Вероятностные тематические модели Лекция 3. Аддитивная регуляризация тематических моделей (ARTM)

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

ВМК МГУ • 1 марта 2018

Содержание

Задача тематического моделирования

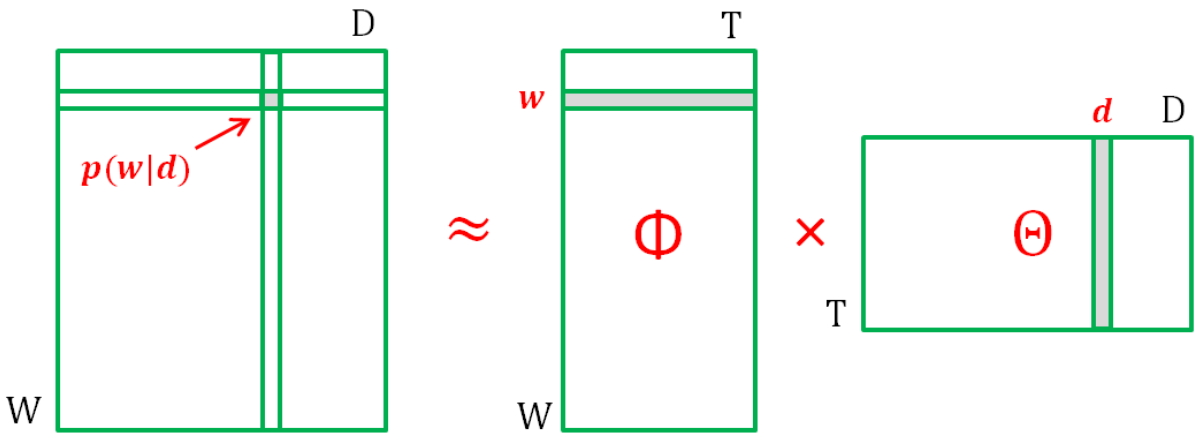
Дано: коллекция текстовых документов, $p(w|d) = \frac{n_{dw}}{n_d}$

Вероятностная тематическая модель:

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

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

Это задача стохастического матричного разложения:



PLSA, Probabilistic Latent Semantic Analysis (1999)

Задача: найти максимум правдоподобия

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

при ограничениях неотрицательности и нормировки

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

E-шаг:
$$n_{dwt} = n_{dw} \frac{\phi_{wt}\theta_{td}}{\sum_s \phi_{ws}\theta_{sd}};$$

M-шаг:
$$\phi_{wt} = \frac{n_{wt}}{n_t}; \quad n_{wt} = \sum_{d\in D} n_{dwt}; \quad n_t = \sum_w n_{wt}$$

$$\theta_{td} = \frac{n_{td}}{n_d}; \quad n_{td} = \sum_{w\in d} n_{dwt}; \quad n_d = \sum_t n_{td}$$

ARTM, Additive Regularization for Topic Modeling (2014)

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

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

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

Комбинирование регуляризаторов в ARTM

Максимизация log правдоподобия с k регуляризаторами R_i:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} + \sum_{i=1}^{k} \tau_i R_i(\Phi,\Theta) \;\to\; \max_{\Phi,\Theta},$$

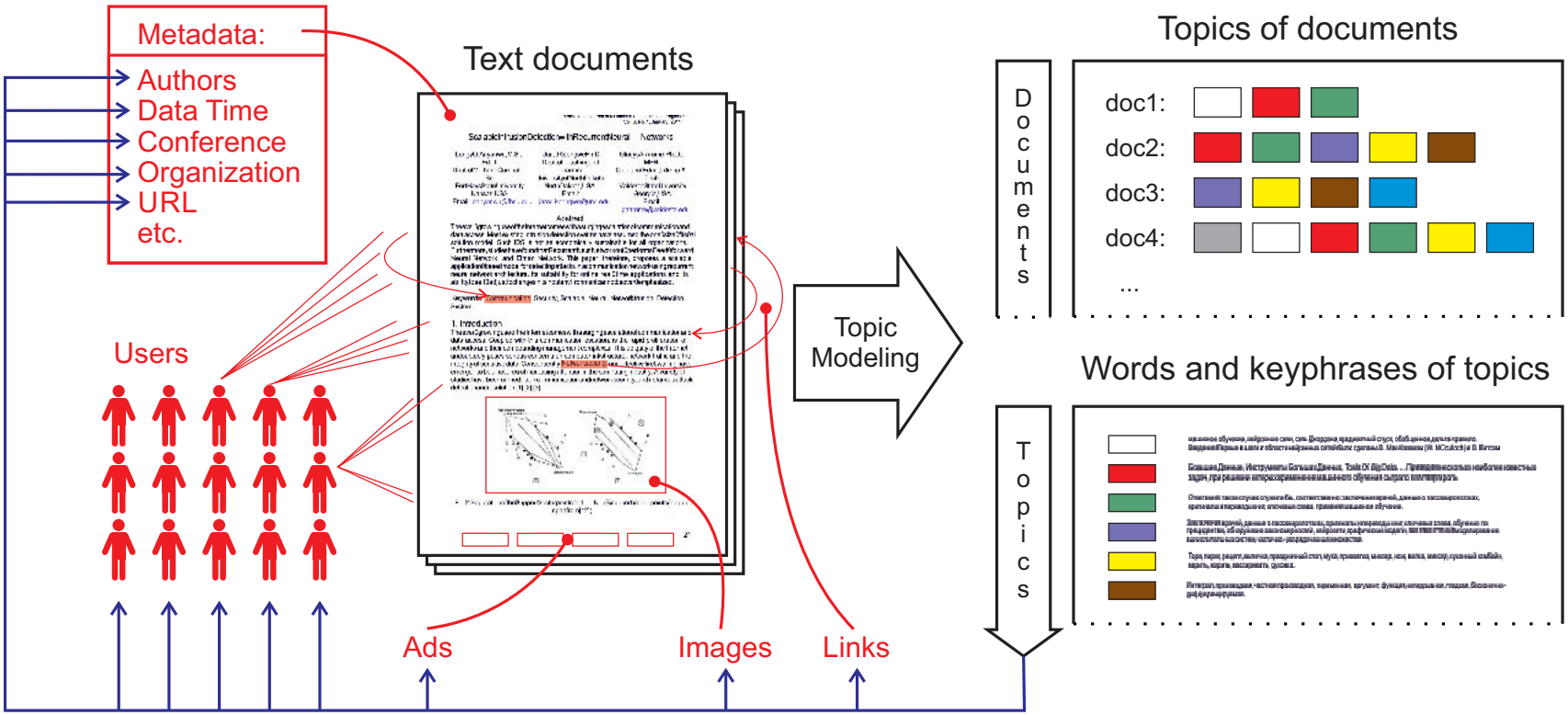
где τ_i — коэффициенты регуляризации.

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \text{M-шаг:} & \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum\limits_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\sum\limits_{i=1}^{k}\tau_i\frac{\partial R_i}{\partial \phi_{wt}}\right) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum\limits_{w\in d} n_{dw}p_{tdw} + \theta_{td}\sum\limits_{i=1}^{k}\tau_i\frac{\partial R_i}{\partial \theta_{td}}\right) \end{cases}$$

Мультимодальная тематическая модель

Документ — универсальный контейнер не только терминов, но и токенов других *модальностей*: $p(t|\text{автор})$, $p(t|\text{время})$, $p(t|\text{ссылка})$, $p(t|\text{баннер})$, $p(t|\text{изображение})$, $p(t|\text{пользователь})$



Мультимодальная ARTM

W^m — словарь токенов m-й модальности, $m \in M$

$W = W^1 \sqcup \dots \sqcup W^M$ — объединённый словарь всех модальностей

Максимизация суммы log правдоподобий с регуляризацией:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

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

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

Напоминания. Дивергенция Кульбака–Лейблера

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

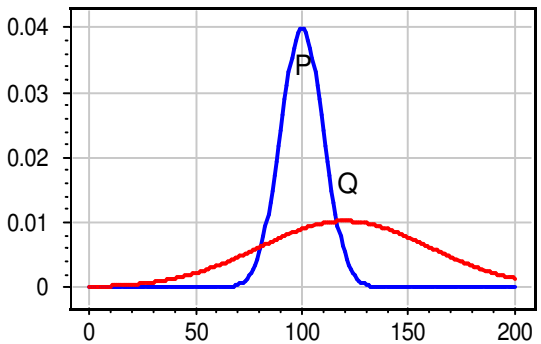
$$\mathrm{KL}(P\|Q) \equiv \mathrm{KL}_i(p_i\|q_i) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

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

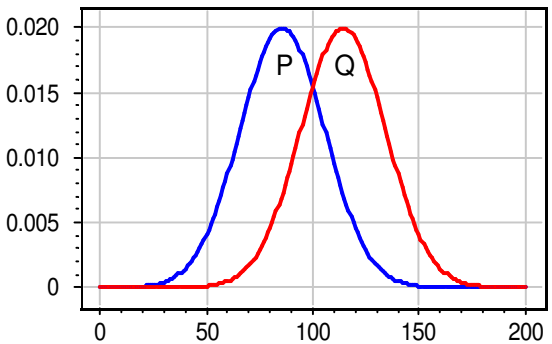
2. Минимизация KL эквивалентна максимизации правдоподобия:

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

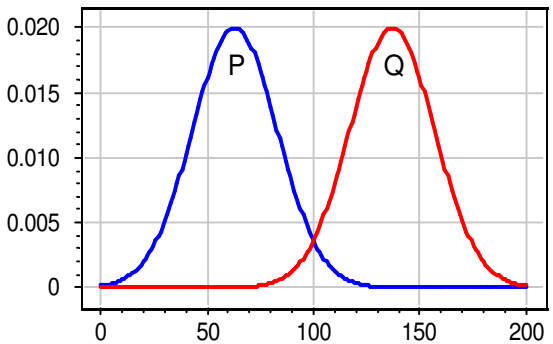
3. Если $\mathrm{KL}(P\|Q) < \mathrm{KL}(Q\|P)$, то P сильнее вложено в Q, чем Q в P:



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



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



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

Регуляризатор сглаживания (LDA)

Гипотеза сглаженности:
распределения ϕ_{wt} близки к заданному распределению β_w;
распределения θ_{td} близки к заданному распределению α_t.

