Переосмысление вероятностных тематических моделей с позиций классической не-байесовской регуляризации

Воронцов Константин Вячеславович
(проф. ВМК МГУ, проф. МФТИ, г.н.с. ФИЦ ИУ РАН)

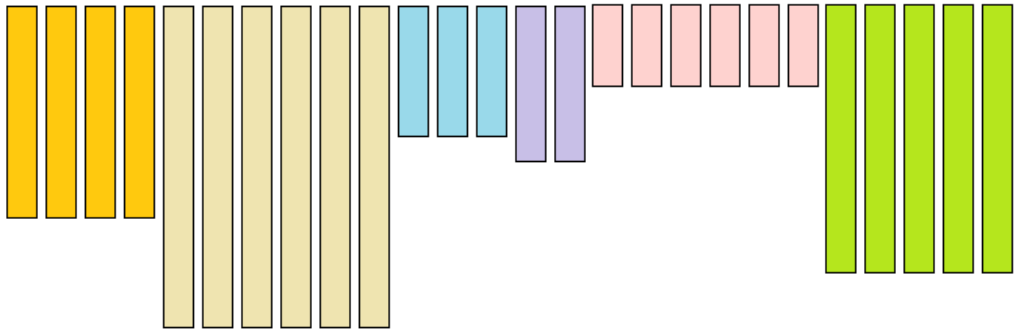
Научная конференция «Анализ данных и оптимизация»
МФТИ • 30 января 2023

Содержание

Задача максимизации функции на единичных симплексах

Пусть $\Omega = (\omega_j)_{j \in J}$ — набор нормированных неотрицательных векторов $\omega_j = (\omega_{ij})_{i \in I_j}$ различных размерностей $|I_j|$:

$$\Omega = \left(\quad \right)$$



Задача максимизации функции $f(\Omega)$ на единичных симплексах:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad j \in J; \\ \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

Необходимые условия экстремума и метод простых итераций

Операция нормировки вектора: $p_i = \underset{i \in I}{\text{norm}}(x_i) = \dfrac{\max(x_i, 0)}{\sum_k \max(x_k, 0)}$

Лемма. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω. Если ω_j — вектор локального экстремума задачи $f(\Omega) \to \max$ и $\exists i$: $\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} > 0$, то ω_j удовлетворяет системе уравнений

$$\omega_{ij} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right).$$

- Численное решение системы — методом простых итераций
- Векторы $\omega_j = 0$ отбрасываются как вырожденные решения
- Итерации похожи на градиентную оптимизацию:

$$\omega_{ij} := \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}},$$

но учитывают ограничения и не требуют подбора шага η

Доказательство леммы о максимизации на симплексах

Задача: $f(\Omega) \to \max\limits_{\Omega}$; $\sum\limits_{i \in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J.$

Функция Лагранжа:

$$\mathscr{L}(\Omega; \mu, \lambda) = f(\Omega) + \sum_{j \in J} \lambda_j \Big(\sum_{i \in I_j} \omega_{ij} - 1 \Big) - \sum_{j \in J} \sum_{i \in I_j} \mu_{ij} \omega_{ij}.$$

Условия Каруша–Куна–Таккера для вектора ω_j:

$$\frac{\partial f(\Omega)}{\partial \omega_{ij}} = \lambda_j - \mu_{ij}, \quad \mu_{ij} \omega_{ij} = 0, \quad \mu_{ij} \geqslant 0.$$

Умножим обе части первого равенства на ω_{ij}:

$$A_{ij} \equiv \omega_{ij} \frac{\partial f(\Omega)}{\partial \omega_{ij}} = \omega_{ij} \lambda_j.$$

Согласно условию леммы $\exists i$: $A_{ij} > 0$. Значит, $\lambda_j > 0$.

Если $\frac{\partial f(\Omega)}{\partial \omega_{ij}} < 0$ для некоторого i, то $\mu_{ij} > 0 \;\Rightarrow\; \omega_{ij} = 0$.

Тогда $\omega_{ij} \lambda_j = (A_{ij})_+$; $\lambda_j = \sum\limits_i (A_{ij})_+ \;\Rightarrow\; \omega_{ij} = \underset{i}{\operatorname{norm}}(A_{ij})$.

■

Теорема о сходимости итерационного процесса

$$\omega_{ij}^{t+1} = \underset{i \in I_j}{\text{norm}} \left(\omega_{ij}^{t} \frac{\partial f(\Omega^t)}{\partial \omega_{ij}^{t}} \right)$$

Теорема. Пусть $f(\Omega)$ — ограниченная сверху, непрерывно дифференцируемая функция, и все Ω^t, начиная с некоторой итерации t^0 обладают свойствами:

- $\forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^{t} = 0 \to \omega_{ij}^{t+1} = 0$ (сохранение нулей)
- $\exists \varepsilon > 0 \;\; \forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^{t} \notin (0, \varepsilon)$ (отделимость от нуля)
- $\exists \delta > 0 \;\; \forall j \in J \;\; \exists i \in I_j \;\; \omega_{ij}^{t} \frac{\partial f(\Omega^t)}{\partial \omega_{ij}} \geqslant \delta$ (невырожденность)

Тогда $f(\Omega^{t+1}) > f(\Omega^t)$ и $\left|\omega_{ij}^{t+1} - \omega_{ij}^{t}\right| \to 0$ при $t \to \infty$.

Ирхин И. А., Воронцов К. В. Сходимость алгоритма аддитивной регуляризации тематических моделей // Труды Института математики и механики УрО РАН. 2020.

Постановка задачи тематического моделирования

Дано:

- W — конечное множество (словарь) термов (слов, токенов)
- D — конечное множество (коллекция) документов
- n_{dw} — частота терма $w \in W$ в документе $d \in D$

Найти: вероятностную тематическую модель

$$p(w|d) = \sum_{t \in T} p(w \mid \cancel{d}, t)\, p(t|d) = \sum_{t \in T} \varphi_{wt} \theta_{td}$$

где $\varphi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$ — параметры модели

Критерий: максимум логарифма правдоподобия

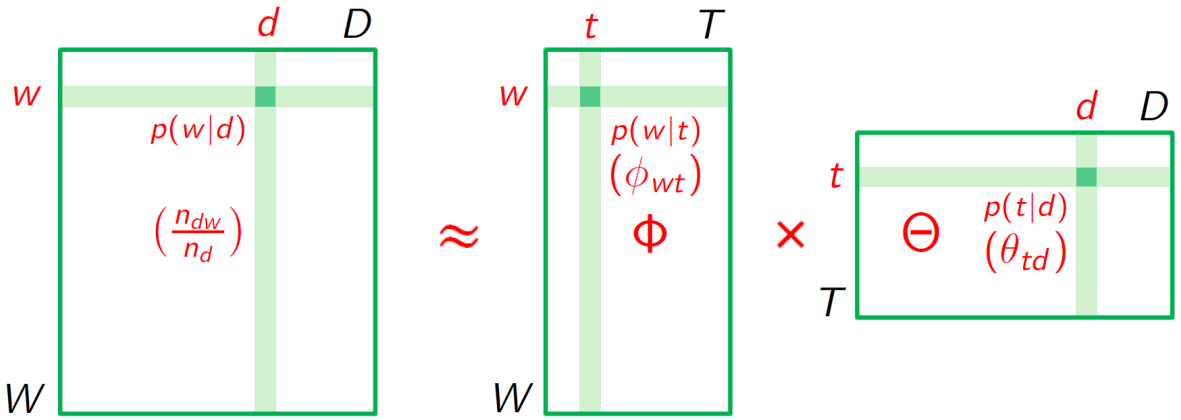
$$L(\Phi, \Theta) = \ln \prod_{d,w} p(w|d)^{n_{dw}} = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

при ограничениях $\varphi_{wt} \geqslant 0$, $\sum_w \varphi_{wt} = 1$, $\theta_{td} \geqslant 0$, $\sum_t \theta_{td} = 1$

Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.

Некорректно поставленная задача матричного разложения

Низкоранговое стохастическое матричное разложение:



Если Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$, $\operatorname{rank} S = |T|$
- $L(\Phi', \Theta') = L(\Phi, \Theta)$ — линейно не зависимые решения
- $L(\Phi', \Theta') \geqslant L(\Phi, \Theta) - \varepsilon$ — приближённые решения

Регуляризация — стандартный приём доопределения решения с помощью добавления дополнительных критериев.

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \qquad R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta)$$

EM-алгоритм: метод простой итерации для системы уравнений

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

$$\text{M-шаг:} \quad \varphi_{wt} = \underset{w \in W}{\text{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} \right), \quad n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}$$

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

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

Доказательство (по лемме о максимизации на симплексах)

Применим лемму к log-правдоподобию с регуляризатором:

$$f(\Phi,\Theta) = \sum_{d,w} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

$$\varphi_{wt} = \operatorname*{norm}_{w\in W}\left(\varphi_{wt}\frac{\partial f}{\partial \varphi_{wt}}\right) = \operatorname*{norm}_{w\in W}\left(\varphi_{wt}\sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right) =$$

$$= \operatorname*{norm}_{w\in W}\left(\sum_{d\in D} n_{dw} p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right);$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\left(\theta_{td}\frac{\partial f}{\partial \theta_{td}}\right) = \operatorname*{norm}_{t\in T}\left(\theta_{td}\sum_{w\in W} n_{dw}\frac{\varphi_{wt}}{p(w|d)} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) =$$

$$= \operatorname*{norm}_{t\in T}\left(\sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right).$$

■

Две наиболее известные модели — частные случаи ARTM

PLSA: probabilistic latent semantic analysis [Hofmann, 1999] (вероятностный латентный семантический анализ):

$$R(\Phi, \Theta) = 0.$$

M-шаг — частотные оценки условных вероятностей:

$$\varphi_{wt} = \underset{w}{\text{norm}}\big(n_{wt}\big), \qquad \theta_{td} = \underset{t}{\text{norm}}\big(n_{td}\big).$$

LDA: latent Dirichlet allocation (латентное размещение Дирихле):

$$R(\Phi, \Theta) = \sum_{t,w} \beta_w \ln \varphi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}.$$

M-шаг — частотные оценки с поправками $\beta_w > -1$, $\alpha_t > -1$:

$$\varphi_{wt} = \underset{w}{\text{norm}}\big(n_{wt} + \beta_w\big), \qquad \theta_{td} = \underset{t}{\text{norm}}\big(n_{td} + \alpha_t\big).$$

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet allocation. NIPS 2001.

