Методы машинного обучения. Регуляризация в тематическом моделировании

Воронцов Константин Вячеславович
www.MachineLearning.ru/wiki?title=User:Vokov
вопросы к лектору: voron@forecsys.ru

материалы курса:
github.com/MSU-ML-COURSE/ML-COURSE-21-22
орг.вопросы по курсу: ml.cmc@mail.ru

ВМК МГУ • 29 марта 2022

Содержание

Напоминание. Задача тематического моделирования

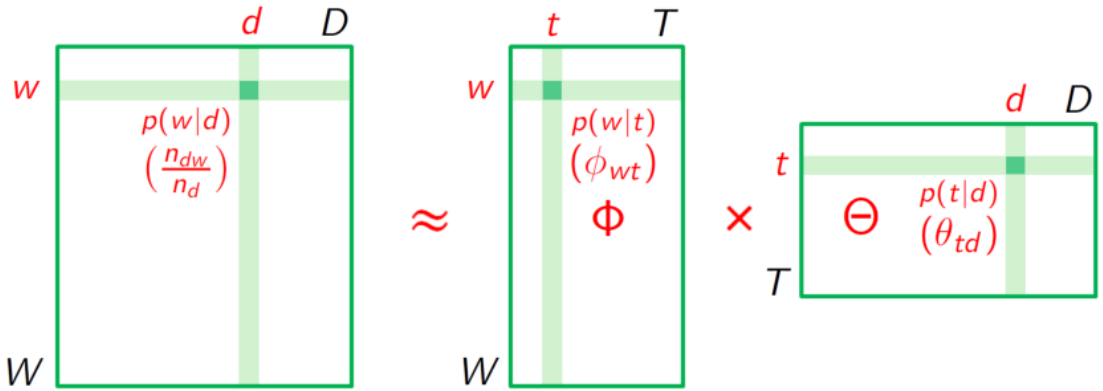
Дано: коллекция текстовых документов, $p(w|d) = \frac{n_{dw}}{n_d}$

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:



Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.

Напоминание. ARTM — аддитивная регуляризация

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

Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН, 2014.

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

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

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

Vorontsov K., Potapenko A. Additive regularization of topic models. 2015.

Обобщение. Латентное вероятностное моделирование

$X = (x_i)_{i=1}^n$ — выборка данных, *наблюдаемые переменные*
$Z = (z_i)_{i=1}^n$ — *скрытые переменные*
Ω — параметры порождающей модели $p(X, Z|\Omega)$

Задача максимизации правдоподобия выборки X:

$$\ln p(X|\Omega) = \ln \sum_Z p(X, Z|\Omega) \;\to\; \max_\Omega$$

Пример 1. Разделение смеси распределений, $z_i \in \{1, \ldots, k\}$:

$$p(X, Z|\Omega) = \prod_{i=1}^n \underbrace{p(x_i|z_i)}_{p(x_i;\theta_j)} \underbrace{p(z_i)}_{w_j,\ j=z_i}, \qquad \Omega = \{w_j, \theta_j : j = 1, \ldots, k\}$$

Пример 2. Вероятностное тематическое моделирование, $z_i \in T$:

$$p(X, Z|\Omega) = \prod_{i=1}^n \underbrace{p(w_i|z_i)}_{\phi_{w_i t}} \underbrace{p(z_i|d_i)}_{\theta_{td_i},\ t=z_i} p(d_i), \qquad \Omega = (\Phi, \Theta)$$

Классическая и байесовская регуляризация

Байесовский вывод апостериорного распределения $p(\Omega|X)$ (обычно приближённый) ради получения точечной оценки Ω:

$$\text{Posterior}(\Omega|X,\gamma) \;\propto\; p(X|\Omega)\,\text{Prior}(\Omega|\gamma)$$
$$\Omega := \arg\max_{\Omega} \text{Posterior}(\Omega|X,\gamma)$$

Максимизация апостериорной вероятности (MAP) даёт точечную оценку Ω напрямую, без вывода Posterior:

$$\Omega := \arg\max_{\Omega} \big(\ln p(X|\Omega) + \ln \text{Prior}(\Omega|\gamma)\big)$$

Многокритериальная аддитивная регуляризация обобщает MAP на любые регуляризаторы и их комбинации:

$$\Omega := \arg\max_{\Omega} \Big(\ln p(X|\Omega) + \sum_{i=1} \tau_i R_i(\Omega)\Big)$$

Напоминание. Общий EM-алгоритм

Теорема. Точка Ω локального максимума регуляризованного *маргинализованного правдоподобия* (Marginal log-Likelihood)

$$\ln \sum_Z p(X, Z|\Omega) + R(\Omega) \;\to\; \max_\Omega \qquad \text{(RML)}$$

удовлетворяет системе уравнений, решение которой методом простых итераций сводится к чередованию двух шагов:

E-шаг: $q(Z) = p(Z|X, \Omega)$

M-шаг: $\sum_Z q(Z) \ln p(X, Z|\Omega) + R(\Omega) \;\to\; \max_\Omega$

Следствие. Значение RML не убывает на каждом EM-шаге

Регуляризованный EM-алгоритм для тематической модели

Для тематической модели: $X=(d_i,w_i)_{i=1}^n$, $Z=(t_i)_{i=1}^n$, $\Omega=(\Phi,\Theta)$

Лемма. Точка (Φ,Θ) локального максимума RML (регуляризованного маргинализованного log-правдоподобия)

$$\ln\sum_Z p(X,Z|\Omega)+R(\Omega)=\sum_{d,w} n_{dw}\ln\sum_t \phi_{wt}\theta_{td}+R(\Phi,\Theta)$$

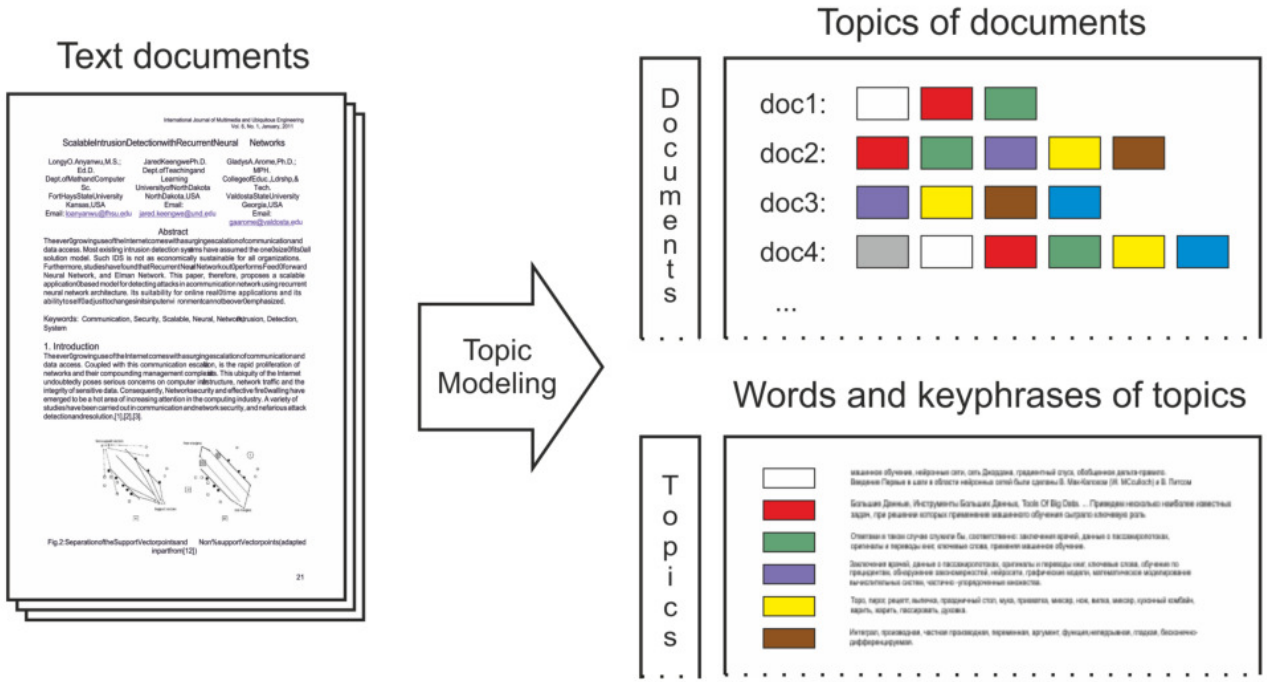
удовлетворяет системе уравнений, решение которой методом простых итераций сводится к чередованию двух шагов:

