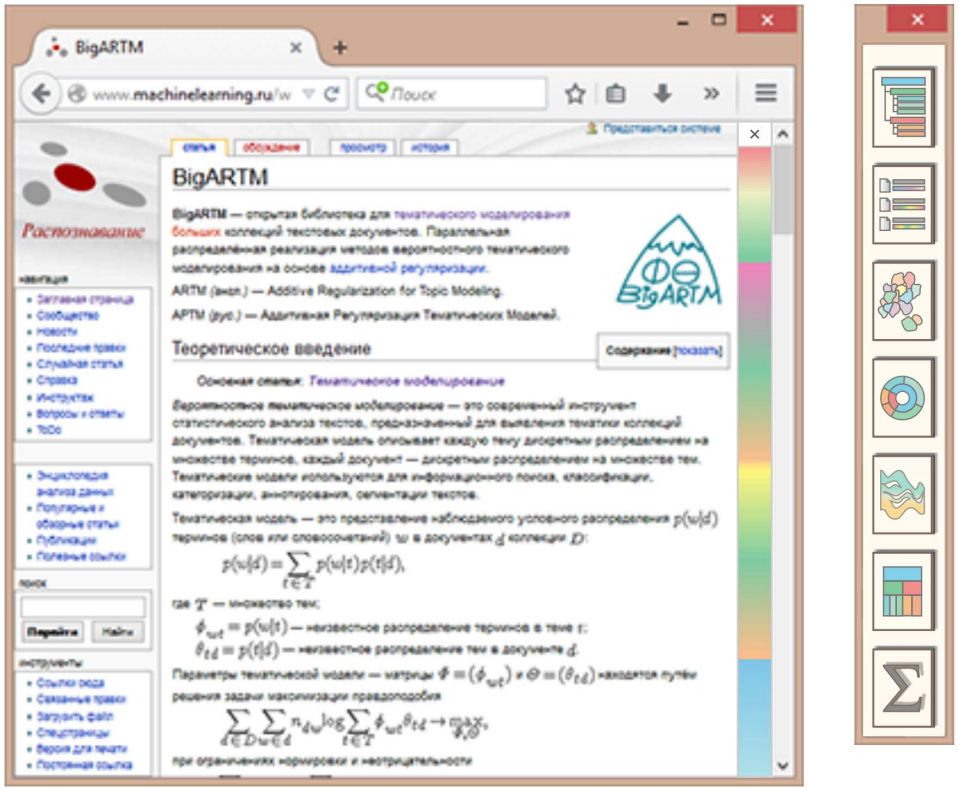
Разведочный поиск: прототип интерфейса

Радужная полоса напоминает, что знания всегда под рукой



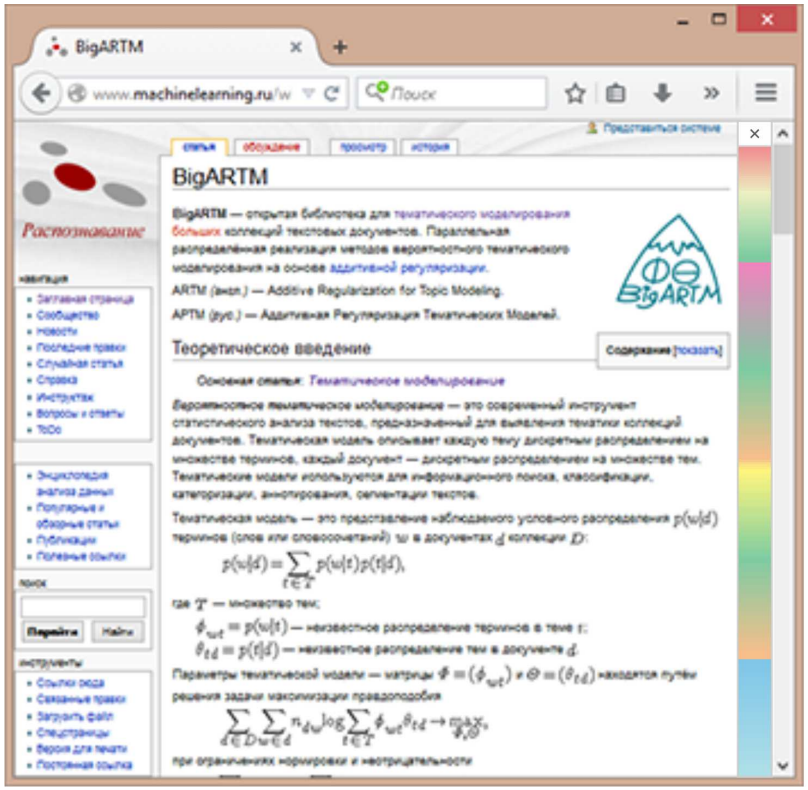
Разведочный поиск: прототип интерфейса

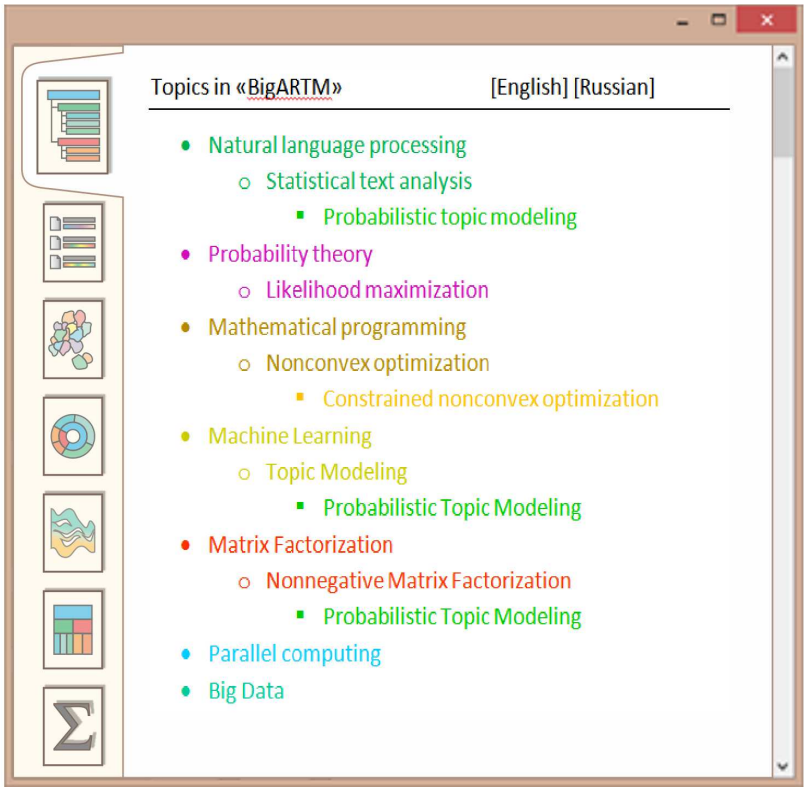
Клик по радужной полосе — тематический поисковый запрос



Разведочный поиск: прототип интерфейса

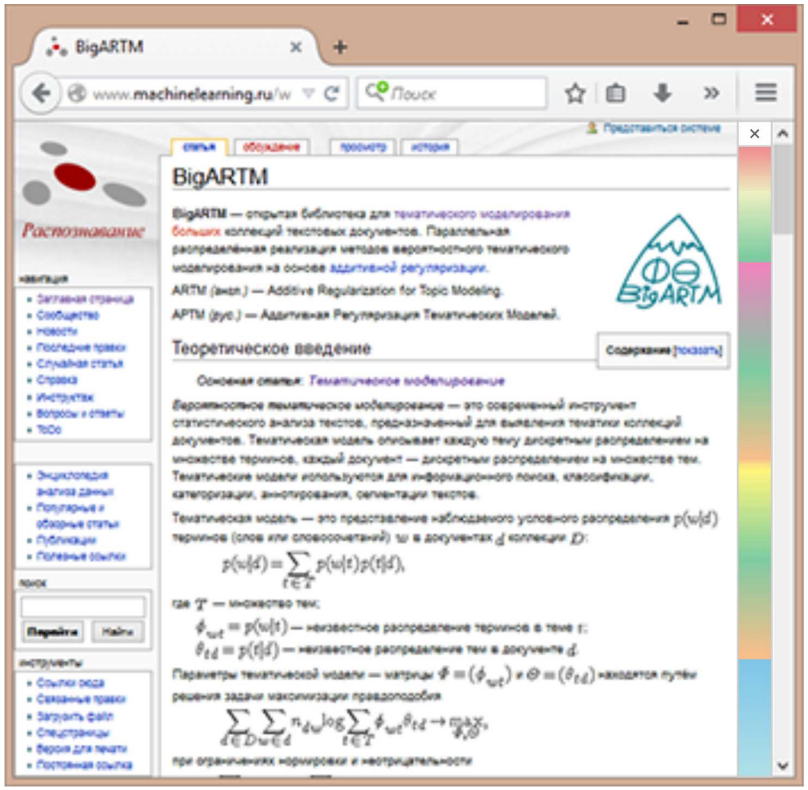
Темы-подтемы выбранного фрагмента текста

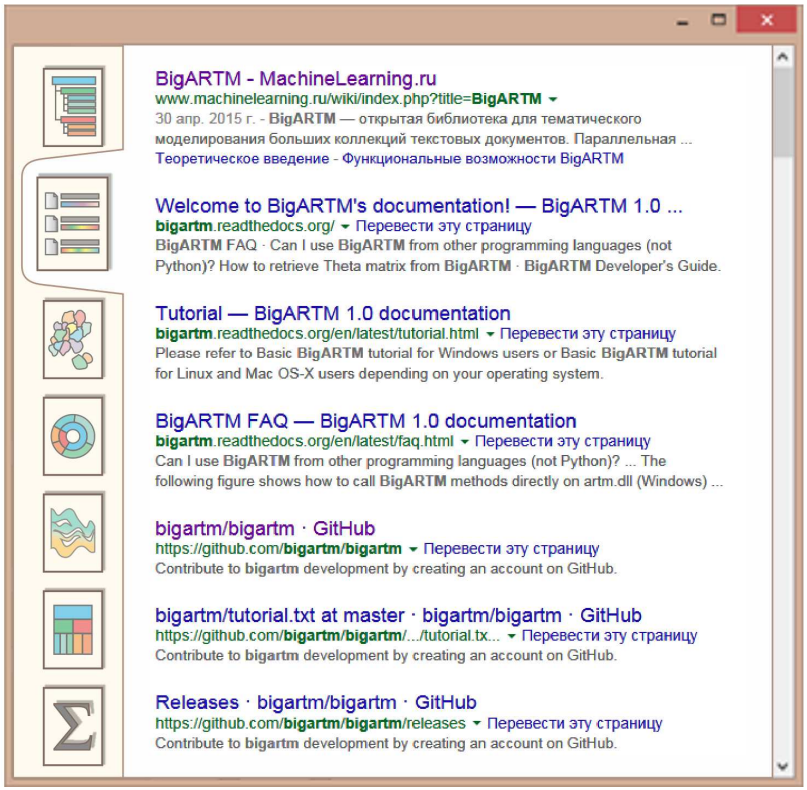




Разведочный поиск: прототип интерфейса

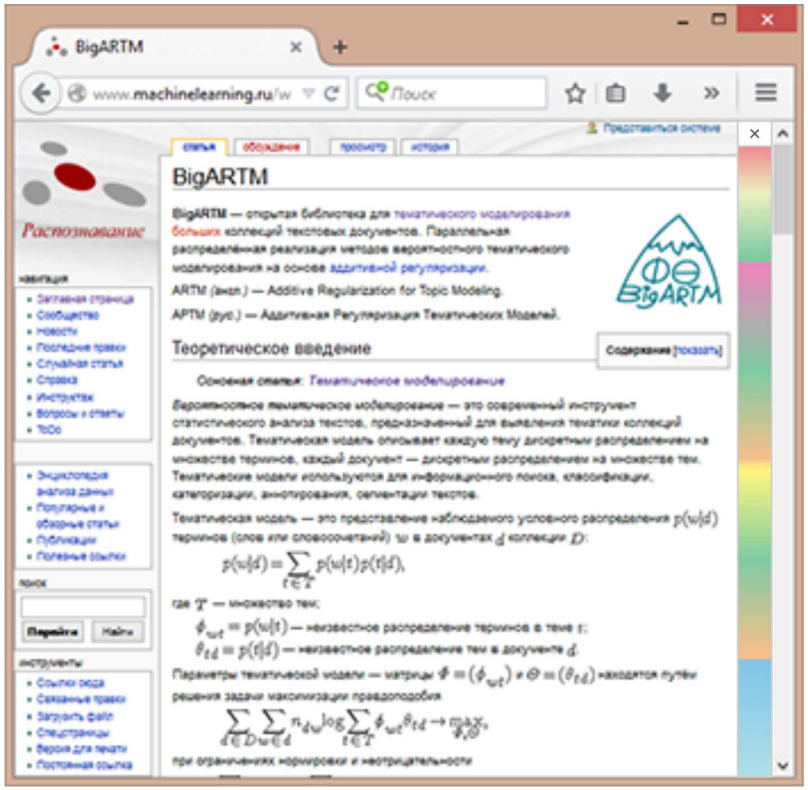
Документы и иные объекты, ранжированные по релевантности

