Вероятностные тематические модели
Лекция 6. Анализ зависимостей

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)»

МФТИ – ФИЦ ИУ РАН • 14 октября 2020

Содержание

Напоминание. Мультимодальная ARTM: постановка задачи

Дано: W^m — словарь термов m-й модальности, $m \in M$,
D — коллекция текстовых документов $d \subset W = \bigsqcup_m W^m$,
n_{dw} — сколько раз терм w встретился в документе d.

Найти: модель $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$ с параметрами $\underset{W^m \times T}{\Phi^m}$ и $\underset{T \times D}{\Theta}$:
$\phi_{wt} = p(w|t)$ — вероятности терма w в каждой теме t,
$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d.

Критерий максимума регуляризованного правдоподобия:

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

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

Vorontsov K.V., Frei O.I., Apishev M.A., Romov P.A., Suvorova M.A., Yanina A.O. Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections // CIKM. Workshop on Topic Models, 2015.

Напоминание. Мультимодальная ARTM: EM-алгоритм

Максимизация log правдоподобия с регуляризатором R:

$$\sum_{d\in D}\sum_{w\in d}\tilde{n}_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\phi,\theta};$$

где $\tilde{n}_{dw} = \tau_{m(w)} n_{dw}$, $m(w)$ — модальность терма w.

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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.

Напоминание. Онлайновый EM-алгоритм для ARTM

Коллекция D разбивается на пакеты D_b, $b = 1, \ldots, B$, которые могут обрабатываться параллельно и/или распределённо.

Вход: коллекция документов D,
параметры $\delta \equiv \texttt{decay\_weight}$, $\alpha \equiv \texttt{apply\_weight}$;
Выход: матрица Φ;

инициализировать ϕ_{wt} для всех $w \in W$, $t \in T$;
$n_{wt} := 0$, $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;
для всех пакетов D_b, $b = 1, \ldots, B$
  $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \textbf{ProcessBatch}(D_b, \Phi)$;
  если пора обновить матрицу Φ **то**
    $n_{wt} := \delta n_{wt} + \alpha \tilde{n}_{wt}$ для всех $w \in W$, $t \in T$;
    $\phi_{wt} := \underset{w \in W^m}{\text{norm}} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ для всех $m \in M$, $w \in W^m$, $t \in T$;
    $\tilde{n}_{wt} := 0$ для всех $w \in W$, $t \in T$;

Напоминание. Онлайновый EM-алгоритм для ARTM

Функция **ProcessBatch** обрабатывает пакет документов D_b, не меняя матрицу Φ, и выдаёт счётчики термов в темах $\tilde{n}_{wt}$.

Вход: пакет документов D_b, матрица $\Phi = (\phi_{wt})$;
Выход: матрица счётчиков $(\tilde{n}_{wt})_{W \times T}$;

$\tilde{n}_{wt} := 0$ для всех $w \in W,\ t \in T$;
для всех $d \in D_b$
  инициализировать $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;
  повторять
    $p_{tdw} := \underset{t \in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$ для всех $w \in d,\ t \in T$;
    $\theta_{td} := \underset{t \in T}{\text{norm}}\left(\sum_{w \in d} \tau_{m(w)} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)$ для всех $t \in T$;
  пока θ_d не сойдётся;
  $\tilde{n}_{wt} := \tilde{n}_{wt} + \tau_{m(w)} n_{dw} p_{tdw}$ для всех $w \in W,\ t \in T$;

Тематическая модель классификации (категоризации)

Обучающие данные: C — множество классов (категорий);
$C_d \subseteq C$ — классы, к которым d относится;
$C'_d \subseteq C$ — классы, к которым d не относится.

$p(c|d) = \sum_{t \in T} \phi_{ct}\theta_{td}$ — линейная модель классификации

Правдоподобие вероятностной модели бинарных данных:

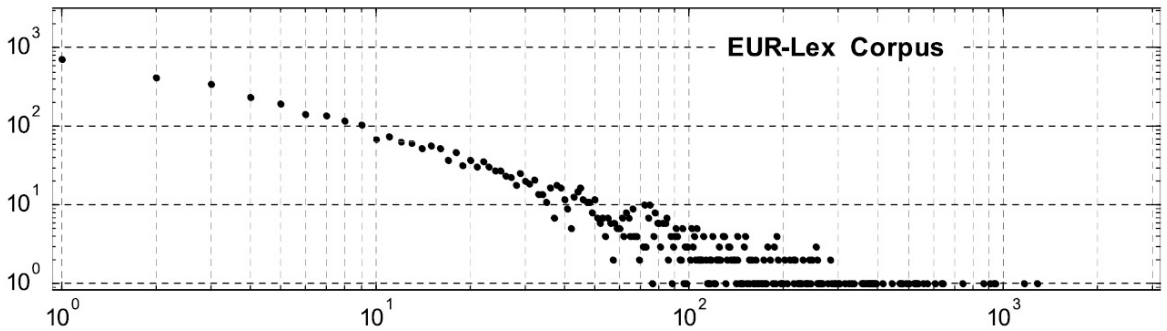
$$R(\Phi, \Theta) = \tau \sum_{d \in D} \sum_{c \in C_d} \ln \sum_{t \in T} \phi_{ct}\theta_{td} + \\ + \tau \sum_{d \in D} \sum_{c \in C'_d} \ln\Big(1 - \sum_{t \in T} \phi_{ct}\theta_{td}\Big) \;\to\; \max$$

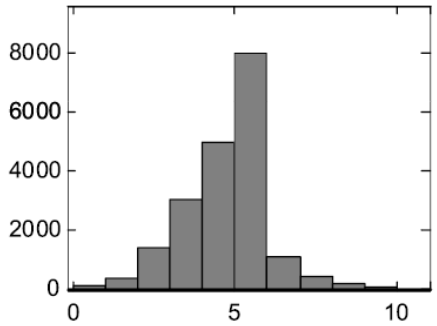
При $C'_d = \varnothing$, $n_{dc} = [c \in C_d]$ это правдоподобие модальности C.

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 88 (1–2).

Эксперимент. Категоризация коллекции EUR-Lex

- EUR-Lex: $|D| = 19\,800$ документов — законы Евросоюза
- Две модальности: W^1 слова (21K), W^2 категории (3 250)
- Категории несбалансированные и пересекающиеся:





- слева: # категорий с заданным # документов в категории
- справа: # документов с заданным # категорий

Rubin T. N., Chambers A., Smyth P., Steyvers M. Statistical topic models for multi-label document classification // Machine Learning, 2012, no. 88 (1–2).