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

M-шаг: $$\sum_{d,w,t} n_{dw}p(t|d,w)\ln\left(\phi_{wt}\theta_{td}\right)+R(\Phi,\Theta)\ \to\ \max_{\Phi,\Theta}$$

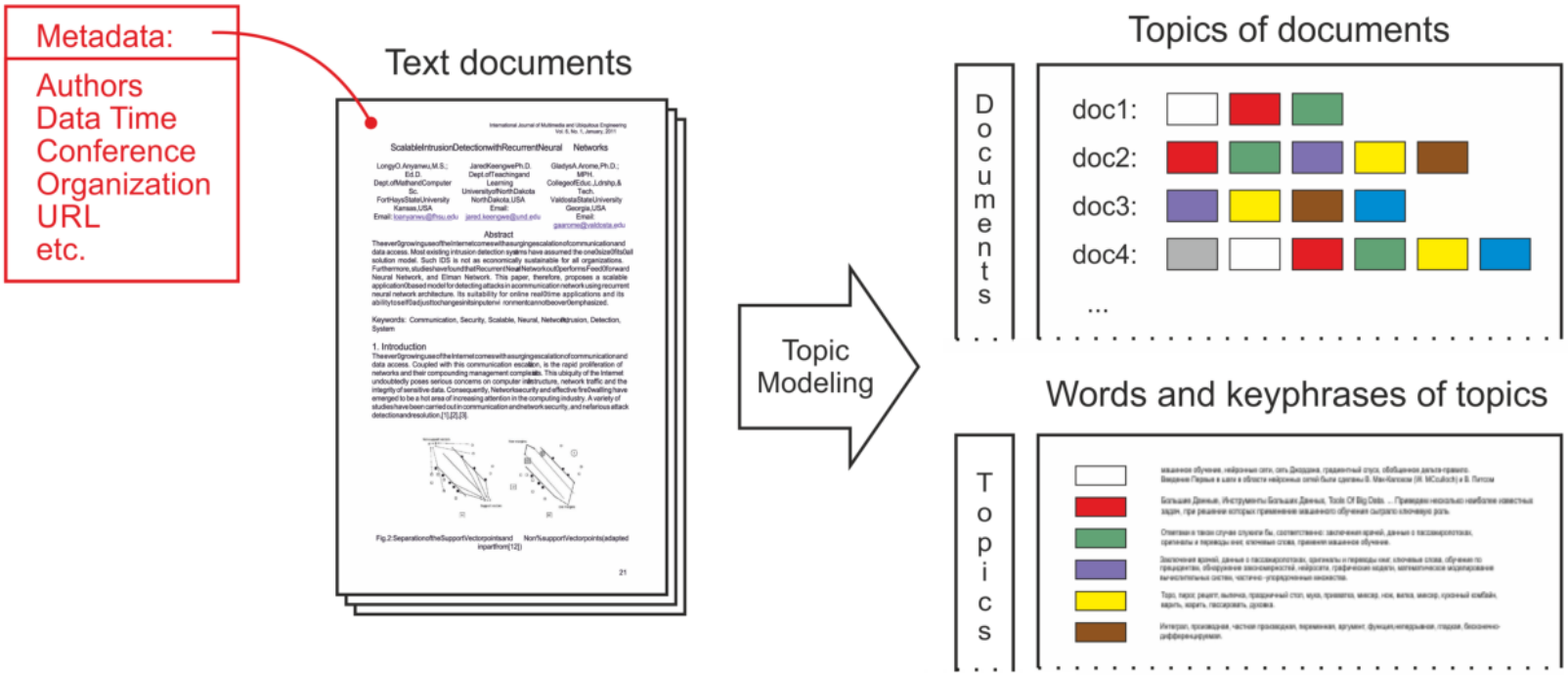
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$,



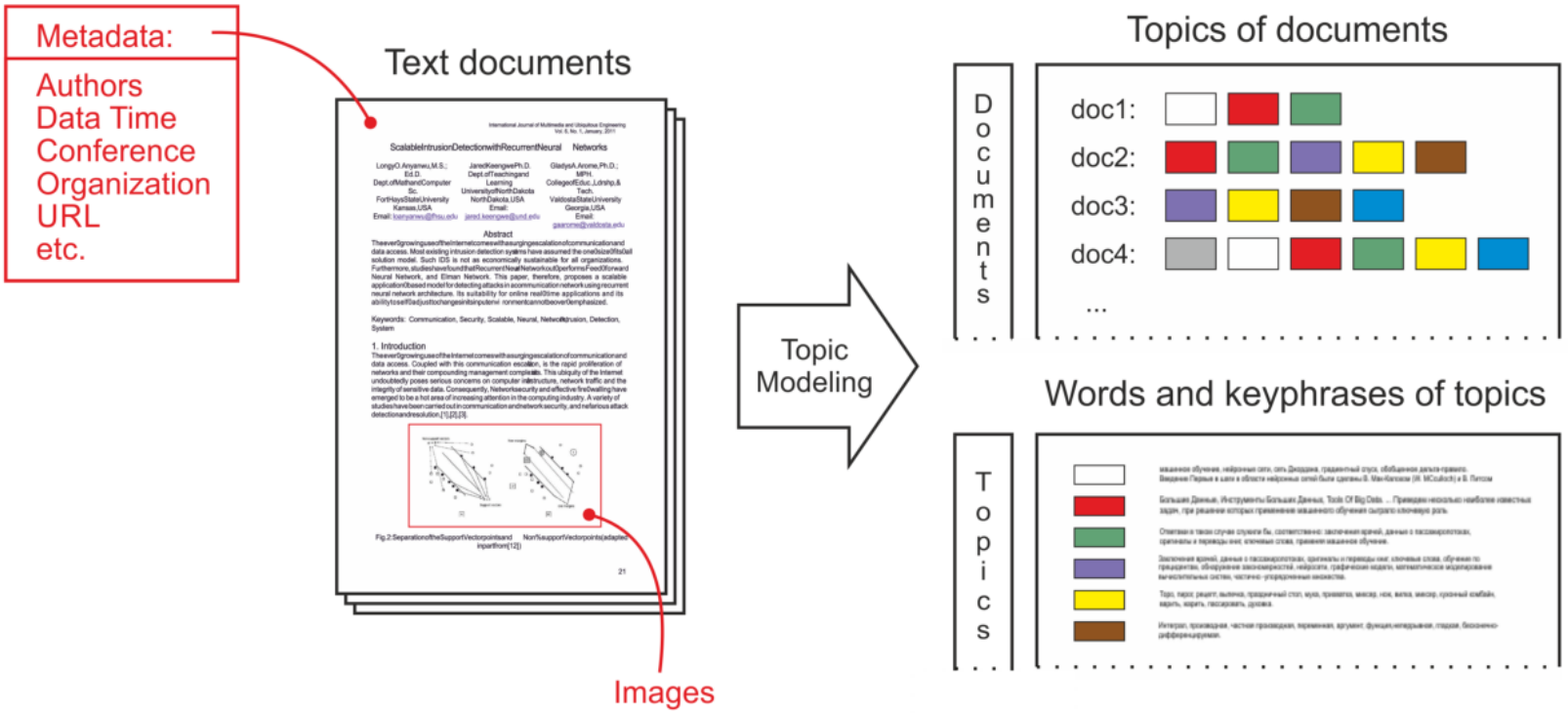
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,