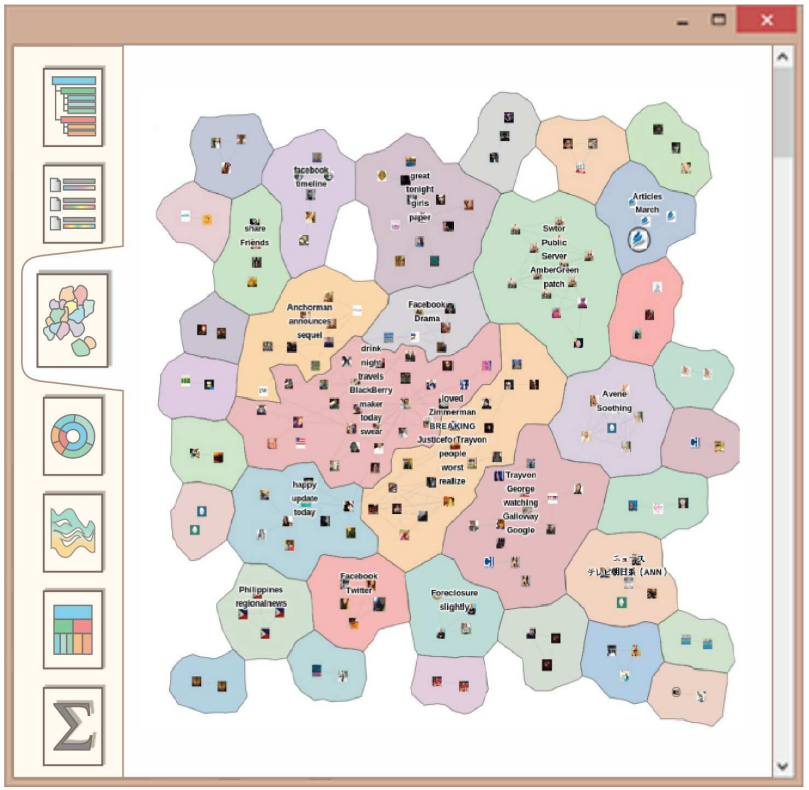




Разведочный поиск: прототип интерфейса

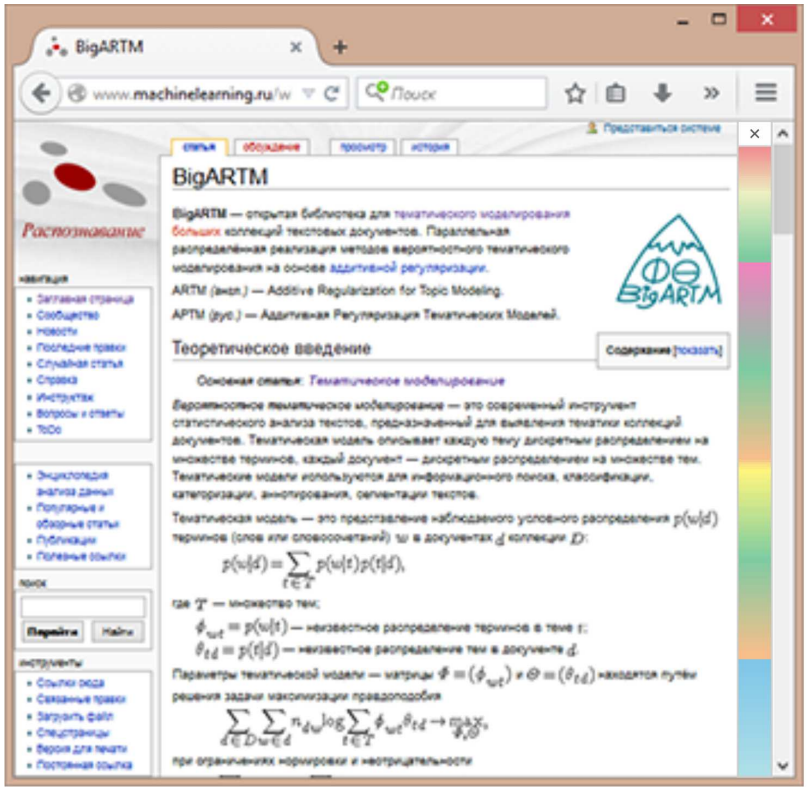
Дорожная карта: кластеризация релевантных документов

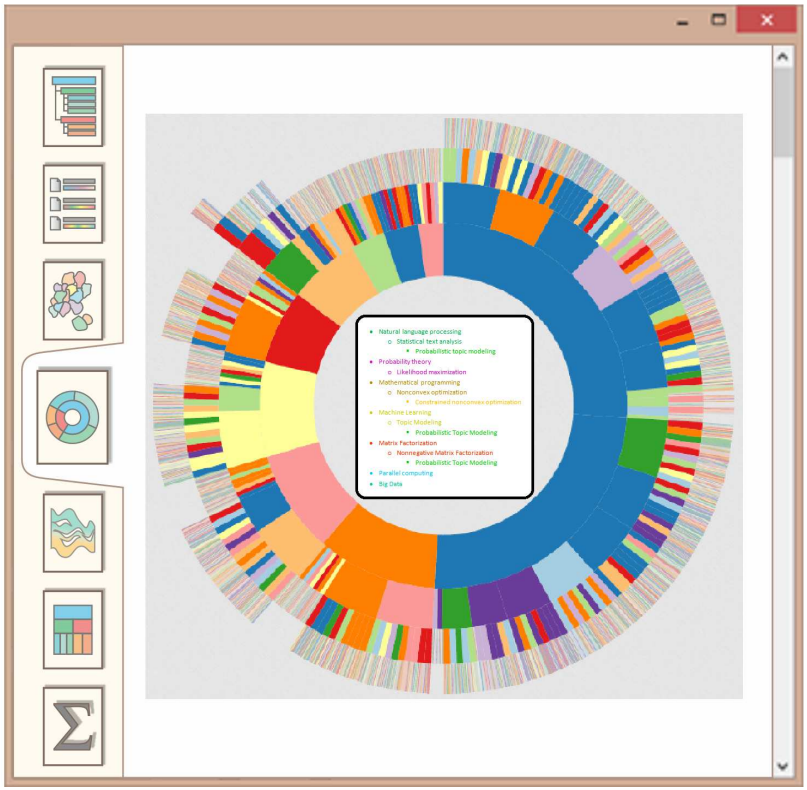




Разведочный поиск: прототип интерфейса

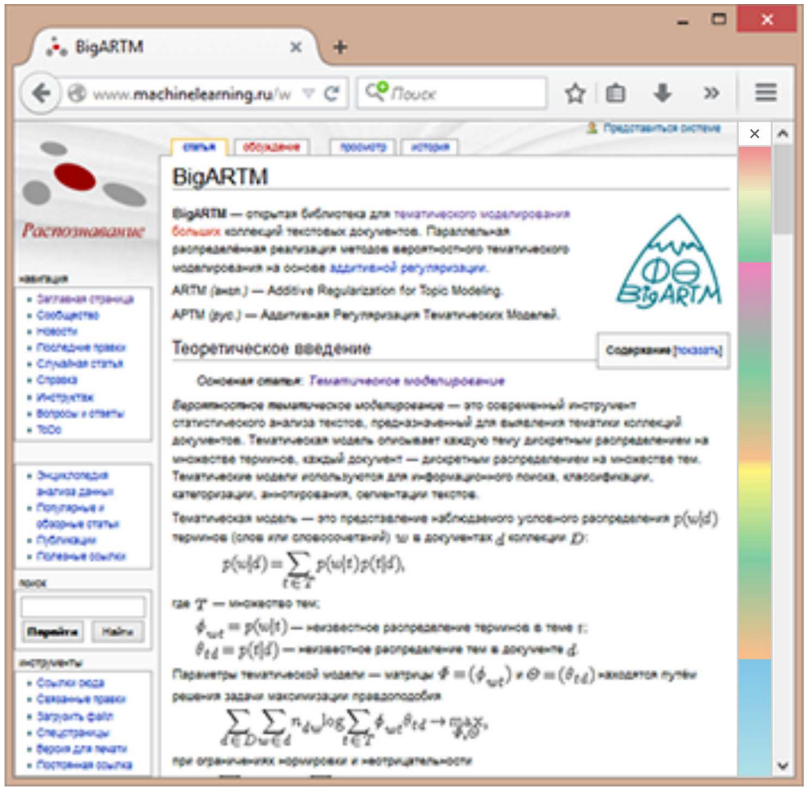
Тематическая иерархия: структура предметной области

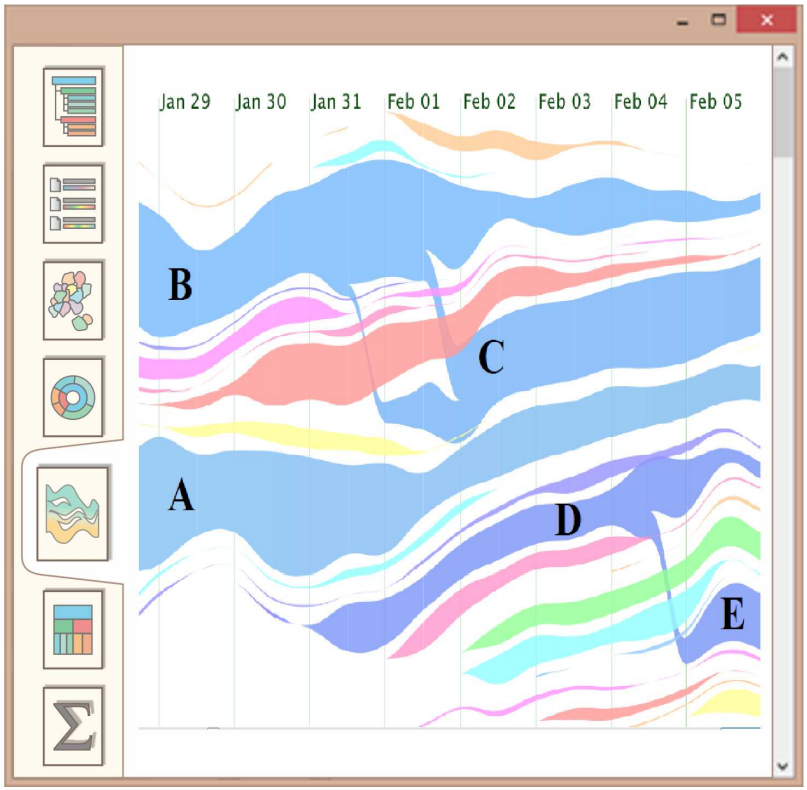




Разведочный поиск: прототип интерфейса

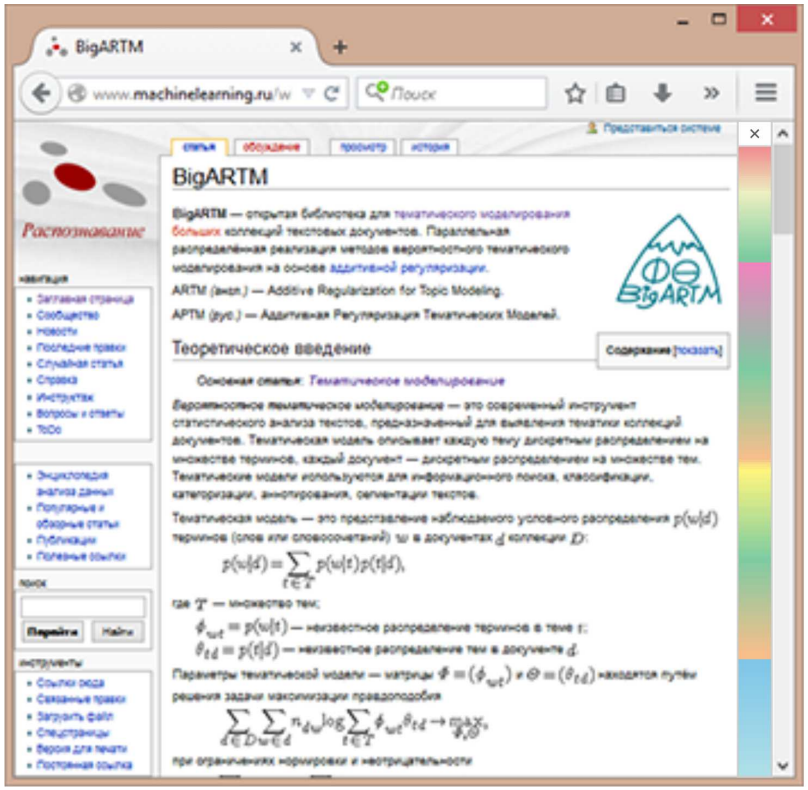
Динамика тем: эволюция предметной области

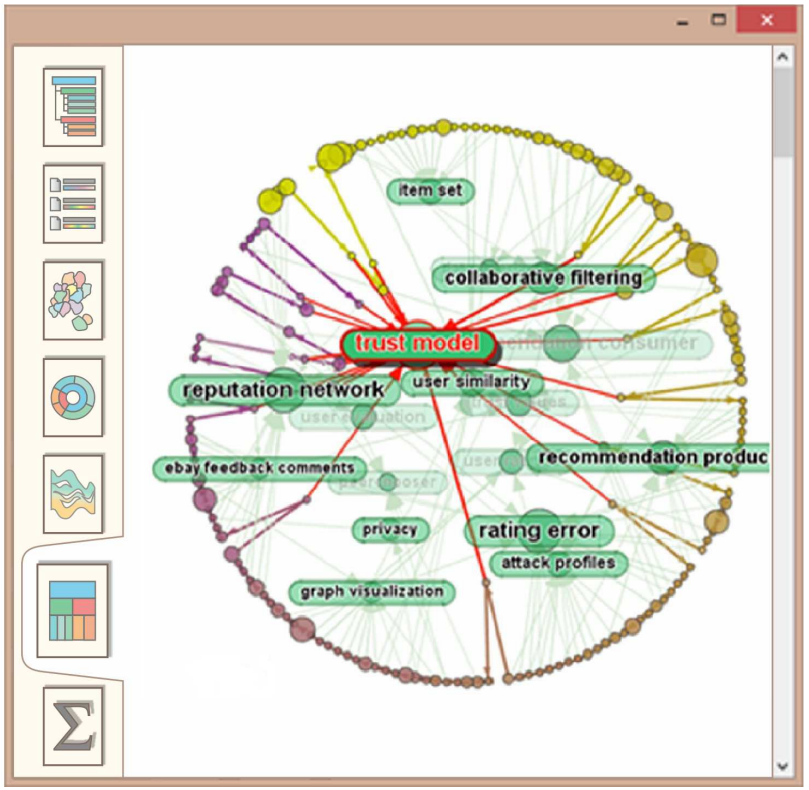




Разведочный поиск: прототип интерфейса

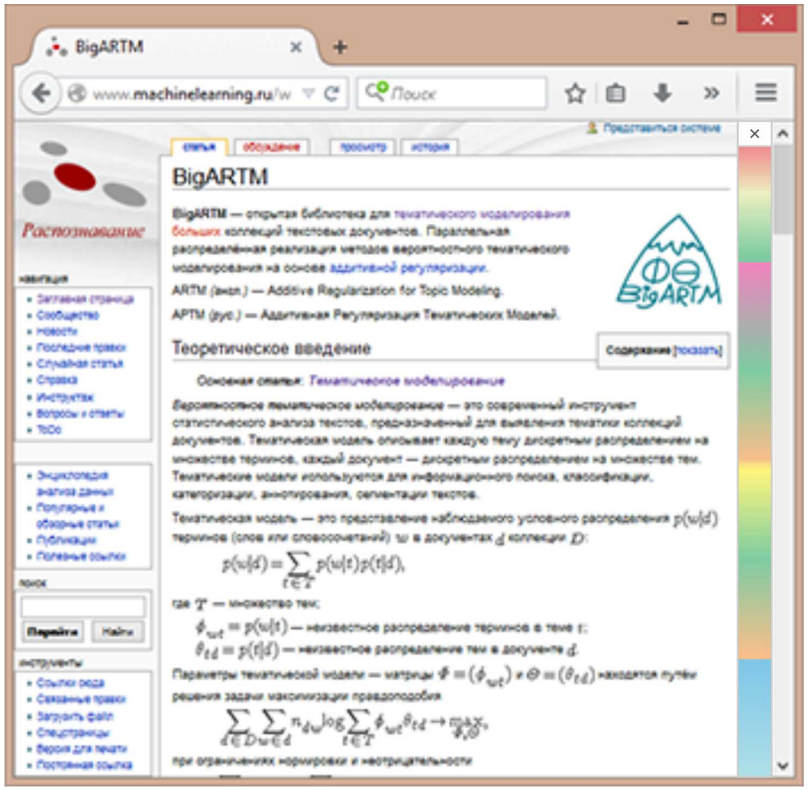
Тематическая сегментация документа запроса