Байесовская и классическая регуляризация

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

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

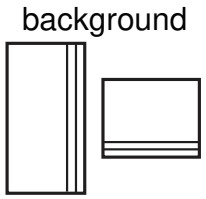
Максимизация апостериорной вероятности (MAP) даёт точечную оценку Ω напрямую, без вывода Posterior:

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

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

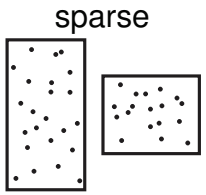
$$\Omega := \arg\max_{\Omega} \Big(\ln p(X|\Omega) + \sum_{i=1} \tau_i R_i(\Omega)\Big)$$

Регуляризаторы для улучшения интерпретируемости тем



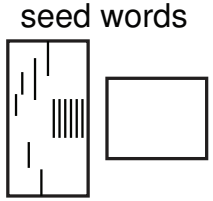
Сглаживание фоновых тем $B \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_w \beta_w \ln \varphi_{wt} + \alpha_0 \sum_d \sum_{t \in B} \alpha_t \ln \theta_{td}$$

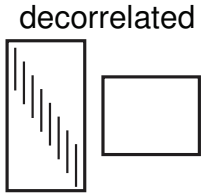


Разреживание предметных тем $S = T \setminus B$:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_w \beta_w \ln \varphi_{wt} - \alpha_0 \sum_d \sum_{t \in S} \alpha_t \ln \theta_{td}$$

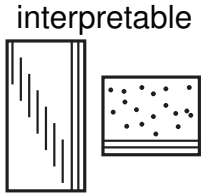


Сглаживание для выделения релевантных тем
с помощью словаря «затравочных» ключевых слов



Декоррелирование для повышения различности тем:

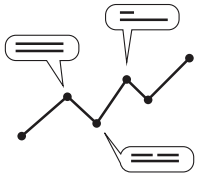
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_w \varphi_{wt} \varphi_{ws}$$



Сглаживание + разреживание + декоррелирование
для улучшения интерпретируемости тем

Регуляризаторы для учёта дополнительной информации

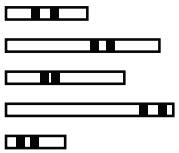
regression



Линейная модель регрессии $\hat{y}_d = \langle v, \theta_d \rangle$ документов:

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

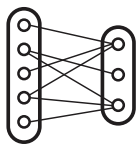
biterm



Связи сочетаемости слов (n_{uv} — частота битерма):

$$R(\Phi) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} n_t \varphi_{ut} \varphi_{vt}$$

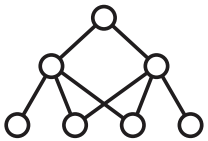
relational



Связи или ссылки между документами:

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

hierarchy

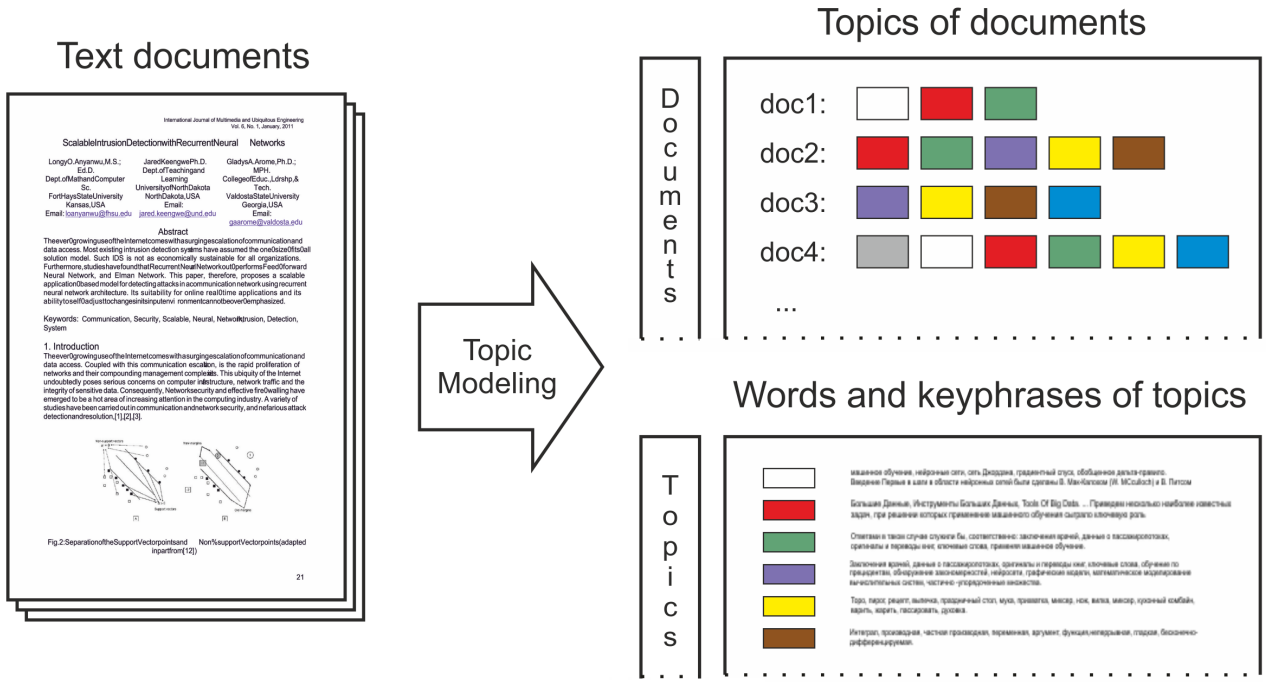


Связи родительских тем t с дочерними подтемами s:

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

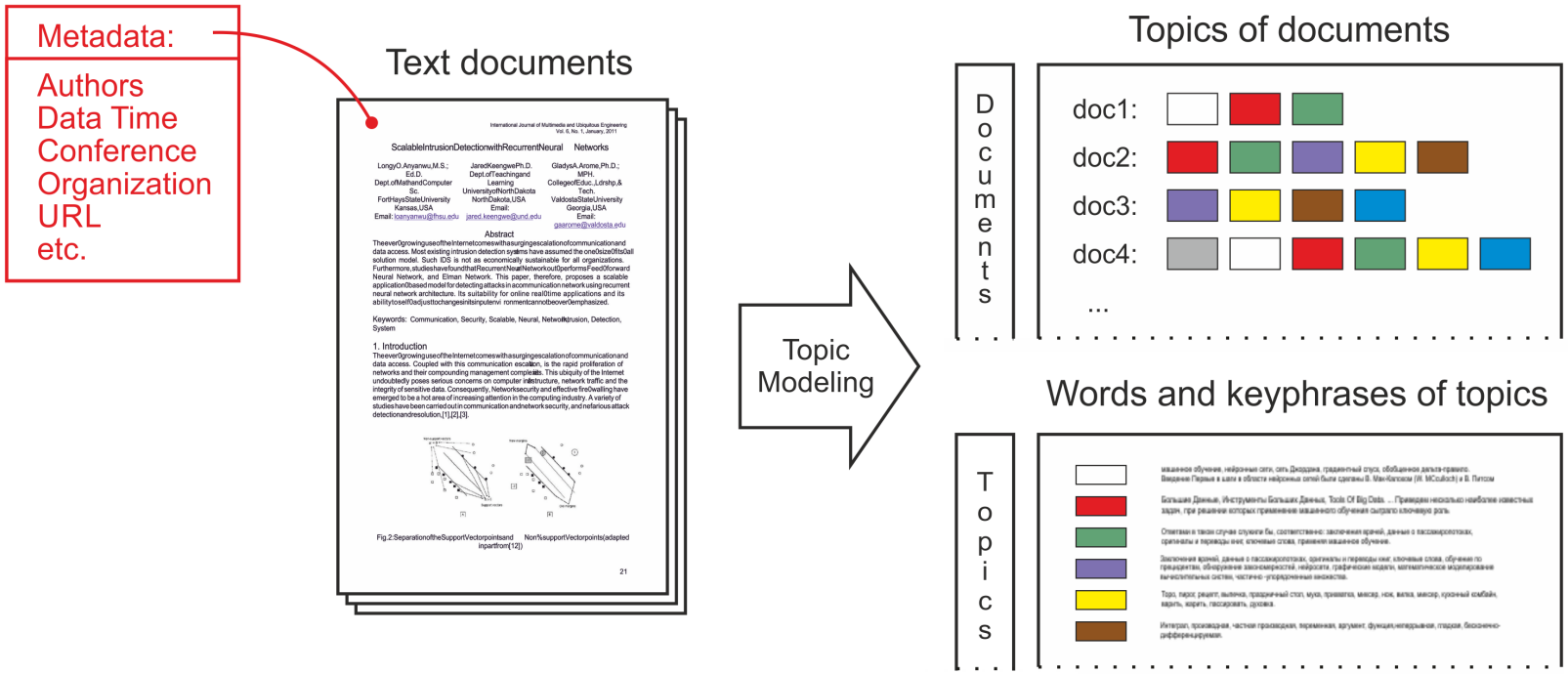
Мультимодальная тематическая модель

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



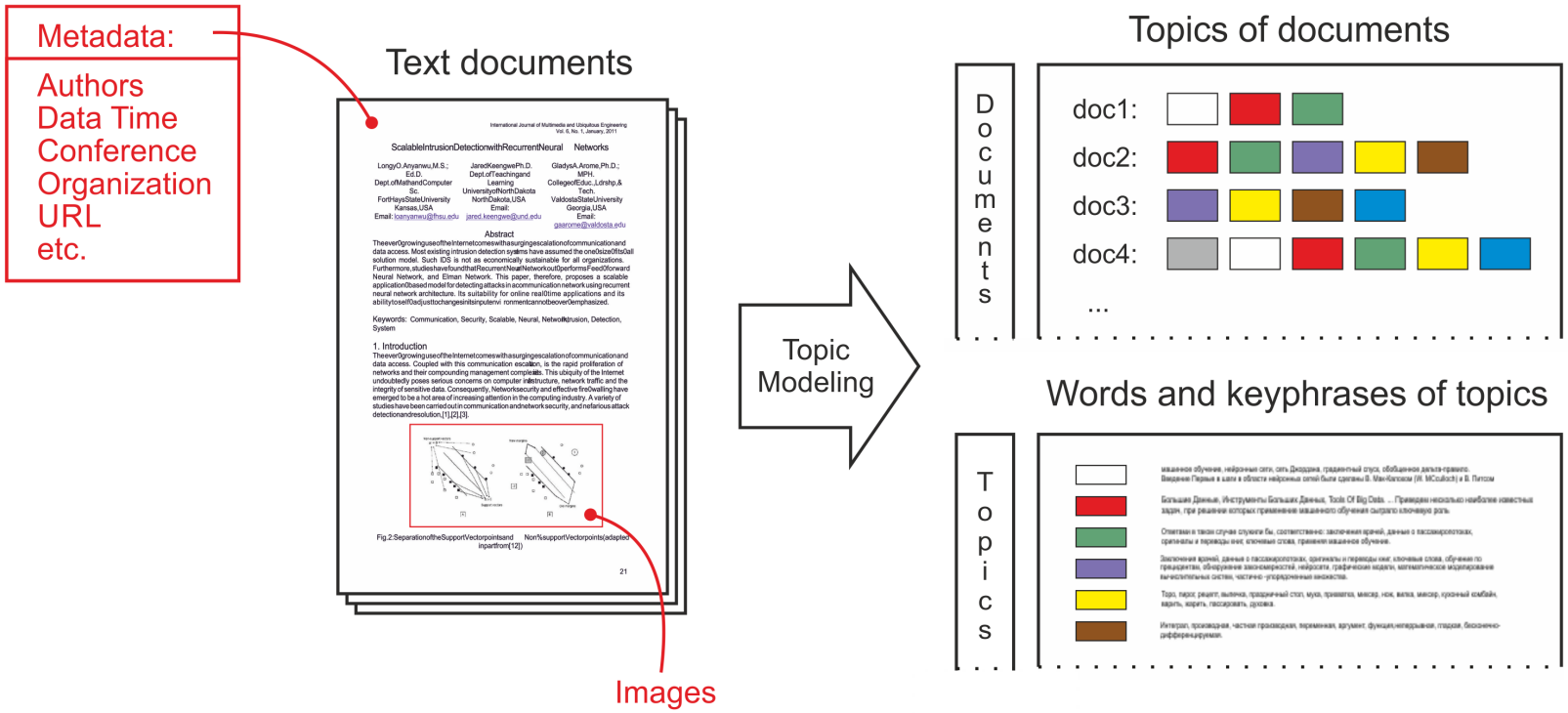
Мультимодальная тематическая модель

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



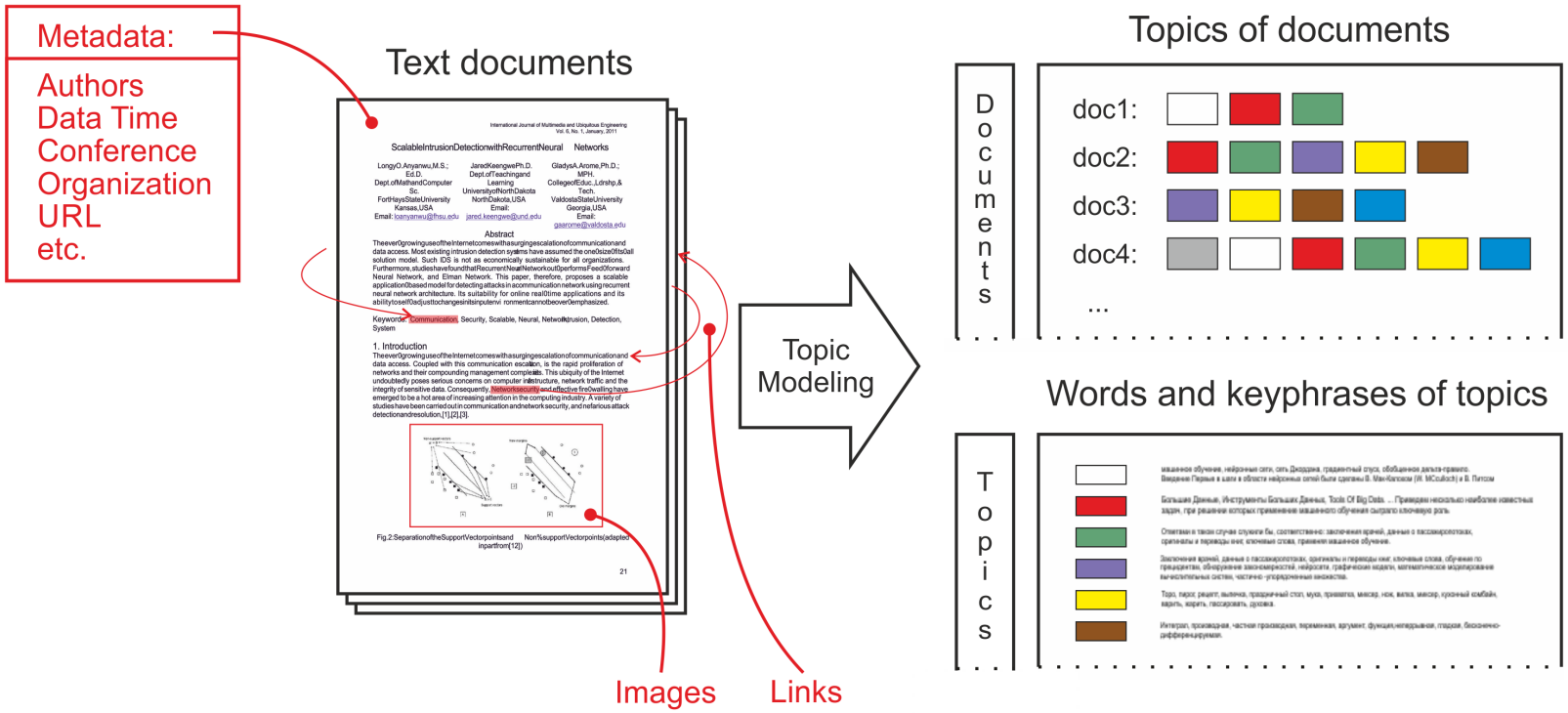
Мультимодальная тематическая модель

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



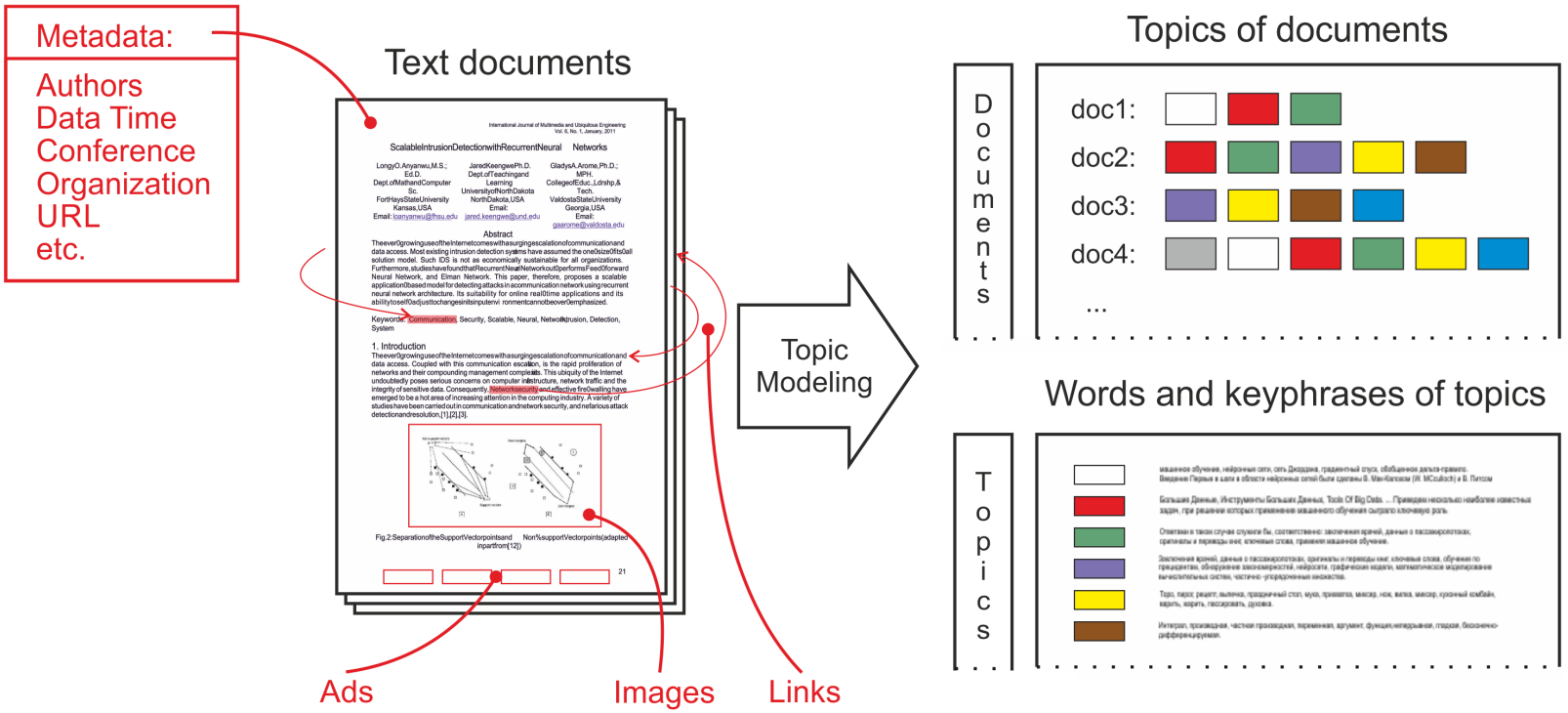
Мультимодальная тематическая модель

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



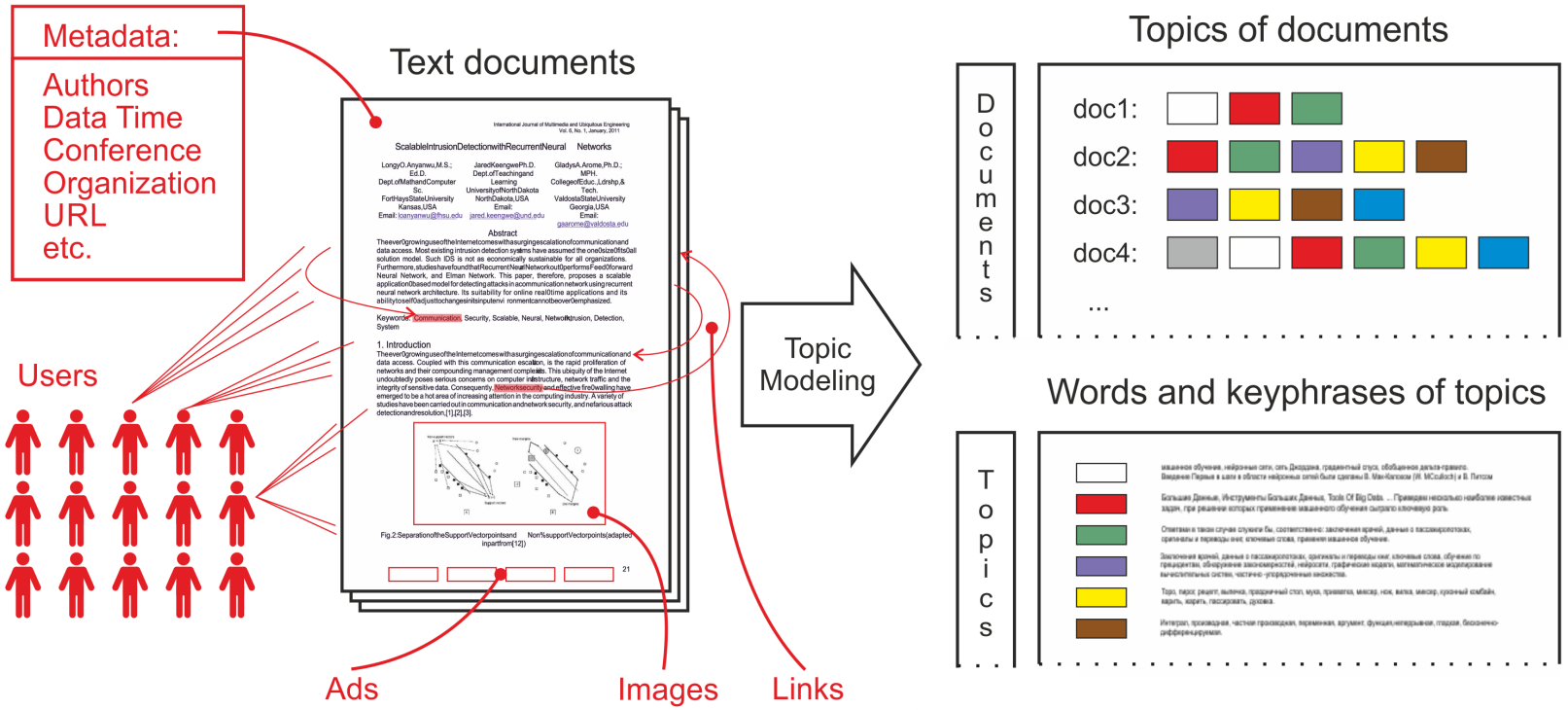
Мультимодальная тематическая модель

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



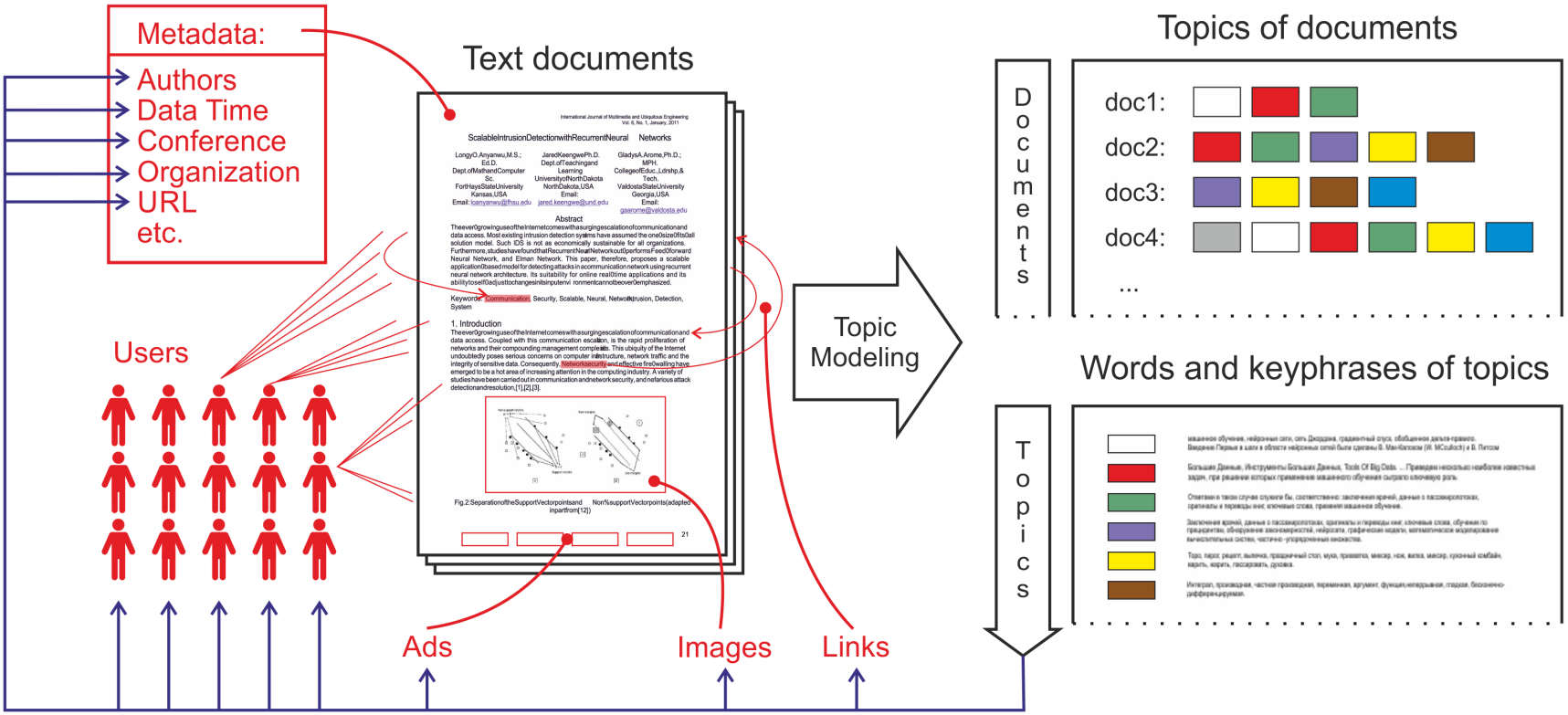
Мультимодальная тематическая модель

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 тематическая модель ARTM

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

Максимизация суммы log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

Пример 1. Мультиязычная тематическая модель Википедии

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

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример 1. Мультиязычная тематическая модель Википедии

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

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример 2. Биграммы улучшают интерпретируемость тем

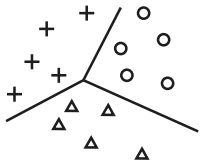
Коллекция 1000 статей конференций ММРО, ИОИ на русском

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

Сергей Стенин. Мультиграммные аддитивно регуляризованные тематические модели // Магистерская диссертация, МФТИ, 2015.

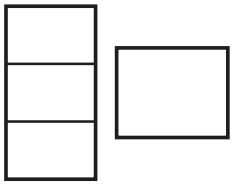
Регуляризаторы для мультимодальных тематических моделей

supervised



Модальности меток классов или категорий для задач классификации и категоризации текстов.

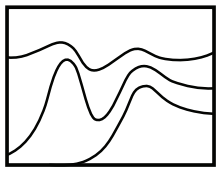
multilanguage



Модальность языков и регуляризация со словарём $\pi_{uwt} = p(u|w,t)$ переводов с языка k на ℓ:

$$R(\Phi,\Pi) = \tau \sum_{u\in W^k} \sum_{t\in T} n_{ut} \ln \sum_{w\in W^\ell} \pi_{uwt}\varphi_{wt}$$

temporal



Темпоральные модели с модальностью времени i:

$$R(\Phi) = -\tau \sum_{i\in I} \sum_{t\in T} \left|\varphi_{it} - \varphi_{i-1,t}\right|.$$

geospatial

Модальность геолокаций g с близостью $S_{gg'}$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g'\in G} S_{gg'} \sum_{t\in T} n_t^2 \left(\frac{\varphi_{gt}}{n_g} - \frac{\varphi_{g't}}{n_{g'}}\right)^2$$

Транзакционные данные

Выборка может содержать не только пары (d, w), но также тройки, четвёрки, ..., n-ки термов разных модальностей.

- **Данные социальной сети:**
  (d, u, w) — пользователь u записал слово w в блоге d
- **Данные сети интернет-рекламы:**
  (u, d, b) — пользователь u кликнул баннер b на странице d
- **Данные рекомендательной системы:**
  (u, f, s) — пользователь u оценил фильм f в ситуации s
- **Данные финансовых организаций:**
  (b, s, g) — покупатель u купил у продавца s товар g
- **Данные о пассажирских авиаперелётах:**
  (u, a, b, c) — перелёт клиента u из a в b авиакомпанией c

Задача: по наблюдаемой выборке рёбер гиперграфа найти латентные тематические векторные представления его вершин.

Гиперграфовая ARTM транзакционных данных

V^m — словарь термов модальности $m \in M$
$V = V^1 \sqcup \cdots \sqcup V^M$ — словарь термов всех модальностей
$\Gamma = \langle V, E \rangle$ — гиперграф, система конечных подмножеств V
(d, x) — ребро из E, где $d \in V$ — вершина-контейнер, $x \subset V$

Дано:
E_k — наблюдаемая выборка рёбер (транзакций) типа k,
n_{kdx} — число вхождений ребра (d, x) в выборку E_k.

Найти: тематическую модель рёбер типа k

$$p(x|d) = \sum_{t \in T} \underbrace{p(t|d)}_{\theta_{td}} \prod_{v \in x} \underbrace{p(v|t)}_{\varphi_{vt}}$$

Критерий: максимум регуляризованного правдоподобия:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} n_{kdx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vt} \; + \; R(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k\in K}\tau_k \sum_{(d,x)\in E_k} n_{kdx} \ln \sum_{t\in T}\theta_{td}\prod_{v\in x}\varphi_{vt} \;+\; R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tdx} = p(t|d,x)$:

$$\text{E-шаг:}\quad p_{tdx} = \operatorname*{norm}_{t\in T}\Big(\theta_{td}\prod_{v\in x}\varphi_{vt}\Big)$$

$$\text{M-шаг:}\quad \begin{cases} \varphi_{vt} = \operatorname*{norm}_{v\in V^m}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)\in E_k}\big[v\in x\big] n_{kdx}p_{tdx} + \varphi_{vt}\dfrac{\partial R}{\partial\varphi_{vt}}\Big) \\ \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum\limits_{k\in K}\tau_k \sum\limits_{(d,x)\in E_k} n_{kdx}p_{tdx} + \theta_{td}\dfrac{\partial R}{\partial\theta_{td}}\Big) \end{cases}$$

Доказательство (по лемме о максимизации на симплексах)

Применим лемму к log-правдоподобию с регуляризатором R:

$$\varphi_{vt} = \operatorname*{norm}_{v\in V_m}\left(\varphi_{vt}\sum_{k\in K}\tau_k\sum_{dx\in E_k} n_{kdx}\frac{\theta_{td}}{p(x|d)}\frac{\partial}{\partial\varphi_{vt}}\prod_{u\in x}\varphi_{ut} + \varphi_{vt}\frac{\partial R}{\partial\varphi_{vt}}\right) =$$

$$= \operatorname*{norm}_{v\in V_m}\left(\sum_{k\in K}\sum_{dx\in E_k}\tau_k n_{kdx}[v\in x]p_{tdx} + \varphi_{vt}\frac{\partial R}{\partial\varphi_{vt}}\right)$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\left(\theta_{td}\sum_{k\in K}\tau_k\sum_{x\in d} n_{kdx}\frac{1}{p(x|d)}\prod_{v\in x}\varphi_{vt} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\right) =$$

$$= \operatorname*{norm}_{t\in T}\left(\sum_{k\in K}\sum_{x\in d}\tau_k n_{kdx}p_{tdx} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\right)$$

■

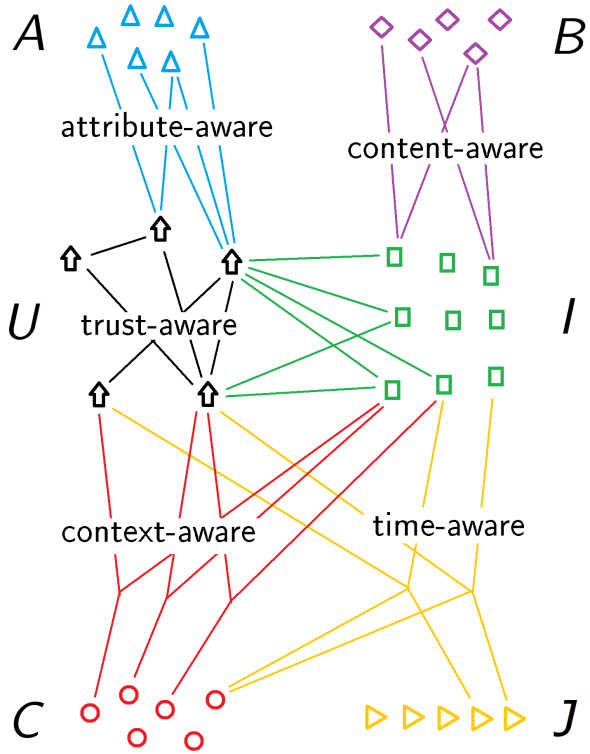
K.Vorontsov. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization // Springer Optimization and Its Applications. 2023

Транзакционные данные в рекомендательных системах

U — конечное множество (словарь) клиентов (users)
I — конечное множество (словарь) объектов (items)
A — словарь атрибутов клиентов (соцдем, регион, хобби...)
B — словарь свойств объектов (слова в текстовых объектах)
C — словарь ситуативных контекстов
J — словарь интервалов времени

Возможные виды данных:

n_{ui} — клиент u выбрал объект i
n_{ua} — клиент u имеет атрибут a
n_{ib} — объект i имеет свойство b
n_{uv} — клиент u доверяет клиенту v
n_{uib} — клиент u отметил i тэгом b
n_{uic} — клиент u выбрал i в контексте c
n_{uicj} — u выбрал i в c в интервале j



Гиперграфовые тематические модели языка

Рёбрами гиперграфа могут быть любые подмножества термов, связанные по смыслу и порождаемые общей темой:

- предложение / фраза / синтагма
- ветка синтаксического дерева / именная группа
- факт «объект, субъект, действие»
- пары синонимов, гипоним–гипероним, мероним–холоним
- лексическая цепочка
- текст комментария и его автор

Модель даёт интерпретируемые тематические эмбединги:

- $p(t|d)$ — каждого контейнера, в частности, документа
- $p(t|w) = \varphi_{wt}\frac{p(t)}{p(w)}$ — каждого терма, в частности, слова
- $p(t|d,x)$ — каждой отдельной транзакции (фразы, факта)

Модели предложений и коротких текстов TwitterLDA, senLDA

S_d — множество предложений документа d

n_{sw} — сколько раз терм w встречается в предложении s

Тематическая модель предложения s:

$$p(s|d) = \sum_{t\in T} p(t|d) \prod_{w\in s} p(w|t)^{n_{sw}} = \sum_{t\in T} \theta_{td} \prod_{w\in s} \varphi_{wt}^{n_{sw}}$$

Максимизация регуляризованного log-правдоподобия

$$\sum_{d\in D}\sum_{s\in S_d} \ln \sum_{t\in T} \theta_{td} \prod_{w\in s} \varphi_{wt}^{n_{sw}} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

это частный случай гиперграфовой модели, предложения являются гипер-рёбрами.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee Peng Lim et al. Comparing Twitter and traditional media using topic models. ECIR 2011.

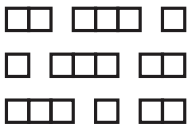
G.Balikas, M.-R.Amini, M.Clausel. On a topic model for sentences. SIGIR 2016.

Регуляризаторы для моделирования последовательного текста

sentence

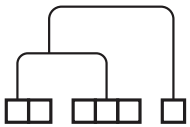
Тематические модели, учитывающие границы предложений, абзацев и секций документов

n-gram



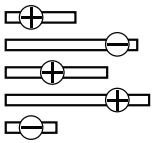
Модели с модальностями *n*-грамм, коллокаций, именованных сущностей (используем TopMine)

syntax



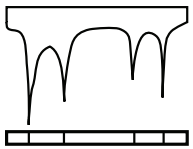
Модели, учитывающие результаты автоматического синтаксического разбора (используем UDPipe)

sentiment



Модели выделения мнений на основе тональностей, фактов, семантических ролей именованных сущностей

segmentation



Тематические модели сегментации с автоматическим определением границ сегментов

Сегментная структура текста и пост-обработка E-шага

Документ $d = \{w_1, \ldots, w_{n_d}\}$, n_d — длина документа d

Тематика термов в документе $p(t|d, w_i)$ — матрица $T \times n_d$:

Регуляризация E-шага как постобработка матриц $p(t|d,w)$

Трёхмерная матрица $\Pi = \left(p_{tdw} = p(t|d,w)\right)_{T\times D\times W}$

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Pi(\Phi,\Theta),\Phi,\Theta) \;\to\; \max_{\Phi,\Theta}.$$

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} = \underset{t\in T}{\text{norm}}\left(\varphi_{wt}\theta_{td}\right) \\ & \tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R}{\partial p_{zdw}}\right)\right) \\ \text{M-шаг:} & \varphi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw}\tilde{p}_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum_{w\in d} n_{dw}\tilde{p}_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

Схема доказательства из трёх шагов

1. Для функции $p_{tdw}(\Phi,\Theta) = \frac{\varphi_{wt}\theta_{td}}{\sum_z \varphi_{wz}\theta_{zd}}$ и любого $z \in T$

$$\varphi_{wt}\frac{\partial p_{zdw}}{\partial \varphi_{wt}} = \theta_{td}\frac{\partial p_{zdw}}{\partial \theta_{td}} = p_{tdw}\big([z=t] - p_{zdw}\big).$$

2. Введём вспомогательную функцию от переменных Π, Φ, Θ:

$$Q_{tdw}(\Pi,\Phi,\Theta) = \frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{tdw}} - \sum_{z\in T} p_{zdw}\frac{\partial R(\Pi,\Phi,\Theta)}{\partial p_{zdw}}.$$

$\tilde{R}(\Phi,\Theta) = R(\Pi(\Phi,\Theta),\Phi,\Theta)$ не зависит от p_{tdw} при $w \notin d$, значит

$$\varphi_{wt}\frac{\partial \tilde{R}}{\partial \varphi_{wt}} = \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}} + \sum_{d\in D} p_{tdw}Q_{tdw};\quad \theta_{td}\frac{\partial \tilde{R}}{\partial \theta_{td}} = \theta_{td}\frac{\partial R}{\partial \theta_{td}} + \sum_{w\in d} p_{tdw}Q_{tdw}.$$

3. Подставляем это в формулы M-шага:

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \sum_{d\in D} Q_{tdw}p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\Big);$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \sum_{w\in d} Q_{tdw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big).$$

■

Любая пост-обработка E-шага эквивалентна регуляризатору $R(\Pi)$

Итак, произвольному гладкому регуляризатору $R(\Pi, \Phi, \Theta)$ однозначно соответствует преобразование $p_{tdw} \to \tilde{p}_{tdw}$.
Верно и обратное:

Теорема. Если на k-й итерации EM-алгоритма для каждого (d, w): $n_{dw} > 0$ в формулах M-шага вместо вектора $(p^k_{tdw})_{t\in T}$ подставить другой вектор $(\tilde{p}^k_{tdw})_{t\in T}$, удовлетворяющий условию нормировки $\sum_t \tilde{p}^k_{tdw} = 1$, то это эквивалентно добавлению регуляризатора сглаживания–разреживания

$$R(\Pi) = \sum_{d\in D}\sum_{w\in d} n_{dw} \sum_{t\in T} (\tilde{p}^k_{tdw} - p^k_{tdw}) \ln p_{tdw}.$$

Общий вывод: пост-обработка E-шага позволяет учитывать порядок термов в документе в обход гипотезы «мешка слов».

Однопроходный по документу EM-алгоритм для ARTM

Максимизация log-правдоподобия при ограничении $\Theta = \Theta(\Phi)$:

$$\sum_{d,w} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td}(\Phi) + R\big(\Phi, \Theta(\Phi)\big) \to \max_{\Phi}$$

EM-алгоритм: метод простой итерации для системы уравнений

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}(\Phi)\big); \qquad n_{td} = \sum_{w \in d} n_{dw} p_{tdw} + \theta_{td}(\Phi)\frac{\partial R}{\partial \theta_{td}}$$

$$\tilde{p}_{tdw} = p_{tdw} + \frac{\varphi_{wt}}{n_{dw}} \sum_{z \in T} \frac{n_{zd}}{\theta_{zd}(\Phi)} \frac{\partial \theta_{zd}}{\partial \varphi_{wt}}$$

$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}}\left(\sum_{d \in D} n_{dw} \tilde{p}_{tdw} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right)$$

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста. КиМ, 2020.

Доказательство (по лемме о максимизации на симплексах)

Оптимизационная задача M-шага относительно Φ:

$$Q(\Phi) = \sum_{d\in D}\sum_{u\in W}\sum_{z\in T} n_{du}p_{zdu}\big(\ln\varphi_{uz} + \ln\theta_{zd}(\Phi)\big) + R\big(\Phi, \Theta(\Phi)\big) \to \max_{\Phi}$$

Применим лемму к регуляризованному log-правдоподобию Q:

$$\varphi_{wt}\frac{\partial Q}{\partial \varphi_{wt}} = \sum_{d\in D} n_{dw}p_{tdw} + \sum_{d,z,u} n_{du}p_{zdu}\frac{\varphi_{wt}}{\theta_{zd}}\frac{\partial\theta_{zd}}{\partial\varphi_{wt}} + \varphi_{wt}\sum_{d,z}\frac{\partial R}{\partial\theta_{zd}}\frac{\partial\theta_{zd}}{\partial\varphi_{wt}} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}} =$$

$$= \sum_{d\in D} n_{dw}\left(p_{tdw} + \frac{\varphi_{wt}}{n_{dw}}\sum_{z\in T}\frac{1}{\theta_{zd}}\underbrace{\left(\sum_{u\in d} n_{du}p_{zdu} + \theta_{zd}\frac{\partial R}{\partial\theta_{zd}}\right)}_{n_{zd}}\frac{\partial\theta_{zd}}{\partial\varphi_{wt}}\right) + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}} =$$

$$= \sum_{d\in D} n_{dw}\underbrace{\left(p_{tdw} + \frac{\varphi_{wt}}{n_{dw}}\sum_{z\in T}\frac{n_{zd}}{\theta_{zd}}\frac{\partial\theta_{zd}}{\partial\varphi_{wt}}\right)}_{\tilde{p}_{tdw}} + \varphi_{wt}\frac{\partial R}{\partial\varphi_{wt}}$$

■

Частный случай $\theta_{td}(\Phi) = \sum_w p_{wd} \operatorname{norm}_t(\varphi_{wt} p_t)$

Частные производные: $\varphi_{wt} \frac{\partial \theta_{zd}}{\partial \varphi_{wt}} = p_{wd} \tilde{\varphi}_{tw} \left(\delta_{zt} - \tilde{\varphi}_{zw}\right)$

EM-алгоритм: метод простой итерации для системы уравнений

$$\tilde{\varphi}_{tw} = \underset{t \in T}{\operatorname{norm}}\left(\varphi_{wt} p_t\right); \qquad \theta_{td} = \sum_{w \in d} p_{wd} \tilde{\varphi}_{tw}$$

$$p_{tdw} = \underset{t \in T}{\operatorname{norm}}\left(\varphi_{wt} \theta_{td}\right); \qquad n_{td} = \sum_{w \in d} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}$$

$$\tilde{p}_{tdw} = p_{tdw} + \frac{\tilde{\varphi}_{tw}}{n_d} \left(\frac{n_{td}}{\theta_{td}} - \sum_{z \in T} \tilde{\varphi}_{zw} \frac{n_{zd}}{\theta_{zd}} \right)$$

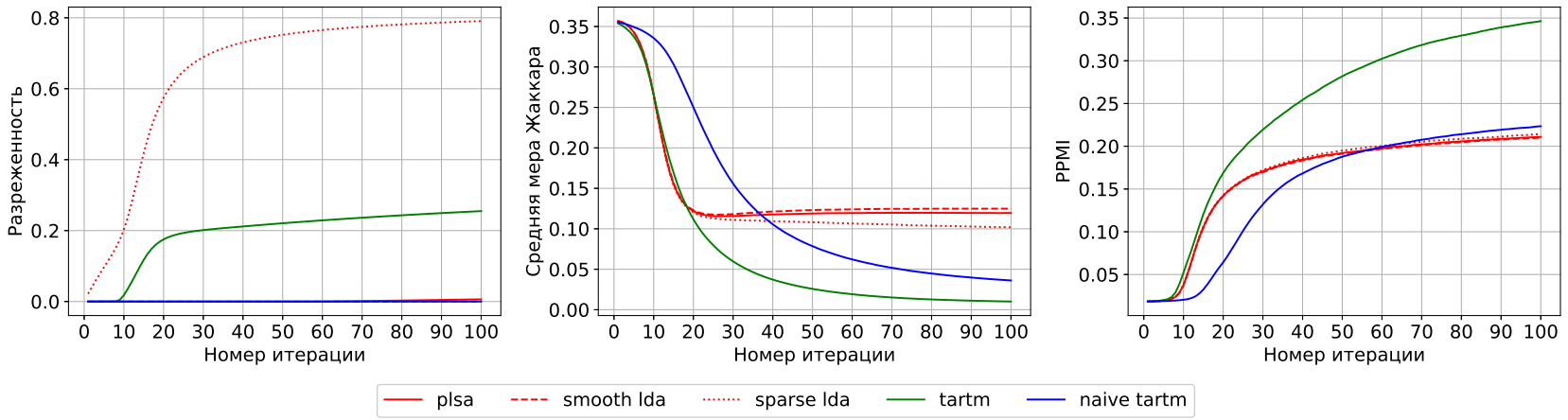
$$\varphi_{wt} = \underset{w \in W}{\operatorname{norm}} \left(\sum_{d \in D} n_{dw} \tilde{p}_{tdw} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right)$$

E-шаг по-прежнему занимает $O(n_d |T|)$ операций для каждого d

Эксперимент. Проверка модифицированного EM-алгоритма

Коллекция NIPS, $|T| = 50$, модели:

- TARTM (Θless ARTM) — модифицированный EM-алгоритм
- naive TARTM — одна итерация обычного EM-алгоритма

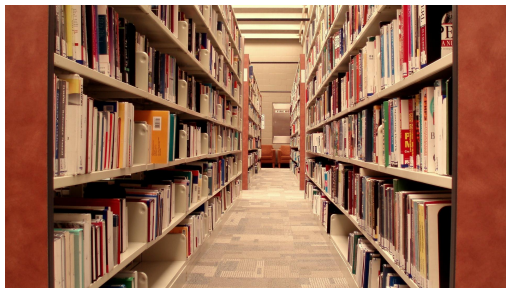


- TARTM очищает темы от общеупотребительных слов,
- улучшает разреженность, различность и когерентность тем

И.А.Ирхин, В.Г.Булатов, К.В.Воронцов. Аддитивная регуляризация тематических моделей с быстрой векторизацией текста, 2020.

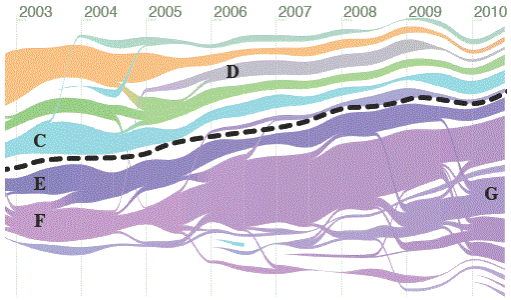
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

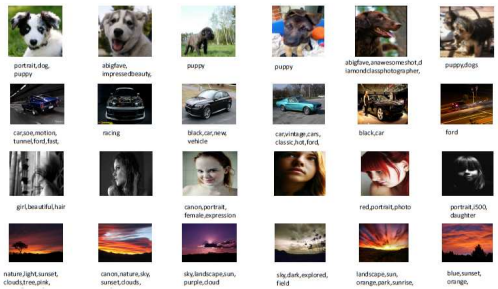


поиск тематического контента в соцсетях

выявление и отслеживание цепочек новостей

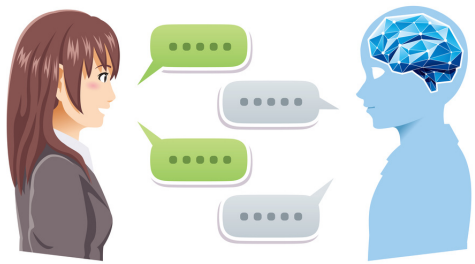


мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

управлением диалогом в разговорном интеллекте

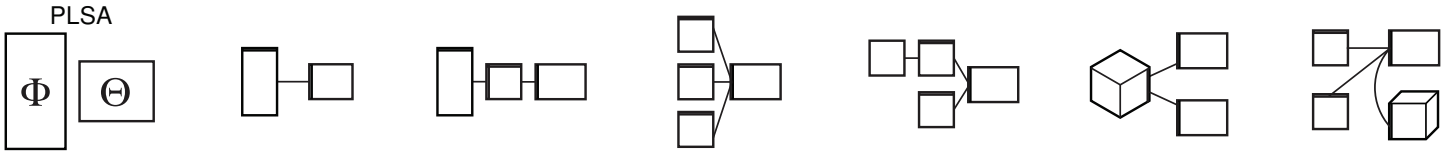


Тематическая модель для разведочного поиска должна быть...

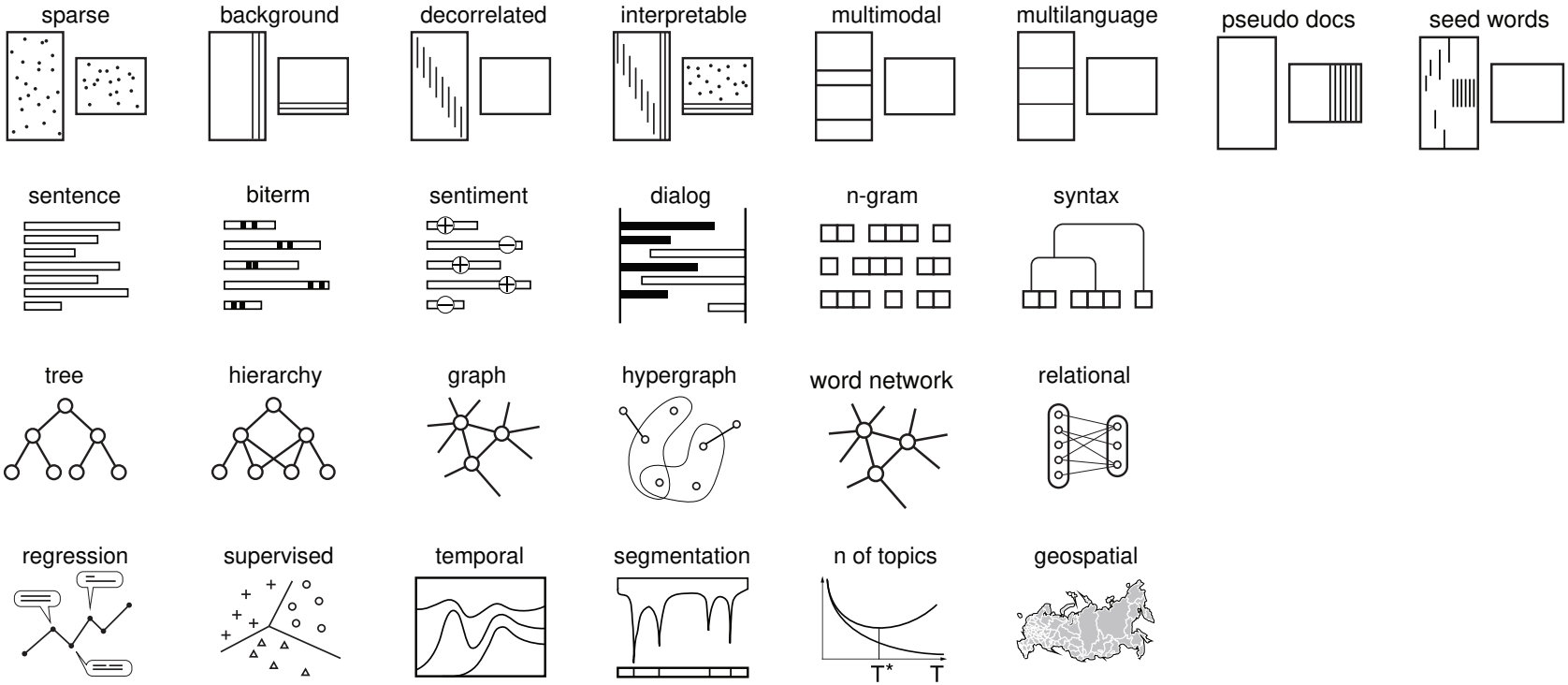
1. **Интерпретируемая**: объяснение смысла каждой темы
2. **Иерархическая**: разделение тем на подтемы
3. **Динамическая**: развитие каждой темы во времени
4. **Мультимодальная**: слова + авторы, категории, связи, теги,...
5. **Мультиграммная**: слова + термины-словосочетания
6. **Мультиязычная**: для кросс- и много-языкового поиска
7. **Сегментирующая** документ на тематические блоки
8. **Обучаемая** по оценкам асессоров и логам пользователей
9. **Определяющая число тем** автоматически
10. **Создающая и именующая новые темы** автоматически
11. **Онлайновая**: обработка коллекции за один проход
12. **Параллельная, распределённая** для больших коллекций

Палитра регуляризаторов в ARTM (список не полон)

Структуры матричных разложений в вероятностных моделях:



Регуляризаторы — дополнительные критерии и ограничения:



Модульный подход к синтезу моделей с заданными свойствами

Для построения композитных моделей в BigARTM не нужны ни математические выкладки, ни программирование «с нуля».

Этапы моделирования	Bayesian TM	ARTM	
	Анализ требований	Анализ требований	
Формализация:	Вероятностная модель порождения данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Единый регуляризованный EM-алгоритм для любых моделей и их композиций	
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

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Онлайновый параллельный мультимодальный ARTM
- Встроенная библиотека регуляризаторов и мер качества

Сообщество:

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

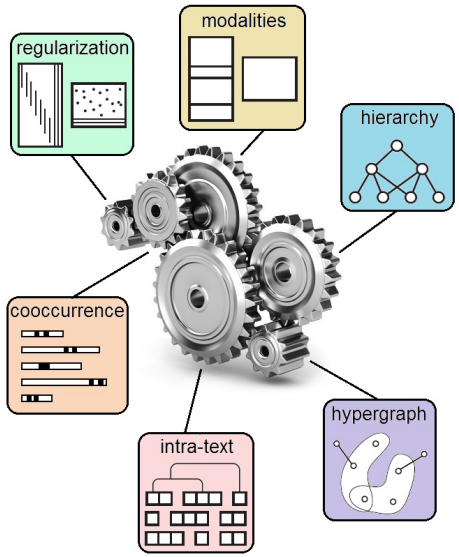
Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Windows, Linux, MacOS (32/64 bit)
- Интерфейсы API: command-line, C++, and Python

Ключевые возможности библиотек BigARTM и TopicNet

BigARTM

- библиотека регуляризаторов
- мультимодальные модели
- иерархические модели
- гиперграфовые модели
- модели связного текста



TopicNet

- Перебор сценариев регуляризации для выбора моделей
- Автоматическое протоколирование экспериментов
- Построение «банка тем» из множества моделей
- Визуализация результатов тематического моделирования

V.Bulatov, E.Egorov, E.Veselova, D.Polyudova, V.Alekseev, A.Goncharov, K.Vorontsov.
TopicNet: making additive regularisation for topic modelling accessible. LREC-2020

Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

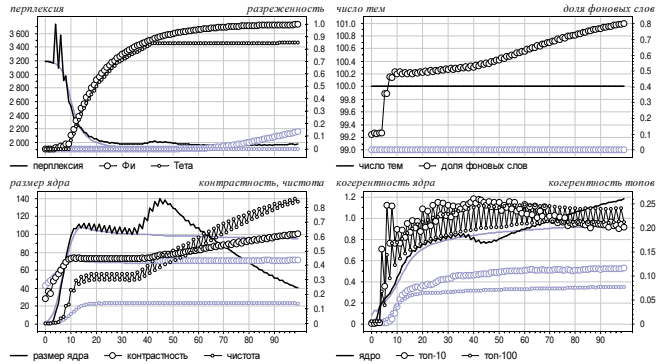
3.7M статей Википедии, 100K слов: время min (перплексия)

проц.	$\|T\|$	Gensim	Vowpal Wabbit	BigARTM	BigARTM асинхрон
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (3297)		47m (3347)	28m (3380)
4	50	88m (5311)		12m (5216)	7m (5353)
4	100	104m (4338)		16m (4233)	10m (4357)
4	200	315m (3583)		26m (3520)	16m (3634)
8	50	88m (6344)		8m (5648)	5m (6220)
8	100	107m (5380)		10m (4660)	6m (5119)
8	200	288m (4263)		15m (3929)	10m (4309)

D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov.
Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

Декоррелирование, сглаживание, разреживание

Цель: найти комбинацию регуляризаторов, улучшающую интерпретируемость тем по совокупности критериев.



Регуляризаторы:

$$\mathscr{L}\left(\underset{\Phi\ \Theta}{\text{PLSA}}\right) + R\left(\underset{}{\text{decorrelated}}\right) + R\left(\underset{}{\text{sparse}}\right) + R\left(\underset{}{\text{background}}\right) \to \max$$

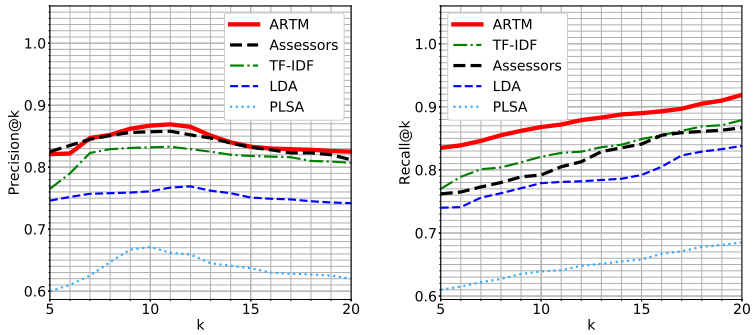
Результаты:

- разреженность $0 \to 95\%$, когерентность $0.25 \to 0.96$, чистота $0.14 \to 0.89$, контрастность $0.43 \to 0.52$,
- без заметного ущерба для перплексии: $1920 \to 2020$
- выработаны рекомендации по стратегии регуляризации

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

Разведочный поиск в технологических блогах — 1

Цель: поиск документов
по длинным текстовым запросам
— Habr.ru (175K документов),
— TechCrunch.com (760K док.).



Регуляризаторы:

$$\mathscr{L}\left(\Phi,\Theta\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) \to \max$$

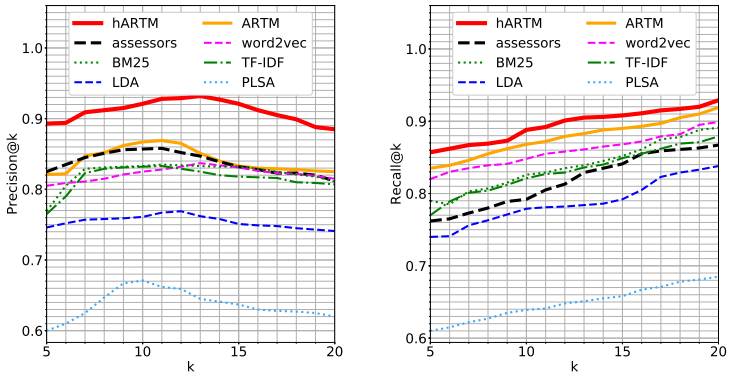
Результаты:

- Точность и полнота 88%, превосходит асессоров и другие методы (tf-idf, word2vec, PLSA, LDA).
- Векторный поиск мгновенный, асессоры тратили 5–65 мин.

A.Ianina, L.Golitsyn, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Разведочный поиск в технологических блогах — 2

Цель: улучшение качества поиска с помощью иерархической тематической модели hARTM и отсечения нерелевантных тем.



Регуляризаторы:

$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{hierarchy}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) \to \max$$

Результаты:

- Точность и полнота 93%, превосходит асессоров и другие методы (tf-idf, BM25, word2vec, PLSA, LDA, ARTM).
- Увеличилась оптимальная размерность векторов: 200 → 1400 (Habr.ru), 475 → 2800 (TechCrunch.com).

A.Ianina, K.Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. FRUCT–ISMW, 2019.

Поиск и рубрикация научных публикаций на 100 языках

Цель: мультиязыковой поиск и классификация научных публикаций по рубрикаторам УДК, ГРНТИ, ОЭСР, ВАК

модель	ср.ч. УДК	ср.% УДК	ср.ч. ГРНТИ	ср.% ГРНТИ
Базовая TM	0.558	0.165	0.536	0.220
XLM-RoBERTa	0.835	0.179	0.832	0.288
ARTM	0.995	0.225	0.852	0.366

Регуляризаторы:

$$\mathscr{L}\left(\underset{\text{PLSA}}{\Phi\,\Theta}\right)+R\left(\underset{\text{interpretable}}{}\right)+R\left(\underset{\text{multimodal}}{}\right)+R\left(\underset{\text{multilanguage}}{}\right)+R\left(\underset{\text{supervised}}{}\right)\to\max$$

Результаты:

- точность мультиязычного поиска 94%
- сокращение модели 128 Гб $\to$ 4.8 Гб при редукции словарей (BPE-токенизация) до 11K токенов на каждый язык.

П.Потапова, А.Грабовой, О.Бахтеев, Е.Егоров, Н.Зиновкин, Ю.Чехович, К.Воронцов и др. Мультиязыковая автоматическая рубрикация научных документов. 2023.

Поиск и классификация этно-релевантных тем в соцсетях

Цель: выявление как можно большего числа тем о национальностях и межнациональных отношениях (по словарю из 300 этнонимов).

(японцы): японский, япония, корея, китайский, жилища, авария, фукусиму, цунами, сообщать, океан, станция, хатико, район, правительство, атомный,
(норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, сын,
(венесуэльцы): куба, кастро, венесуэла, чавес, президент, уго, мадуро, боливия, фидель, глава, латинский, венесуэльский, лидер, боливарианской, президентский, альенде, гевару,
(китайцы): китайский, россия, производство, китай, продукция, страна, предприятие, компания, технология, военный, регион, производить, производственный, промышленность, российский, экономический, кнр,
(азербайджанцы): русский, азербайджан, азербайджанец, россия, азербайджанский, таксист, диаспора, анапа, народ, москва, страна, армянин, слово, рынок,
(грузины): грузинский, спецназ, военный, август, баташева, российский, спецназовец, миротворец, операция, румын, бригада, миротворческий, абхазия, группа, войска, русский, цхинвале,
(осетины): конституция, осетия, аминат, русский, осетинский, южный, северный, россия, война, республика, вопрос, алахай, российский, население, конфликт,
(цыгане): наркотик, цыган, цыганка, хороший, место, страна, деньга, время, работать, жизнь, жить, рука, дом, цыганский, наркоманка,

Регуляризаторы:

$$\mathscr{L}\left(\underset{\Phi}{\square}\,\underset{\Theta}{\square}\right) + R\left(\square\right) + R\left(\square\right) + R\left(\square\right) + R\left(\square\right) + R\left(\square\right) + R\left(\square\right) \to \max$$

PLSA; seed words; interpretable; multimodal; temporal; geospatial; sentiment

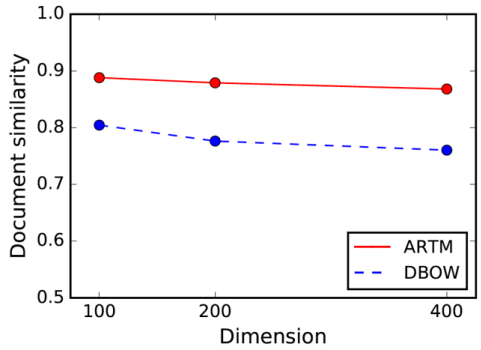
Результаты: число релевантных тем: 45 (LDA) $\to$ 83 (ARTM).

M.Apishev, S.Koltcov, O.Koltsova, S.Nikolenko, K.Vorontsov. Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI, 2016.

Mining ethnic content online with additively regularized topic models. 2016.

Тематические модели коротких текстов

Цель: интерпретируемые разреженные тематические эмбединги на основе дистрибутивной семантики, аналоги word2vec и WNTM.



Регуляризаторы:

$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{co-occurence}\right) + R\left(\text{interpretable}\right) + R\left(\text{multimodal}\right) \to \max$$

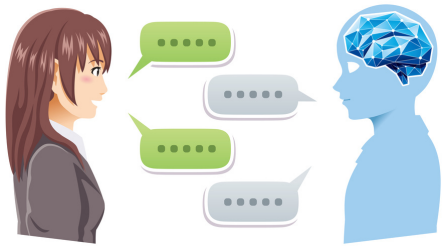
Результаты:

- Точность поиска схожих документов: $0.8 \to 0.9$
- Когерентность тем: $0.08 \to 0.33$
- Семантическая близость слов: $0.53 \to 0.58$, $0.38 \to 0.61$

A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.

Выявление намерений клиентов для построения чат-ботов

Цель: выявить тематику и интенты (намерения) клиентов по коллекции обращений в контактный центр. Построить рубрикатор интентов для последующей разметки диалогов.