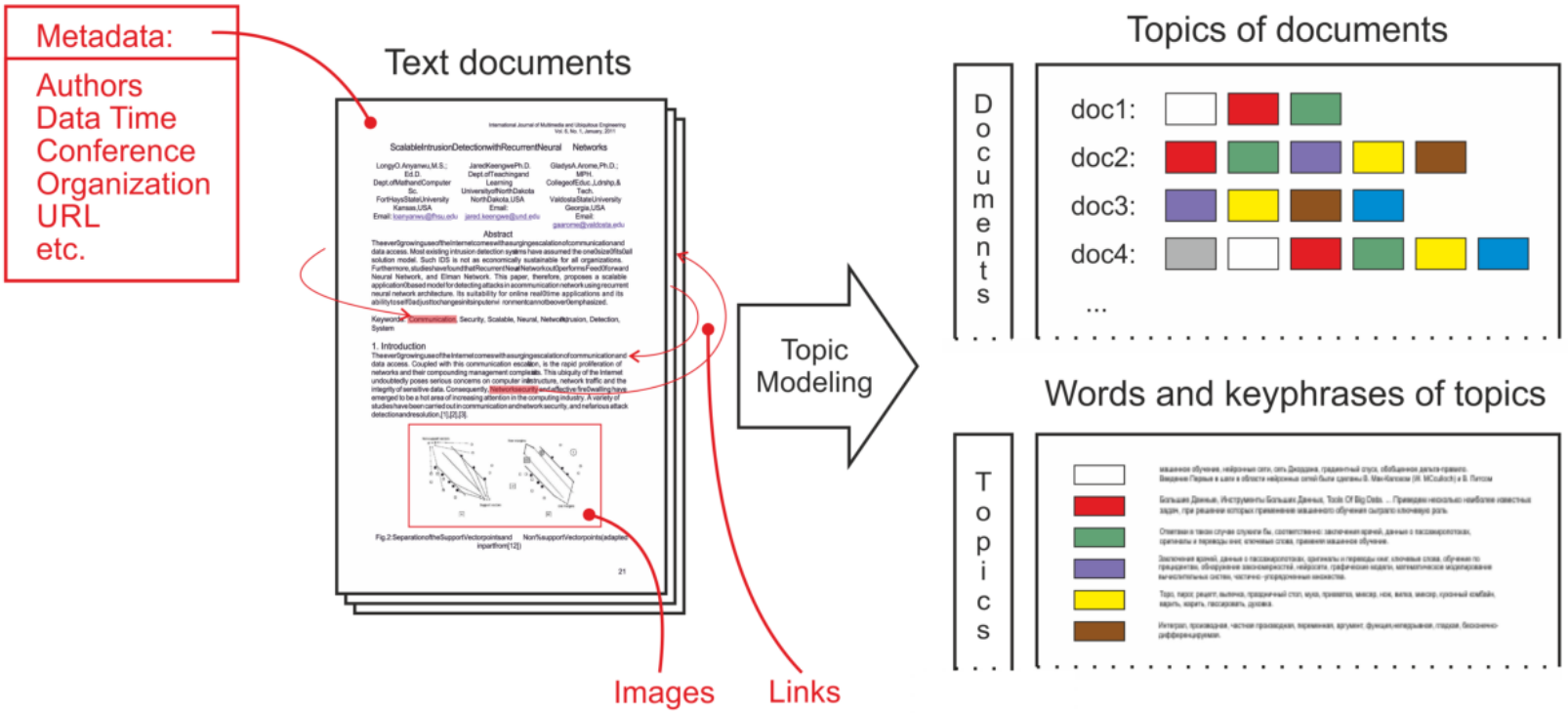
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$,



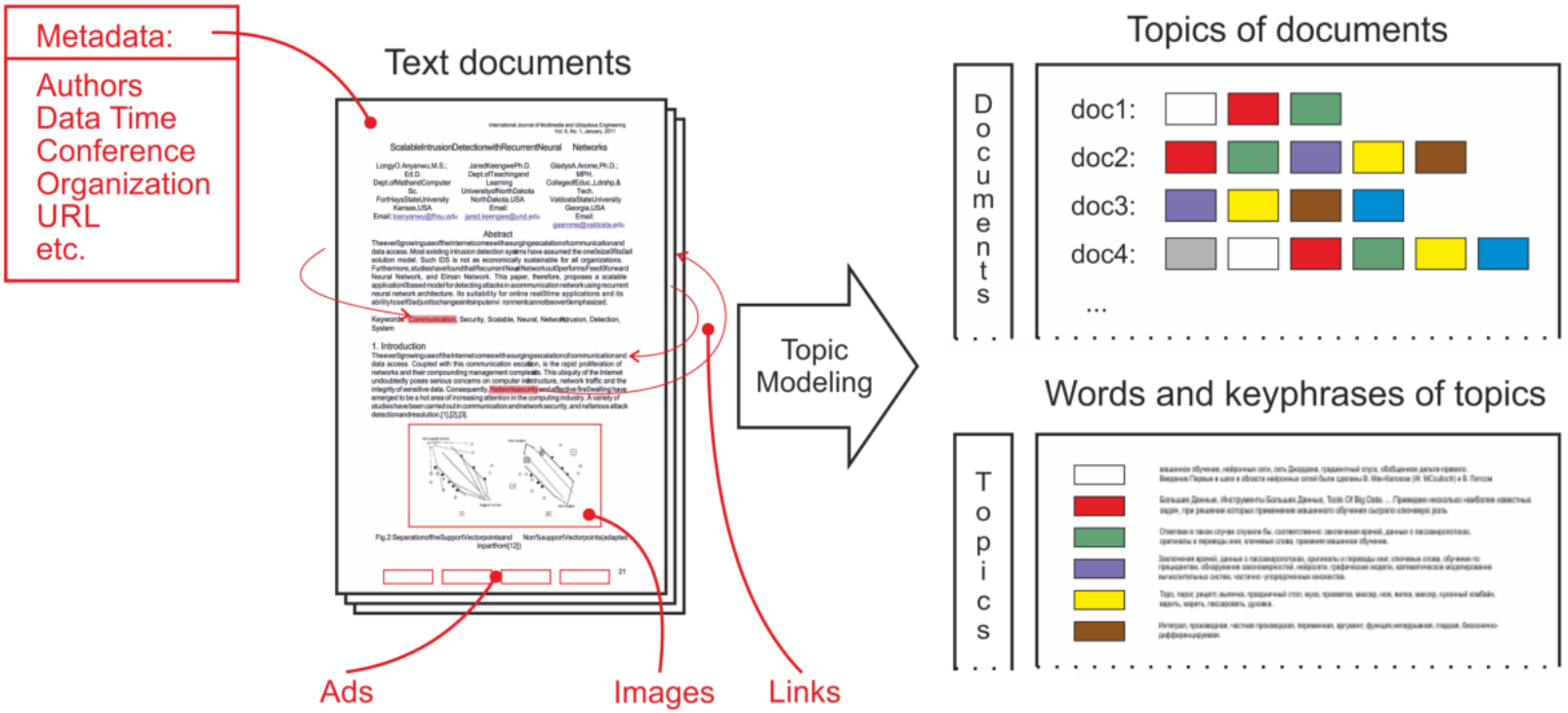
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$,