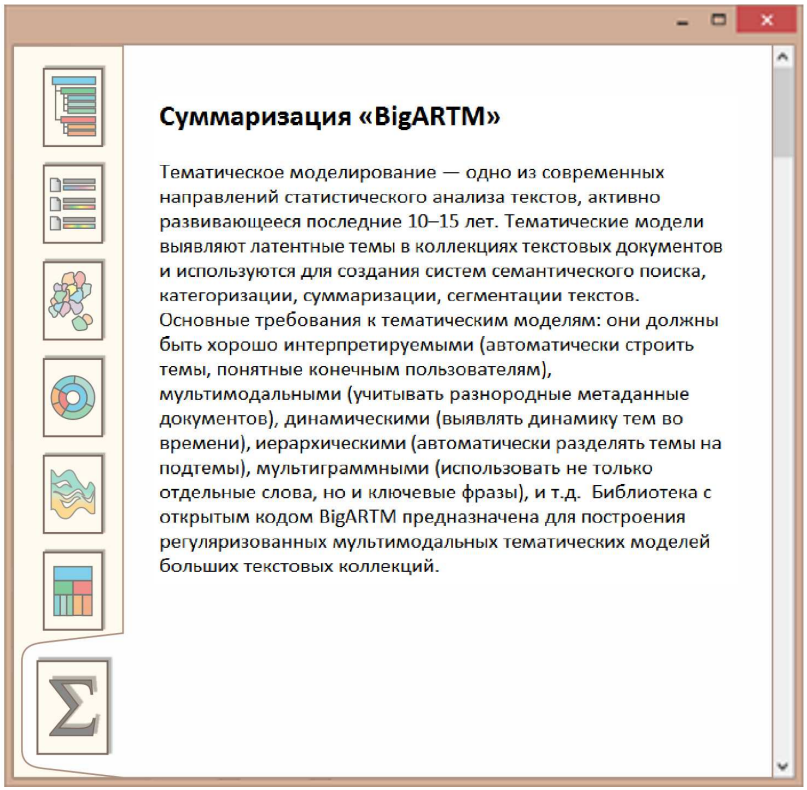




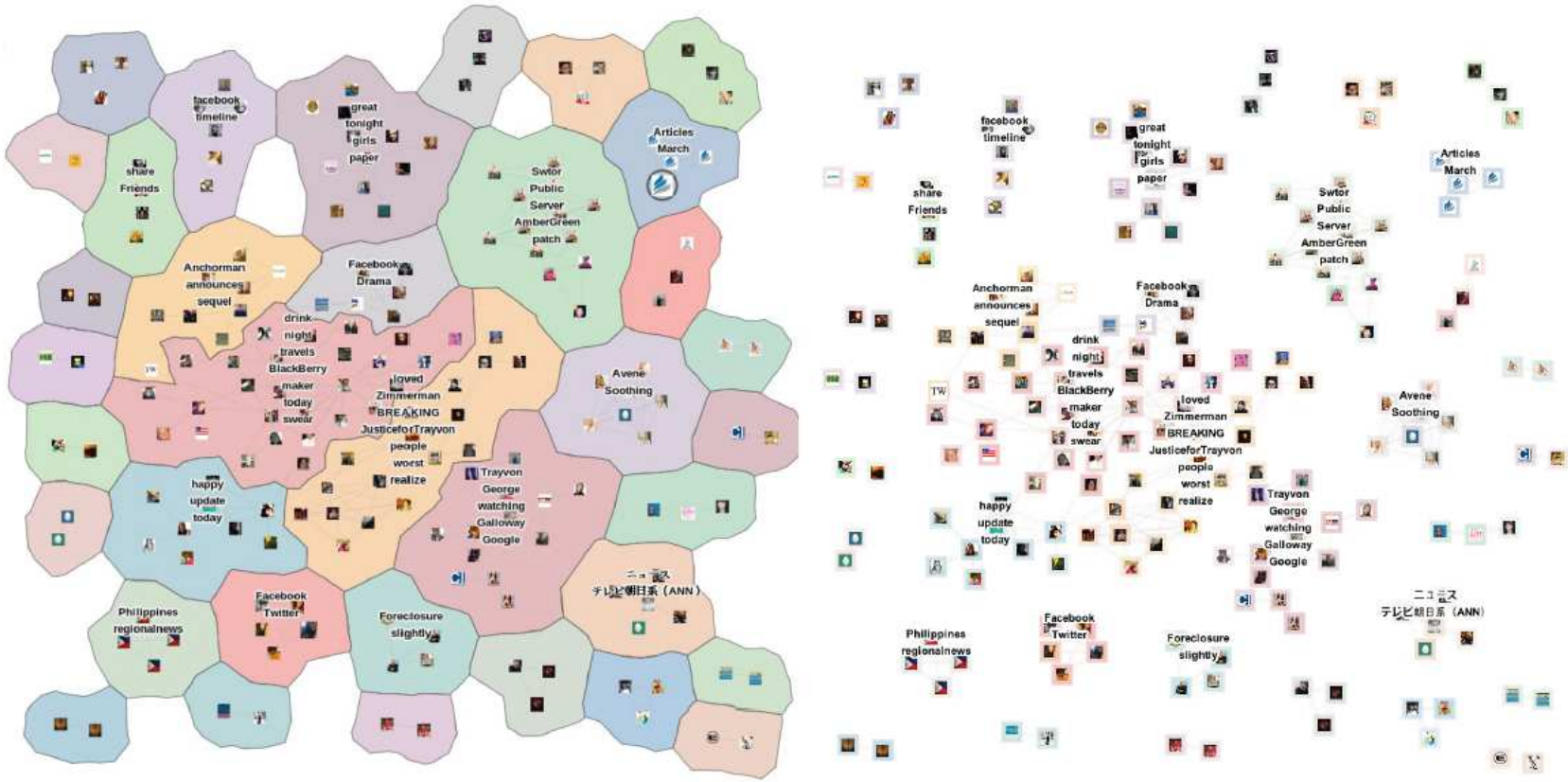
Разведочный поиск: прототип интерфейса

Суммаризация документа запроса





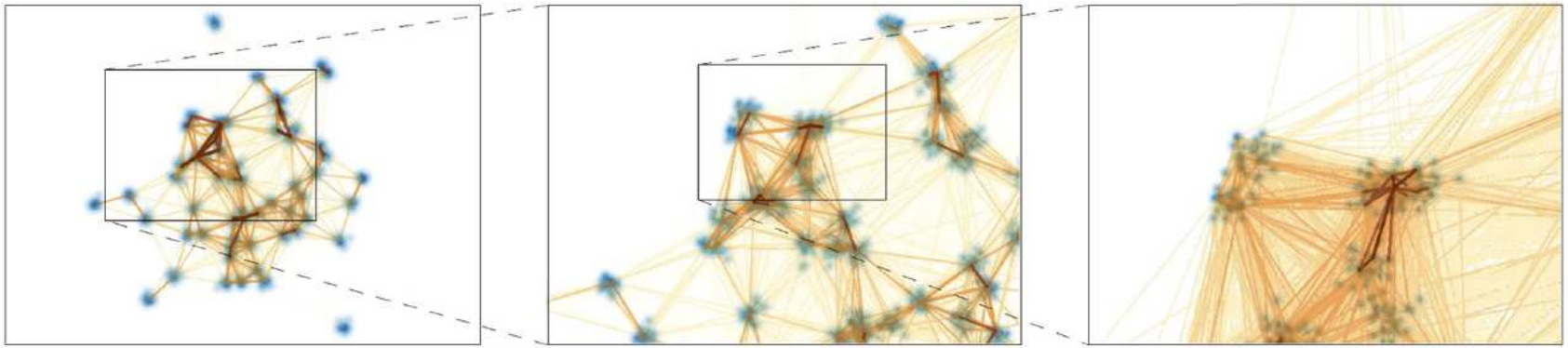
Дорожная карта: кластеризация релевантных документов



«A map metaphor visualization (left) seems more appealing than a plain graph layout (right), and clusters seem easier to identify.»

E.R.Gansner, Y.Hu, S.North. Visualizing Streaming Text Data with Dynamic Maps. 2012.

Дорожная карта: кластеризация релевантных документов



- Кластеры
  - кластеров
    - кластеров
      - кластеров...

M.Zinsmaier, *U.Brandes*, *O.Deussen*, *H.Strobelt*. Interactive level-of-detail rendering of large graphs. IEEE Trans. Vis. Comput. Graph. 2012.

http://textvis.lnu.se

Интерактивный обзор 170 средств визуализации текстов

Технологические элементы разведочного поиска

1. Интернет-краулинг ..... имеются готовые решения
2. Фильтрация контента ..... имеются готовые решения
3. Тематическое моделирование ..... **математика здесь**
4. Инвертированный индекс ..... имеются готовые решения
5. Ранжирование ..... имеются готовые решения
6. Визуализация ..... имеются готовые решения

Тематическая модель для разведочного поиска должна быть...

1. **Интерпретируемая**: каждая тема понятна людям
2. **Мультиграммная**: термины-словосочетания неразрывны
3. **Мультиязычная**: кросс-языковой и много-языковой поиск
4. **Мультимодальная**: авторы, связи, тэги, пользователи,. . .
5. **Динамическая**: развитие тем во времени
6. **Иерархическая**: настраиваемая гранулярность тем
7. **Сегментирующая**: границы тем внутри документа
8. **Обучаемая**: учёт экспертных оценок

Обратная задача — восстановление $p(w|t)$ и $p(t|d)$ по коллекции

Дано: W — словарь терминов

D — коллекция текстовых документов $d = \{w_1 \dots w_{n_d}\}$

n_{dw} = сколько раз термин w встречается в документе d

Найти параметры модели $p(w|d) = \sum_t \phi_{wt}\theta_{td}$:

$\phi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t

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

Задача максимизации логарифма правдоподобия:

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

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

Решение данной задачи удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} \equiv p(t|d,w)$, n_{wt}, n_{td}:

$$\text{E-шаг:} \quad p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \dfrac{n_{wt}}{n_t}; \quad n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw}; \quad n_t = \sum\limits_{w \in W} n_{wt} \\ \theta_{td} = \dfrac{n_{td}}{n_d}; \quad n_{td} = \sum\limits_{w \in d} n_{dw} p_{tdw}; \quad n_d = \sum\limits_{t \in T} n_{td} \end{cases}$$

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!*

EM-алгоритм. Элементарная интерпретация

EM-алгоритм — это чередование E и M шагов до сходимости.

E-шаг: условные вероятности тем $p(t|d,w)$ для всех t, d, w вычисляются через ϕ_{wt}, θ_{td} по формуле Байеса:

$$p(t|d,w) = \frac{p(w,t|d)}{p(w|d)} = \frac{p(w|t)p(t|d)}{p(w|d)} = \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}}.$$

M-шаг: частотные оценки условных вероятностей вычисляются путём суммирования счётчика $n_{dwt} = n_{dw}p(t|d,w)$:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \qquad n_{wt} = \sum_{d \in D} n_{dwt}, \qquad n_t = \sum_{w \in W} n_{wt};$$

$$\theta_{td} = \frac{n_{td}}{n_d}, \qquad n_{td} = \sum_{w \in d} n_{dwt}, \qquad n_d = \sum_{t \in T} n_{td}.$$

LDA — Latent Dirichlet Allocation [Blei 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:

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

Различие проявляется только при малых n_{wt}, n_{td}

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

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.

Тематическое моделирование на основе байесовского обучения

1. Чисто вероятностные модели порождения текста.
2. Байесовский вывод нестандартен для каждой модели.

$$p(Z,W|\boldsymbol{\alpha},\boldsymbol{\beta}) = p(W|Z,\boldsymbol{\beta})p(Z|\boldsymbol{\alpha})$$
$$p(W|Z,\boldsymbol{\beta}) = \int p(W|Z,\boldsymbol{\Phi})p(\boldsymbol{\Phi}|\boldsymbol{\beta})\mathrm{d}\boldsymbol{\Phi}$$
$$p(\boldsymbol{\Phi}|\boldsymbol{\beta}) = \prod_{k=1}^{K} p(\boldsymbol{\phi}_k|\boldsymbol{\beta}) = \prod_{k=1}^{K}\frac{1}{\mathrm{B}(\boldsymbol{\beta})}\prod_{v=1}^{V}\phi_{k,v}^{\beta_v-1}$$
$$p(W|Z,\boldsymbol{\Phi}) = \prod_{i=1}^{N}\phi_{z_i,w_i} = \prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{k,v}^{\Psi(k,v)}$$
$$\Psi(k,v) = \sum_{i=1}^{N}\mathrm{I}\{w_i = v \wedge z_i = k\}$$
$$p(W|Z,\boldsymbol{\beta}) = \int\prod_{k=1}^{K}\frac{1}{\mathrm{B}(\boldsymbol{\beta})}\prod_{v=1}^{V}\phi_{k,v}^{\Psi(k,v)+\beta_v-1}\mathrm{d}\boldsymbol{\phi}_k$$
$$\int\prod_{k=1}^{K} f_k(\phi_k)\,\mathrm{d}\phi_1\ldots\mathrm{d}\phi_K = \prod_{k=1}^{K}\int f_k(\phi_k)\,\mathrm{d}\phi_k$$
$$p(W|Z,\boldsymbol{\beta}) = \prod_{k=1}^{K}\left(\int\frac{1}{\mathrm{B}(\boldsymbol{\beta})}\prod_{v=1}^{V}\phi_{k,v}^{\Psi(k,v)+\beta_v-1}\mathrm{d}\boldsymbol{\phi}_k\right)$$
$$= \prod_{k=1}^{K}\left(\frac{1}{\mathrm{B}(\boldsymbol{\beta})}\int\prod_{v=1}^{V}\phi_{k,v}^{\Psi(k,v)+\beta_k-1}\mathrm{d}\boldsymbol{\phi}_k\right)$$
$$p(W|Z,\boldsymbol{\beta}) = \prod_{k=1}^{K}\frac{\mathrm{B}(\Psi_k+\boldsymbol{\beta})}{\mathrm{B}(\boldsymbol{\beta})}$$