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

Максимизируем сумму 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:

$$\phi_{wt} = \underset{w\in W}{\mathrm{norm}}(n_{wt} + \beta_0\beta_w), \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}(n_{td} + \alpha_0\alpha_t).$$

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 разреживания (обобщение LDA)

Гипотеза разреженности: среди ϕ_{wt}, θ_{td} много нулей;
распределения ϕ_{wt} далеки от заданного распределения β_w;
распределения θ_{td} далеки от заданного распределения α_t.

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

Максимизируем сумму 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 «анти-LDA»:

$$\phi_{wt} = \underset{w\in W}{\mathrm{norm}}\big(n_{wt} - \beta_0\beta_w\big), \qquad \theta_{td} = \underset{t\in T}{\mathrm{norm}}\big(n_{td} - \alpha_0\alpha_t\big).$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010.

Объединение сглаживания и разреживания

Общий вид регуляризаторов сглаживания и разреживания:

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

где $\beta_0 > 0$, $\alpha_0 > 0$ — коэффициенты регуляризации,
β_{wt}, α_{td} — параметры, задаваемые пользователем:

- $\beta_{wt} > 0$, $\alpha_{td} > 0$ — сглаживание
- $\beta_{wt} < 0$, $\alpha_{td} < 0$ — разреживание

Частичное обучение (semi-supervised learning) темы t:

- $\beta_{wt} = [w \in W_t]$ — *белый список* W_t терминов темы t
- $\alpha_{td} = [d \in D_t]$ — *белый список* D_t документов темы t
- $\beta_{wt} = -[w \in W_t]$ — *чёрный список* W_t терминов темы t
- $\alpha_{td} = -[d \in D_t]$ — *чёрный список* D_t документов темы t

Проблема $\ln 0$ в дивергенции Кульбака–Лейблера

Почему в регуляризаторе сглаживания/разреживания

$$R(\Phi) = \beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln \phi_{wt} \to \max$$

не возникает проблем с $\ln \phi_{wt}$ при $\phi_{wt} \to 0$?

Подправим регуляризатор, при сколь угодно малом ε:

$$R(\Phi) = \beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln(\phi_{wt} + \varepsilon) \to \max$$

Подставив в формулу M-шага, получим для всех $t \in S$:

$$\phi_{wt} \propto \left(n_{wt} + \beta_0 \beta_w \frac{\phi_{wt}}{\phi_{wt} + \varepsilon} \right)_+$$

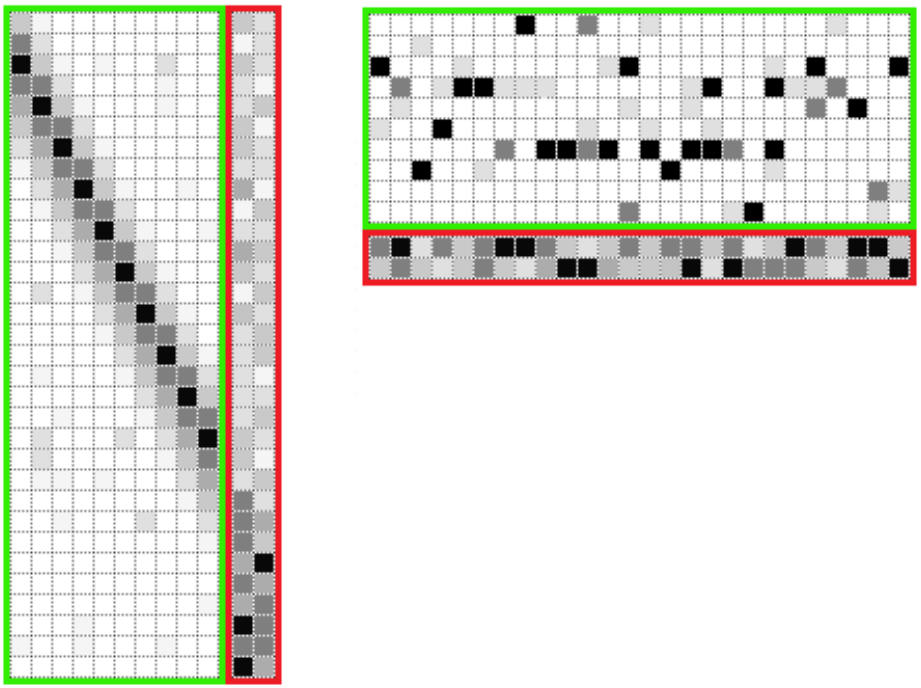
Если $\phi_{wt} = 0$, то разреживания не будет, но оно и не нужно.

Разделение тем на предметные и фоновые

Предметные темы S содержат термины предметной области,
$p(w|t)$, $p(t|d)$, $t \in S$ — разреженные, существенно различные

Фоновые темы B содержат слова общей лексики,
$p(w|t)$, $p(t|d)$, $t \in B$ — существенно отличные от нуля

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



Регуляризатор декоррелирования тем

Цель — выделить *лексическое ядро* каждой темы, набор терминов, отличающий её от других тем.

Минимизируем ковариации между вектор-столбцами ϕ_t:

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

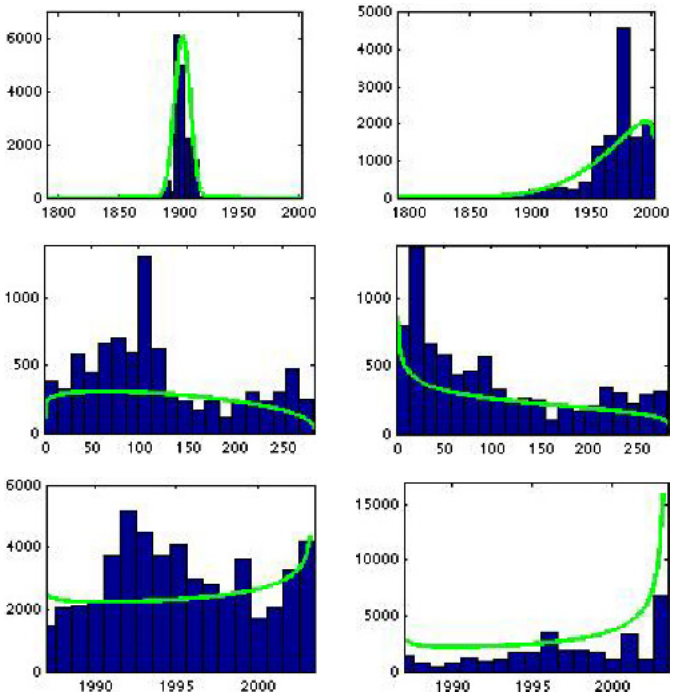
Подставляем, получаем ещё один вариант разреживания — постепенное контрастирование строк матрицы Φ:

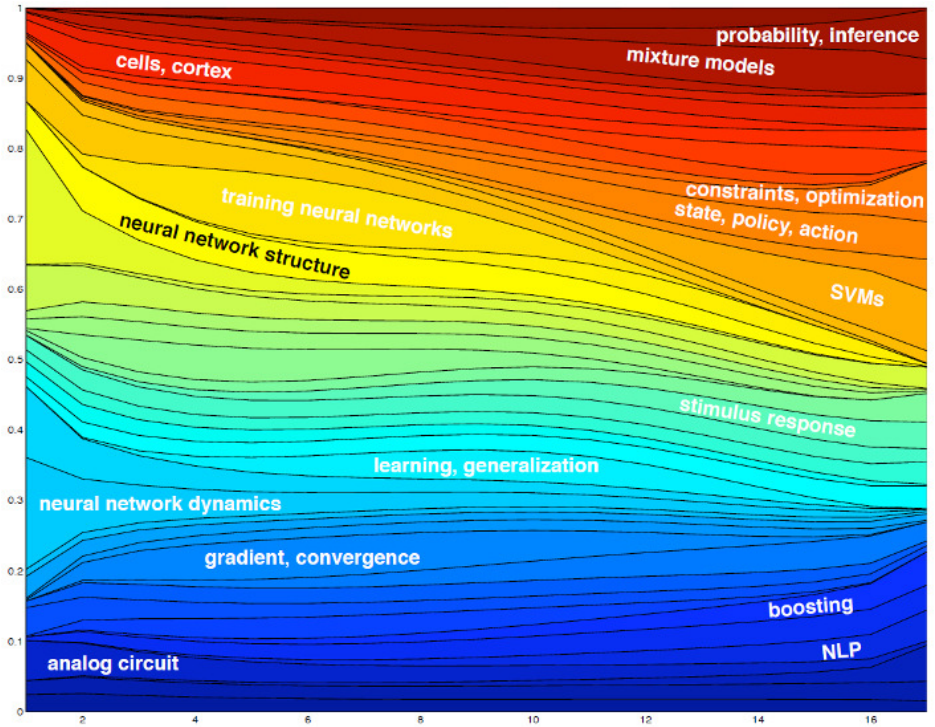
$$\phi_{wt} = \underset{w \in W}{\text{norm}} \left(n_{wt} - \tau \phi_{wt} \sum_{s \in T \setminus t} \phi_{ws} \right).$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation. 2010.

Модель TOT (Topics over Time)

1. Каждая тема имеет непрерывное β-распределение во времени
2. Каждое слово имеет метку времени





Xuerui Wang, *Andrew McCallum*. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends // ACM SIGKDD-2006

Темпоральные тематические модели

Неадекватность TOT очевидна даже по картинкам из статьи!