Эксперимент. Категоризация коллекции EUR-Lex

Регуляризаторы:

- Равномерное сглаживание Θ
- Равномерное сглаживание матрицы слова–темы Φ^1
- *Label regularization* для матрицы категории–темы Φ^2:

$$R(\Phi^2) = \tau \sum_{c \in W^2} \hat{p}_c \ln p(c) \to \max,$$

где $p(c) = \sum_{t \in T} \phi_{ct} p(t)$ — распределение на категориях c,

$p(t) = \frac{n_t}{n}$ — распределение на темах,

$\hat{p}_c$ — доля документов категории c в обучающей выборке.

Mann G. S., McCallum A. Simple, robust, scalable semi-supervised learning via expectation regularization // ICML 2007, Pp. 593–600.

Эксперимент. Категоризация коллекции EUR-Lex

DLDA (Dependency LDA) [Rubin 2012] — среди байесовских моделей ближайший аналог ARTM для классификации

Критерии качества [Rubin 2012]:

- AUC-PR (%, $\Uparrow$) — Area under precision-recall curve
- AUC (%, $\Uparrow$) — Area under ROC curve
- OneErr (%, $\Downarrow$) — One error (most ranked label is not relevant)
- IsErr (%, $\Downarrow$) — Is error (no perfect classification)

Результаты сравнения:

	AUC-PR$\Uparrow$	AUC$\Uparrow$	OneErr$\Downarrow$	IsErr$\Downarrow$
BigARTM	**52.9**	98.0	**27.1**	**94.2**
DLDA [Rubin 2012]	49.2	**98.2**	32.0	97.2
SVM	43.5	97.5	31.6	98.1

Тематическая модель регрессии

Обучающие данные: $y_d \in \mathbb{R}$ для всех документов $d \in D$.

$E(y|d) = \sum_{t \in T} v_t \theta_{td}$ — линейная модель регрессии, $v \in \mathbb{R}^{|T|}$.

Регуляризатор — среднеквадратичная ошибка (МНК):

$$R(\Theta, v) = -\tau \sum_{d \in D} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right)^2 \to \max$$

Подставляем, получаем формулы M-шага:

$$\theta_{td} = \operatorname*{norm}_t \left(n_{td} + \tau v_t \theta_{td} \left(y_d - \sum_{s \in T} v_s \theta_{sd} \right) \right);$$

$$v = (\Theta \Theta^T)^{-1} \Theta y.$$

Sokolov E., Bogolubsky L. Topic Models Regularization and Initialization for Regression Problems // CIKM-2015 Workshop on Topic Models. ACM, pp. 21–27.

Примеры задач регрессии на текстах

MovieReview [Pang, Lee, 2005]

d — текст отзыва на фильм

y_d — рейтинг фильма (1..5), поставленный автором отзыва

Salary (kaggle.com: *Adzuna Job Salary Prediction*)

d — описание вакансии, предлагаемой работодателем

y_d — годовая зарплата

Yelp (kaggle.com: *Yelp Recruiting Competition*)

d — отзыв (на ресторан, отель, сервис и т.п.)

y_d — число голосов «useful», которые получит отзыв

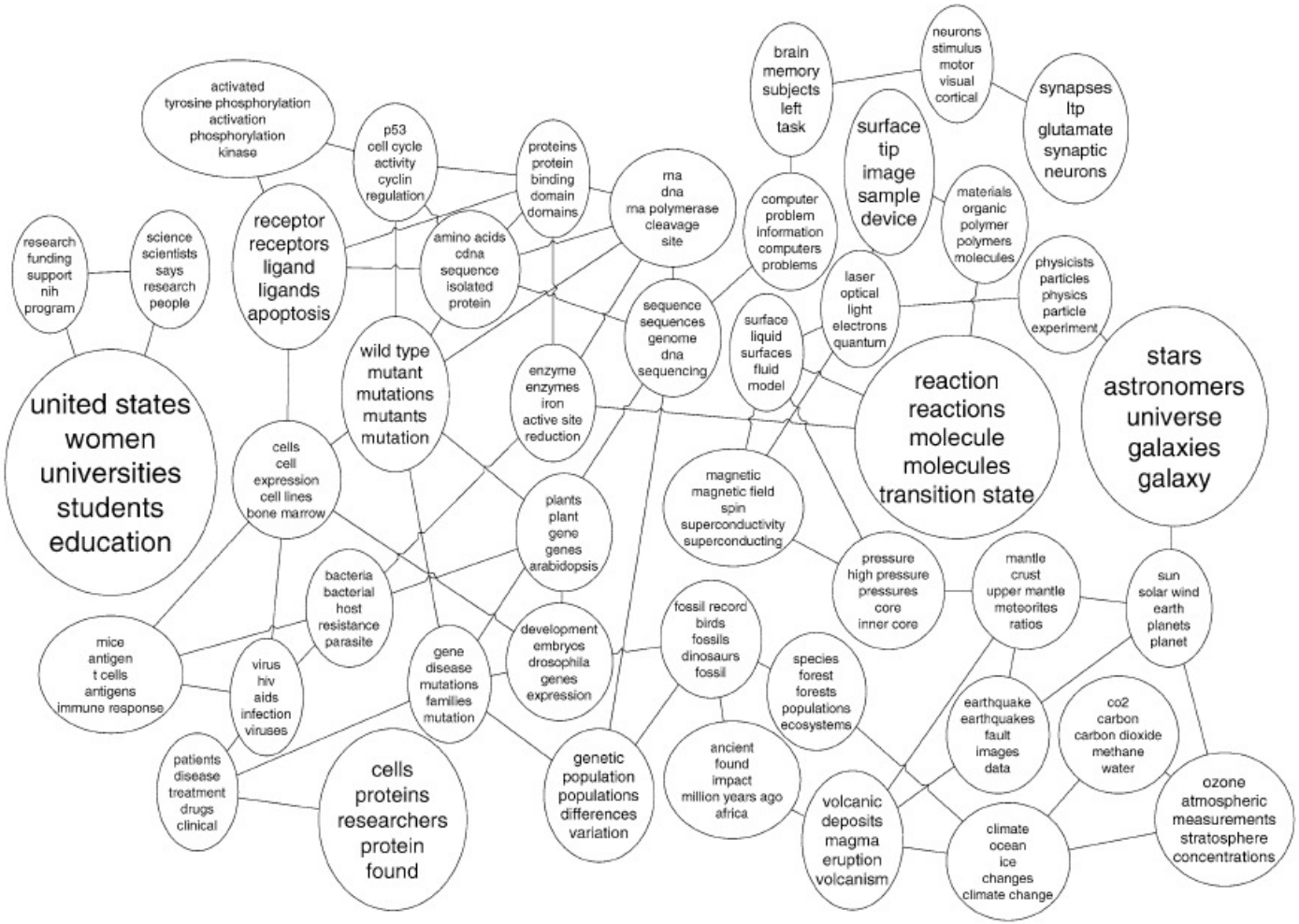
Прогнозирование скачков цен на финансовых рынках

d — текст новости

y_d — изменение цены в последующие 10–60 минут

B. Pang, L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales // ACL, 2005.

CTM: модель коррелированных тем



David Blei, John Lafferty. A Correlated Topic Model of SCIENCE // Annals of Applied Statistics, 2007. Vol. 1, Pp. 17-35.