$$p(\boldsymbol{\Theta}|\boldsymbol{\alpha}) = \prod_{d=1}^{D}p(\boldsymbol{\theta}_m|\boldsymbol{\alpha}) = \prod_{d=1}^{D}\frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_{d,k}^{\alpha_k-1}$$
$$p(Z|\boldsymbol{\Theta}) = \prod_{i=1}^{N}\theta_{d_i,z_i} = \prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{d,k}^{\Omega(d,k)}$$
$$p(Z|\boldsymbol{\alpha}) = \int p(Z|\boldsymbol{\Theta})p(\boldsymbol{\Theta}|\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\Theta}$$
$$= \prod_{d=1}^{D}\left(\int\frac{1}{\mathrm{B}(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_{d,k}^{\Omega(d,k)+\alpha_k-1}\mathrm{d}\boldsymbol{\theta}_k\right)$$
$$= \prod_{d=1}^{D}\frac{\mathrm{B}(\Omega_d+\boldsymbol{\alpha})}{\mathrm{B}(\boldsymbol{\alpha})},$$
$$\Omega(d,k) = \sum_{i=1}^{N}\mathrm{I}\{d_i = m \wedge z_i = z\}$$
$$p(Z,W|\boldsymbol{\alpha},\boldsymbol{\beta}) = p(W|Z,\boldsymbol{\beta})p(Z|\boldsymbol{\alpha})$$
$$= \prod_{k=1}^{K}\frac{\mathrm{B}(\Psi_k+\boldsymbol{\beta})}{\mathrm{B}(\boldsymbol{\beta})}\cdot\prod_{d=1}^{D}\frac{\mathrm{B}(\Omega_d+\boldsymbol{\alpha})}{\mathrm{B}(\boldsymbol{\alpha})}$$
$$p(z_i = k|Z_{-i},W,\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{p(z_i=k,Z_{-i},W|\boldsymbol{\alpha},\boldsymbol{\beta})}{p(Z_{-i},W|\boldsymbol{\alpha},\boldsymbol{\beta})}$$
$$p(z_i|Z_{-i},W,\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{p(Z,W|\boldsymbol{\alpha},\boldsymbol{\beta})}{p(Z_{-i},W_{-i}|\boldsymbol{\alpha},\boldsymbol{\beta})}$$
$$p(Z,W|\boldsymbol{\alpha},\boldsymbol{\beta}) = p(W|Z,\boldsymbol{\beta})p(Z|\boldsymbol{\alpha})$$
$$= \prod_{k=1}^{K}\frac{\mathrm{B}(\Psi_k^{-i}+\boldsymbol{\beta})}{\mathrm{B}(\boldsymbol{\beta})}\cdot\prod_{d=1}^{D}\frac{\mathrm{B}(\Omega_d^{-i}+\boldsymbol{\alpha})}{\mathrm{B}(\boldsymbol{\alpha})}$$

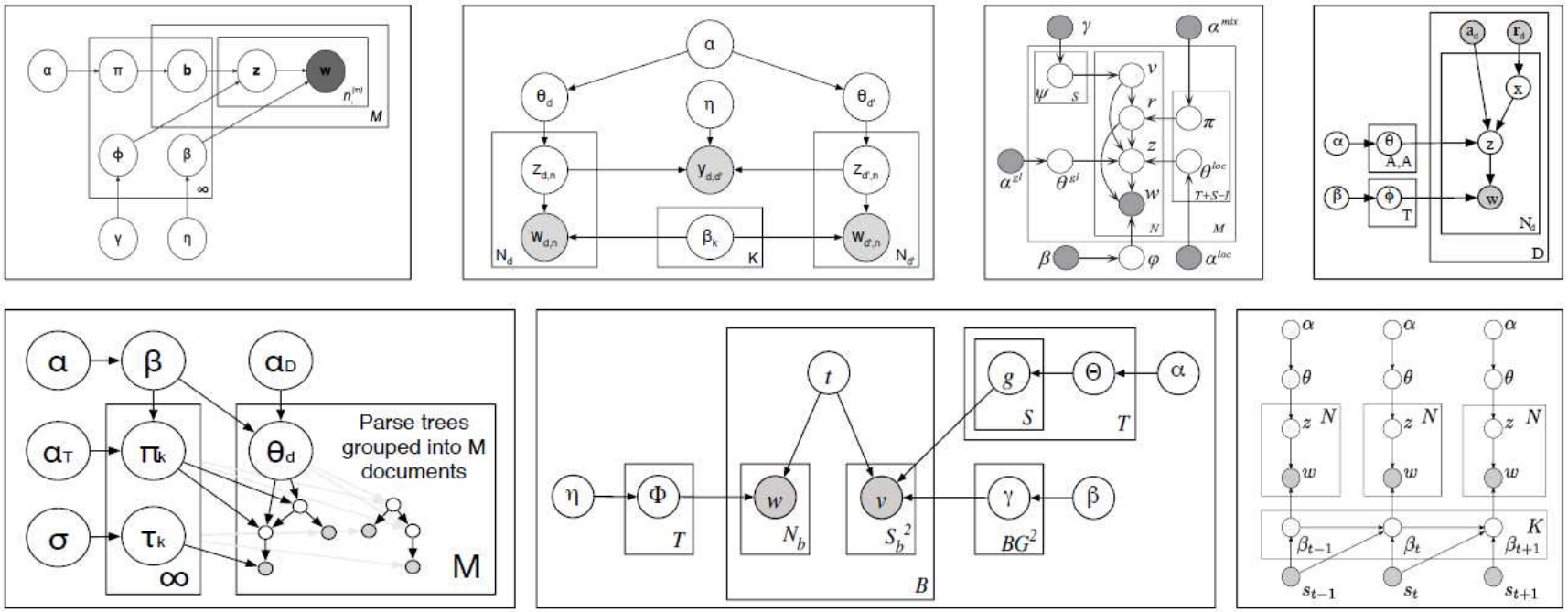
$$\Psi^{-i}(k,v) = \sum_{\substack{1\le j\le N\\ j\ne i}}\mathrm{I}\{w_j = v \wedge z_j = k\}$$
$$\Omega^{-i}(d,k) = \sum_{\substack{1\le j\le N\\ j\ne i}}\mathrm{I}\{d_j = d \wedge z_j = k\}$$
$$\Psi(k,v) = \begin{cases}\Psi^{-i}(k,v)+1 & \text{if } v = w_i \text{ and } k = z_i;\\ \Psi^{-i}(k,v) & \text{all other cases.}\end{cases}$$
$$\Omega(d,k) = \begin{cases}\Omega^{-i}(d,k)+1 & \text{if } d = d_i \text{ and } k = z_i;\\ \Omega^{-i}(d,k) & \text{all other cases.}\end{cases}$$
$$\sum_{v=1}^{V}n(v;z_i) = 1+\sum_{v=1}^{V}n_{-i}(v|z_i)$$
$$\sum_{k=1}^{K}n(z;d_i) = 1+\sum_{k=1}^{K}n_{-i}(z|d_i)$$
$$\mathrm{B}(\boldsymbol{x}) = \frac{\prod_{k=1}^{\dim\boldsymbol{x}}\Gamma(x_k)}{\Gamma(\sum_{k=1}^{\dim\boldsymbol{x}}x_k)}$$
$$p(z_i|Z_{-i},W,\boldsymbol{\alpha},\boldsymbol{\beta}) = \frac{\mathrm{B}(n(\cdot;z_i)+\boldsymbol{\beta})}{\mathrm{B}(n_{-i}(\cdot|z_i)+\boldsymbol{\beta})}\cdot\frac{\mathrm{B}(n(\cdot;m_i)+\boldsymbol{\alpha})}{\mathrm{B}(n_{-i}(\cdot|m_i)+\boldsymbol{\alpha})}$$
$$\frac{\prod_{v=1}^{V}\Gamma(n(t;z_i)+\beta_t)}{\Gamma(\sum_{v=1}^{V}n(t;z_i)+\beta_t)}\cdot\frac{\prod_{k=1}^{K}\Gamma(n(z;d_i)+\alpha_z)}{\Gamma(\sum_{k=1}^{K}n(z;d_i)+\alpha_z)}$$
$$\frac{\prod_{v=1}^{V}\Gamma(n_{-i}(t|z_i)+\beta_t)}{\Gamma(\sum_{v=1}^{V}n_{-i}(t|z_i)+\beta_t)}\cdot\frac{\prod_{k=1}^{K}\Gamma(n_{-i}(z|d_i)+\alpha_z)}{\Gamma(\sum_{k=1}^{K}n_{-i}(z|d_i)+\alpha_z)}$$
$$\frac{\Gamma(n(w_i;z_i)+\beta_{w_i})}{\Gamma(\sum_{v=1}^{V}n(t;z_i)+\beta_t)}\cdot\frac{\Gamma(n(z_i;d_i)+\alpha_{z_i})}{\Gamma(\sum_{k=1}^{K}n(z;d_i)+\alpha_z)}$$
$$\frac{\Gamma(n_{-i}(w_i|z_i)+\beta_{w_i})}{\Gamma(\sum_{v=1}^{V}n_{-i}(t|z_i)+\beta_t)}\cdot\frac{\Gamma(n_{-i}(z_i|d_i)+\alpha_{z_i})}{\Gamma(\sum_{k=1}^{K}n_{-i}(z|d_i)+\alpha_z)}$$

$$p(z_i|Z_{-i},W,\boldsymbol{\alpha},\boldsymbol{\beta}) =$$
$$\frac{n(w_i;z_i)+\beta_{w_i}-1}{\left[\sum_{v=1}^{V}n(t;z_i)+\beta_t\right]-1}\cdot\frac{n(z_i;d_i)+\alpha_{z_i}-1}{\left[\sum_{k=1}^{K}n(z;d_i)+\alpha_z\right]-1}$$
$$p(z_i|Z_{-i},W,\boldsymbol{\alpha},\boldsymbol{\beta}) =$$
$$\frac{n(w_i;z_i)+\beta_{w_i}-1}{\left[\sum_{v=1}^{V}n(t;z_i)+\beta_t\right]-1}\cdot\left[n(z_i;d_i)+\alpha_{z_i}-1\right]$$
$$\phi_{k,t} = p(w=t|z=k,W,Z,\boldsymbol{\beta})$$
$$\theta_{m,k} = p(z=k|Z,\boldsymbol{\alpha})$$
$$\phi_{k,t}\cdot\theta_{m,k} = p(w=t\mid z=k,W,Z,\boldsymbol{\beta})\cdot p(z=k\mid Z,\boldsymbol{\alpha})$$
$$= p(w=t,z=k\mid W,Z,\boldsymbol{\alpha},\boldsymbol{\beta})$$
$$= \frac{p(\hat{W},\hat{Z}\mid\boldsymbol{\alpha},\boldsymbol{\beta})}{p(W,Z\mid\boldsymbol{\alpha},\boldsymbol{\beta})},$$
$$= \frac{\frac{\Gamma(n(t;k)+1+\beta_t)}{\Gamma(\sum_{v=1}^{V}n(t;k)+1+\beta_t)}}{\frac{\Gamma(n_{-i}(t|k)+\beta_t)}{\Gamma(\sum_{v=1}^{V}n_{-i}(t|k)+\beta_t)}}\cdot\frac{\frac{\Gamma(n(k;m)+1+\alpha_k)}{\Gamma(\sum_{k=1}^{K}\Omega(d,k)+1+\alpha_z)}}{\frac{\Gamma(n_{-i}(k|m)+\alpha_k)}{\Gamma(\sum_{k=1}^{K}\Omega^{-i}(d,k)+\alpha_z)}}$$
$$\phi_{k,t}\cdot\theta_{m,k} = \frac{n(t;k)+\beta_t}{\left(\sum_{v=1}^{V}n(t;k)+\beta_t\right)}\cdot\frac{n(k;m)+\alpha_k}{\left(\sum_{k=1}^{K}\Omega(d,k)+\alpha_z\right)}$$
$$\phi_{k,t} = \frac{n(t;k)+\beta_t}{\left(\sum_{t'=1}^{V}n(t';k)+\beta_{t'}\right)}$$
$$\theta_{m,k} = \frac{n(k;m)+\alpha_k}{\left(\sum_{k=1}^{K}\Omega(d,k)+\alpha_z\right)}$$

Yi Wang. Distributed Gibbs Sampling of Latent Dirichlet Allocation: The Gritty Details. 2008.

Тематическое моделирование на основе байесовского обучения

Графические модели упрощают понимание, но байесовский вывод всё равно остаётся нестандартным для каждой модели.



David M. Blei. Probabilistic topic models // Communications of the ACM, 2012. Vol. 55, No. 4., Pp. 77–84.

Тематическое моделирование на основе байесовского обучения

Недостатки байесовского обучения как доминирующей теоретической основы тематического моделирования:

- сложность понимания,
- сложность байесовского вывода,
- сложность унификации и сравнения моделей,
- невозможность комбинирования моделей,
- сотни моделей в литературе ... но не в открытом коде,
- всё это создаёт барьеры вхождения для прикладников,
- в итоге все пользуются старыми PLSA и LDA ...
- ... и плюются

ARTM — 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 максимизации регуляризованного правдоподобия:

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

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

Решение данной задачи удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t|d,w)$, n_{wt}, n_{td}:

$$\text{E-шаг:} \quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right);$$

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

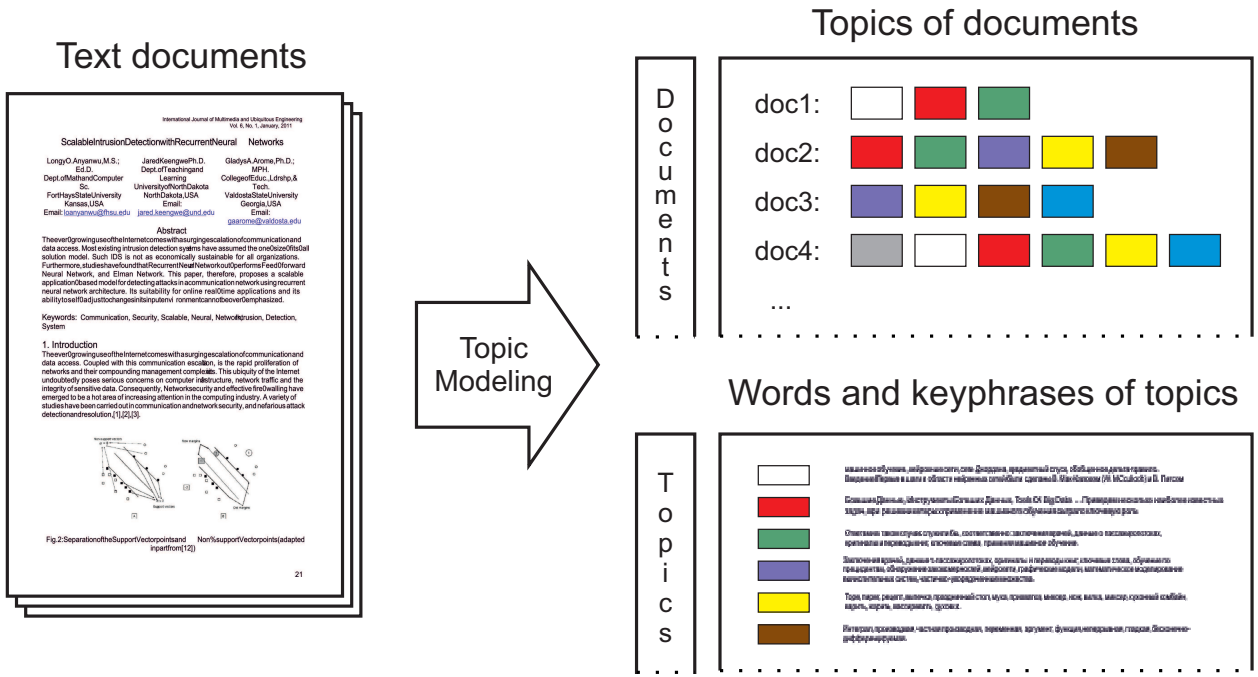
где $\underset{t\in T}{\text{norm}}\, x_t = \dfrac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ — операция нормировки вектора.