Наши предположения:

- Время дискретно, $i \in I$ — интервалы времени
- Как и в TOT, темы $p(w|t)$ не меняются во времени
- *Перманентные* темы имеют медленно меняющиеся $p(i|t)$
- *Событийные* темы имеют $p(i|t) = 0$ почти всё время
- Метки времени приписываются документам, а не словам
- Параметрические модели не используются

Цели моделирования:

- Выделить событийные и перманентные темы.
- Проследить развитие тем во времени.
- Выделить тренды (в новостях, в научных публикациях).

Регуляризаторы Θ для темпоральных тематических моделей

I — интервалы времени (например, годы публикаций),
$D_i \subset D$ — все документы, относящиеся к интервалу $i \in I$.
$n_i = \sum\limits_{d \in D_i} n_d$ — доля коллекции, относящаяся к интервалу i.

1. Разреживание $p(t|i) = \sum\limits_{d \in D_i} \theta_{td} \frac{n_d}{n_i}$ в каждом интервале i:

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

2. Сглаживание $p(i|t) = \sum\limits_{d \in D_i} \theta_{td} \frac{n_d}{n_t}$ в соседних интервалах i, $i-1$:

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

Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Время как модальность. Регуляризатор Φ

Проблема регуляризатора Θ в онлайновом EM-алгоритме: соседние по времени документы могут попасть в разные пакеты.

Документы содержат слова $w \in W^1$ и время $i \in W^2 = I$
W^2 — модальность интервалов времени (time stamps)

1. Разреживание $p(t|i)$ эквивалентно разреживанию $p(i|t) = \phi_{it}$:

$$R_1(\Phi^2) = -\tau_1 \sum_{i \in I} \sum_{t \in T} \ln \phi_{it} \to \max$$

2. Сглаживание $p(i|t) = \phi_{it}$ в соседних интервалах i, $i-1$:

$$R_2(\Phi^2) = -\tau_2 \sum_{i \in I} \sum_{t \in T} \left|\phi_{it} - \phi_{i-1,t}\right| \to \max$$

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.

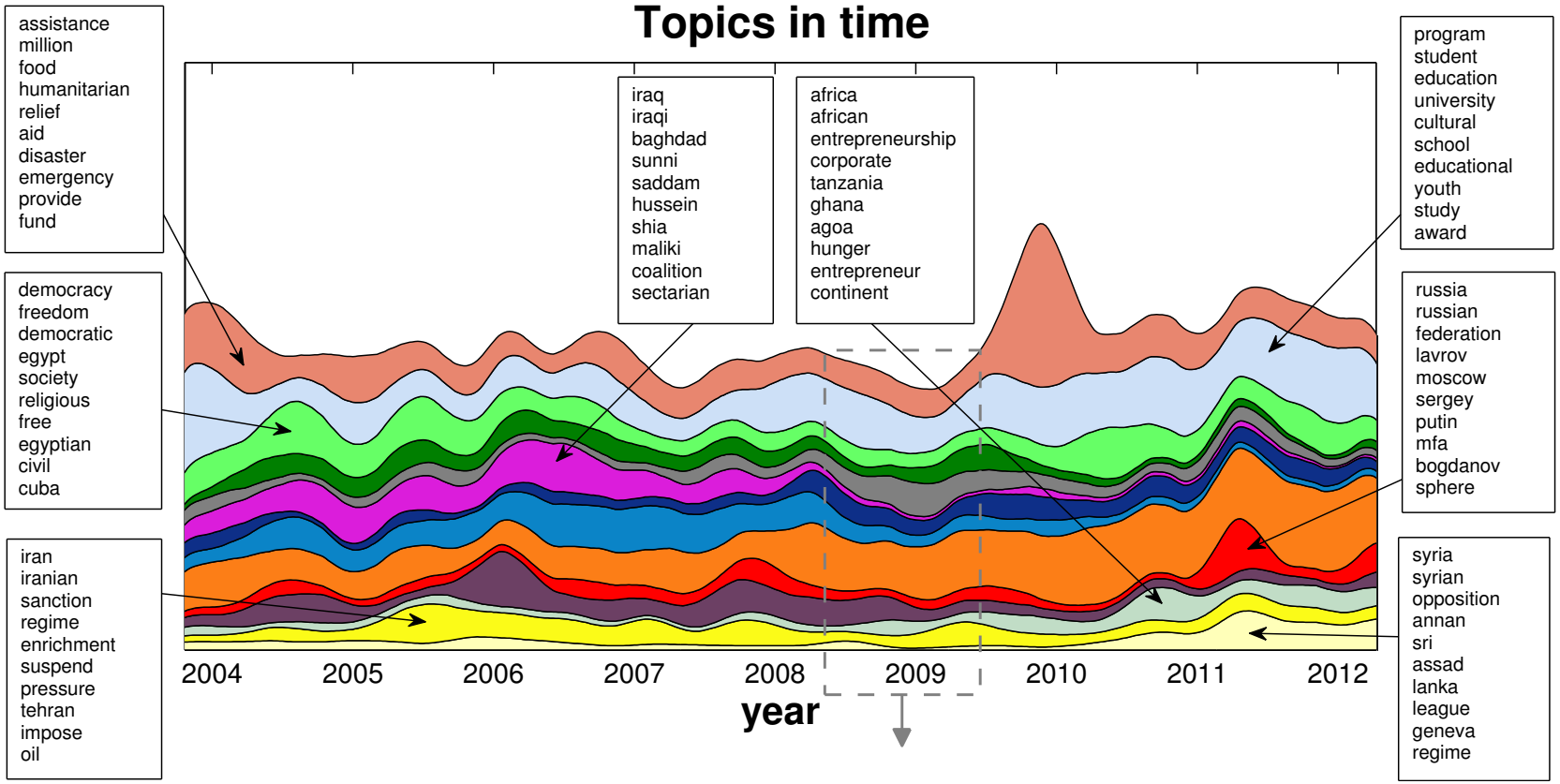
Цели исследования:

- какие темы общие, какие специфичны для источников?
- какие темы событийные, какие перманентные?
- какие темы и когда коррелируют с заданной темой?

Модальности и регуляризаторы:

- две модальности: источники, интервалы времени
- разреживание, сглаживание, декоррелирование
- сглаживание тем во времени