Многомерное лог-нормальное распределение

Мотивация. Темы могут коррелировать: «статьи по археологии чаще связаны с историей и геологией, чем с генетикой».

Гипотеза. Вектор-столбцы θ_d порождаются $|T|$-мерным лог-нормальным распределением с ковариационной матрицей S:

$$p(\eta_d|\mu,S) = \frac{1}{(2\pi)^{\frac{n}{2}}|S|^{\frac{1}{2}}}\exp\Big(-\tfrac{1}{2}(\eta_d-\mu)^{\mathsf{T}}S^{-1}(\eta_d-\mu)\Big),$$

где $\eta_d = (\eta_{td})_{t\in T}$, $\eta_{td} = \ln\theta_{td} + C_d$ — векторы документов, определённые с точностью до константы C_d, не зависящей от t,

μ, S — параметры гауссовского распределения.

David Blei, John Lafferty. A Correlated Topic Model of SCIENCE // Annals of Applied Statistics, 2007. Vol. 1, Pp. 17-35.

Регуляризатор модели коррелированных тем CTM

Максимизация правдоподобия выборки векторов $\eta_d = (\eta_{td})$:

$$\sum_{d \in D} \ln p(\eta_d | \mu, S) \to \max .$$

Регуляризатор с параметрами μ, S:

$$R(\Theta) = -\frac{\tau}{2} \sum_{d \in D} (\eta_d - \mu)^{\mathsf{T}} S^{-1} (\eta_d - \mu) \to \max .$$

Формулы M-шага (S, μ можно обновлять в конце итерации):

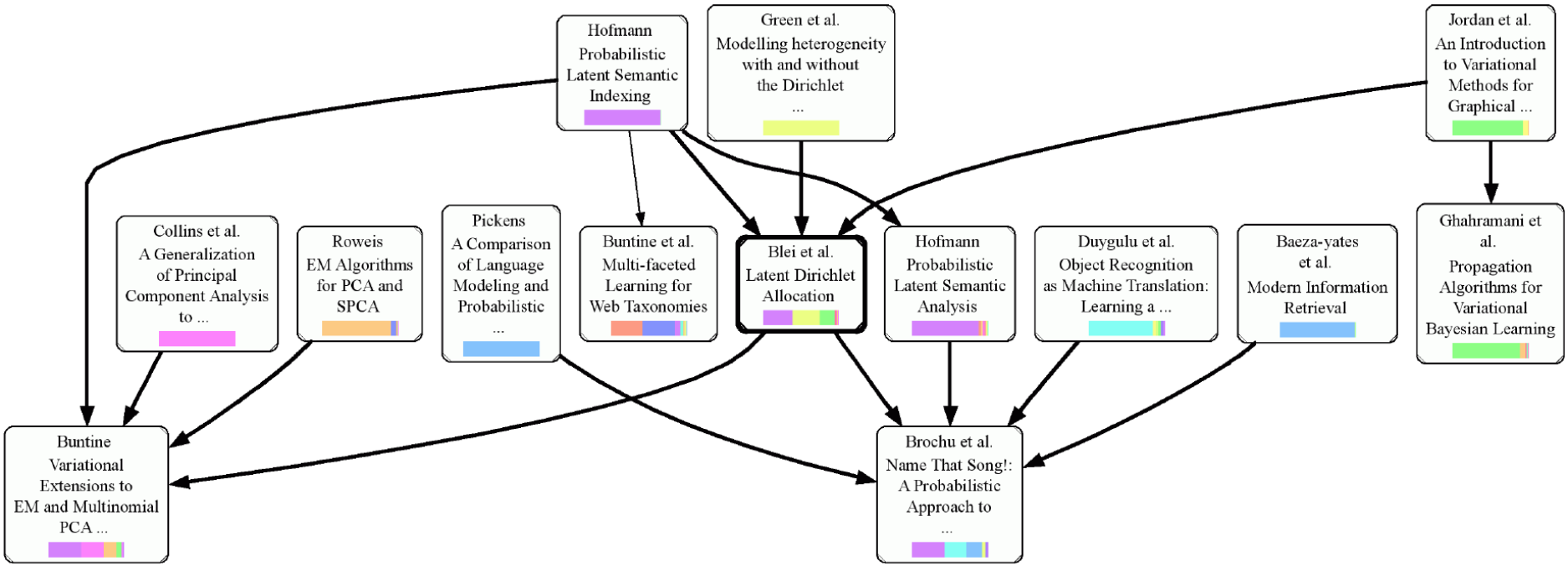
$$\theta_{td} = \operatorname*{norm}_{t \in T} \Big(n_{td} - \tau \sum_{s \in T} S^{-1}_{ts} \big(\ln \theta_{sd} - \mu_s \big) \Big);$$

$$\mu = \frac{1}{|D|} \sum_{d \in D} \ln \theta_d;$$

$$S = \frac{1}{|D|} \sum_{d \in D} \big(\ln \theta_d - \mu \big) \big(\ln \theta_d - \mu \big)^{\mathsf{T}}.$$

Модели, учитывающие цитирования или гиперссылки

- Учёт ссылок уточняет тематическую модель
- Тематическая модель выявляет влиятельные ссылки



Laura Dietz, Steffen Bickel, Tobias Scheffer. Unsupervised prediction of citation influences. ICML-2007.

Регуляризатор Θ для учёта связей между документами

Цель: улучшить темы, используя ссылки или цитирования (если документы ссылаются друг на друга, то их темы близки):

n_{dc} — число ссылок из d на c.

Максимизируем ковариации тематических векторных представлений связанных документов θ_d, θ_c:

$$R(\Theta) = \tau \sum_{d,c \in D} n_{dc} \sum_{t \in T} \theta_{td}\theta_{tc} \to \max.$$

Подставляем, получаем ещё один вариант сглаживания:

$$\theta_{td} = \underset{t}{\text{norm}}\left(n_{td} + \tau\theta_{td} \sum_{c \in D} n_{dc}\theta_{tc} \right).$$

Laura Dietz, Steffen Bickel, Tobias Scheffer. Unsupervised prediction of citation influences. ICML-2007.