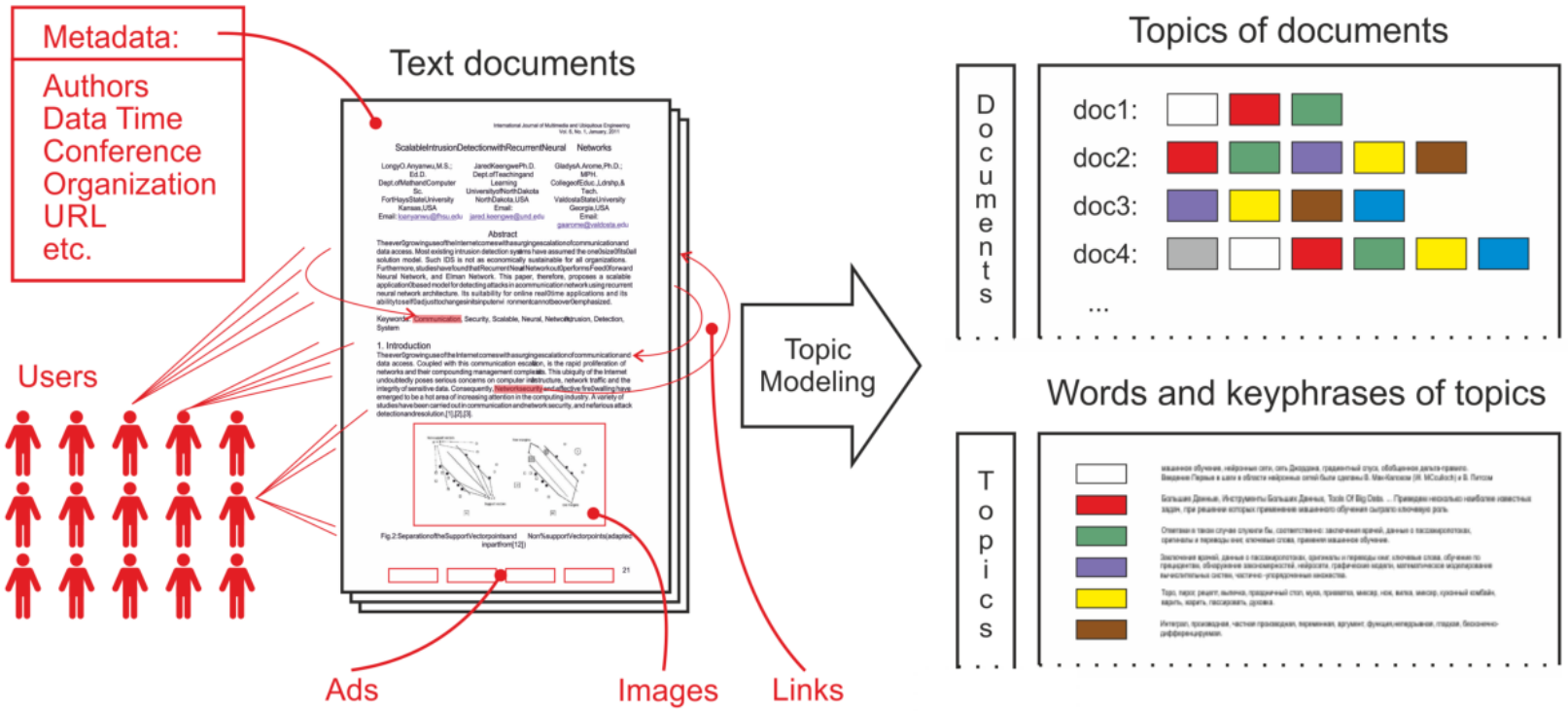
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$,



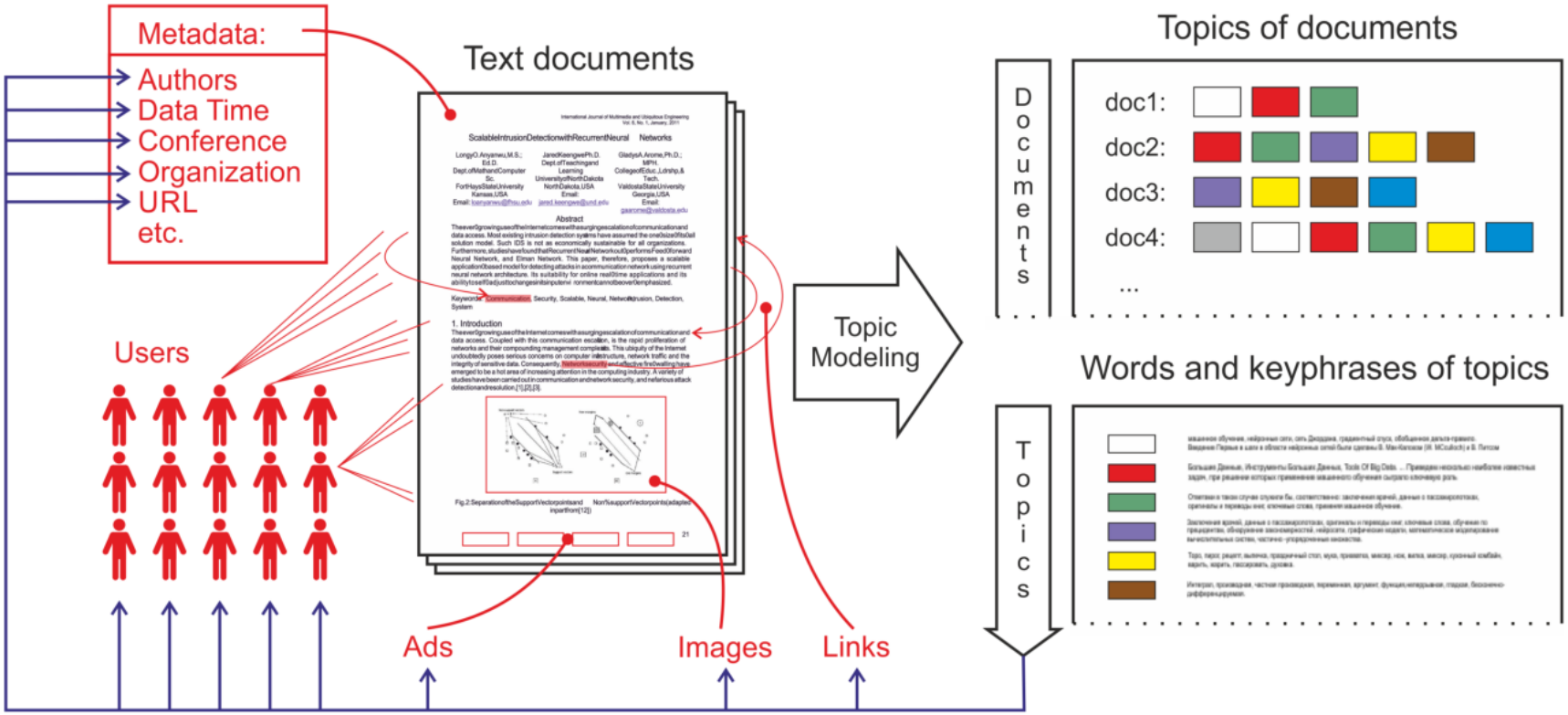
Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{пользователь}|t)$