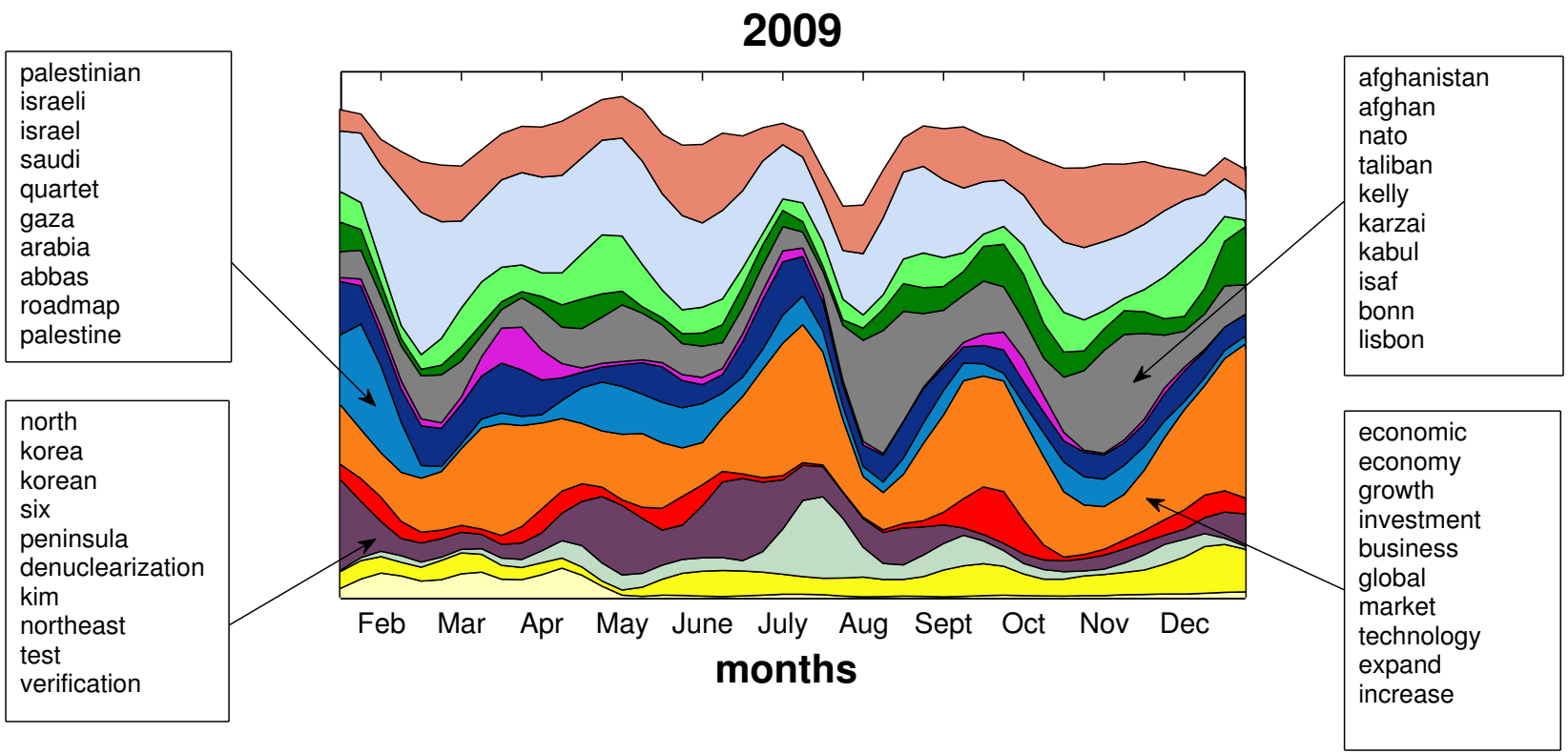
Динамика тем во времени



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Динамика тем во времени

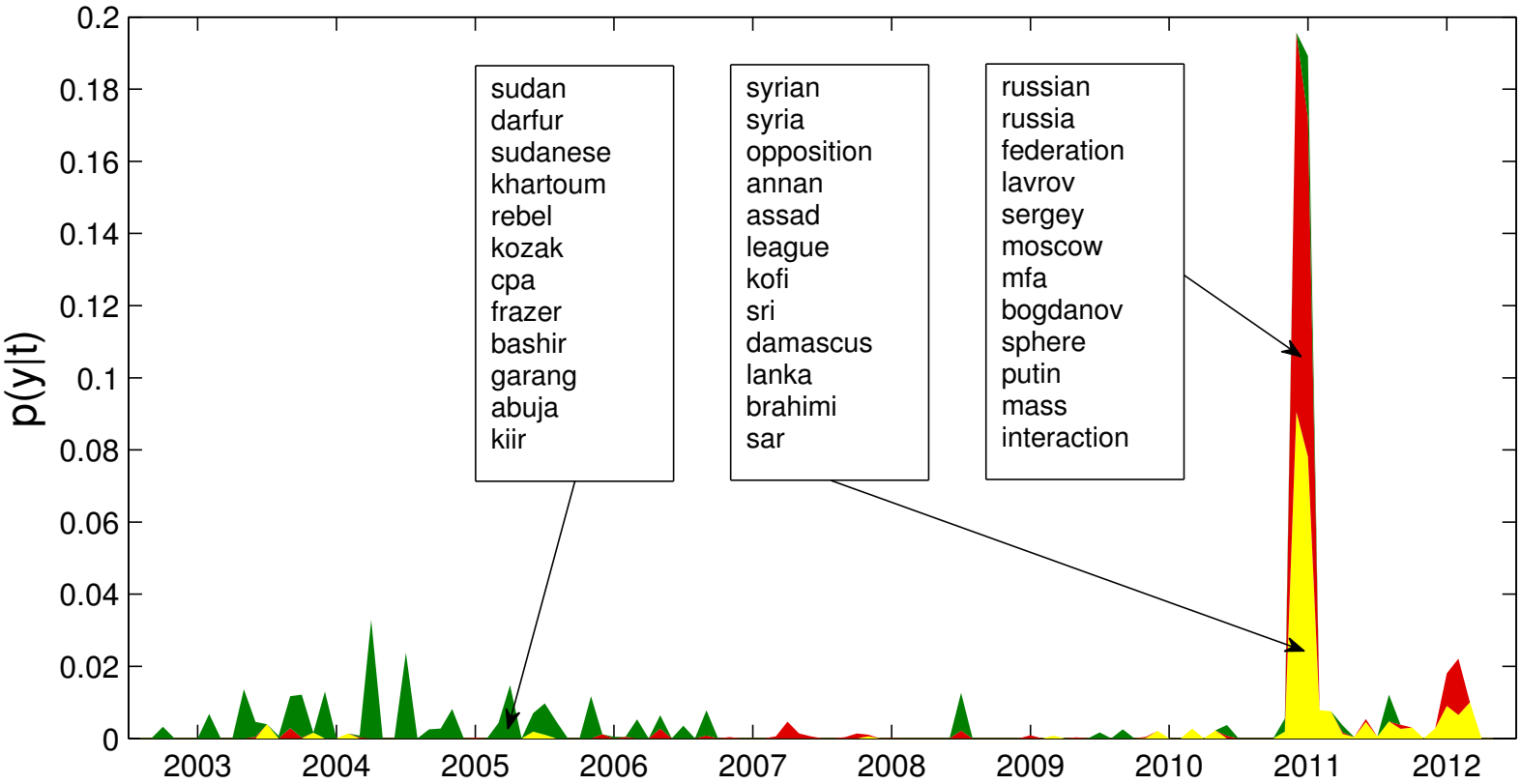
Укрупнение масштаба времени



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Динамика тем во времени

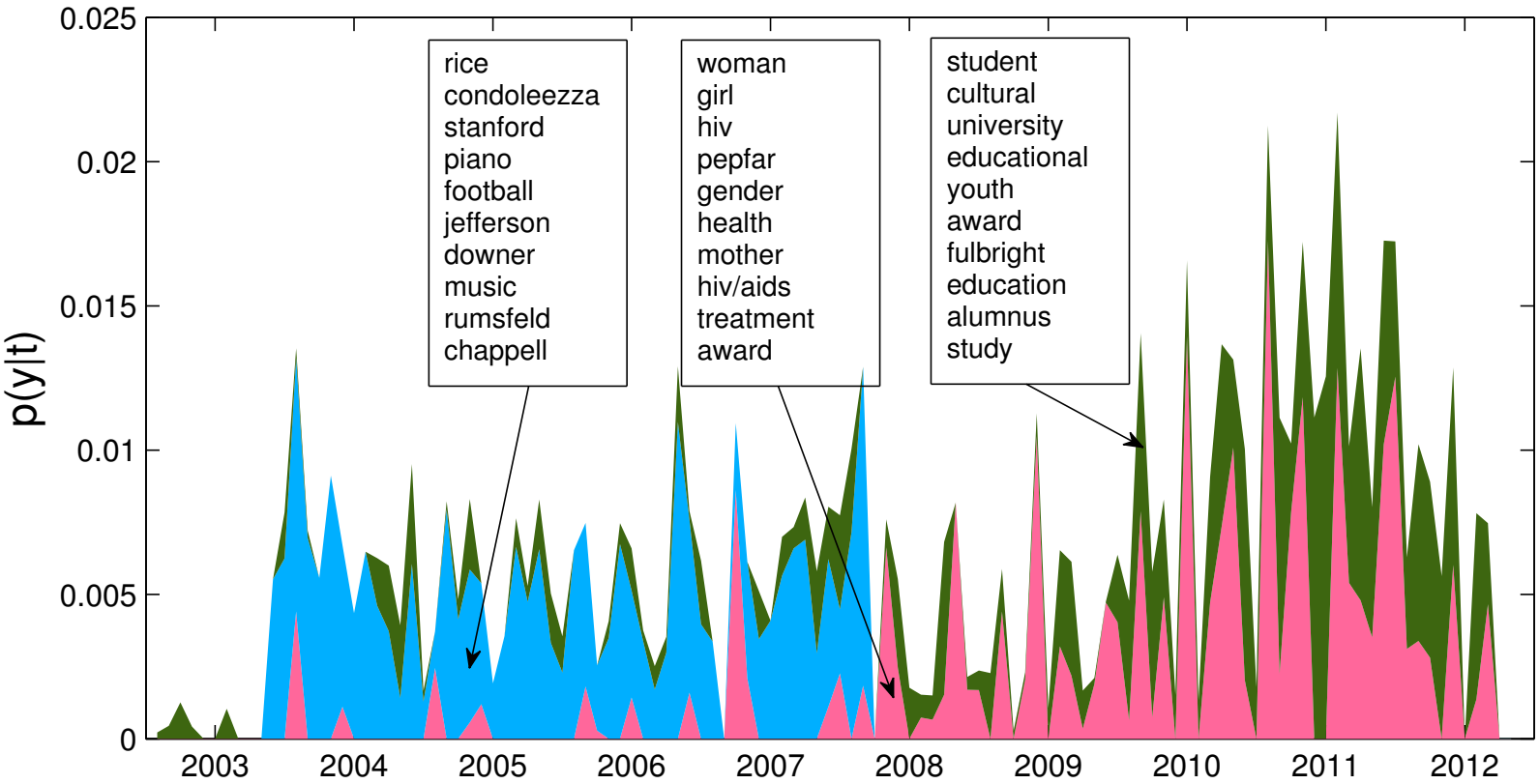
Пример: событийные темы и момент их совместного всплеска



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Динамика тем во времени

Примеры перманентных тем (сглаживание отключено)



Никита Дойков. Адаптивная регуляризация вероятностных тематических моделей // ВКР бакалавра, 2015. ВМК МГУ.

Гео-пространственные модели

Данные: $\ell_d = (x_d, y_d)$ — геолокация (GPS) документа d

Цели исследования:

- какие темы общие, какие специфичны для региона?
- какие есть похожие темы в других регионах?

Регуляризатор:

$$R(\Theta) = -\frac{\tau}{2} \sum_{(c,d)} w_{cd} \sum_{t \in T} (\theta_{td} - \theta_{tc})^2 \to \max,$$

w_{cd} — вес пары (c, d), близость геолокаций (x_c, y_c) и (x_d, y_d)

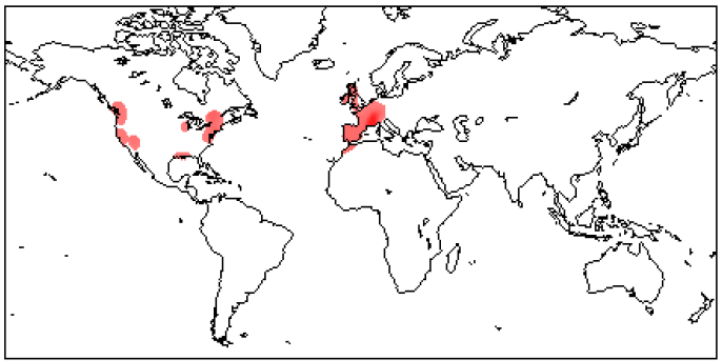
Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, Thomas Huang. Geographical Topic Discovery and Comparison // WWW 2011.