Связи как модальность. Регуляризатор Ф

Проблема учёта связей в пакетном EM-алгоритме: связанные документы могут оказаться в разных пакетах.

Документы содержат термы $w \in W^1$ и ссылки $c \in W^2 \subseteq D$

W^2 — модальность документов, на которые есть ссылки

Регуляризатор — log-правдоподобие модальности W^2:

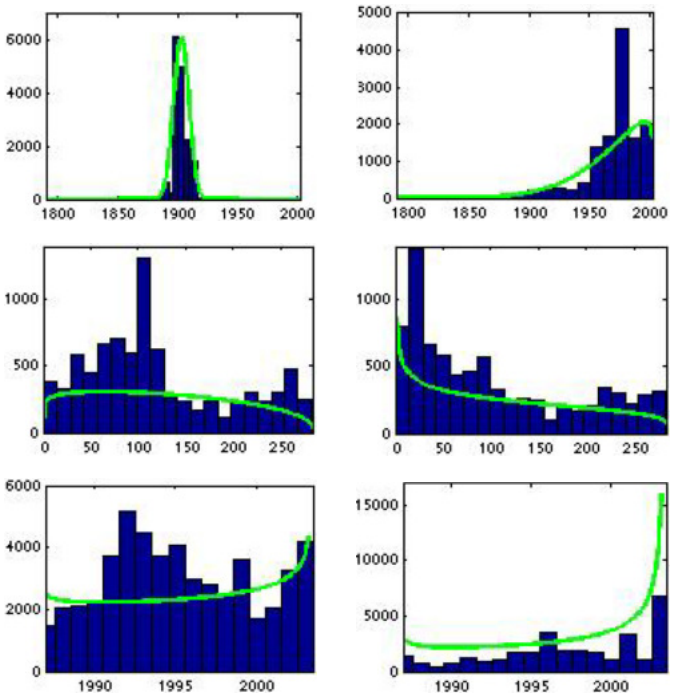
$$R(\Phi^2, \Theta) = \tau \sum_{d \in D} \sum_{c \in W^2} n_{dc} \ln \sum_{t \in T} \phi_{ct} \theta_{td} \to \max .$$

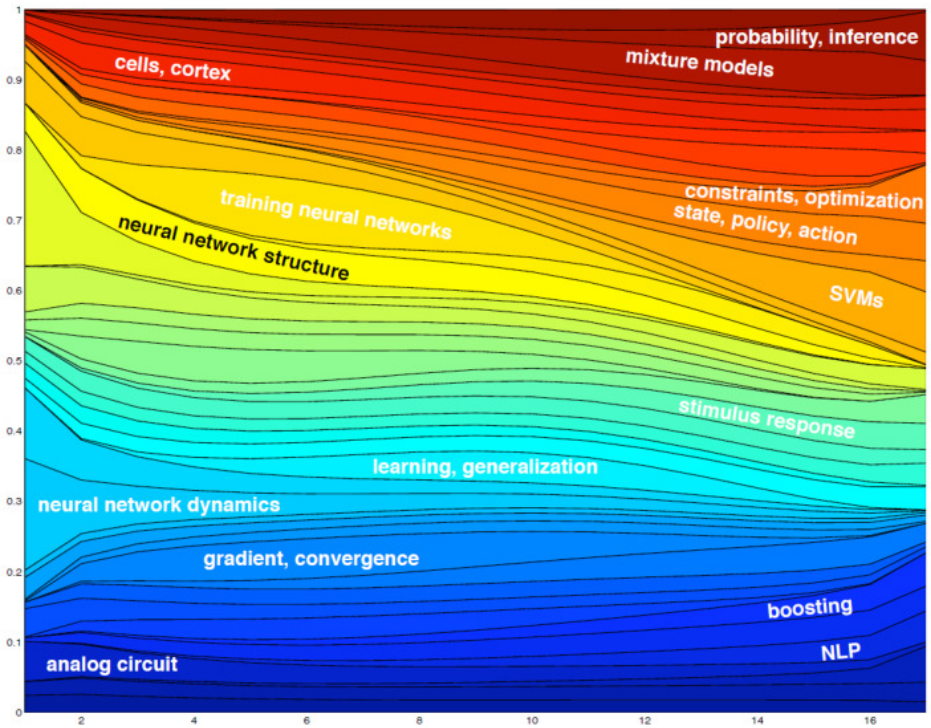
Другой вариант — сумма ковариационных регуляризаторов:

$$R(\Phi^2, \Theta) = \tau \sum_{d,c} n_{dc} \sum_{t \in T} \phi_{ct} \theta_{td} \to \max .$$

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 пакетном 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.

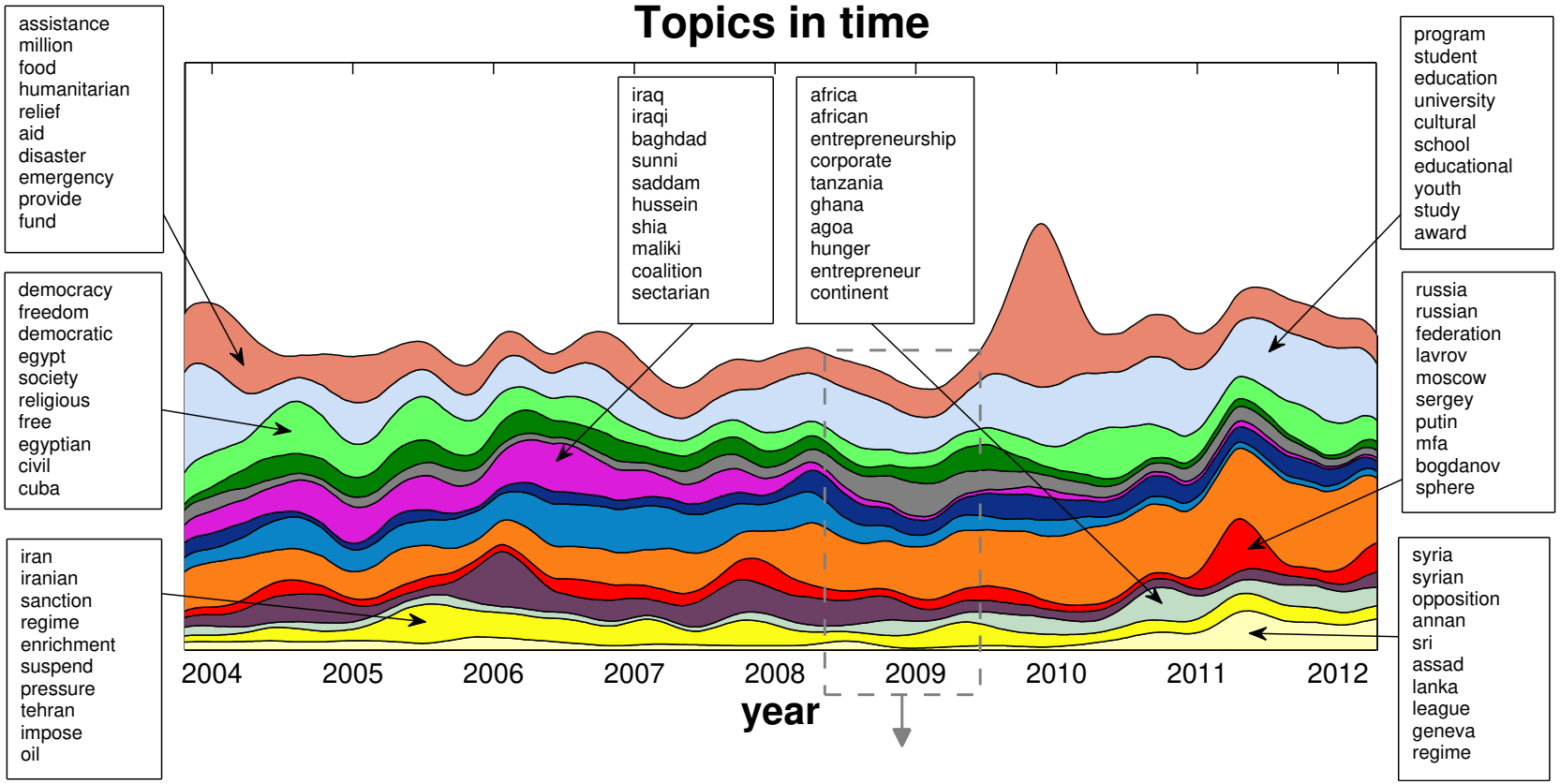
Цели исследования:

- 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

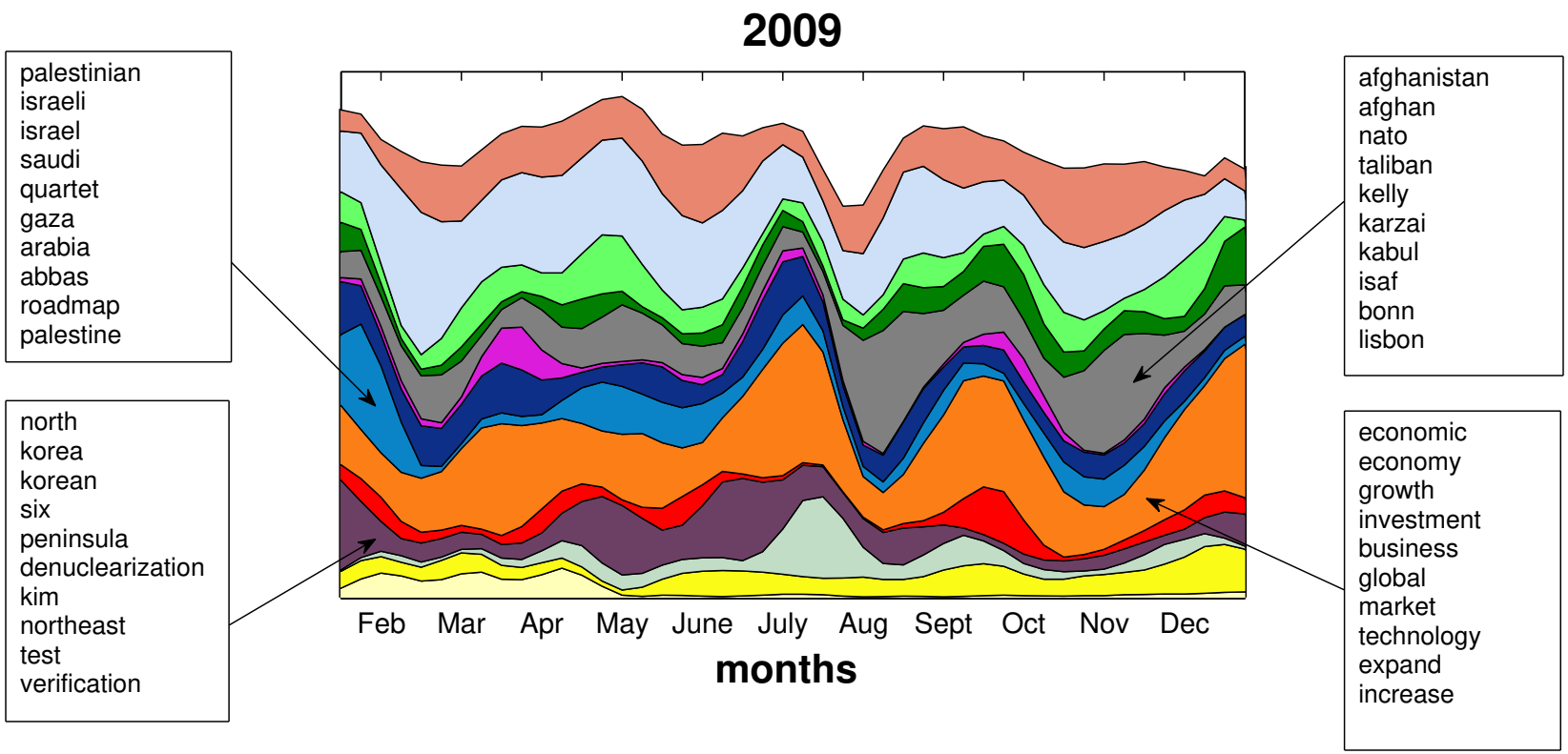
Динамика тем во времени



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

Динамика тем во времени

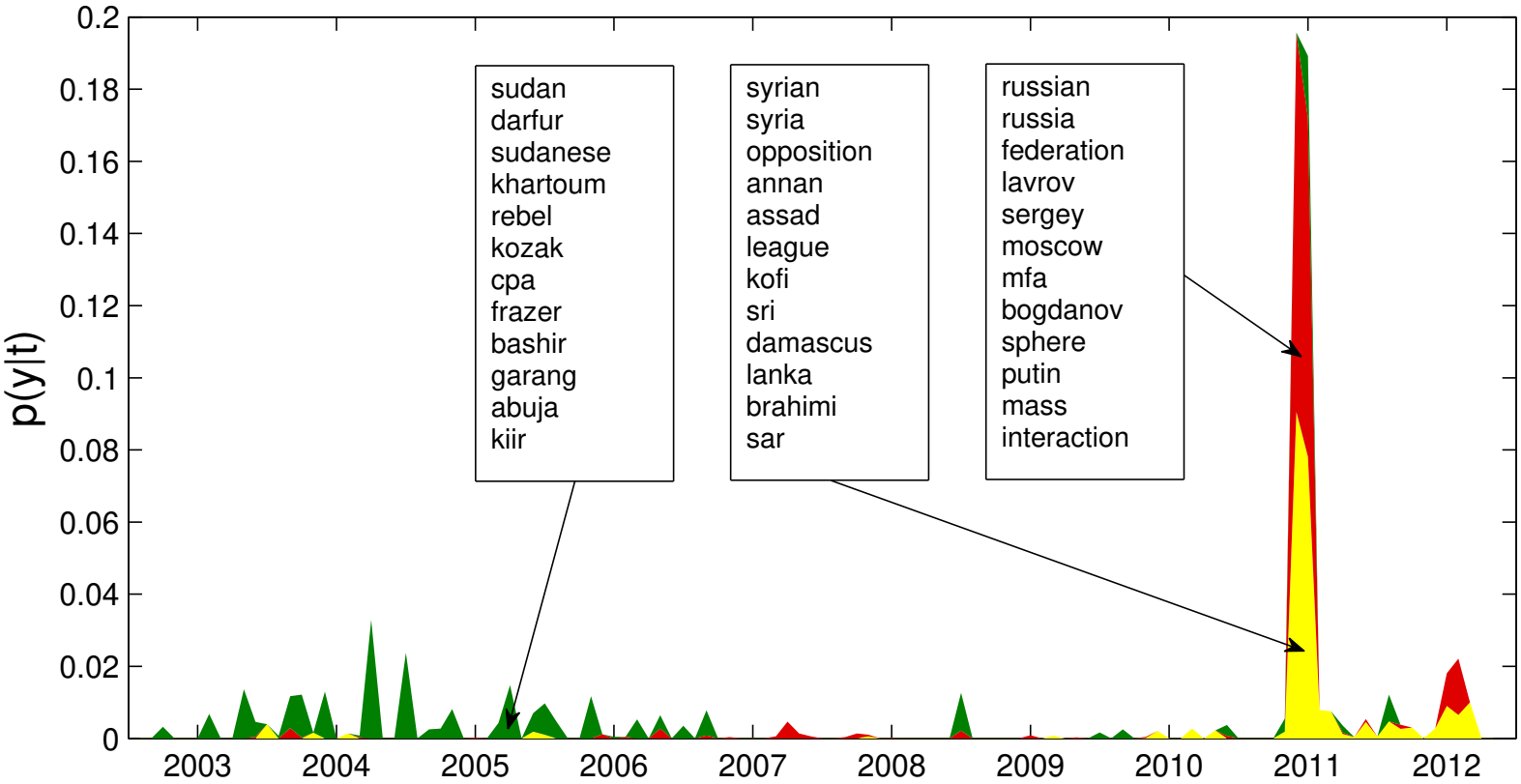
Укрупнение масштаба времени



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

Динамика тем во времени

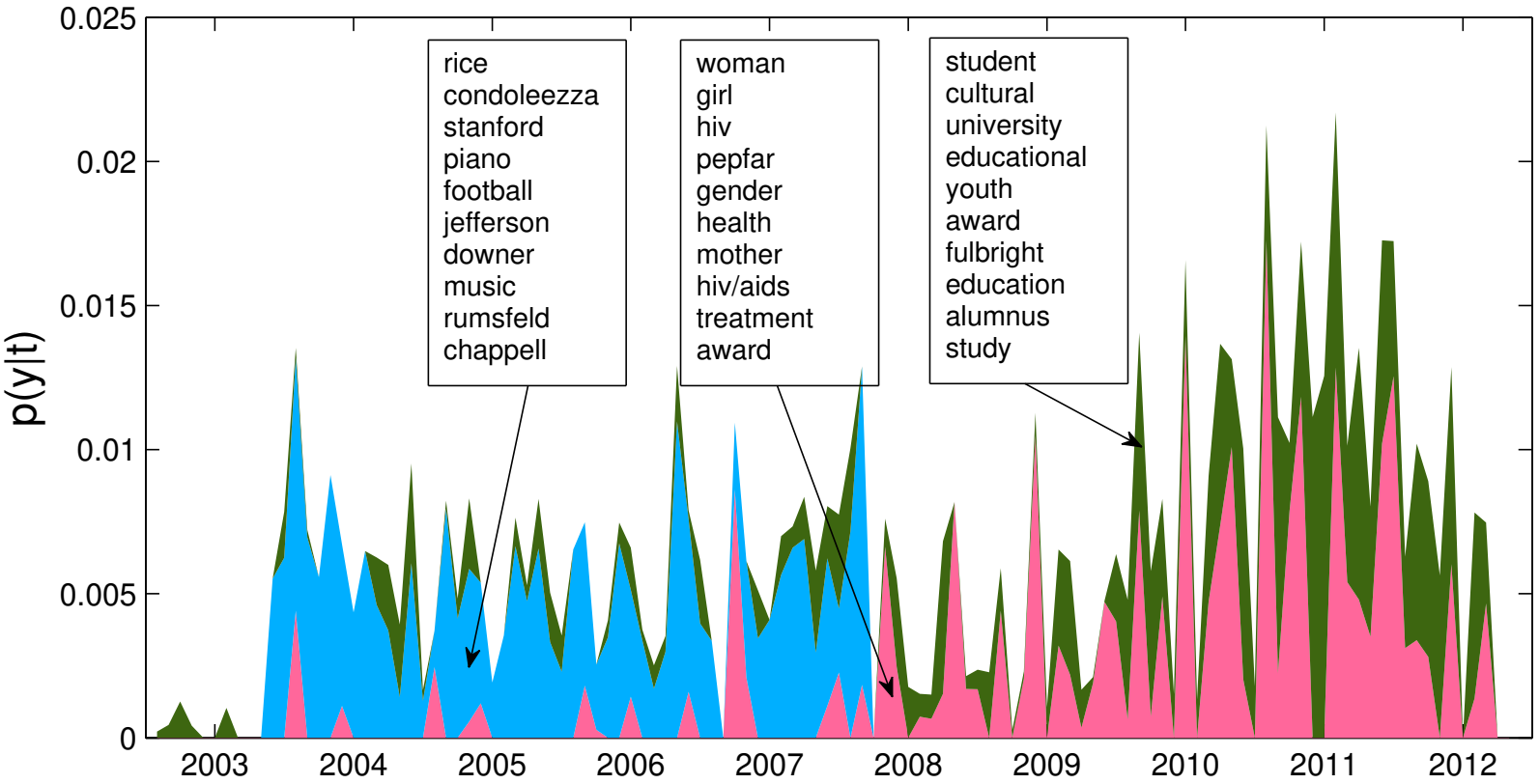
Пример: событийные темы и момент их совместного всплеска



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 регионов?
- есть ли похожие темы в разных регионах?