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

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

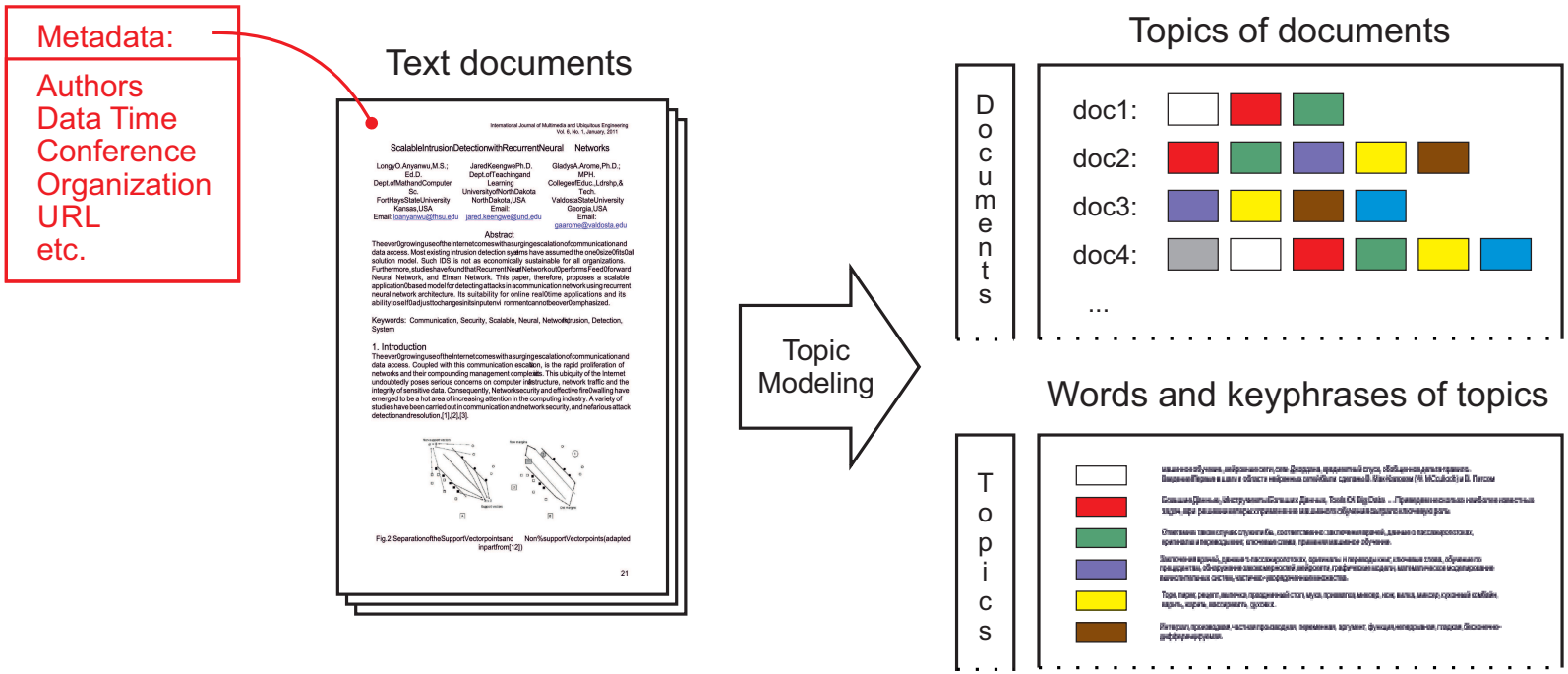
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$,...



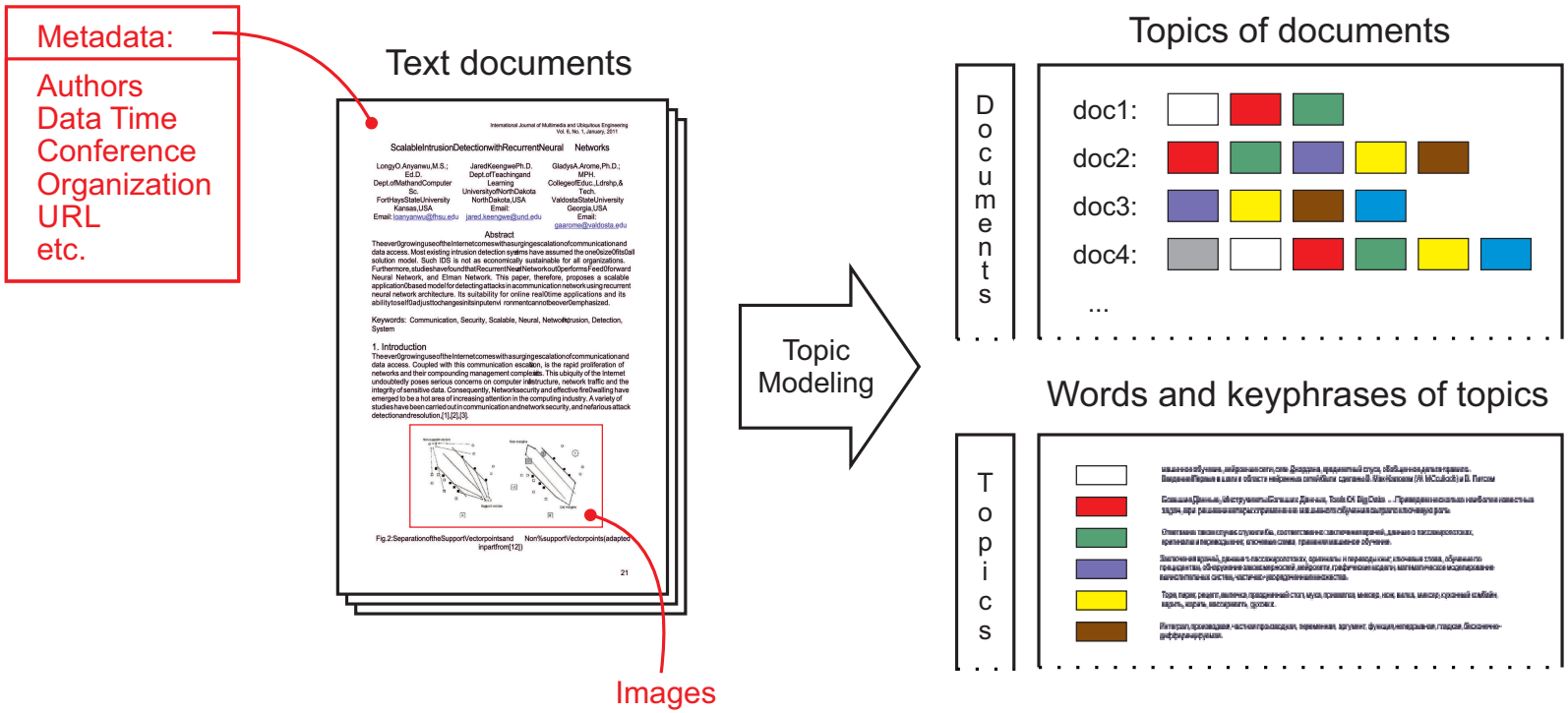
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$,
авторов $p(t|a)$, времени $p(t|a)$,...



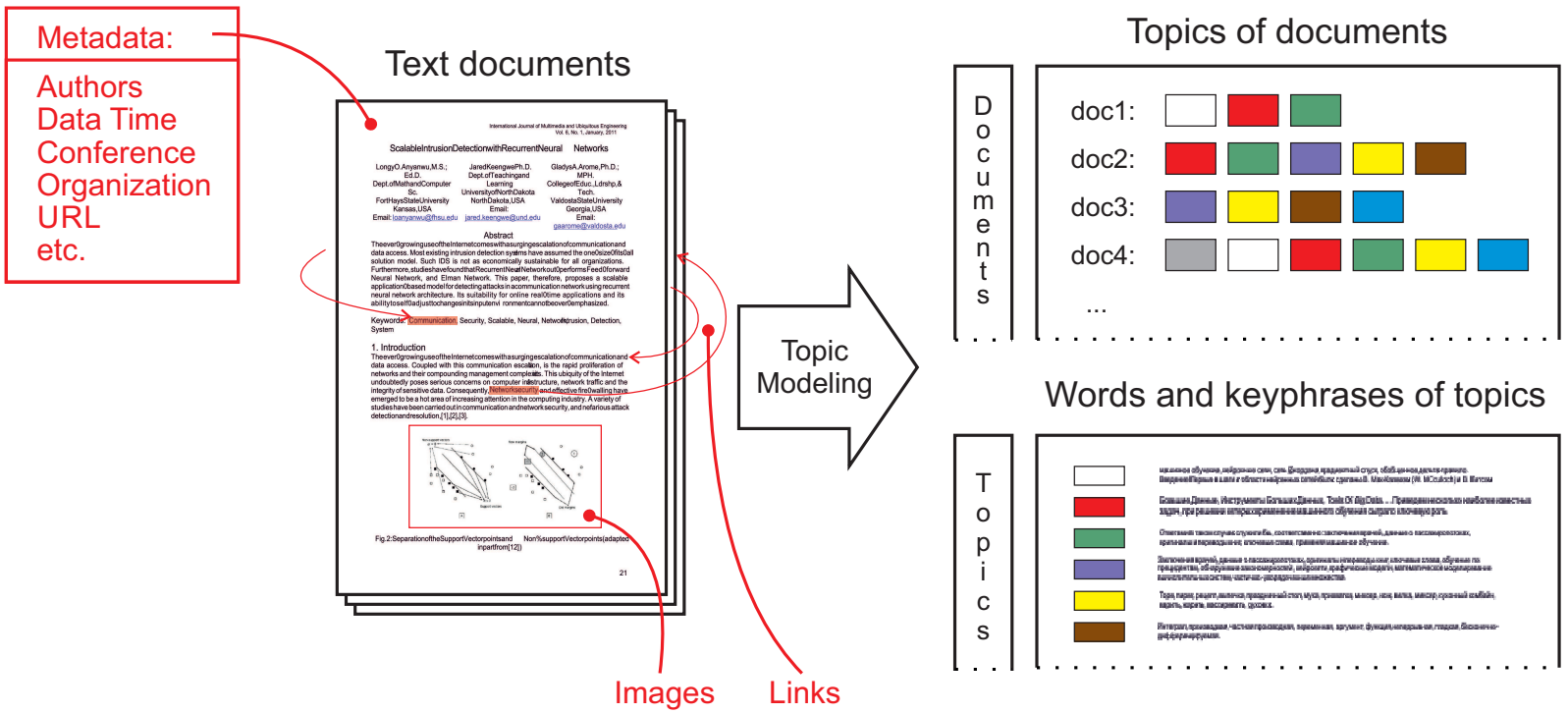
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$, авторов $p(t|a)$, времени $p(t|a)$, элементов изображений $p(t|e)$,...



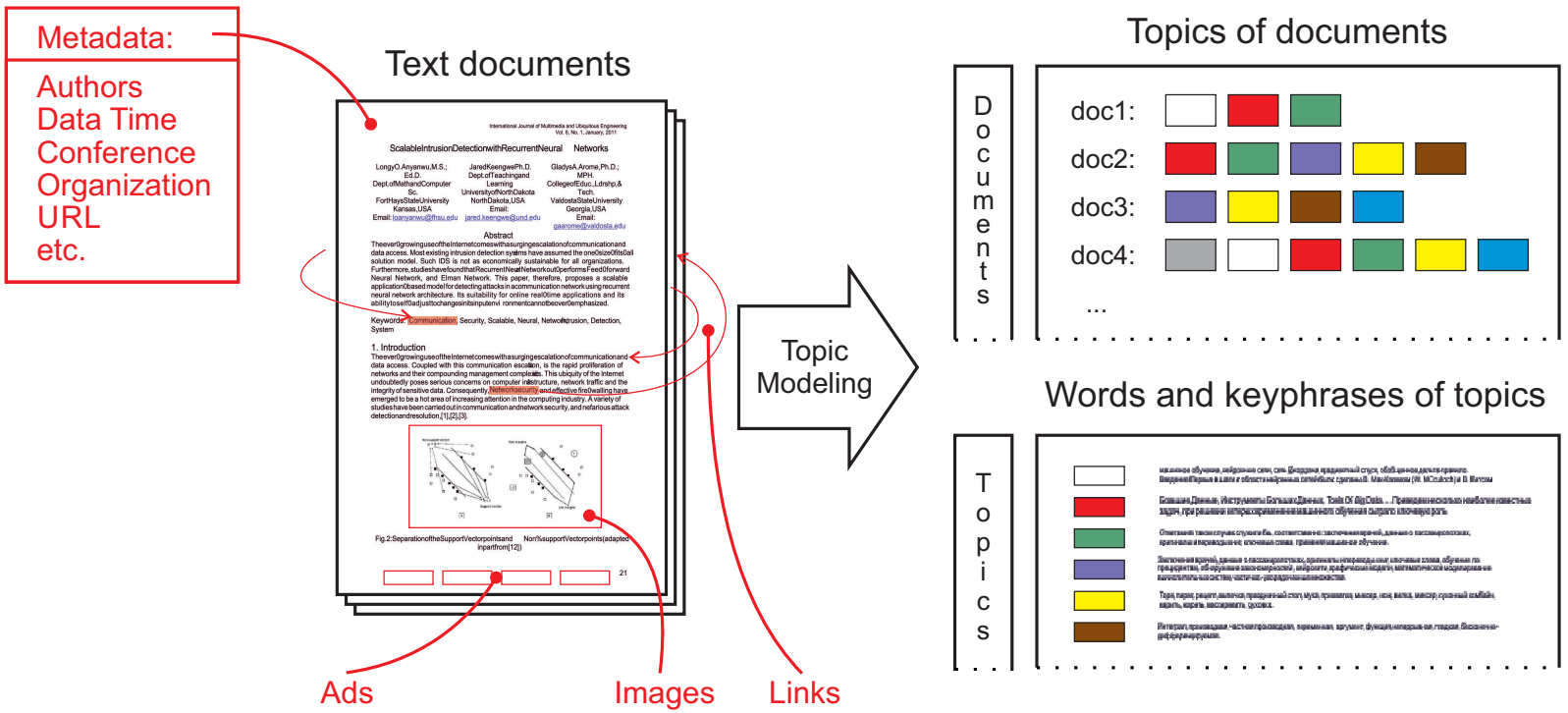
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$, авторов $p(t|a)$, времени $p(t|a)$, элементов изображений $p(t|e)$, ссылок $p(d'|r)$,...



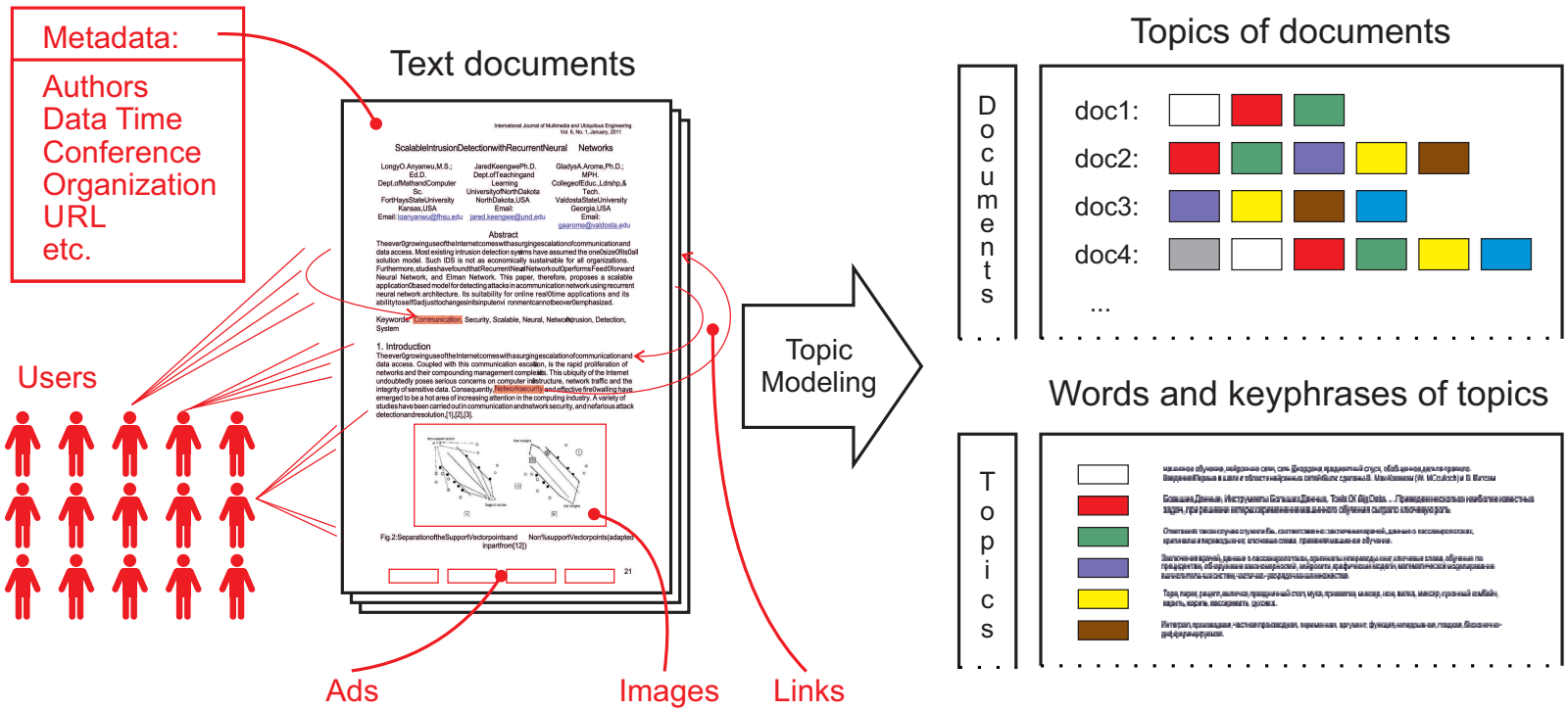
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$, авторов $p(t|a)$, времени $p(t|a)$, элементов изображений $p(t|e)$, ссылок $p(d'|r)$, баннеров $p(t|b)$,...



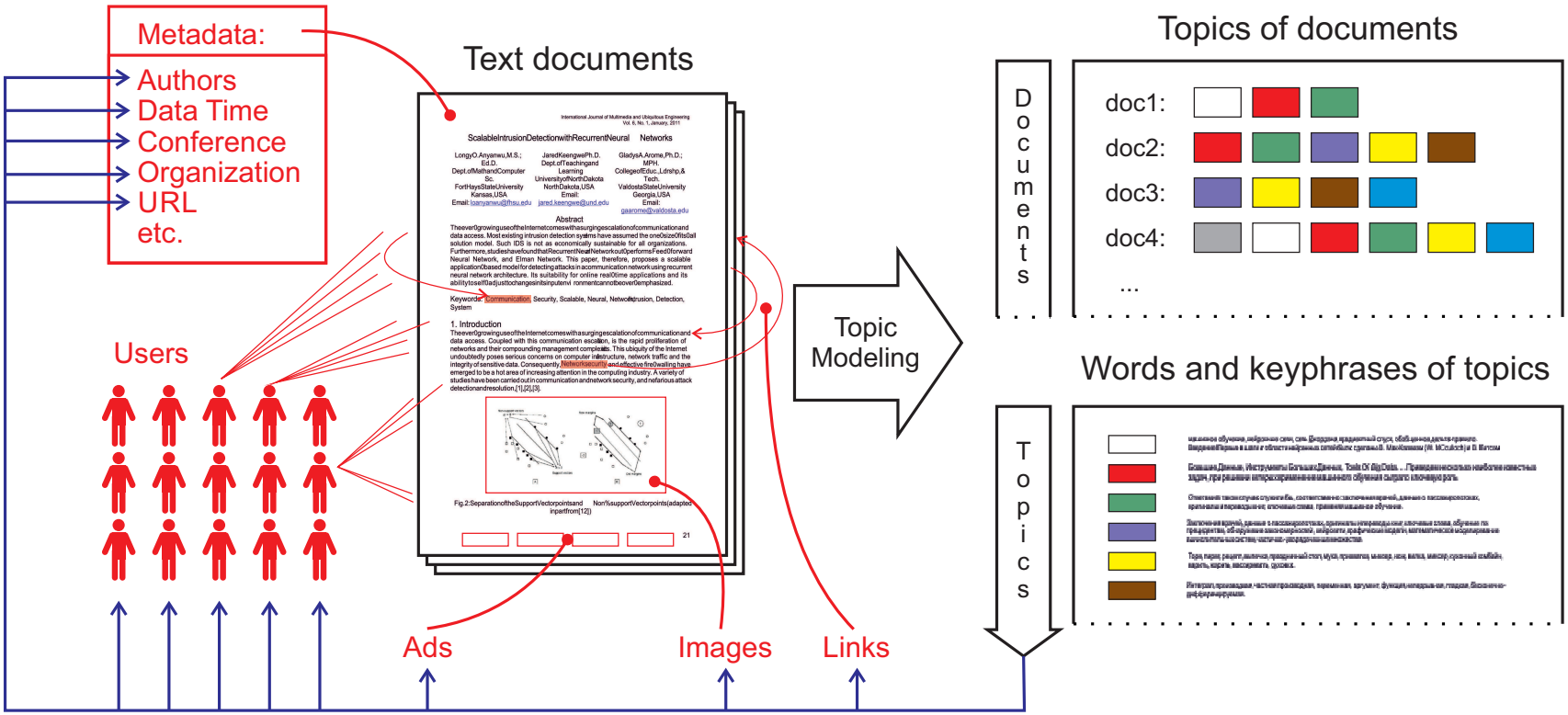
Мультимодальная тематическая модель

находит тематику документов $p(t|d)$, терминов $p(t|w)$, авторов $p(t|a)$, времени $p(t|a)$, элементов изображений $p(t|e)$, ссылок $p(d'|r)$, баннеров $p(t|b)$, пользователей $p(t|u)$,...



Мультимодальная тематическая модель

Каждая модальность $m \in M$ описывается своим словарём W^m, документы могут содержать элементы разных модальностей, каждая тема имеет своё распределение $p(w|t)$, $w \in W^m$



MultiARTM — мультимодальная ARTM

Каждая модальность $m \in M$ описывается своим словарём W^m, документы могут содержать элементы разных модальностей, каждая тема имеет своё распределение $p(w|t)$, $w \in W^m$

Задача максимизации регуляризованного правдоподобия:

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

при ограничениях неотрицательности и нормировки

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

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

EM-алгоритм для мультимодальной ARTM

Теорема

Решение данной задачи удовлетворяет системе уравнений со вспомогательными переменными $p_{tdw} = p(t|d,w)$, n_{wt}, n_{td}:

$$p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right);$$

$$\phi_{wt} = \underset{w \in W^m}{\text{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right); \quad n_{wt} = \sum_{d \in D} \tau_{m(w)} n_{dw} p_{tdw};$$

$$\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \quad n_{td} = \sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw};$$

где $m(w)$ — модальность термина w, т.е. $w \in W^{m(w)}$.

EM-алгоритм = метод простых итераций для системы уравнений

ARTM — альтернатива байесовскому подходу

ARTM унифицирует разработку моделей с заданными свойствами

Этапы моделирования	Bayesian TM	ARTM	
	Анализ требований	Анализ требований	
Формализация:	Вероятностная порождающая модель данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Общий регуляризованный EM-алгоритм для любых моделей	
Реализация:	Исследовательский код (Matlab, Python, R)	Промышленный код BigARTM (C++, Python API)	
Оценивание:	Исследовательские метрики, исследовательский код	Стандартные метрики	Свои метрики
	Внедрение	Внедрение	

-- нестандартизуемые этапы, уникальная разработка для каждой задачи

-- стандартизуемые этапы

ARTM: зоопарк регуляризаторов

- сглаживание тем общей лексики (LDA)
- разреживание предметных тем
- декоррелирование предметных тем
- энтропийное разреживание для отбора тем
- максимизация согласованности (когерентности)
- обучение с учителем для классификации и регрессии
- частичное (semi-supervised) обучение
- динамическое (темпоральное) моделирование
- многоязычное тематическое моделирование
- и др.

Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning. Special Issue "Data Analysis and Intelligent Optimization with Applications". Springer, 2014.

Справочные сведения. Дивергенция Кульбака–Лейблера

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

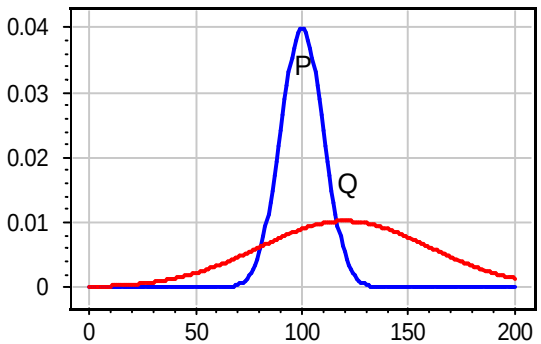
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

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

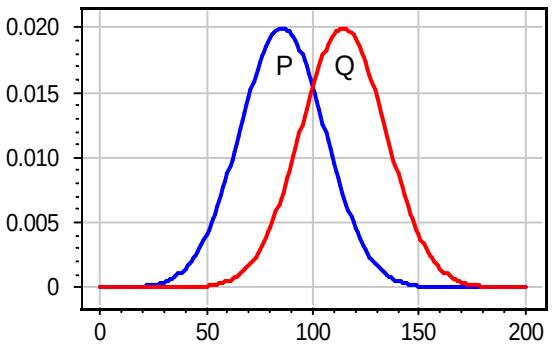
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

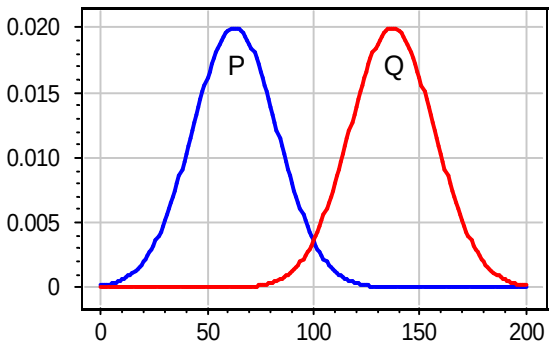
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



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



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



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

Регуляризатор сглаживания (переосмысление LDA)

Гипотеза сглаженности фоновых тем $t \in B \subset T$:
распределения ϕ_{wt} близки к β_w, распределения θ_{td} близки к 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].

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003. — No. 3. — Pp. 993–1022.

Регуляризатор для разреживания предметных тем

Гипотеза разреженности предметных тем $t \in S \subset T$:
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 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.

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:

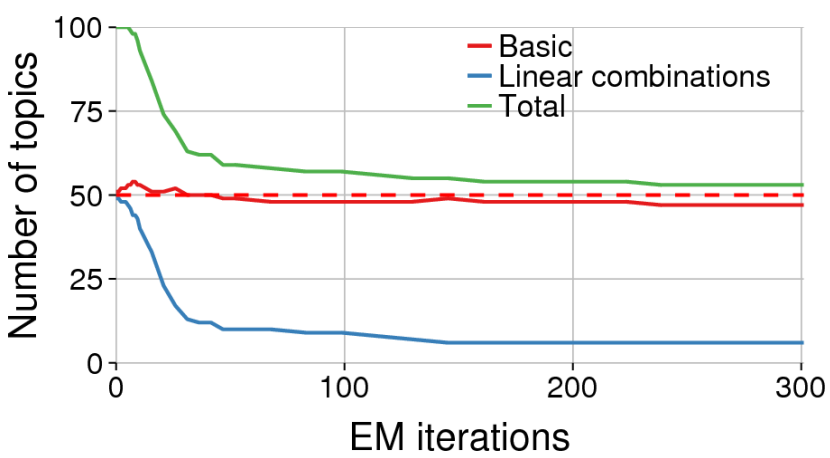
$$\theta_{td} \propto \left(n_{td} - \tau \frac{n_d}{n_t} \theta_{td} \right)_+.$$

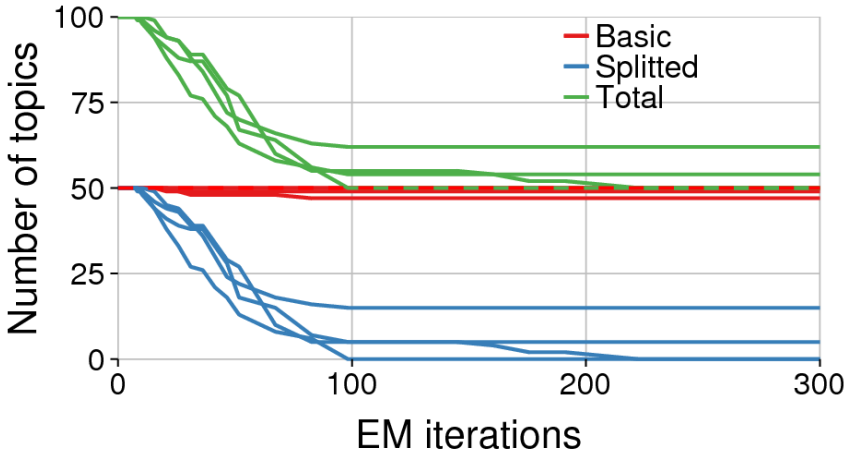
Эффект: строки матрицы Θ могут целиком обнуляться для тем t, собравших мало слов по коллекции, $n_t = \sum_d \sum_w n_{dwt}$.

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // SLDS 2015.

Удаление линейно зависимых и расщеплённых тем

- Синтетическая коллекция NIPS-PLSA, $|T| = 50$.
- Добавили 50 линейных комбинаций тем в модельную Ф.
- Расщепили 50 тем, каждую на две подтемы в модельной Ф.





- Удаляются линейно зависимые и расщеплённые темы
- Остаются более различные темы исходной модели.

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновый параллельный MultiARTM
- Большие данные: коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов

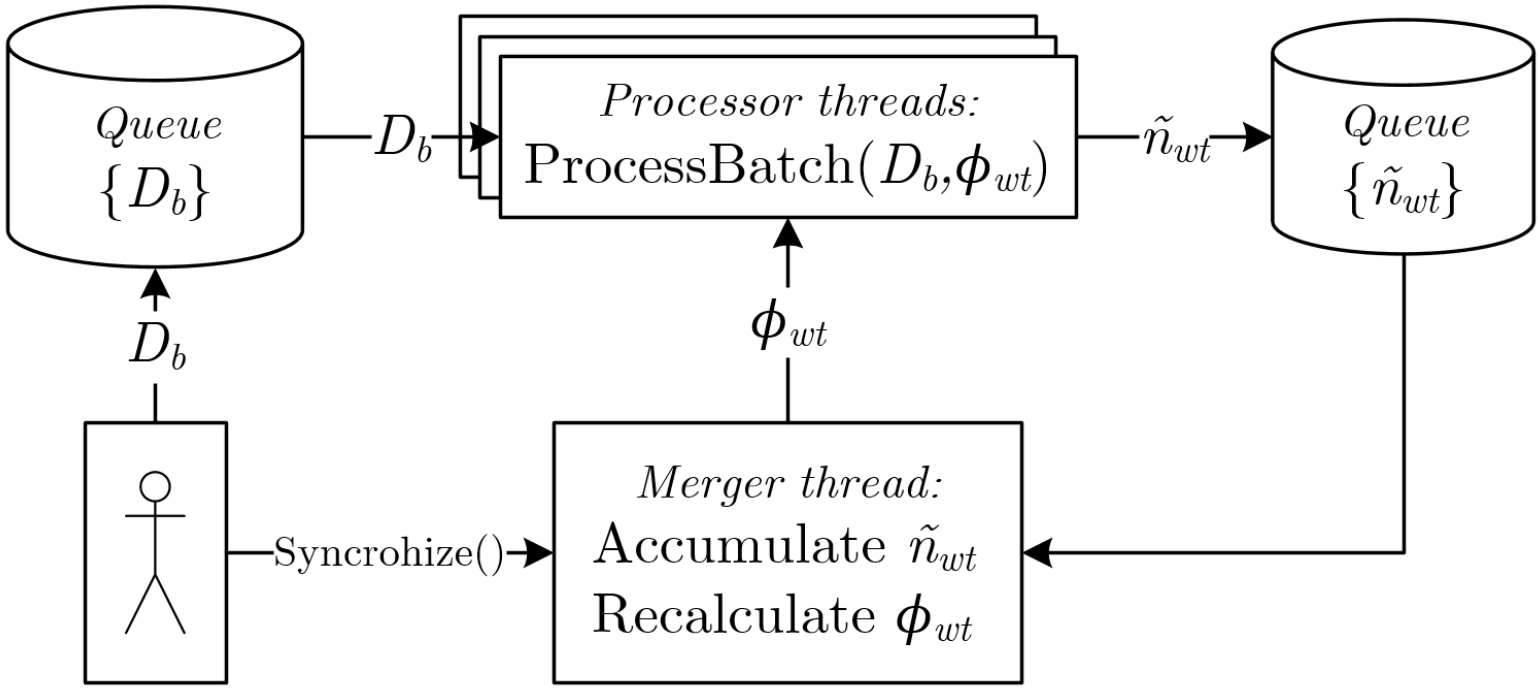
Сообщество:

- Открытый код https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Документация http://bigartm.org

Лицензия и среда разработки:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

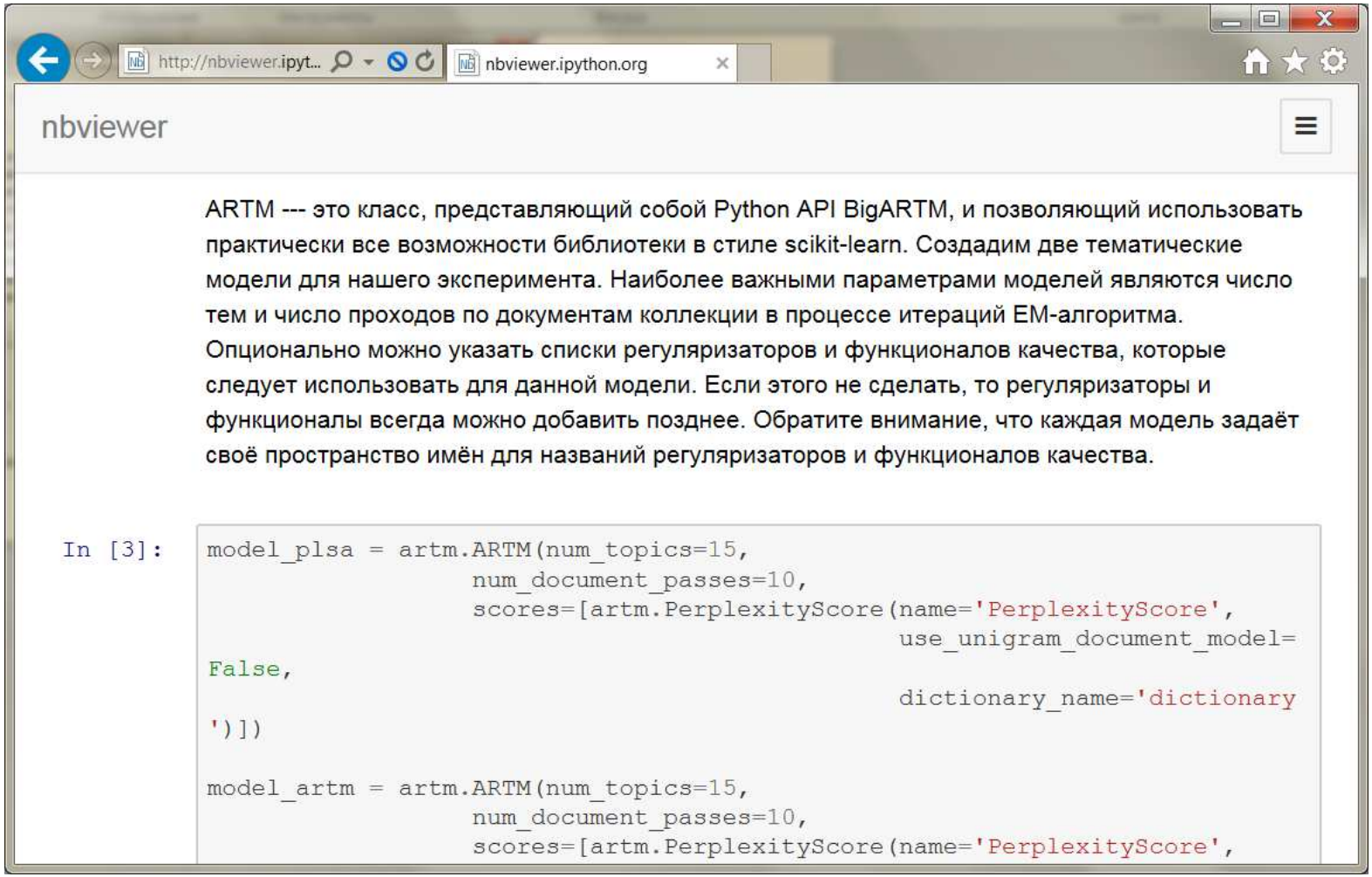
Параллельная архитектура



- коллекция разбивается на пакеты $D = D_1 \sqcup \cdots \sqcup D_B$
- простой однопоточный *ProcessBatch*
- пользователь определяет моменты обновлений модели
- гарантируется воспроизводимость от запуска к запуску

Разработка тематических моделей в среде IPython Notebook

http://nbviewer.ipython.org/github/bigartm/bigartm-book/blob/master/BigARTM_example_RU.ipynb



Эксперимент 1. Обгоняем конкурентов по скорости

- 3.7M статей английской Вики, 100K уникальных слов

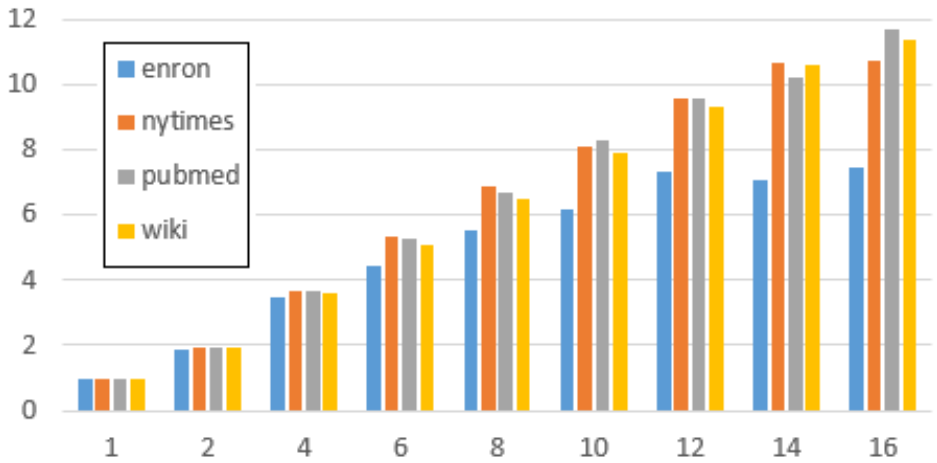
	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

- *procs* = число параллельных потоков
- *inference* = время тематизации 100K тестовых документов
- *perplexity* вычислена на тестовой выборке документов

Эксперимент 1. Масштабируемость по числу потоков

коллекция	$\|W\|$, 10^3	$\|D\|$, 10^6	n, 10^6	размер, Гб
enron	28	0.04	6.4	0.07
nytimes	103	0.3	100	0.13
pubmed	141	8.2	738	1.0
wiki	100	3.7	1009	1.2

ускорение

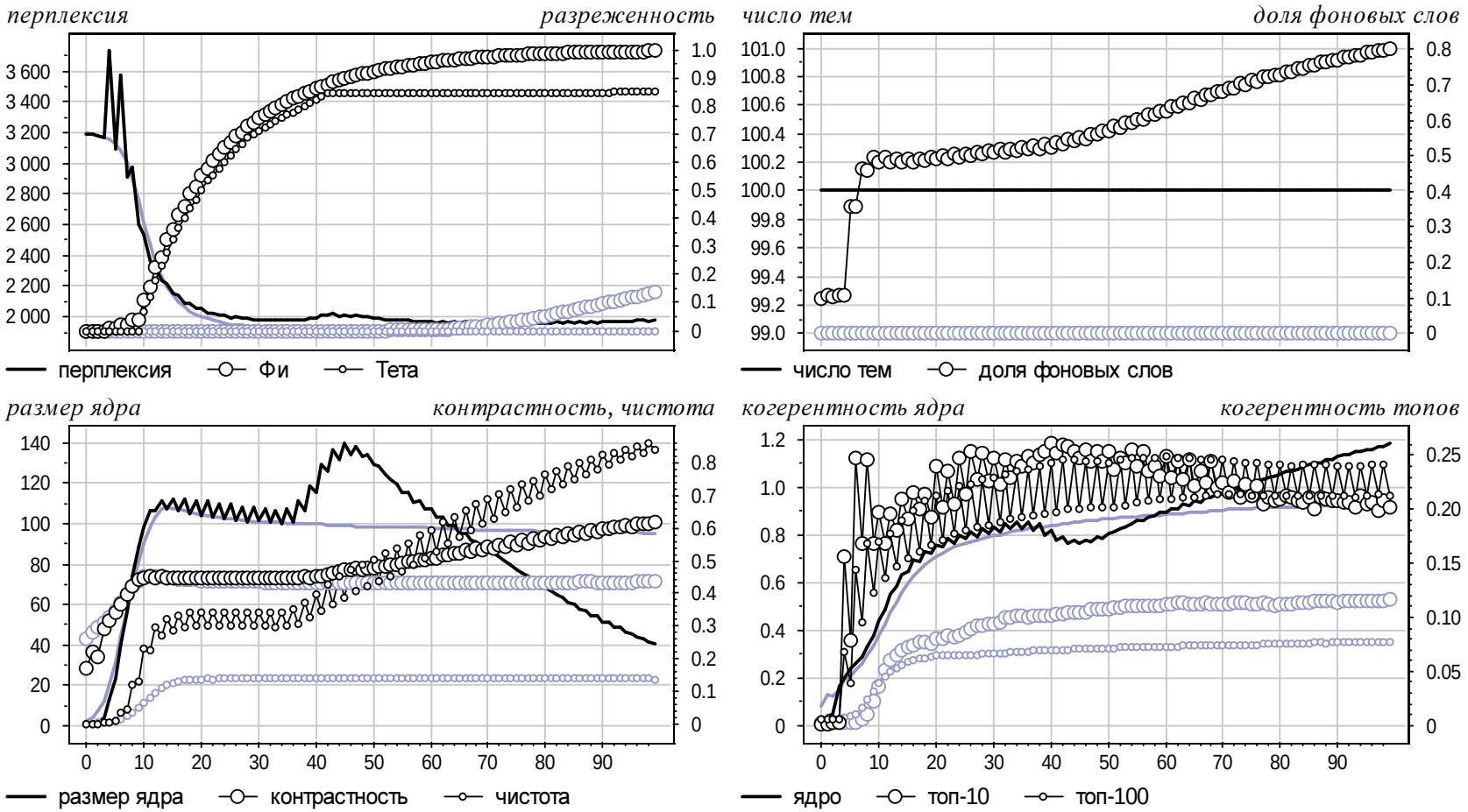


процессоров

Amazon EC2 cc2.8xlarge instance:
16 cores + hyperthreading, Intel® Xeon® CPU E5-2670 2.6GHz.