Мультимодальная тематическая модель

Тема может порождать термы различных *модальностей*:
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
$p(\text{объект}|t)$, $p(\text{ссылка}|t)$, $p(\text{баннер}|t)$, $p(\text{пользователь}|t)$



Мультимодальная ARTM

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

Максимизация суммы log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

K.Vorontsov, O.Frei, M.Apishev et al. Non-Bayesian additive regularization for multimodal topic modeling of large collections. CIKM TM workshop, 2015.

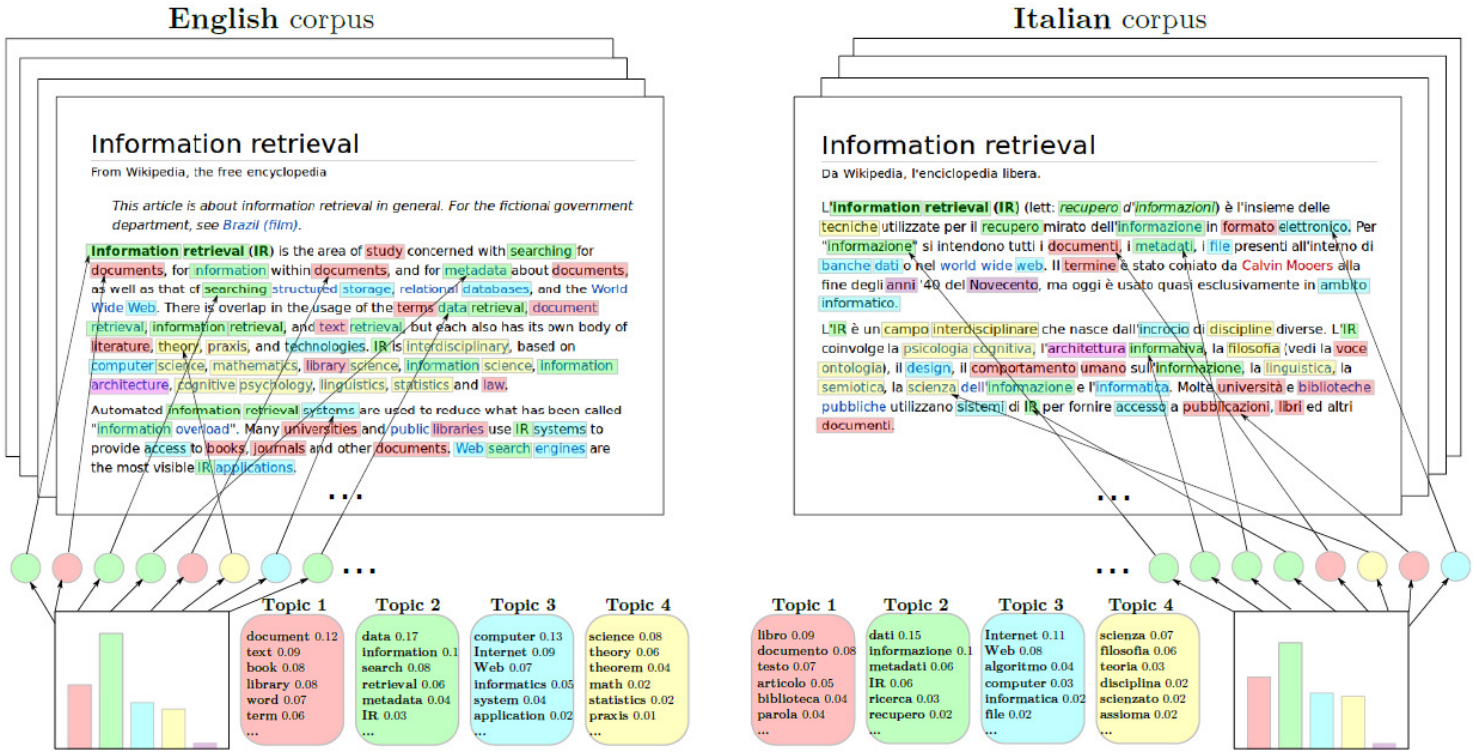
Модальность биграмм улучшает интерпретируемость тем

Коллекция 850 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

С.Стенин. Мультиграммные аддитивно регуляризованные тематические модели. 2015.

Многоязычные модели параллельных коллекций



Для построения многоязычных тем достаточно иметь парные документы, без выравнивания, без двуязычных словарей!

I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. 2015

Пример. Многоязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их вероятности $p(w|t)$ в %:

Тема №68				Тема №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

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open source library for regularized multimodal topic modeling of large collections. AIST-2015.

Пример. Многоязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их вероятности $p(w|t)$ в %:

Тема №88				Тема №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

Асессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open source library for regularized multimodal topic modeling of large collections. AIST-2015.

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. 2012.

Регуляризатор для задач регрессии

$y_d \in \mathbb{R}$ для всех документов d — обучающие данные.

$E(y|d) = \sum_{t \in T} v_t \theta_{td}$ — 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:

$$\theta_{td} = \underset{t \in T}{\text{norm}} \left(n_{td} + \tau v_t \theta_{td} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right) \right);$$

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

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

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.

Проблема коротких текстов

Короткие тексты (short text):

- Twitter и другие микроблоги
- социальные медиа
- заголовки статей и новостных сообщений

Тривиальные подходы:

- считать каждое сообщение отдельным документом
- разреживать $p(t|d)$ вплоть до единственной темы
- объединить сообщения по автору/времени/региону/и т. п.
- объединить посты с комментариями
- дополнить коллекцию длинными текстами (Википедия и др.)