Регуляризатор:

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

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

$w_{cd} = K\big(\rho(\ell_c, \ell_d)\big)$, $K(\rho)$ — убывающая функция расстояния

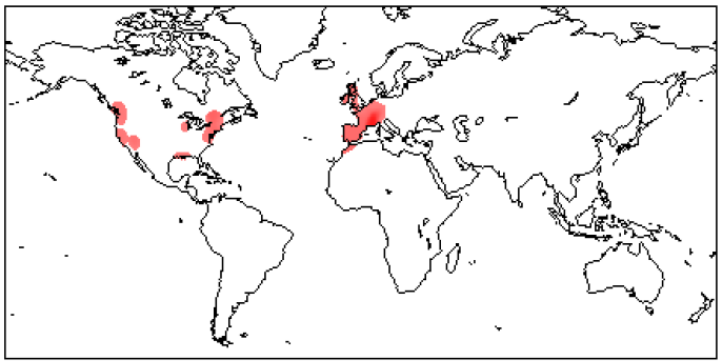
Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, Thomas Huang. Geographical Topic Discovery and Comparison. WWW 2011.

Пример: Food dataset

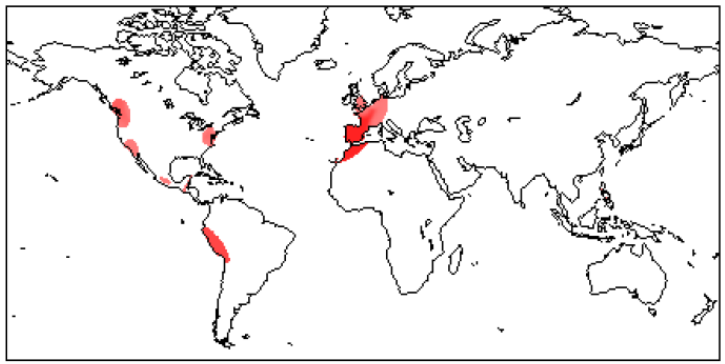
Где и что едят пользователи Flickr?

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.

Задача выявления тематических сообществ

Граф $\langle V, E \rangle$, вершины v — подмножества $D_v \subset D$, например:
D_v — отдельный документ $v \equiv d$
D_v — все статьи одного автора v
D_v — все посты из одной геолокации v

Тематика вершины:

$$p(t|v) = \sum_{d \in D_v} p(t|d)p(d|v) = \frac{1}{|D_v|} \sum_{d \in D_v} \theta_{td}$$

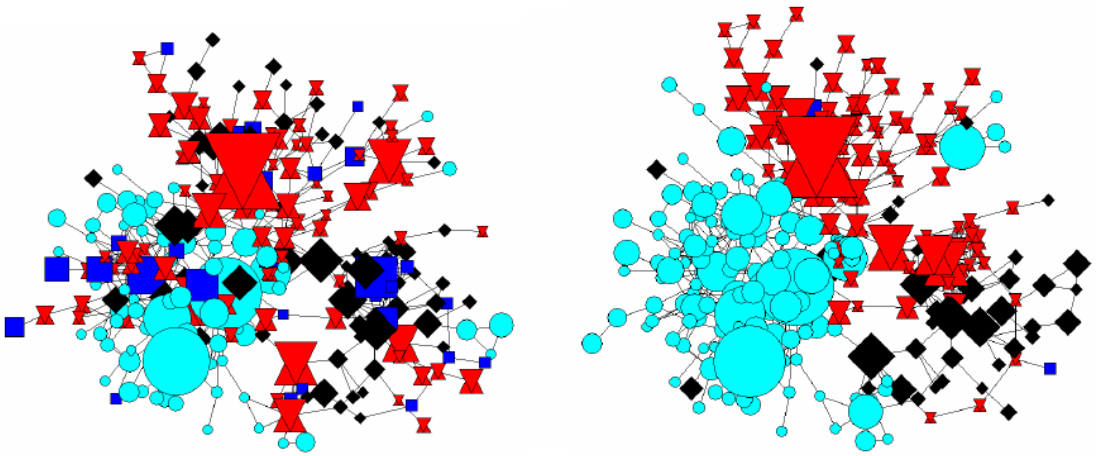
Регуляризатор NetPLSA, при заданных весах рёбер w_{uv}:

$$R(\Theta) = -\frac{\tau}{2} \sum_{(u,v) \in E} w_{uv} \sum_{t \in T} \big(p(t|v) - p(t|u)\big)^2 \to \max_{\Theta}$$

Qiaozhu Mei, Deng Cai, Duo Zhang, ChengXiang Zhai. Topic Modeling with Network Regularization. WWW-2008.

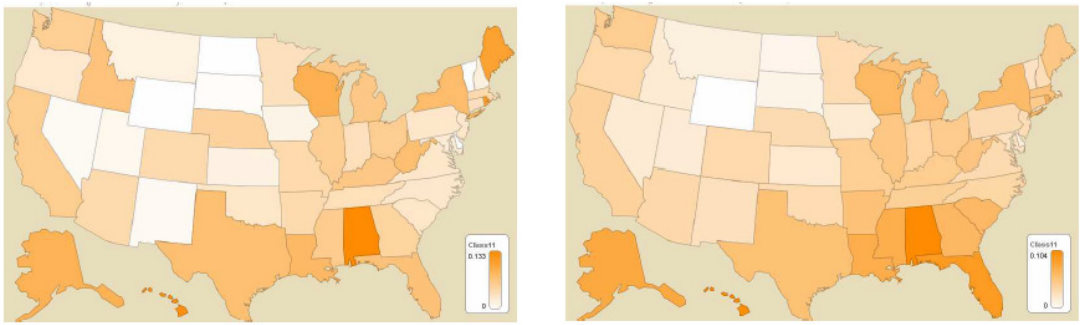
Примеры тематических сообществ

D_v — все статьи автора v на четырёх конференциях:



PLSA NetPLSA

D_v — все посты про ураган Катрина из штата v:



With PLSA With NetPLSA

От NetPLSA к модальности вершин графа

Проблема регуляризатора Θ в пакетном EM-алгоритме: связанные документы могут попасть в разные пакеты.