Пример: Food dataset

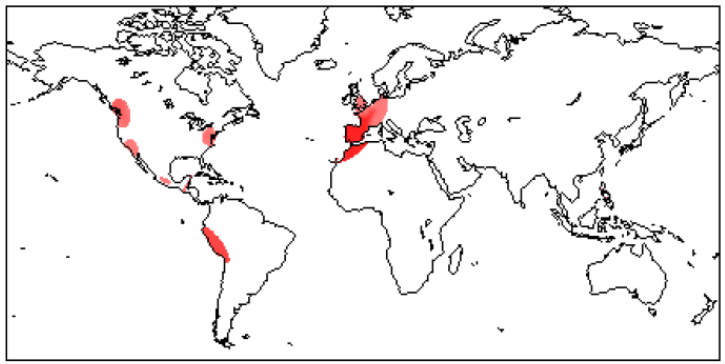
Где и что едят пользователи Flickr?

Chinese Food

Japanese Food



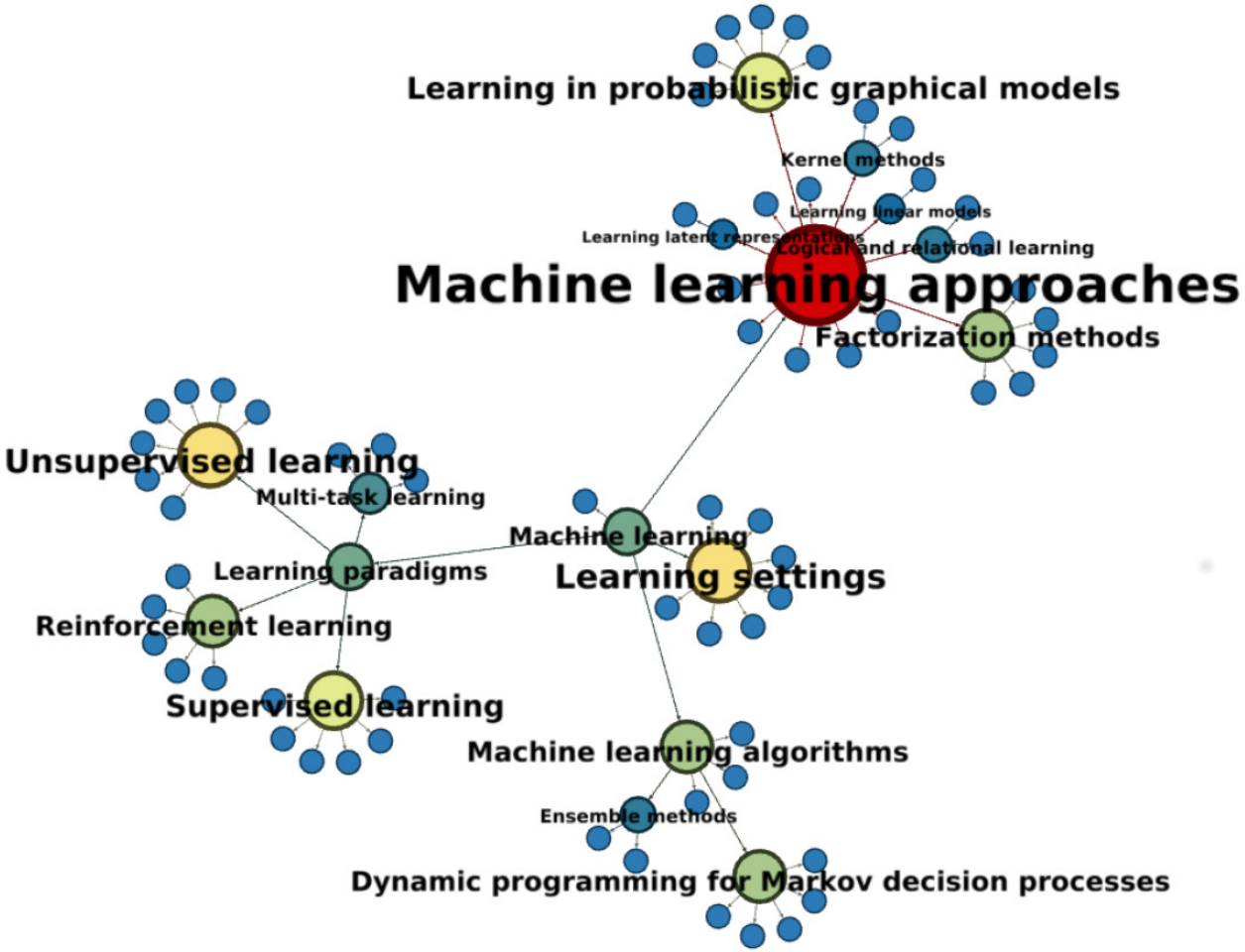
French Food



Spanish Food

Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, Thomas Huang. Geographical Topic Discovery and Comparison // WWW 2011.

Пример тематической иерархии



Georgeta Bordea. Domain adaptive extraction of topical hierarchies for Expertise Mining. 2013.

Иерархические тематические модели

- структура иерархии: дерево / многодольный граф
- направление: снизу вверх / сверху вниз / одновременно
- наращивание: поверхинное / послойное

Открытые проблемы:

- "Despite recent activity in the field of HPTMs, determining the hierarchical model that best fits a given data set, in terms of the structure and size of the learned hierarchy, still remains a challenging task and an open issue."
- "The evaluation of hierarchical PTMs is also an open issue."

Zavitsanos E., Paliouras G., Vouros G. A. Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes. 2011.

Регуляризатор Φ: родительские темы как псевдо-документы

Шаг 1. Строим модель с небольшим числом тем.

Шаг *k*. Пусть модель с множеством тем T уже построена. Строим множество дочерних тем S (subtopics), $|S| > |T|$.

Родительские темы приближаются смесями дочерних тем:

$$\sum_{t\in T} n_t \,\mathrm{KL}_w\Big(p(w|t) \,\Big\|\, \sum_{s\in S} p(w|s)p(s|t)\Big) \;\to\; \min_{\Phi,\tilde{\Psi}},$$

где $p(s|t) = \tilde{\psi}_{st}$, $\tilde{\Psi} = (\tilde{\psi}_{st})_{S\times T}$ — *матрица связей*, .

Родительская $\Phi^p \approx \Phi\tilde{\Psi}$, отсюда регуляризатор матрицы Φ:

$$R(\Phi,\tilde{\Psi}) = \tau \sum_{t\in T}\sum_{w\in W} n_{wt} \ln \sum_{s\in S} \phi_{ws}\tilde{\psi}_{st} \;\to\; \max.$$

Родительские темы t — «документы» с частотами слов n_{wt}.

Регуляризатор Θ: родительские темы как модальность

Шаг 1. Строим модель с небольшим числом тем.

Шаг k. Пусть модель с множеством тем T уже построена. Строим множество дочерних тем S (subtopics), $|S| > |T|$.

Родительские темы приближаются смесями дочерних тем:

$$\sum_{d\in D} n_d \,\mathrm{KL}_t\Big(p(t|d) \,\Big\|\, \sum_{s\in S} p(t|s)p(s|d)\Big) \;\to\; \min_{\Theta,\Psi},$$

где $\psi_{ts} = p(t|s)$, $\Psi = (\psi_{ts})_{T\times S}$ — матрица связей.

Родительская $\Theta^p \approx \Psi\Theta$, отсюда регуляризатор матрицы Θ:

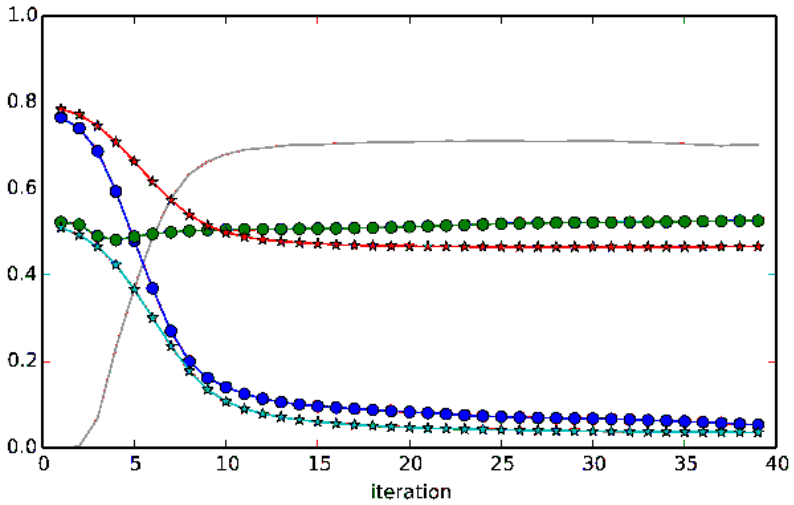
$$R(\Theta,\Psi) = \tau \sum_{d\in D}\sum_{t\in T} n_{td} \ln \sum_{s\in S} \psi_{ts}\theta_{sd} \;\to\; \max.$$

Родительские темы t — модальность с частотами токенов n_{td}.