Регуляризаторы:

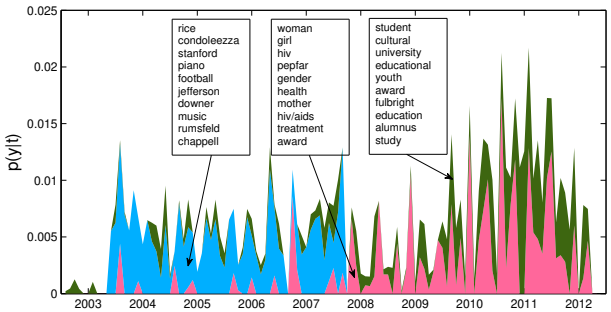
$$\mathscr{L}\left(\text{PLSA}\right) + R\left(\text{interpretable}\right) + R\left(\text{hierarchy}\right) + R\left(\text{segmentation}\right)$$
$$+ R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) + R\left(\text{syntax}\right) \to \max$$

Результаты: точность классификации интентов 60% → 66%.

A.Popov, V.Bulatov, D.Polyudova, E.Veselova. Unsupervised dialogue intent detection via hierarchical topic model. RANLP, 2019.

Выявление динамики тем в новостных потоках

Цель: выделение тем в коллекции пресс-релизов МИДов 4х стран, с привязкой ко времени.



Регуляризаторы:

$$\mathscr{L}\left(\Phi,\Theta\right) + R\left(\text{interpretable}\right) + R\left(\text{temporal}\right) + R\left(\text{multimodal}\right)$$
$$+ R\left(\text{n-gram}\right) + R\left(\text{multilanguage}\right) \to \max$$

Результаты:

- разделение тем на событийные и перманентные
- когерентность тем: 5.5 $\to$ 6.5

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

Выделение поляризованных мнений в политических новостях

Цель: найти признаки, по которым событийная тема разделяется на кластеры-мнения

Modalities	*Pr*	*Rec*	*F1*
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
All	**0.77**	**0.97**	**0.86**

Регуляризаторы:

$$\mathscr{L}\left(\underset{\Phi\ \Theta}{\text{PLSA}}\right)+R\left(\text{interpretable}\right)+R\left(\text{multimodal}\right)+R\left(\text{n-gram}\right)+R\left(\text{syntax}\right)\to\max$$

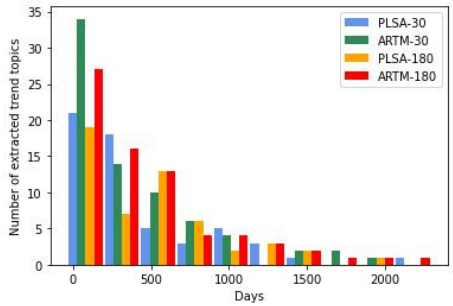
Результаты:

- выделение мнений внутри тем: F1-мера = 0.86%
- совместное использование трёх модальностей: факты «субъект–предикат–объект», семантические роли слов по Филлмору, тональности именованных сущностей

D.Feldman, T.Sadekova, K.Vorontsov. Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. Dialogue 2020.

Выявление трендов в коллекции научных публикаций

Цель: ранее обнаружение трендовых тем с начальным экспоненциальным ростом в области AI/ML 2009–2021 гг.



Регуляризаторы:

$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{interpretable}\right) + R\left(\text{dynamic}\right) + R\left(\text{multimodal}\right) + R\left(\text{n-gram}\right) \to \max$$

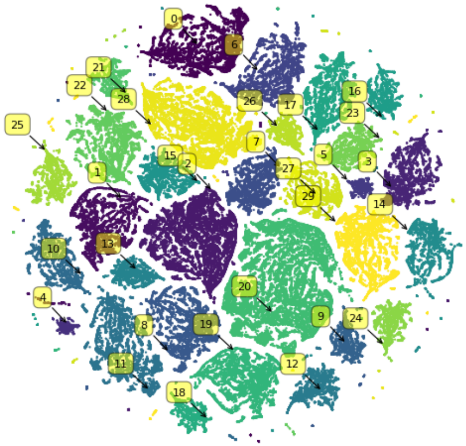
Результаты:

- выделение 90 из 91 тренда в области машинного обучения
- 63% тем выделяется за год, 79% за два года

Н.Герасименко, А.Чернявский, М.Никифорова, М.Никитин, К.Воронцов.
Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях. Доклады РАН, 2022.

Тематическая модель банковских транзакционных данных

Цель: Выявление паттернов потребительского поведения клиентов банка;
документы = клиенты,
слова = MCC-коды продавцов.



Регуляризаторы:

$$\mathscr{L}\Big(\underset{\Phi\ \Theta}{\text{PLSA}}\Big) + R\Big(\underset{}{\text{interpretable}}\Big) + R\Big(\underset{}{\text{multimodal}}\Big) + R\Big(\underset{}{\text{supervised}}\Big) \to \max$$

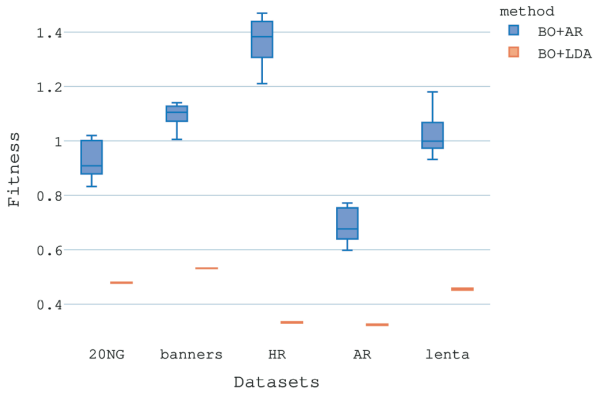
Результаты:

- темы — паттерны потребительского поведения
- предсказание пола, возраста, достатка клиентов

E.Egorov, F.Nikitin, A.Goncharov, V.Alekseev, K.Vorontsov. Topic modelling for extracting behavioral patterns from transactions data. 2019.

Автоматический подбор коэффициентов регуляризации

Цель: AutoARTM — автоподбор гиперпараметров (коэффициентов регуляризации, числа итераций) по критерию когерентности тем



Регуляризаторы:

$$\mathscr{L}\left(\Phi\,\Theta\right) + R\left(\text{decorrelated}\right) + R\left(\text{sparse}\right) + R\left(\text{background}\right) \to \max$$

Результаты:

- Значимое улучшение когерентности тем на 5 датасетах
- Генетический алгоритм показал лучшие результаты

M.Khodorchenko, S.Teryoshkin, T.Sokhin, N.Butakov. Optimization of learning strategies for ARTM-based topic models. LNCS, 2020.

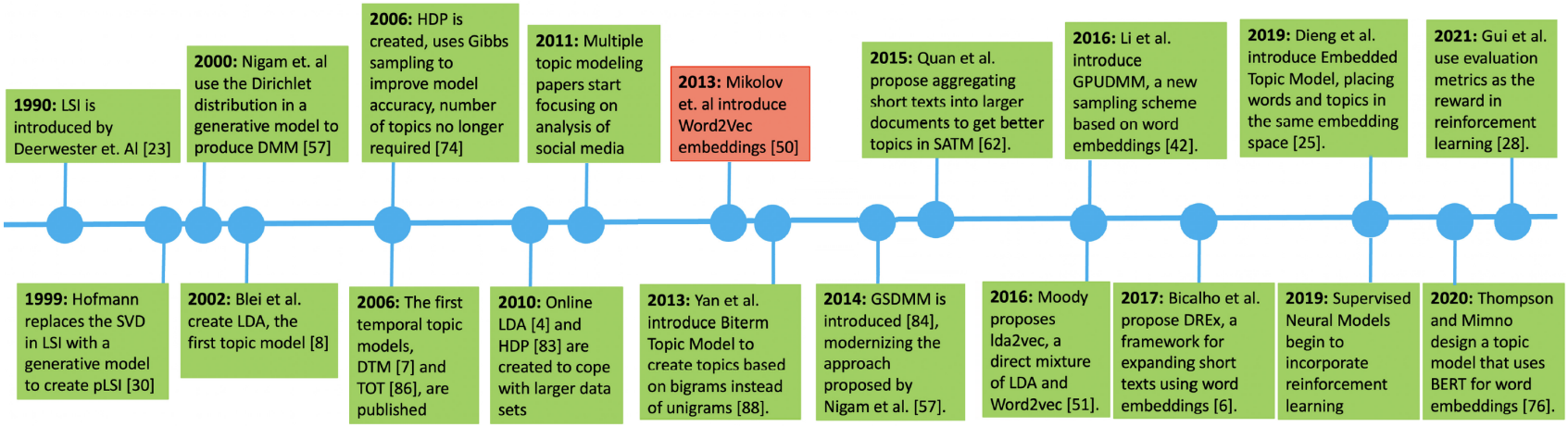
Резюме

- 20 лет развития PTM, сотни моделей, тысячи публикаций в рамках избыточно сложного байесовского обучения
- Был пропущен естественный этап развития теории PTM в рамках классической не-байесовской регуляризации
- ARTM — запоздалая попытка восполнить этот пробел
- ARTM — это «теория одной леммы»
- Если бы сообщество PTM знало об этой лемме, развитие PTM вряд ли пошло бы по пути байесовского обучения
- Эта же лемма применима для обучения нейронных сетей с неотрицательными нормированными векторами
- Нейросетевые тематические модели — основной тренд TM
- Неотрицательность и нормированность векторов — путь к интерпретируемости нейросетевых моделей?

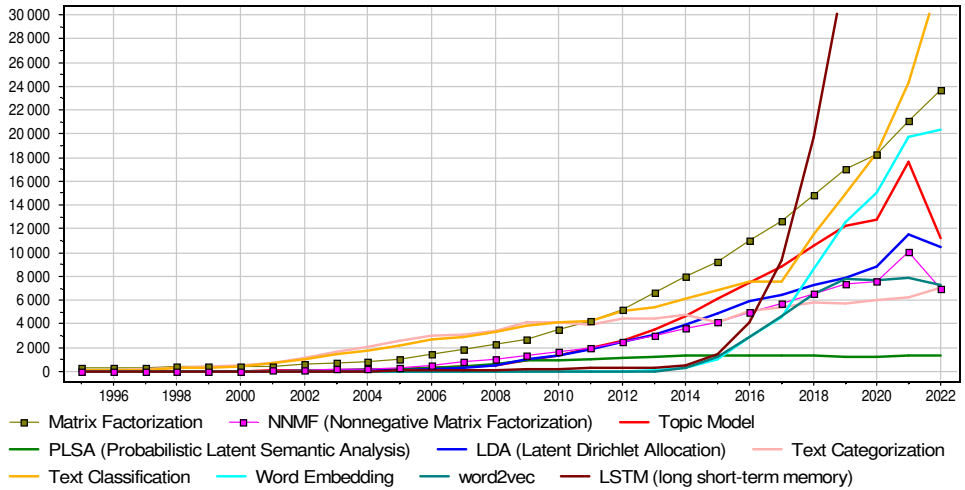
К.Воронцов. Вероятностное тематическое моделирование: теория ARTM и проект BigARTM. 2022.
http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022.

Дополнение. Эволюция тематических моделей



Динамика цитирования:
Topic Modeling и смежные области исследований (по данным Google Scholar)



Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022.