$W^2 = V$ — модальность вершин графа $\langle V, E \rangle$.
В каждый документ $d \in D_v$ добавляется терм-вершина v.

Тематика вершины:

$$p(t|v) = p(v|t)\frac{p(t)}{p(v)} = \phi_{vt}\frac{n_t}{n_v}$$

Регуляризатор NetPLSA, при заданных весах рёбер w_{uv}:

$$R(\Phi^2) = -\frac{\tau}{2}\sum_{(u,v)\in E} w_{uv}\sum_{t\in T} n_t^2\left(\frac{\phi_{vt}}{n_v} - \frac{\phi_{ut}}{n_u}\right)^2 \to \max_{\Phi}$$

Виктор Булатов. Использование графовой структуры в тематическом моделировании // Магистерская диссертация, ФИВТ МФТИ, 2016.

Направленные связи

Проблема: квадратичный регуляризатор NetPLSA игнорирует направленность связей $u \to v$.

Предположение: направление связи $u \to v$ означает, что распределение $p(t|v)$ «подчиняется» распределению $p(t|u)$, т. е. тематика вершины u шире, чем тематика вершины v.

Модель iTopicModel. В отличие от NetPLSA, минимизируется не квадратичный критерий, а Дивергенция $\mathrm{KL}\big(p(t|v) \,\big\|\, p(t|u)\big)$:

$$R(\Theta \text{ или } \Phi^2) = \frac{\tau}{2} \sum_{(u,v)\in E} w_{uv} \sum_{t\in T} p(t|v) \ln p(t|u) \to \max,$$

причём $p(t|v)$ можно выразить и через Θ, и через Φ^2.

Yizhou Sun, Jiawei Han, Jing Gao, Yintao Yu. iTopicModel: Information Network-Integrated Topic Modeling. 2009.

Создатель или распространитель контента?

Документ $a \in D$ — все твиты, созданные пользователем a
Документ $b \in D$ — все ретвиты пользователя b
n_a — число сообщений пользователя a
r_b — число ретвитов пользователя b
r_{ab} — сколько раз b сделал ретвит сообщения пользователя a
$\theta_{ta} = p(t|a)$ — тематика a в роли создателя контента
$\theta'_{tb} = p'(t|b)$ — тематика b в роли распространителя контента

Предположения:

- если b ретвитит a, то тематики θ_{ta} и θ'_{tb} близки
- если c ретвитит a и b, то тематики θ_{ta} и θ_{tb} близки
- если a и b ретвитят c, то тематики θ'_{ta} и θ'_{tb} близки

Wayne Xin Zhao, Jinpeng Wang, Yulan He, Jian-Yun Nie, Xiaoming Li. Originator or Propagator? Incorporating Social Role Theory into Topic Models for Twitter Content Analysis. CIKM 2013.

Создатель или распространитель контента?

Меры близости пар пользователей a и b:

$\text{sim}_1(a,b) = \frac{r_{ab}}{n_a + r_b - r_{ab}}$ — как непосредственно взаимодействующих

$\text{sim}_2(a,b) = \frac{\sum_c r_{ac}\, r_{bc}}{\left(\sum_c r_{ac}^2\right)^{1/2}\left(\sum_c r_{bc}^2\right)^{1/2}}$ — как создателей контента

$\text{sim}_3(a,b) = \frac{\sum_c r_{ca}\, r_{cb}}{\left(\sum_c r_{ca}^2\right)^{1/2}\left(\sum_c r_{cb}^2\right)^{1/2}}$ — как распространителей контента

Регуляризаторы:

$$R_1(\Theta) = \tau_1 \sum_{(a,b)} \text{sim}_1(a,b) \sum_{t\in T} \left(\theta_{ta} - \theta'_{tb}\right)^2 \to \max;$$

$$R_2(\Theta) = \tau_2 \sum_{(a,b)} \text{sim}_2(a,b) \sum_{t\in T} \left(\theta_{ta} - \theta_{tb}\right)^2 \to \max;$$

$$R_3(\Theta) = \tau_3 \sum_{(a,b)} \text{sim}_3(a,b) \sum_{t\in T} \left(\theta'_{ta} - \theta'_{tb}\right)^2 \to \max;$$

Переход к модальностям создателей и распространителей

Проблема регуляризатора Θ в пакетном EM-алгоритме: связанные пользователи могут попасть в разные пакеты.

Документ $d \in D$ — отдельный твит, содержащий:
$a_d \in A$ — один терм модальности Φ^A создателя,
$b \in B_d \subset B$ — термы модальности Φ^B распространителей,
$A \equiv B$ — множество всех пользователей социальной сети.

Регуляризаторы над $p(t|a) = \phi^A_{at} \frac{n_a}{n_t}$ и $p(t|b) = \phi^B_{bt} \frac{n_b}{n_t}$:

$$R_1(\Phi) = \tau_1 \sum_{(a,b)} \text{sim}_1(a,b) \sum_{t \in T} \left(\phi^A_{at} \frac{n_a}{n_t} - \phi^B_{bt} \frac{n_b}{n_t} \right)^2 \to \max;$$

$$R_2(\Phi) = \tau_2 \sum_{(a,b)} \text{sim}_2(a,b) \sum_{t \in T} \left(\phi^A_{at} \frac{n_a}{n_t} - \phi^A_{bt} \frac{n_b}{n_t} \right)^2 \to \max;$$

$$R_3(\Phi) = \tau_3 \sum_{(a,b)} \text{sim}_3(a,b) \sum_{t \in T} \left(\phi^B_{at} \frac{n_a}{n_t} - \phi^B_{bt} \frac{n_b}{n_t} \right)^2 \to \max;$$

Резюме

- Регуляризаторы позволяют нацелить тематическую модель на классификацию, регрессию, выявление связей
- Связи могут быть различными:
  — между темами
  — между документами
  — между токенами одной модальности
  — между токенами разных модальностей
- Классы, категории, время, геотеги можно представлять модальностями, а данные о связях — частотами или индикаторами токенов этих модальностей
- Регуляризаторы Θ, не удобные в пакетном алгоритме, можно превращать в регуляризаторы Φ