Эксперимент 2. Комбинирование регуляризаторов

Сравнение PLSA (серый) и ARTM со сглаживанием, разреживанием и декоррелированием (чёрный)



Эксперимент 3. Мультиязычная модель

Модальности — это разные языки.

216 175 русско-английских пар статей Вики.
Первые 10 слов и их вероятностями $p(w|t)$ в %:

Topic 68				Topic 79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Эксперимент 3. Мультиязычная модель

216 175 русско-английских пар статей Вики.
Первые 10 слов и их вероятностями $p(w|t)$ в %:

Topic 88				Topic 251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Независимый асессор оценил 396 тем из $|T| = 400$ как хорошо интерпретируемые.

Эксперимент 4. Интерпретируемость мультиграммной модели

Две модальности — униграммы и биграммы.

Коллекция 1000 статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

Эксперимент 5. 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:

$$R_1(\Theta) = -\tau_1 \sum_{y \in Y} \mathrm{KL}\Big(\tfrac{1}{|T|} \| p(t|y)\Big) \to \max.$$

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

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

Эксперимент 5. Задача анализа потока пресс-релизов

Коллекция официальных пресс-релизов внешнеполитических ведомств ряда стран на английском языке.
Более 20 тыс. сообщений за 10 лет, 180Мб текста.

Найти:

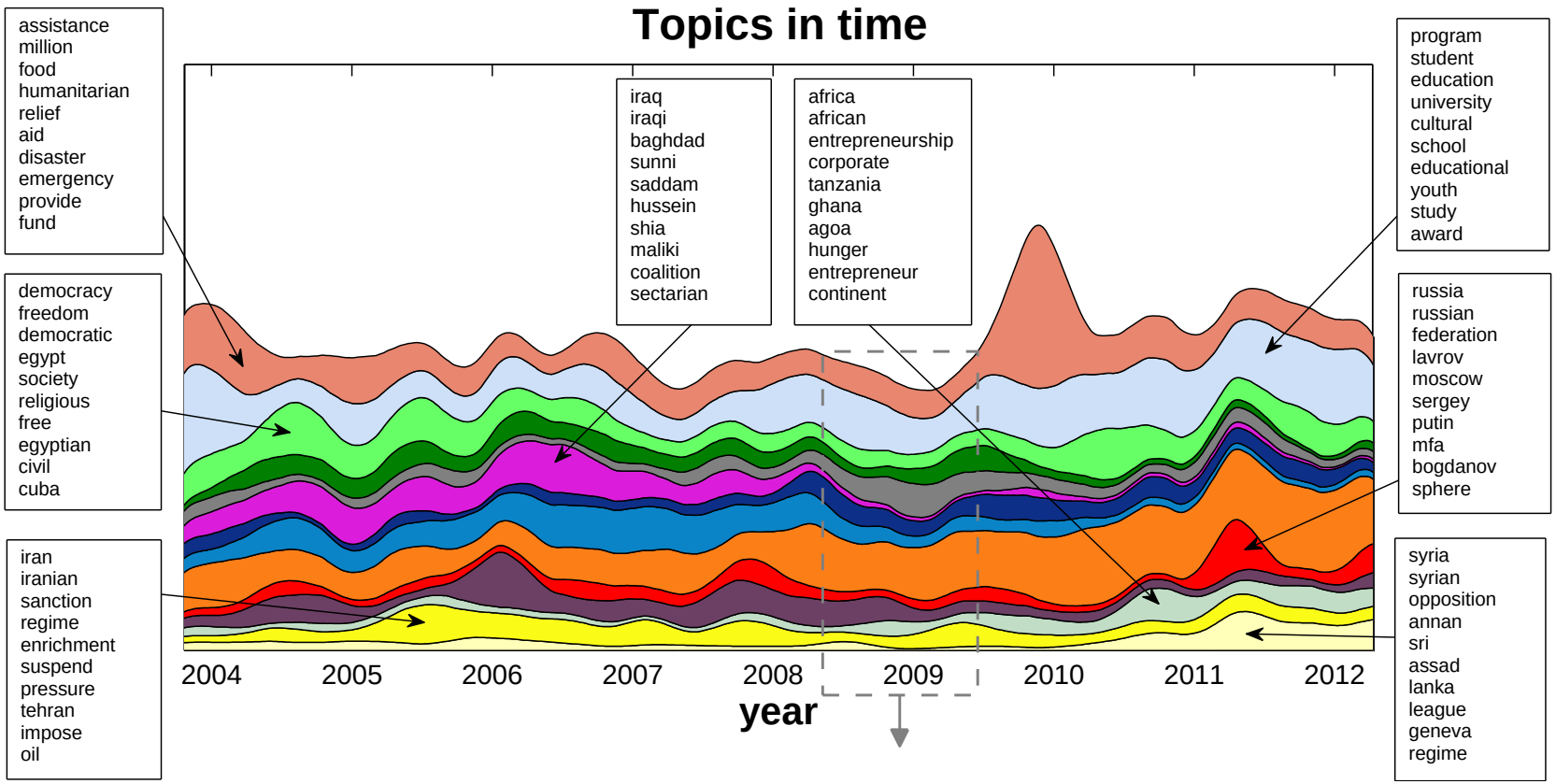
- какие темы перманентные?
- какие темы привязаны к событиям?
- какие темы и в какие моменты коррелируют?

Регуляризаторы:

- разреживание, сглаживание, декоррелирование
- разреживание тем $p(t|y)$ в каждый момент времени y
- сглаживание тем $p(y|t)$ в соседние моменты времени

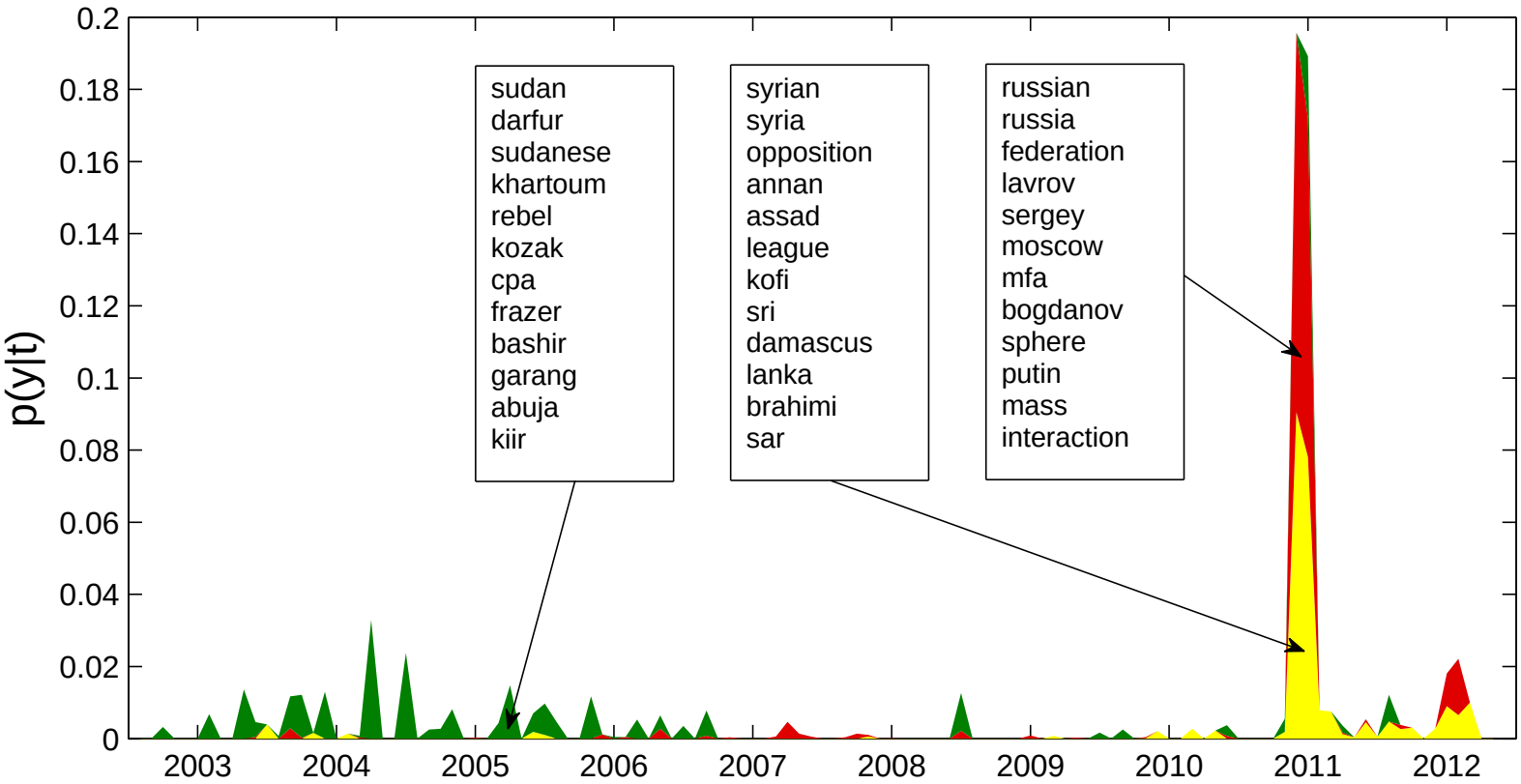
Эксперимент 5. Задача анализа потока пресс-релизов

Примеры хорошо интерпретируемых тем



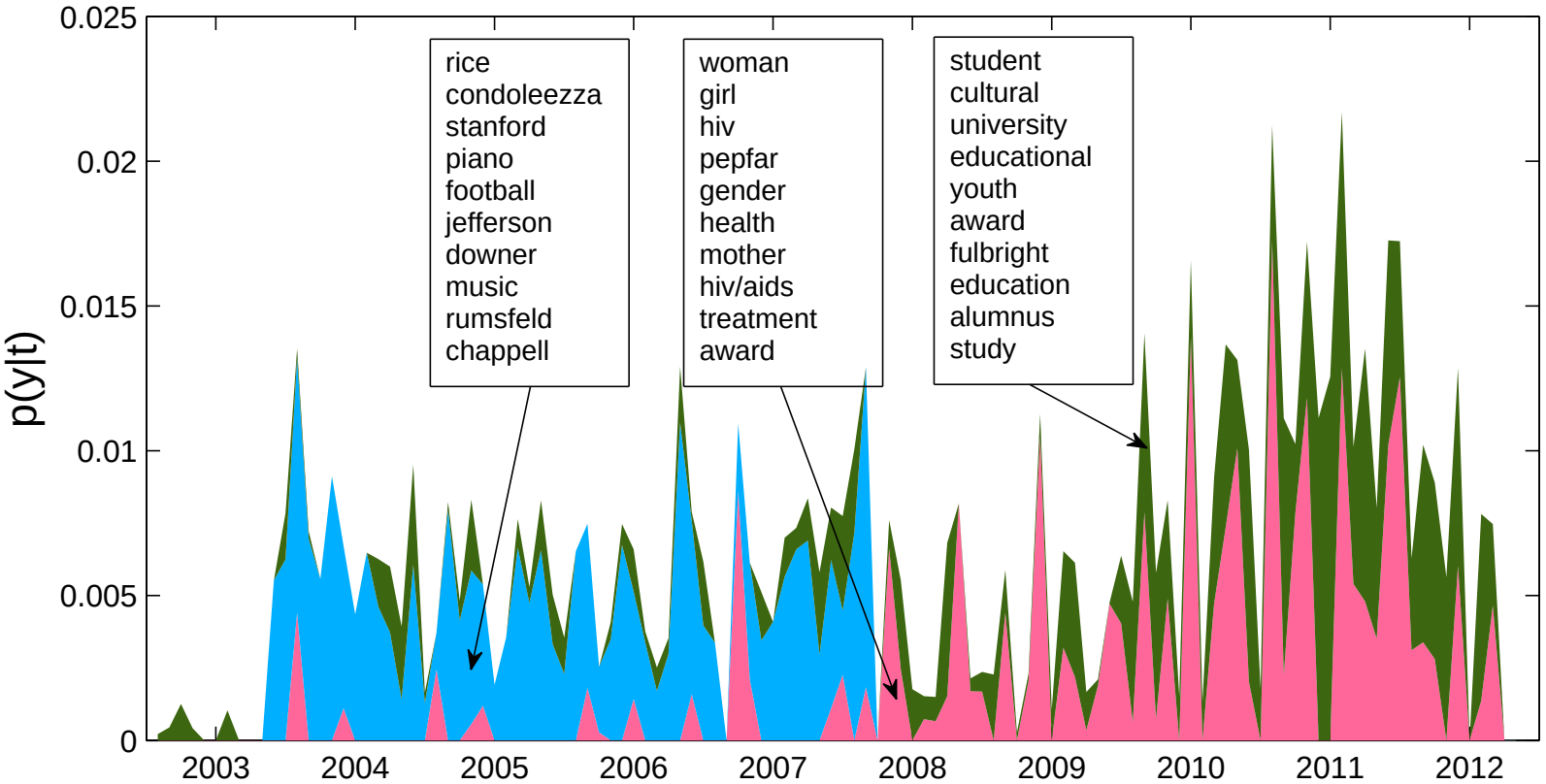
Эксперимент 5. Задача анализа потока пресс-релизов

Примеры событийных тем и момента их совместного всплеска



Эксперимент 5. Задача анализа потока пресс-релизов

Примеры перманентных тем



Направления дальнейших исследований

- большое число тем
- мелкозернистая иерархия
- лингвистическая регуляризация
- гиперграфовые обобщения
- автоматическое именование тем
- визуализация
- тематический разведочный поиск

http://bigartm.org

Join BigARTM community!