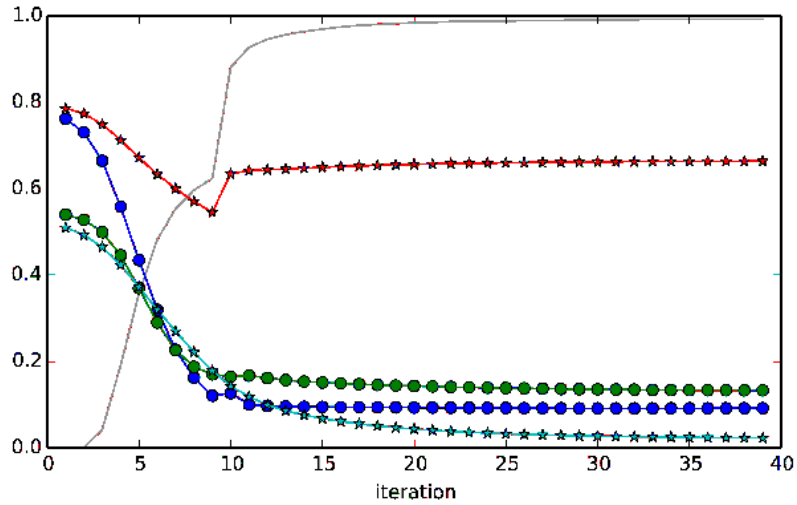
Эксперимент на коллекции ММРО-ИОИ

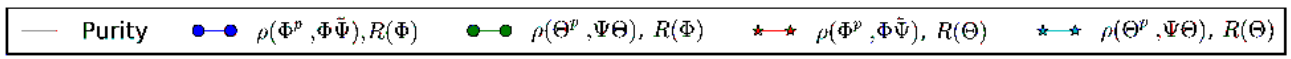
Среднее расстояние Хеллингера $\rho(\Phi^p, \Phi\tilde{\Psi})$ и $\rho(\Theta^p, \Psi\Theta)$ для регуляризаторов Φ и Θ при переходе между уровнями $1 \to 2$:

Разреживание Φ с 1-й итерации



Разреживание Φ с 10-й итерации





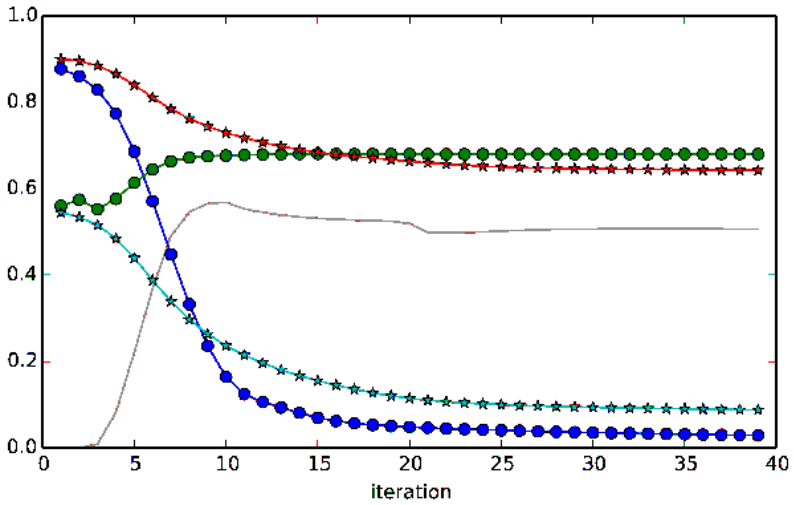
Вывод. Регуляризатор Θ плохо приближает Φ^p.

Chirkova N. A., Vorontsov K. V. Additive regularization for hierarchical multimodal topic modeling // JMLDA, 2016.

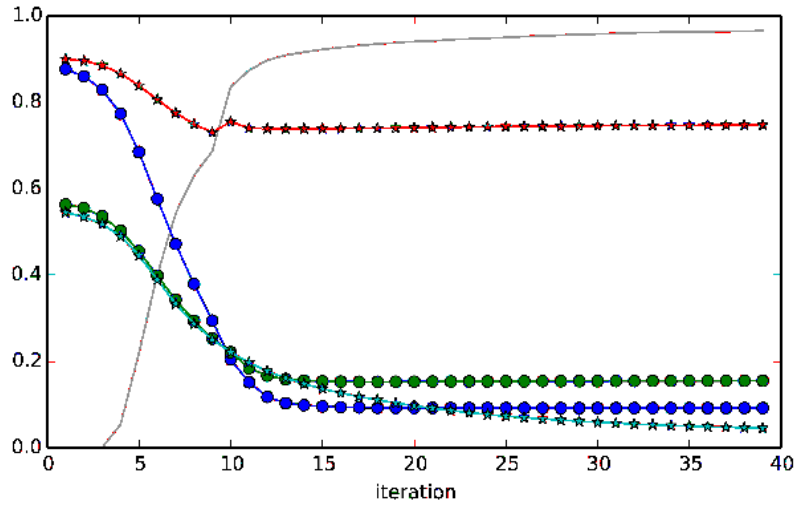
Эксперимент на коллекции ММРО-ИОИ

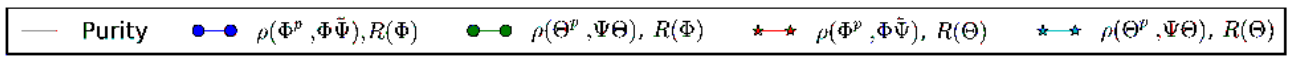
Среднее расстояние Хеллингера $\rho(\Phi^p, \Phi\tilde{\Psi})$ и $\rho(\Theta^p, \Psi\Theta)$ для регуляризаторов Φ и Θ при переходе между уровнями $2 \to 3$:

Разреживание Φ с 1-й итерации



Разреживание Φ с 10-й итерации





Вывод. Регуляризатор Θ плохо приближает Φ^p.

Chirkova N. A., Vorontsov K. V. Additive regularization for hierarchical multimodal topic modeling // JMLDA, 2016.

Выводы

- Регуляризатор Φ приближает $\Phi^p \approx \Phi\tilde{\Psi}$ и $\Theta^p \approx \Psi\Theta$.
- Регуляризатор Θ приближает только $\Theta^p \approx \Psi\Theta$.
- Максимальное разреживание $\psi_{ts} \in \{0, 1\}$ даёт иерархию–дерево.
- Нельзя допускать вырождения $\psi_{ts} = p(t|s) \equiv 0$.

Дальнейшие задачи:

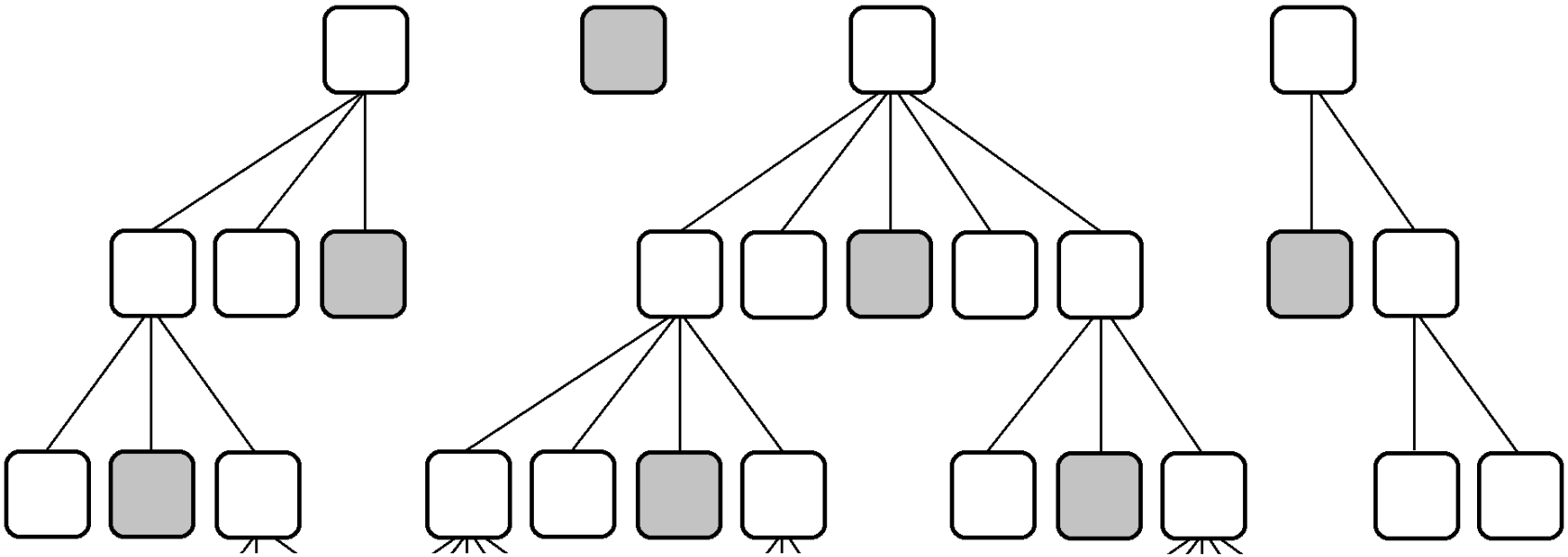
- Согласованная регуляризация: $\tilde{\psi}_{st} p(t) = \psi_{ts} p(s)$

$$\tau_1 \sum_{t,w} n_{wt} \ln \sum_s \phi_{ws}\psi_{ts}\frac{n_s}{n_t} + \tau_2 \sum_{d,t} n_{td} \ln \sum_s \psi_{ts}\theta_{sd} \;\to\; \max_{\Phi,\Psi,\Theta}$$

- Наращивание уровня для заданного подмножества $T' \subseteq T$
- Критерий неоднородности темы для включения её в T'
- Иерархии с темами различной глубины

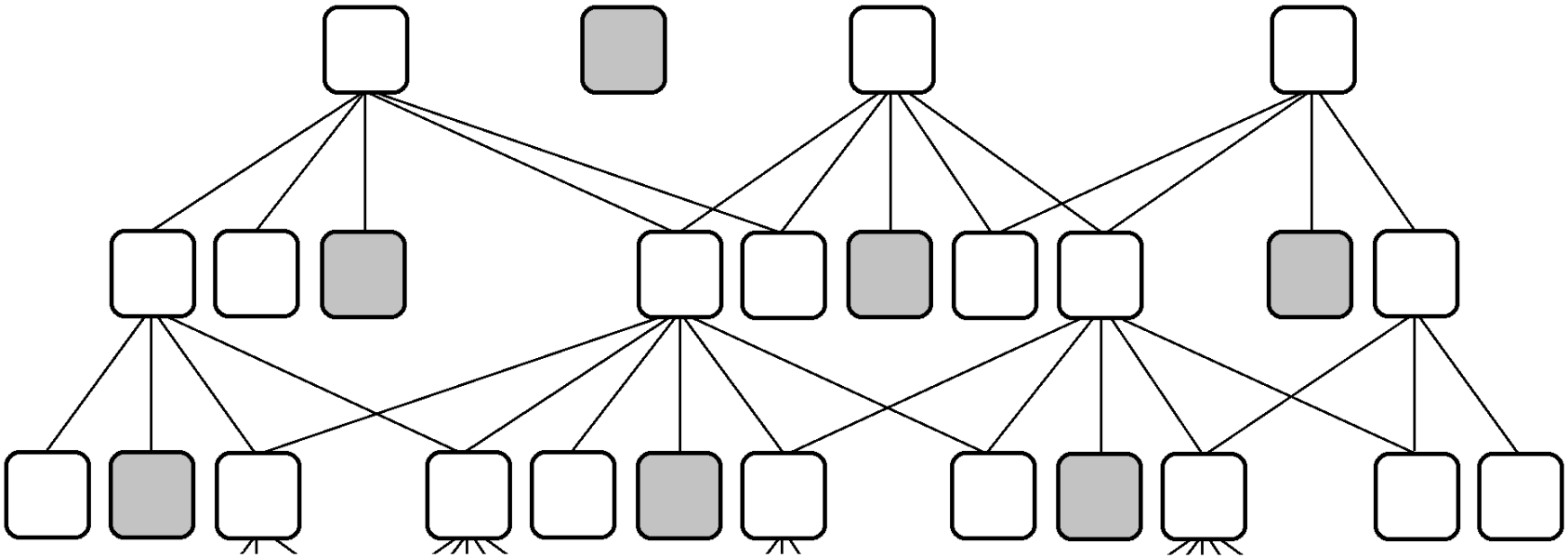
Иерархии с темами различной глубины

- На каждом уровне расщепляются не все темы (допускается вырожденность: $p(s|t) \equiv 0$ для некоторых t)
- Расщепляемая тема может иметь дочернюю фоновую, в которой собирается общая лексика родительской темы
- При максимальном разреживании $p(t|s) \in \{0, 1\}$ иерархия является деревом (корень не показан)



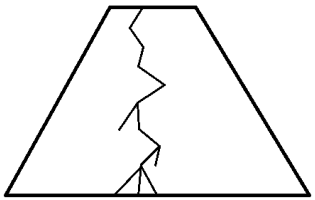
Иерархии с темами различной глубины

- На каждом уровне расщепляются не все темы (допускается вырожденность: $p(s|t) \equiv 0$ для некоторых t)
- Расщепляемая тема может иметь дочернюю фоновую, в которой собирается общая лексика родительской темы
- При умеренном разреживании $p(t|s)$ у вершины может быть несколько родителей (корень не показан)

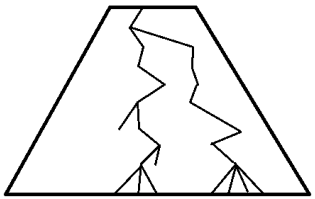


Иерархии с темами различной глубины

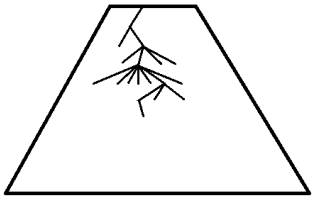
След документа в тематической иерархии определяет степень его специализации, назначение, аудиторию



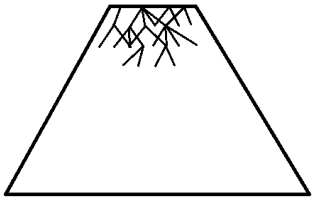
узко специализированный,
для профессионалов



междисциплинарное исследование,
для профессионалов



обзорный,
для ознакомления с предметной областью

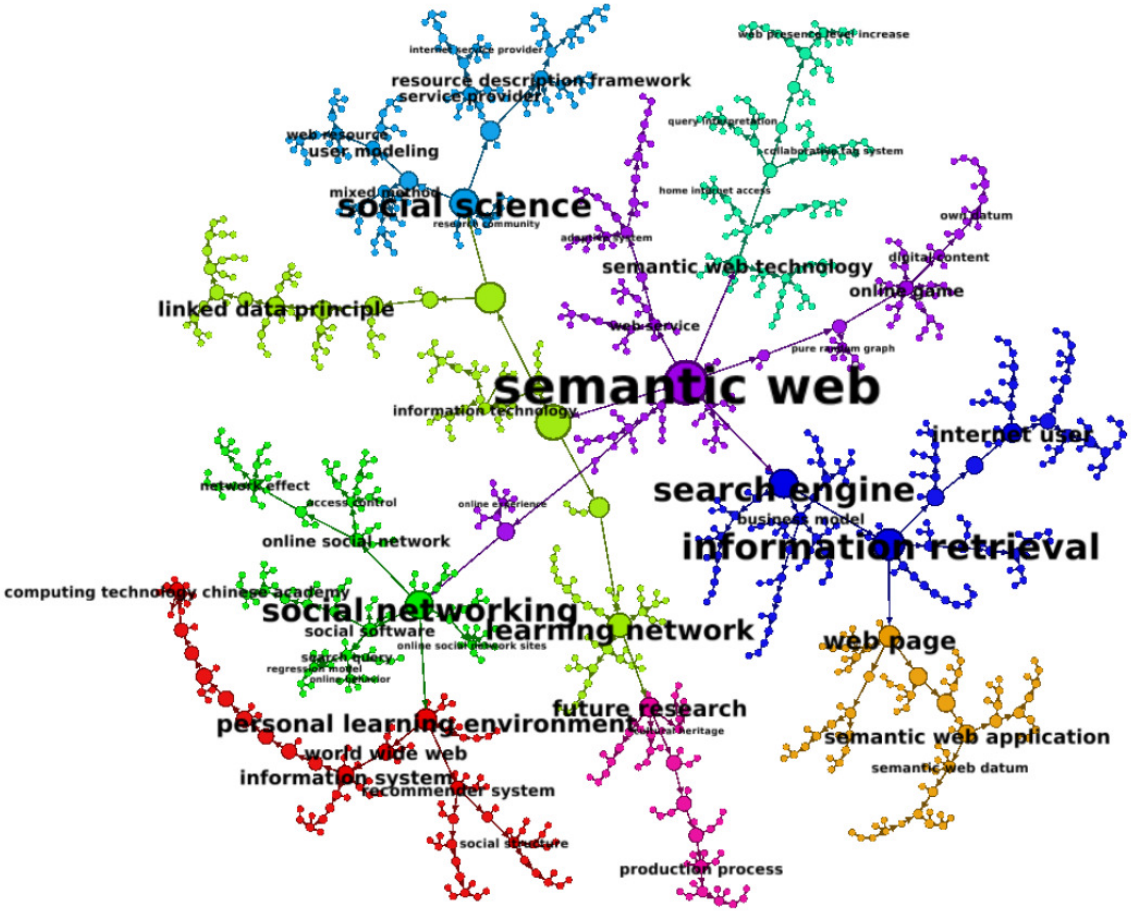


популярный или энциклопедиченский,
для расширения кругозора

Способы оценивания качества тематических иерархий

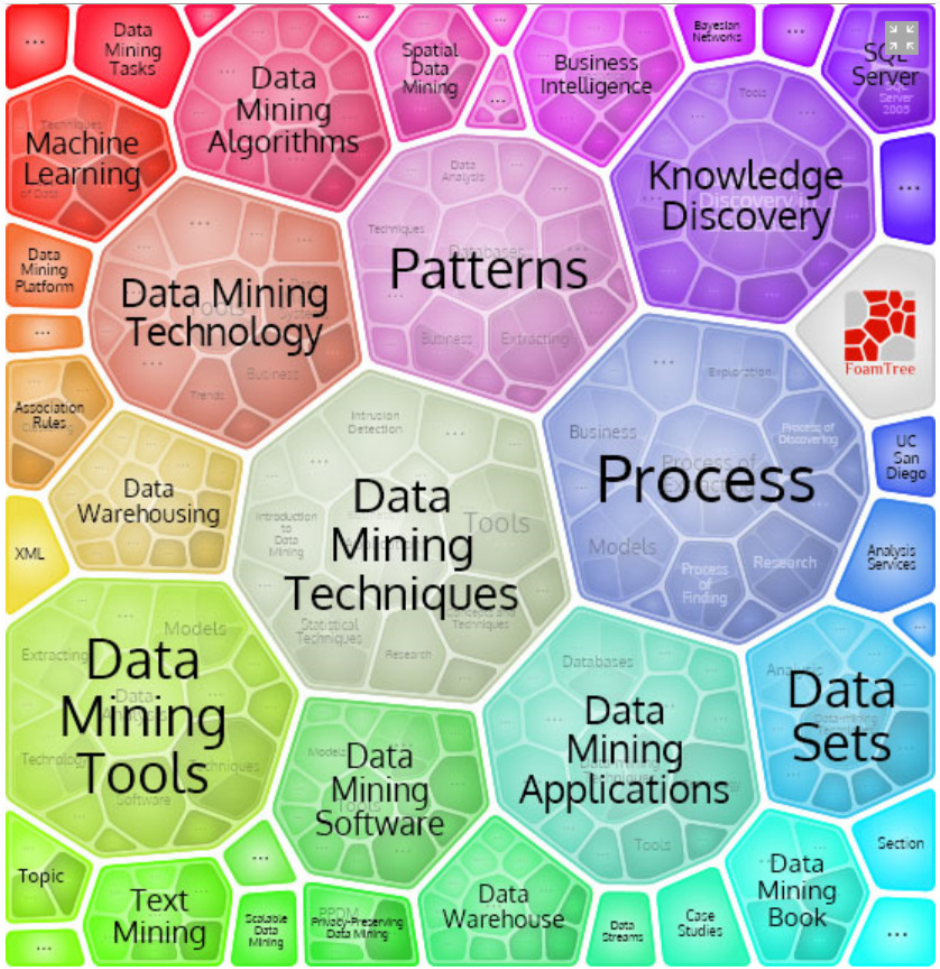
- *Перплексия* или правдоподобие: приводит ли постепенное дробление тем к более точному разложению
- *Устойчивость*: получаются ли схожие иерархии при различных начальных условиях
- *Полезность*: сколько шагов делает пользователь, чтобы найти документ по иерархии
- *Метод интрузий*: правильно ли асессоры определяют чужую тему, внедрённую в список дочерних тем
- *Сравнение с «золотым стандартом»*: насколько иерархия похожа на имеющуюся категоризацию документов

Визуализация древовидных иерархий



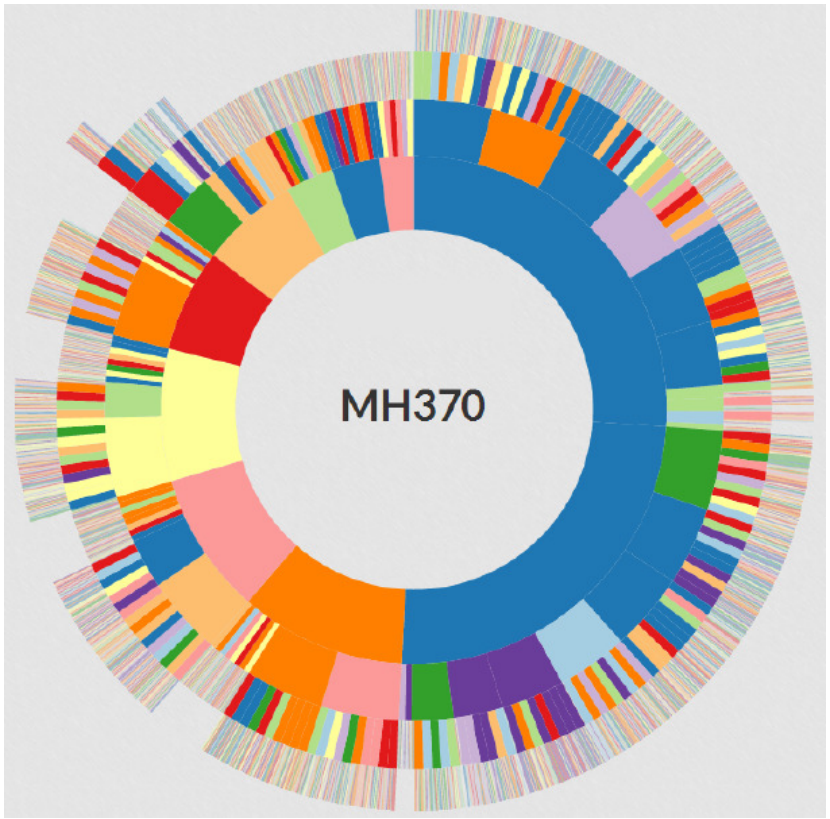
Georgeta Bordea. Domain adaptive extraction of topical hierarchies for Expertise Mining. 2013.

Визуализация древовидных иерархий в проекте FoamTree



https://carrotsearch.com/foamtree-overview

Визуализация древовидных иерархий



Smith A., Hawes T., Myers M.. Hiérarchie: interactive visualization for hierarchical topic models. Workshop on Interactive Language Learning, Visualization, and Interfaces, ACL, 2014.

1. Новостной мониторинг для медиапланирования

- **Дано:**
  поток новостей СМИ ($\sim 100\,\text{K}/\text{день}$) и социальных медиа.
- **Найти:**
  1) иерархическая тематическая модель,
  2) спектр тем и спектр мнений по заданному тексту,
  3) фильтр потока по заданным спектрам тем и мнений,
  4) оценки разнообразия тем и мнений в потоке.
- **Критерий:**
  1) интерпретируемость и различность тем,
  2) интерпретируемость разделения тем на подтемы,
  3) асессорские оценки качества (около 10 критериев):
     — точность отнесения пары новостей к одной теме,
     — точность распознавания новых тем,
     — точность распознавания слов общей лексики и др.

2. Новостной мониторинг для поиска проблемных компаний

- **Дано:**
  1) поток новостных сообщений СМИ,
  2) семантические ядра тем по компаниям,
  3) семантические ядра тем по проблемным ситуациям,
  4) выборка известных случаев проблемных ситуаций.
- **Найти:**
  1) сообщения о проблемных ситуациях по компаниям,
  2) все темы по каждой компании,
  3) новые типы проблемных ситуаций.
- **Критерий:**
  1) интерпретируемость всех тем,
  2) точность и полнота поиска по известным случаям.

3. Сценарный анализ записей разговоров контакт-центра

- **Дано:**
  1) коллекция текстов разговоров,
  2) семантические ядра (обучающие тексты) тем,
  3) сегментная разметка подвыборки разговоров.
- **Найти:**
  1) граф сценариев разговоров,
  2) вероятность успешного исхода в любой точке разговора,
  3) онлайновые подсказки оператору,
  4) автоматические оценки качества работы операторов,
  5) рекомендации операторам.
- **Критерий:**
  1) точность выделения тем в разговорах,
  2) точность сегментации на размеченной подвыборке.

4. Агрегатор русскоязычного научно-популярного контента

- **Дано:**
  1) коллекции статей научно-популярных порталов,
  2) коллекция Википедии на русском языке.
- **Найти:**
  1) общая тематическая иерархия,
  2) контекстные рекомендации по заданному тексту,
  3) тематический разведочный поиск по заданному тексту,
  4) интерактивная графическая «карта знаний».
- **Критерий:**
  1) полнота и точность поиска,
  2) интерпретируемость и различность тем,
  3) интерпретируемость разделения тем на подтемы,
  4) точность асессорского поиска документа по иерархии,
  5) экспертные оценки «интересности» рекомендаций.

5. Тематический разведочный поиск по коллективному блогу

- **Дано:**
  1) коллекция Habrahabr.ru или TechCrunch.com,
  2) выборка тематических запросов (длинные тексты),
  3) асессорские оценки релевантности документов запросам.
- **Найти:**
  1) тематическая модель для разведочного поиска,
  2) признаки сходства тематических векторов,
  3) функции ранжирования документов по запросу.
- **Критерий:**
  1) точность и полнота поиска,
  2) качество ранжирования (MAP или NDCG).

Янина А.О., Воронцов К.В. Мультимодальные тематические модели для разведочного поиска в коллективном блоге. JMLDA. 2016.

6. Кросс-язычный разведочный поиск по патентной базе

- **Дано:**
  1) коллекция патентов США на английском языке,
  2) коллекция их машинных переводов на русский язык,
  3) коллекция двуязычных статей Википедии,
  4) выборка русскоязычных запросов (длинные тексты),
  5) асессорские оценки релевантности документов запросам.
- **Найти:**
  1) тематическая иерархия научно-технической информации,
  2) признаки сходства тематических векторов,
  3) функции ранжирования документов по запросу.
- **Критерий:**
  1) точность и полнота кросс-язычного поиска,
  2) качество ранжирования (MAP или NDCG).

7. Построение продуктовой иерархии по текстам госзакупок

- **Дано:**
  1) описания объектов закупок (~ 200 млн.),
  2) общероссийский классификатор продукции ОКПД.
- **Найти:**
  иерархический тематический классификатор продуктов.
- **Критерий:**
  1) интерпретируемость и различность тем,
  2) интерпретируемость разделения тем на подтемы,
  3) согласованность верхних уровней с ОКПД,
  4) точность асессорского поиска товара по иерархии.

Чиркова Н.А., Воронцов К.В. Аддитивная регуляризация мультимодальных иерархических тематических моделей JMLDA. 2016.

8. Выявление структуры отрасли по транзакционным данным

- **Дано:**
  1) база банковских транзакций между компаниями,
  2) коды ОКВЭД для компаний.
- **Найти:**
  1) латентные темы — виды экономической деятельности,
  2) их соответствие ОКВЭДам,
  3) граф товарно-денежных потоков отрасли,
  4) типовые бизнес-схемы компаний — лидеров отрасли.
- **Критерий:**
  1) точность описания транзакционных данных,
  2) интерпретируемость графа отрасли.

9. Диагностика заболеваний по электрокардиограмме

- **Дано:**
  1) электрокардиограммы, закодированные в символьные последовательности методом В.М.Успенского,
  2) диагнозы по 32 заболеваниям для каждой ЭКГ.
- **Найти:**
  1) диагностические эталоны каждого заболевания,
  2) решающее правило по каждому заболеванию.
- **Критерий:**
  1) чувствительность и специфичность диагностики,
  2) качество ранжирования диагнозов.

Темы исследований, где есть открытые проблемы

- Устойчивость и полнота набора тем.
- Оптимизация параметров онлайнового алгоритма.
- Адаптивная оптимизация коэффициентов регуляризации.
- Эффективная инициализация матрицы Ф.
- Создание новых тем в потоке новостей.
- Автоматическое выделение терминов-словосочетаний.
- Тематическая сегментация.
- Тематические модели дистрибутивной семантики.
- Суммаризация тем.
- Автоматическое именование тем.
- Интеграция с анализом тональности и выявлением мнений.
- Интеграция с синтаксическими анализаторами.

Резюме

- Задача тематического моделирования некорректно поставлена, её решение не единственно и не устойчиво.
- Регуляризация — стандартный приём решения таких задач.
- Подход ARTM позволяет комбинировать регуляризаторы и строить тематические модели с требуемыми свойствами.
- Реализация — в проекте с открытым кодом BigARTM.
- Модель LDA — слишком слабый регуляризатор, не решает проблему неединственности и неустойчивости.
- Модель LDA лучше описывает вероятности редких слов, но для выявления тематики они как раз и не важны.
- Регуляризаторы и модальности — в следующих лекциях.