Более интересная идея:

- использовать сочетаемость пар слов в сообщениях

Битермы: модель сочетаемости слов в коротких текстах

Битерм — пара слов, встречающихся рядом:
в одном коротком сообщении / предложении / окне $\pm h$ слов.

Тематическая модель битермов (Biterm Topic Model):

$$p(u,v) = \sum_{t\in T} p(u|t)p(v|t)p(t) = \sum_{t\in T} \phi_{ut}\phi_{vt}\pi_t,$$

где $\phi_{wt} = p(w|t)$, $\pi_t = p(t)$ — параметры модели.

Критерий максимума логарифма правдоподобия:

$$\sum_{u,v} n_{uv} \ln \sum_t \phi_{ut}\phi_{vt}\pi_t \;\to\; \max_{\Phi,\pi}$$

$$\phi_{vt} \geqslant 0; \quad \textstyle\sum_v \phi_{vt} = 1; \qquad \pi_t \geqslant 0; \quad \textstyle\sum_t \pi_t = 1$$

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng. A Biterm Topic Model for Short Texts. WWW 2013.

Модель сети слов WNTM для коротких текстов

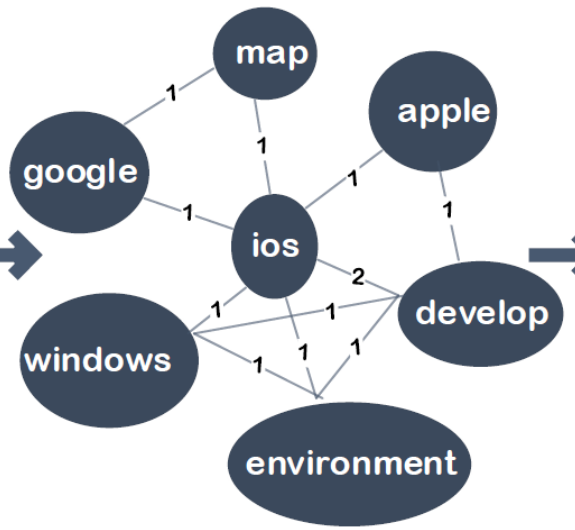
Идея: моделировать не документы, а связи между словами.

d_u — псевдо-документ, объединение всех контекстов слова u.

n_{uw} — число вхождений слова w в псевдо-документ d_u.

Контекст — короткое сообщение / предложение / окно $\pm h$ слов.

doc1	Google Map for iOS
doc2	iOS is developed by Apple
doc3	iOS develop environment for Windows



word	pseudo-document
map	google ios
google	map ios
apple	develop ios
ios	google map apple windows environment develop develop
windows	environment ios develop
develop	apple ios ios windows environment
environment	ios windows develop

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Модели WNTM (Word Network) и WTM (Word Topic Model)

Тематическая модель контекстов, разложение $W \times W$-матрицы:

$$p(w|d_u) = \sum_{t \in T} p(w|t)p(t|d_u) = \sum_{t \in T} \phi_{wt}\theta_{tu},$$

где d_u — псевдо-документ слова u.

Максимизация логарифма правдоподобия:

$$\sum_{u,w \in W} n_{uw} \log \sum_{t \in T} \phi_{wt}\theta_{tu} \to \max_{\Phi,\Theta},$$

где n_{uw} — частота сочетания пары слов (w, u).

Отличие: BitermTM симметрична, WNTM несимметрична

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Berlin Chen. Word Topic Models for spoken document retrieval and transcription. ACM Trans., 2009.

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.

Повышаем сходство (скалярные произведения) тематических векторных представлений связанных документов θ_d, θ_c:

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

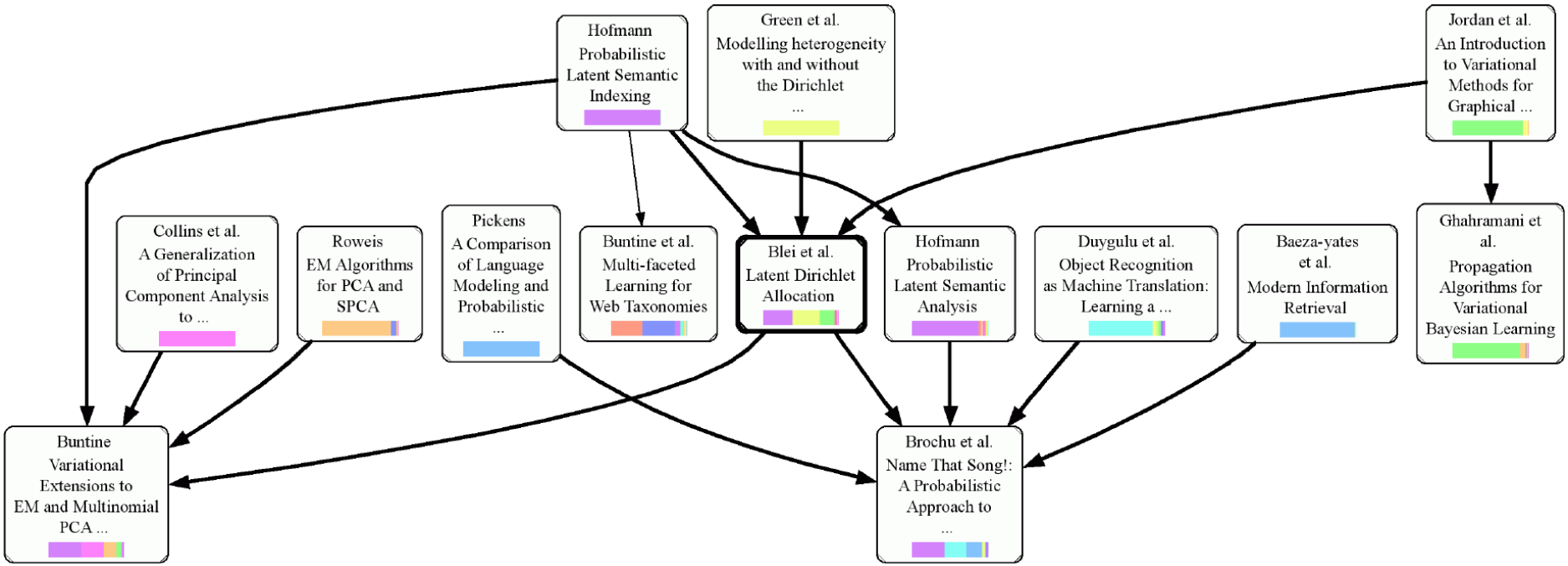
Подставляем, получаем ещё один вариант сглаживания:

$$\theta_{td} = \operatorname*{norm}_t \left(n_{td} + \tau \theta_{td} \sum_{c \in D} n_{dc} \theta_{tc} \right).$$

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

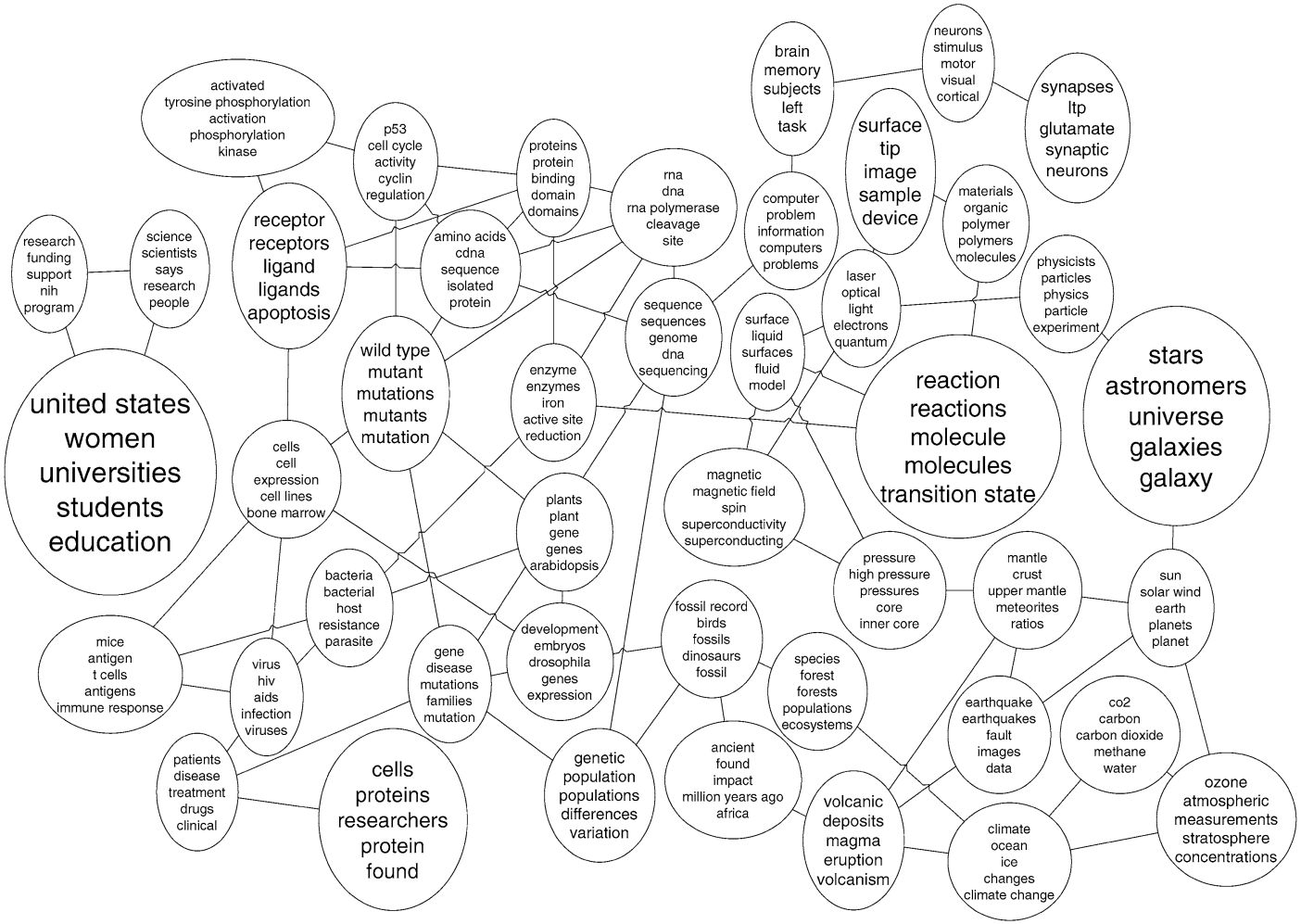
Модели, учитывающие цитирования или гиперссылки

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



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

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».

Выявление корреляций полезно для понимания структуры тем и может улучшать распределения $p(t|d)$.

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 $\eta_d = (\eta_{td})_{t\in T}$ — векторы документов, $\eta_{td} = \ln\theta_{td}$.

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):

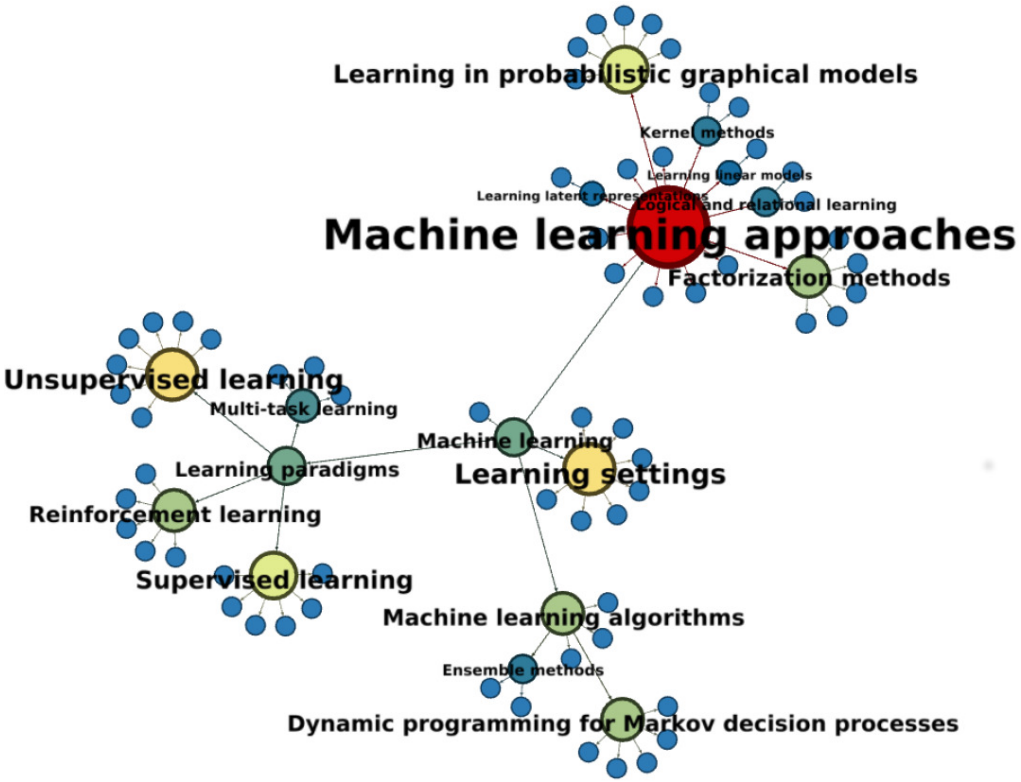
$$\theta_{td} = \underset{t \in T}{\operatorname{norm}} \Big(n_{td} - \tau \sum_{s \in T} S^{-1}_{ts} \big(\ln \theta_{sd} - \mu_s \big) \Big);$$

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

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

Иерархические тематические модели

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



Послойное построение тематической иерархии

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:

$$\sum_{t\in T} n_{wt} \ln p(w|t) = \sum_{t\in T} n_{wt} \ln \sum_{s\in S} p(w|s)p(s|t) \;\to\; \max_{\Phi,\Psi},$$

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

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

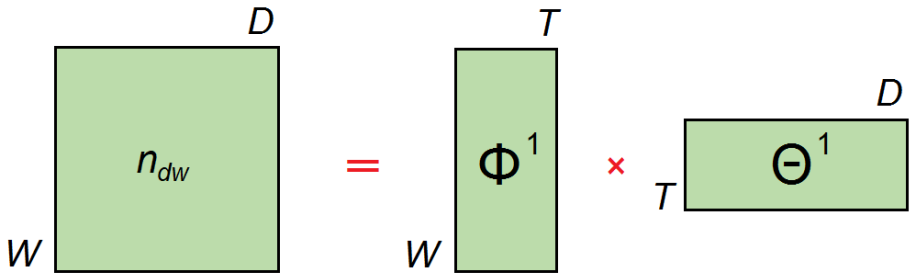
$$R(\Phi,\Psi) = \tau \sum_{t\in T}\sum_{w\in W} n_{wt} \ln \sum_{s\in S} \phi_{ws}\psi_{st} \;\to\; \max.$$

Родительские темы t — *псевдо-документы* с частотами слов n_{wt}.

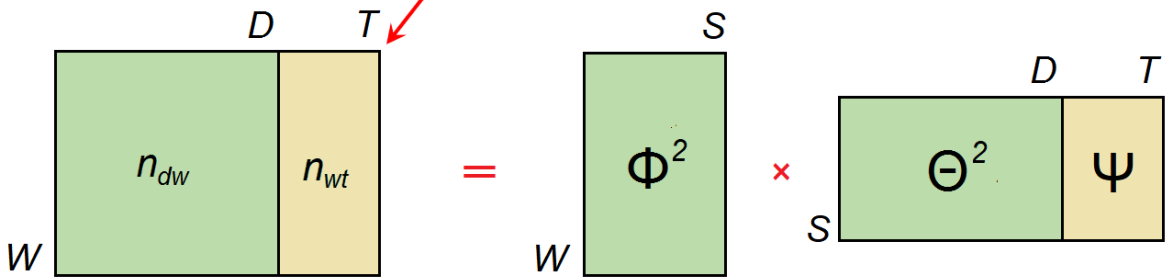
Построение второго уровня иерархии с подтемами S

В коллекцию добавляются $|T|$ псевдодокументов родительских тем с частотами термов $n_{wt} = \tau n_t \phi_{wt}$, $t \in T$

родительский уровень



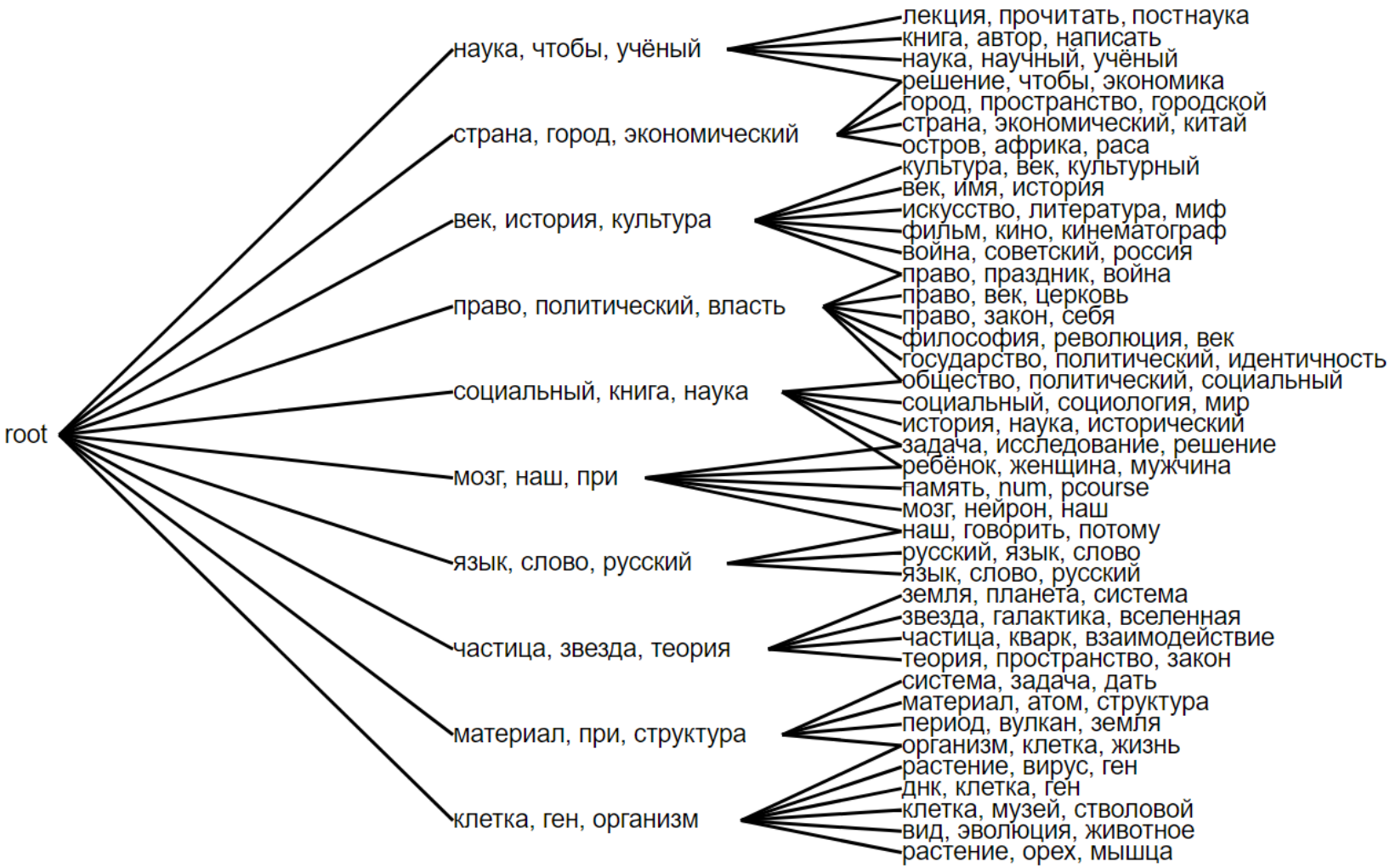
дочерний уровень



Матрица связей тем с подтемами $\Psi = \big(p(s|t)\big)$ образуется в столбцах матрицы Θ, соответствующих псевдодокументам.

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

Иерархический спектр тем (пример на коллекции postnauka.ru)



Д.Федоряка. Технология интерактивной визуализации тематических моделей. 2017.

Резюме

- *EM-алгоритм* — мощный инструмент вероятностного моделирования с латентными (скрытыми) переменными
- *EM-алгоритм* — основной в тематическом моделировании
- *Регуляризация* вводит в модель разнообразные дополнительные требования и/или источники данных
- *Аддитивная регуляризация* — комбинирование моделей
- *Байесовское обучение* часто используется в TM, но на практике оно избыточно: нужны точечные оценки, а не апостериорные распределения
- *Лемма о максимизации на симплексах* применима за пределами TM для оптимизации моделей с дискретными вероятностными распределениями

Asuncion A. et al. On smoothing and inference for topic models. 2009.

Jordan Boyd-Graber. Applications of Topic Models. 2017.

Воронцов К.В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM. 2017–2022.
http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf