Вычислительные методы обработки естественного языка для разведочного информационного поиска, накопления, анализа и систематизации предметных знаний (РФФИ 20-07-00936)

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

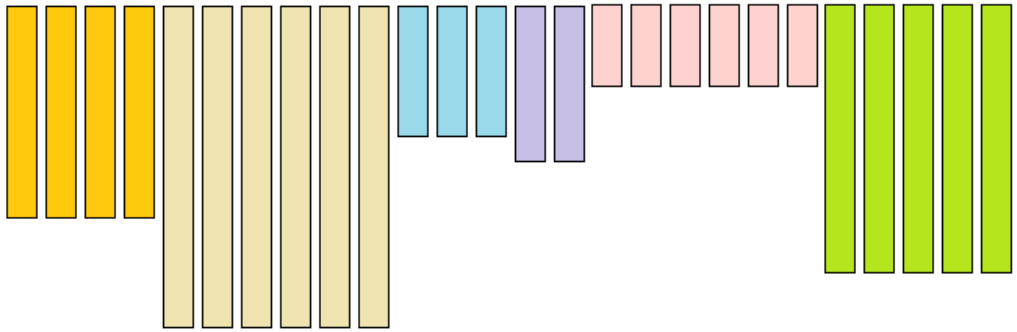
Научный семинар отделения 1 ФИЦ ИУ РАН • 19 января 2023

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

■

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 языковую модель

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

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

Критерий: максимум логарифма правдоподобия

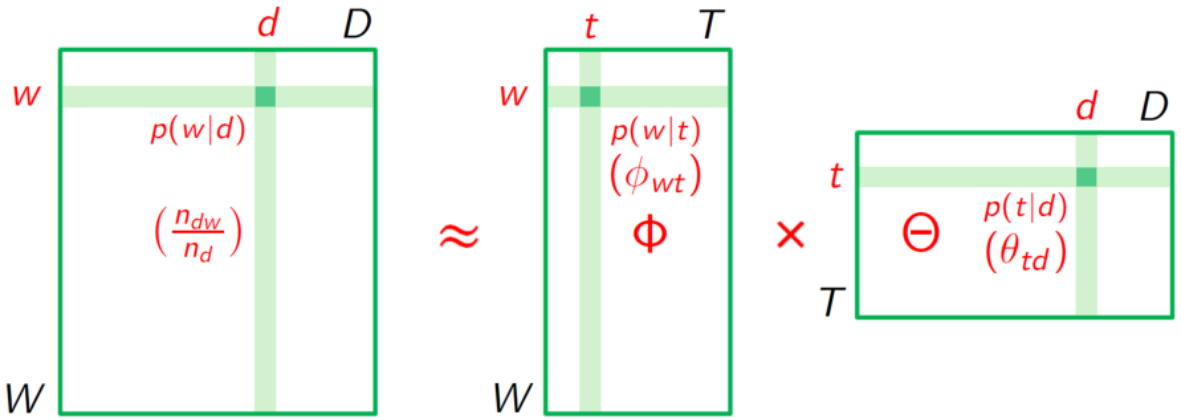
$$L(\Phi, \Theta) = \ln \prod_{d,w} p(w|d)^{n_{dw}} = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi,\Theta}$$

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

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

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

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

ARTM: аддитивная регуляризация тематических моделей

Максимизация логарифма правдоподобия с регуляризатором:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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:

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

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

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

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

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

■

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:

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

LDA: latent Dirichlet allocation (латентное размещение Дирихле):

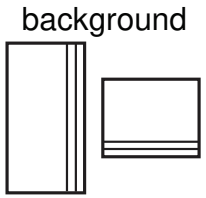
$$R(\Phi, \Theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}.$$

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

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

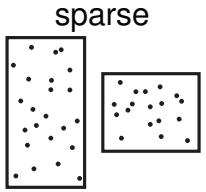
Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet allocation. NIPS 2001.

Регуляризаторы для улучшения интерпретируемости тем



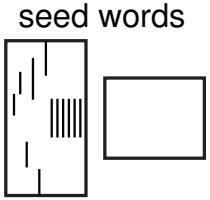
Сглаживание фоновых тем $B \subset T$:

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

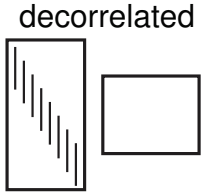


Разреживание предметных тем $S = T \setminus B$:

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

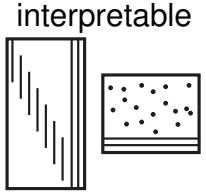


Сглаживание для выделения релевантных тем
с помощью словаря «затравочных» ключевых слов



Декоррелирование для повышения различности тем:

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



Сглаживание + разреживание + декоррелирование
для улучшения интерпретируемости тем

Свойство интерпретируемости тематических моделей

Тематические векторные представления (эмбединги) текста:

- $p(t|d) = \theta_{td}$ для документа d
- $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ для терма w
- $p(t|d,w)$ для локального контекста (d,w)
- $p(t|x)$ для нетекстового объекта x

Интерпретируемость тематических векторов:

- каждая тема t описывается *семантическим ядром* — частотным словарём слов $\left\{w\colon p(w|t) > \gamma p(w)\right\}$, встречающихся в данной теме в γ раз чаще обычного
- любой объект x с вектором $p(t|x)$ описывается частотным словарём слов $\left\{w\colon p(w|x) = \sum\limits_{t\in T} p(w|t)p(t|x) > \gamma p(w)\right\}$

Цели и не-цели тематического моделирования

Цели:

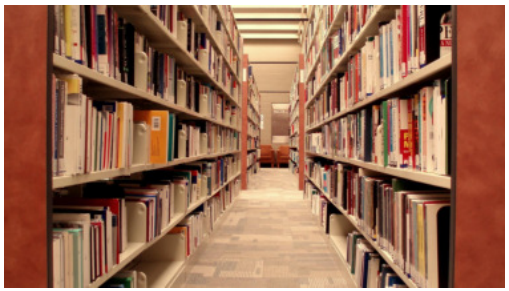
- Выяснить тематическую кластерную структуру текстовой коллекции, сколько в ней тем и какие они
- Получать интерпретируемые тематические векторные представления (эмбединги) документов, фрагментов, слов $p(t|d)$, $p(t|w)$, $p(t|d,w)$ и нетекстовых объектов $p(t|x)$
- Решать задачи поиска, категоризации, сегментации, суммаризации с помощью тематических эмбедингов

Не-цели:

- Угадывать следующие слова (TM — слабые модели языка)
- Генерировать связный текст
- Понимать смысл текста

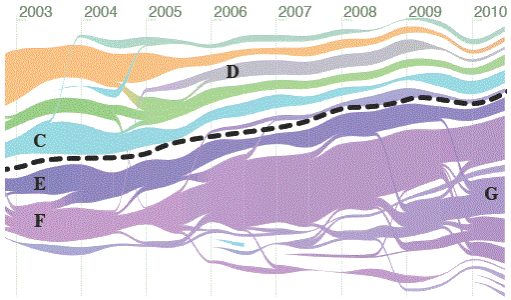
Некоторые приложения тематического моделирования

разведочный поиск в электронных библиотеках

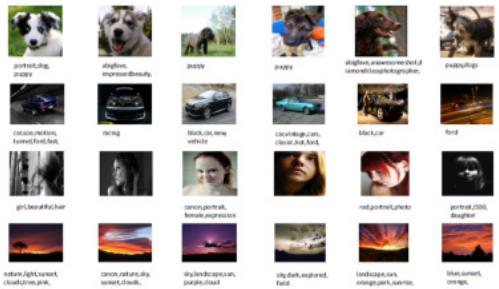


поиск тематического контента в соцсетях

выявление и отслеживание цепочек новостей

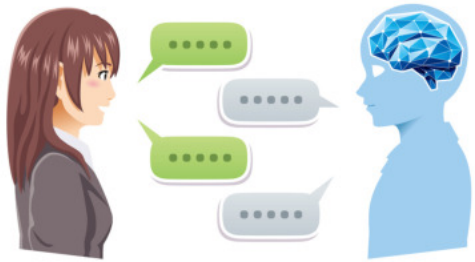


мультимодальный поиск текстов и изображений



анализ банковских транзакционных данных

управлением диалогом в разговорном интеллекте



Байесовская и классическая 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 (ARTM) обобщает MAP на любые регуляризаторы и их комбинации:

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

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:

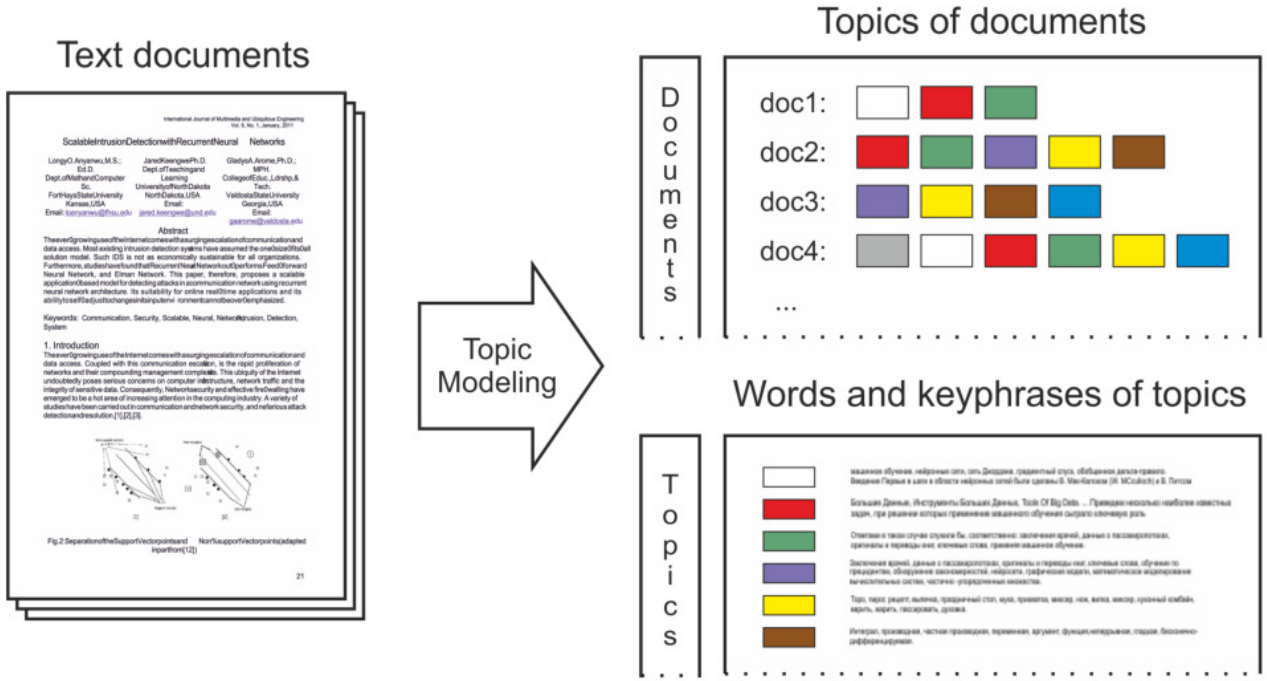
- Открытый код https://github.com/bigartm
  (discussion group, issue tracker, pull requests)
- Документация http://bigartm.org

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

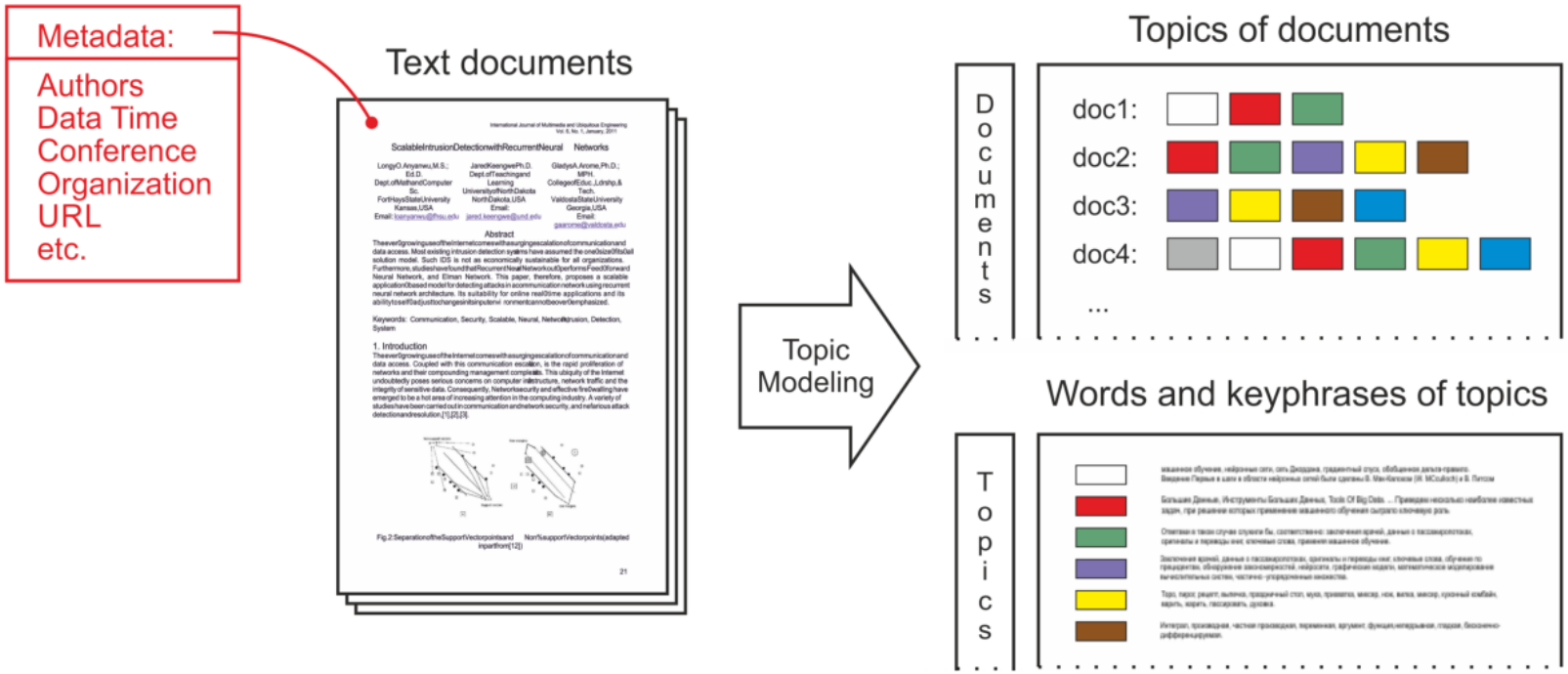
Мультимодальная тематическая модель

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



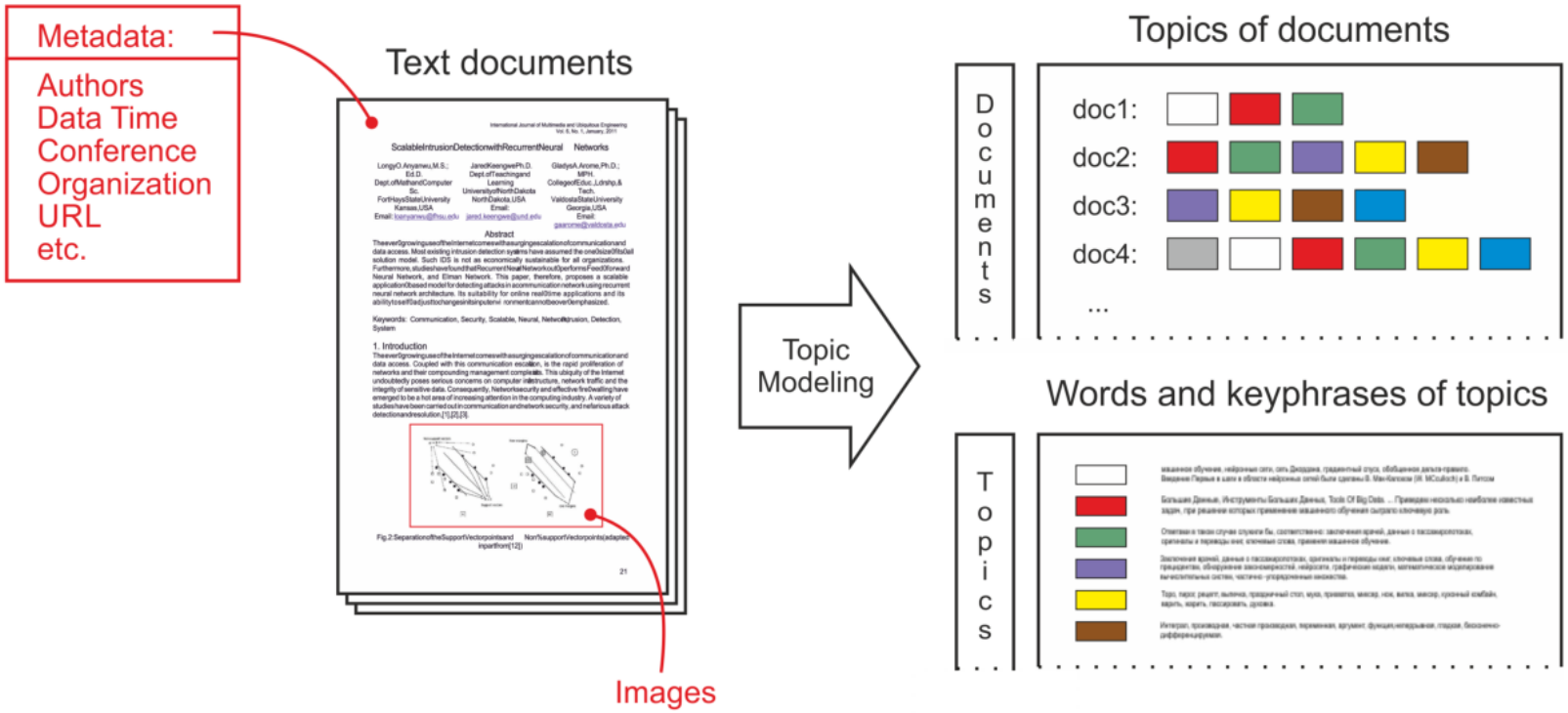
Мультимодальная тематическая модель

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



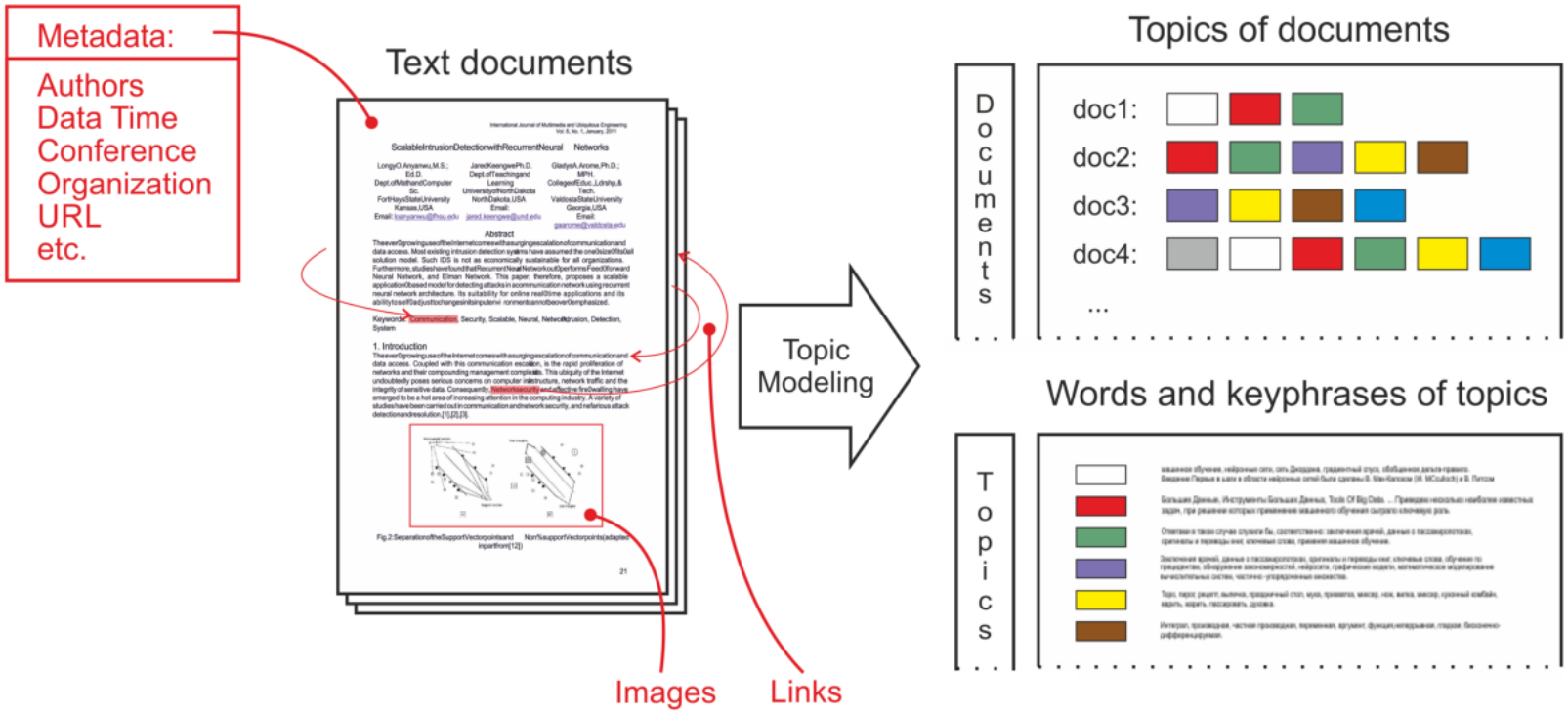
Мультимодальная тематическая модель

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



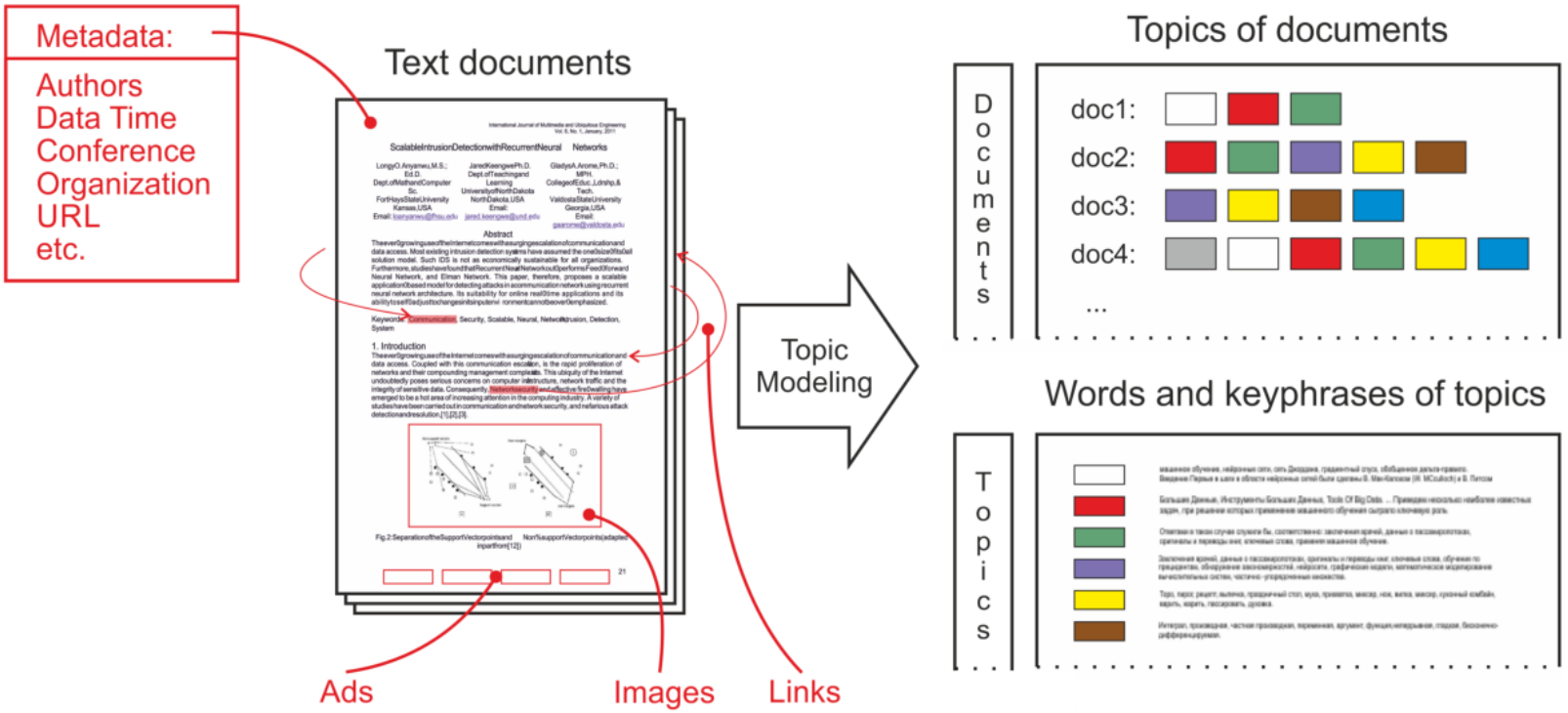
Мультимодальная тематическая модель

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



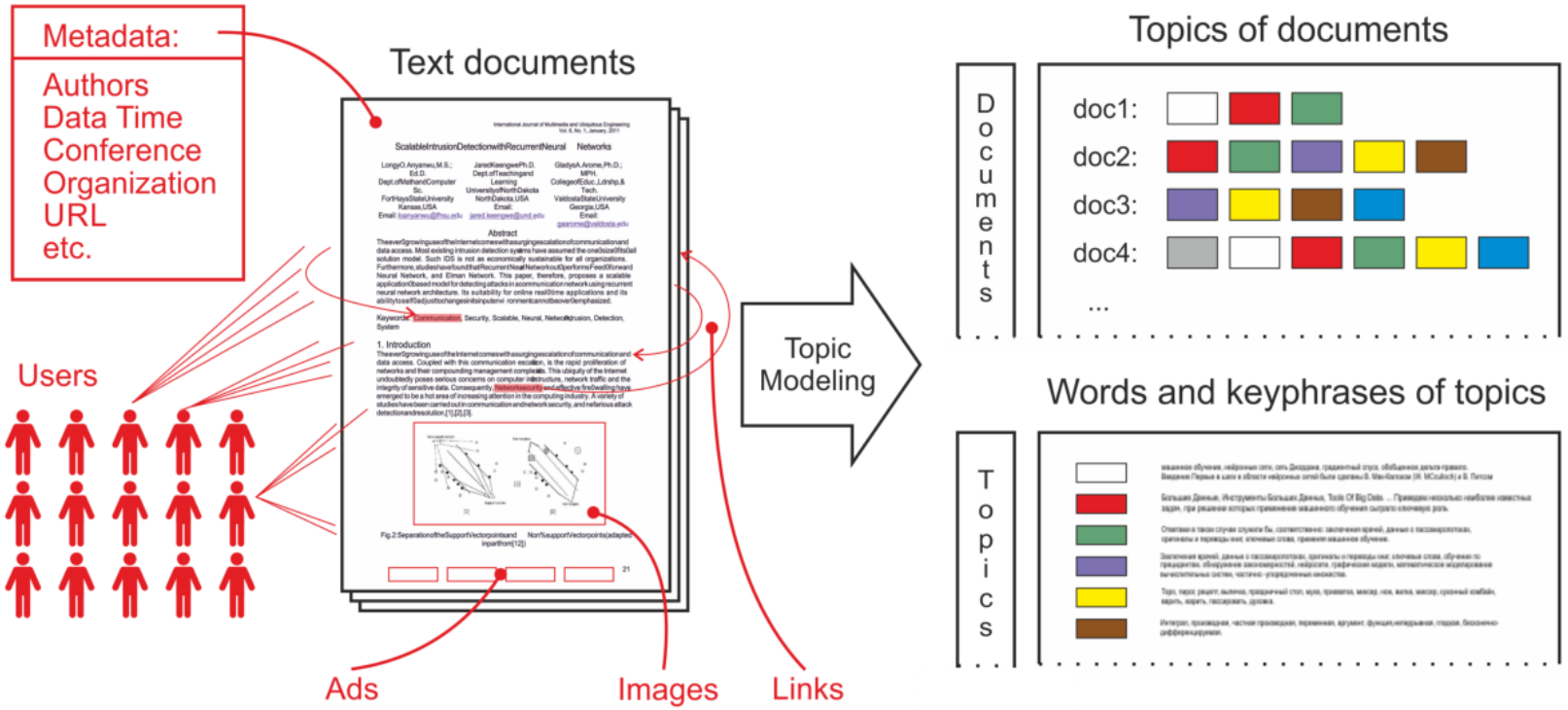
Мультимодальная тематическая модель

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



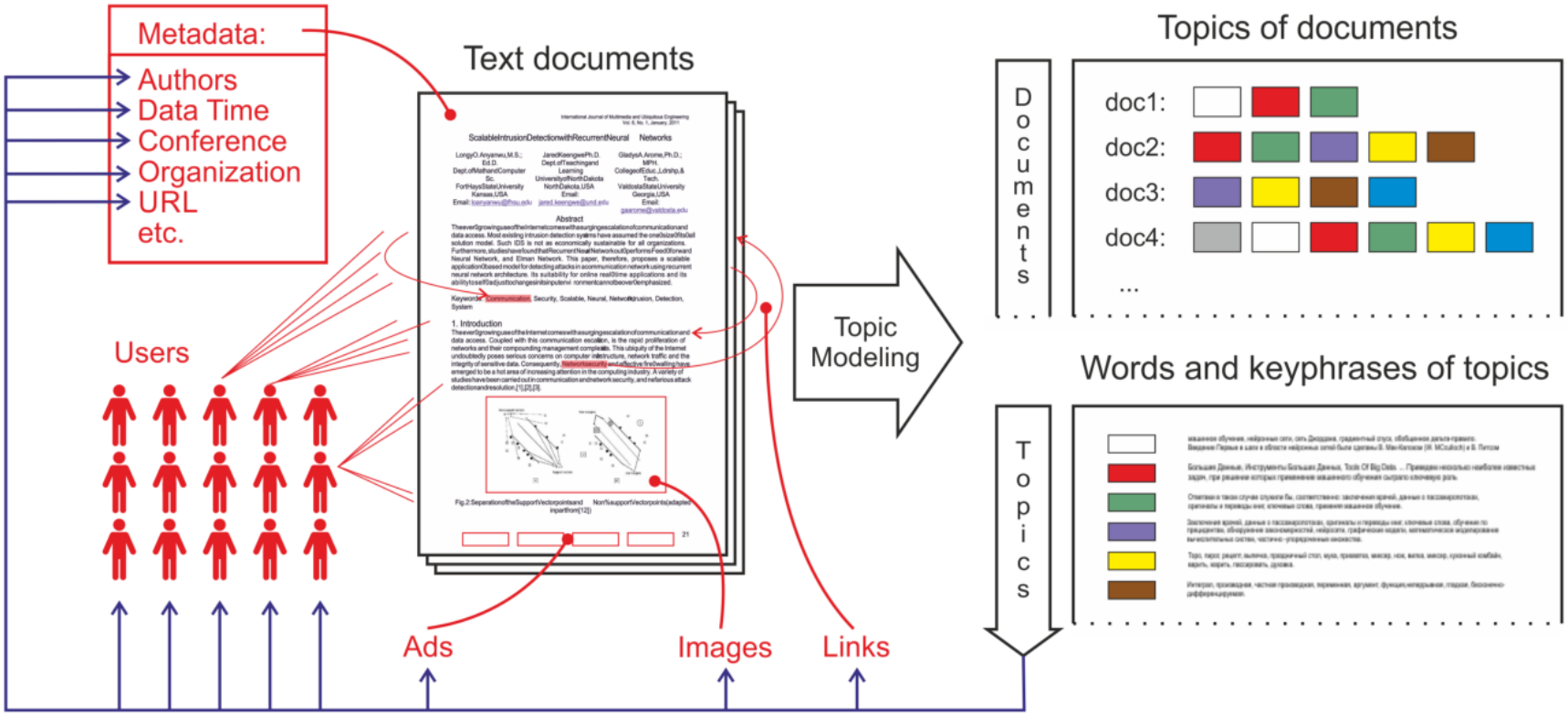
Мультимодальная тематическая модель

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 для документов с метаданными

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

Максимизация суммы log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

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

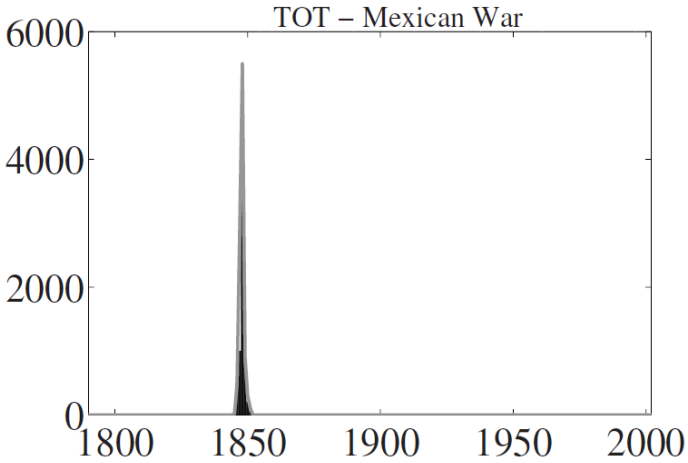
Коллекция 1000 статей конференций ММРО, ИОИ на русском

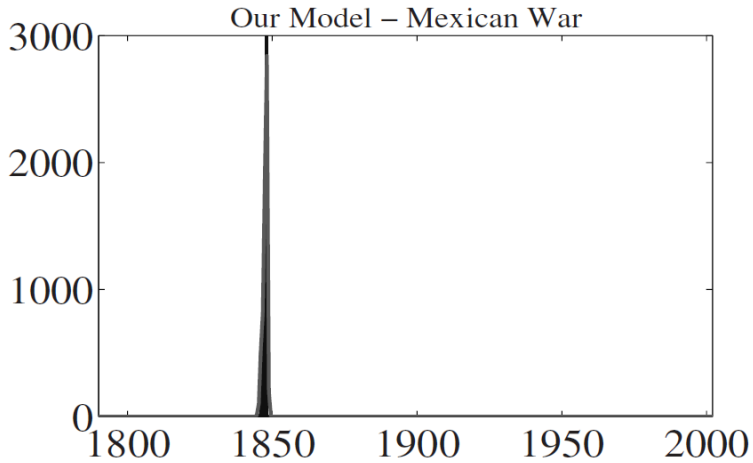
распознавание образов в биоинформатике		теория вычислительной сложности	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	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.

Пример 3. Совмещение темпоральной и n-граммной модели

По коллекции выступлений президентов США





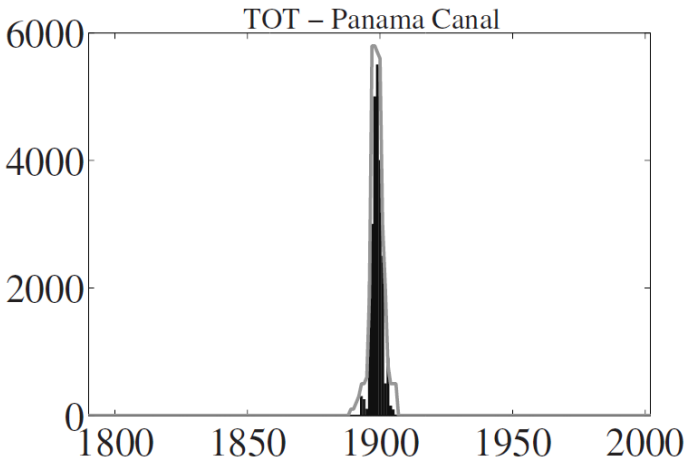
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

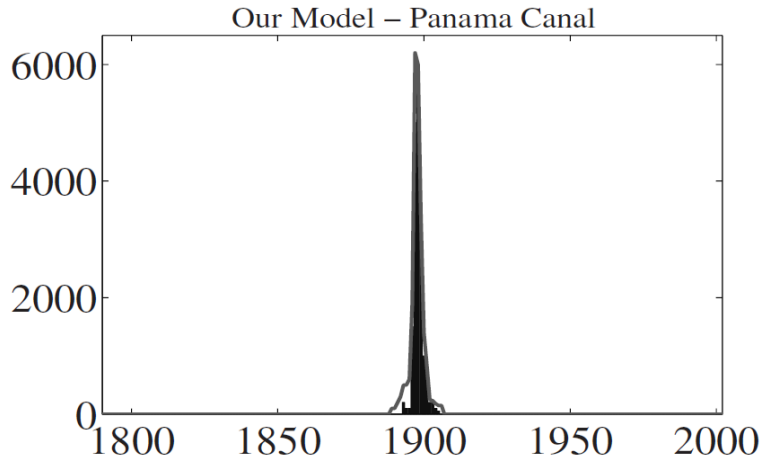
1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents. ECIR 2013.

Пример 3. Совмещение темпоральной и n-граммной модели

По коллекции выступлений президентов США





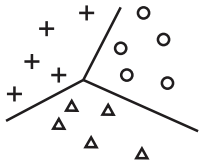
1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

Shoaib Jameel, Wai Lam. An *N*-Gram Topic Model for Time-Stamped Documents. ECIR 2013.

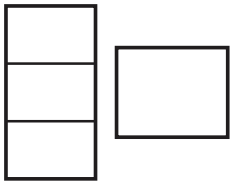
Регуляризаторы для мультимодальных тематических моделей

supervised



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

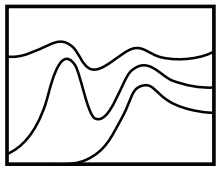
multilanguage



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

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

temporal



Темпоральные модели с модальностью времени i:

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

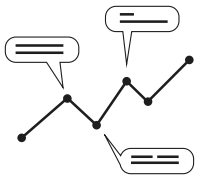
geospatial

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

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

Регуляризаторы для учёта дополнительной информации

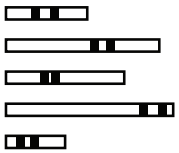
regression



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

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

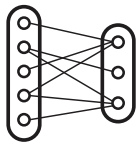
biterm



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

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

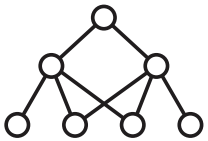
relational



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:

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

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

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

Задача максимизации регуляризованного правдоподобия:

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

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

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

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

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

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

Доказательство (по лемме о максимизации на симплексах)

Применим лемму к log-правдоподобию с регуляризатором R:

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

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

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

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

■

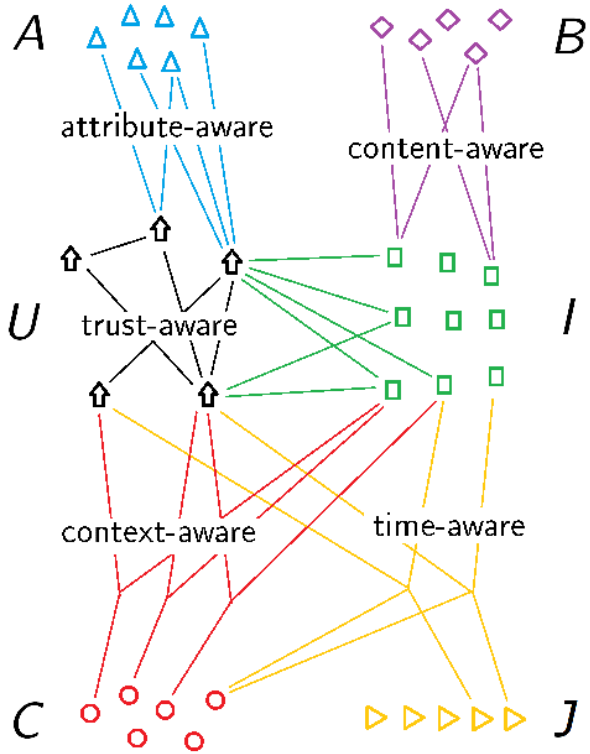
K.Vorontsov. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization // Springer Optimization and Its Applications. 2023

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
- $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ — 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:

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

Максимизация регуляризованного log-правдоподобия

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

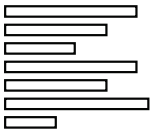
это частный случай гиперграфовой модели, предложения являются гипер-рёбрами.

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.

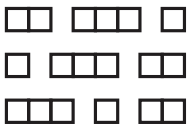
Регуляризаторы для моделирования последовательного текста

sentence



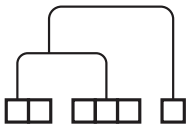
Тематические модели, учитывающие границы предложений, абзацев и секций документов

n-gram



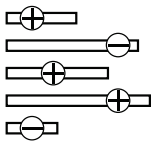
Модели с модальностями *n*-грамм, коллокаций, именованных сущностей (используем TopMine)

syntax



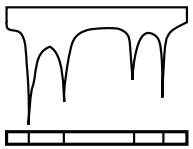
Модели, учитывающие результаты автоматического синтаксического разбора (используем UDPipe)

sentiment



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

segmentation

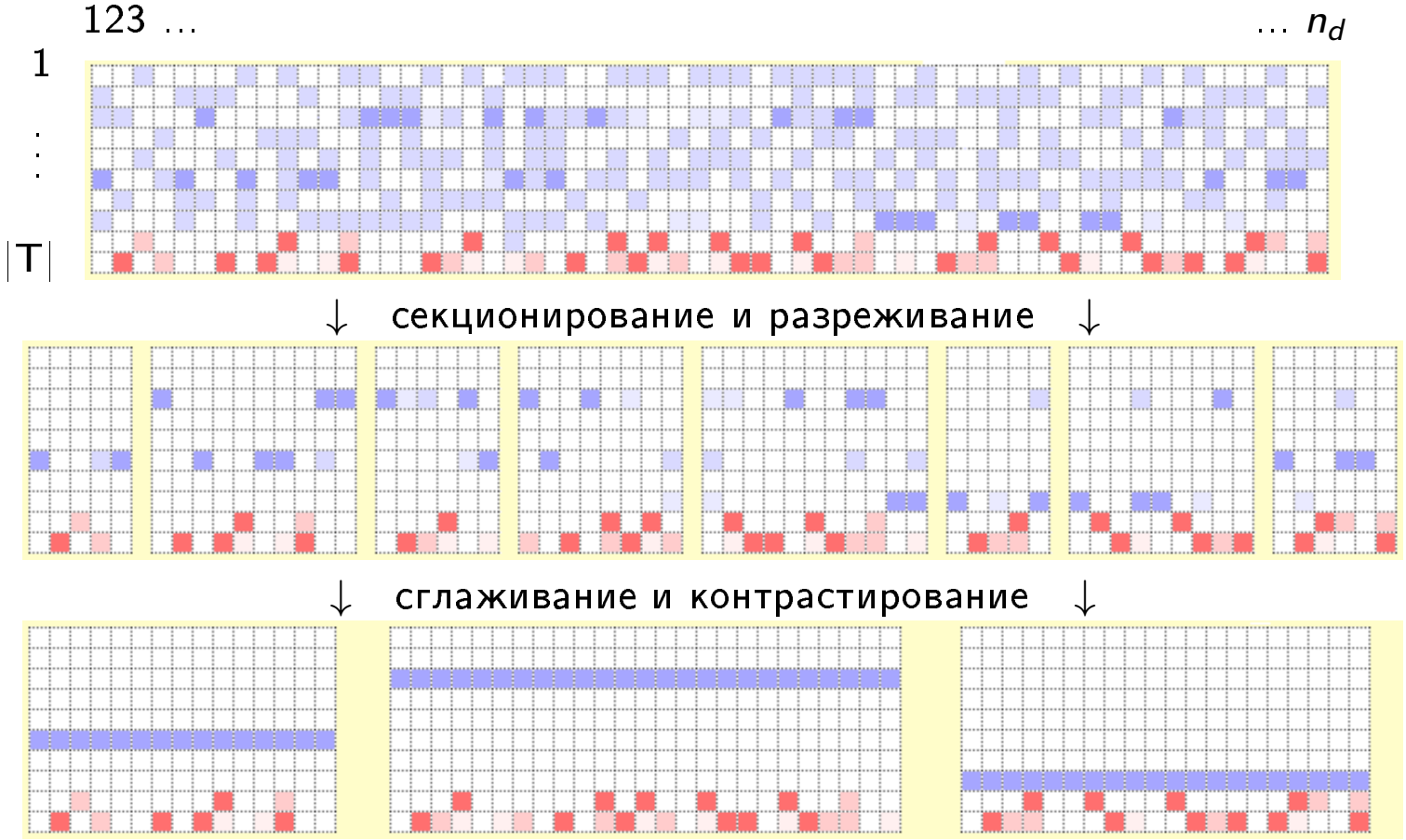


Тематические модели сегментации с автоматическим определением границ сегментов

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

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

EM-алгоритм: метод простой итерации для системы уравнений

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

Набросок доказательства: три леммы

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

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

Введём вспомогательную функцию от переменных Π, Φ, Θ:

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

Лемма 2. Если $R(\Pi,\Phi,\Theta)$ не зависит от p_{tdw} при $w \notin d$, то

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

Лемма 3. Формулы M-шага:

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

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

Гипотеза о пост-обработке E-шага

Между E- и M-шагом добавляется обработка матрицы (p_{tdw}) тематических векторов последовательности термов документа:

$$\tilde{p}_{tdw} = p_{tdw}\left(1 + \frac{1}{n_{dw}}\left(\frac{\partial R(\Pi)}{\partial p_{tdw}} - \sum_{z \in T} p_{zdw}\frac{\partial R(\Pi)}{\partial p_{zdw}}\right)\right) \qquad (1)$$

Пост-обработка E-шага позволяет учитывать порядок термов в документе в обход гипотезы «мешка слов».

Гипотеза

Любое «разумное» преобразование $p_{tdw} \to \tilde{p}_{tdw}$ эквивалентно некоторому регуляризатору $R(\Pi(\Phi, \Theta))$.

Открытый вопрос: при каких условиях по заданным p_{tdw} и $\tilde{p}_{tdw}$ возможно подобрать функцию $R(\Pi)$ так, чтобы выполнялось уравнение пост-обработки (1)?

Однопроходная тематизация текста

Дано: q — фрагмент текста, Φ — готовая тематическая модель
Найти: $p(t|q)$ — тематический вектор фрагмента текста

Проблемы:

- если текст короткий, то определение $p(t|q)$ не надёжно
- согласование $p(t|q)$ с $p(t|w) = \phi_{wt}\frac{p(t)}{p(w)}$ отдельных слов
- согласование $p(t|q)$ с более широким контекстом $d \supset q$

Наводящие соображения:

- первая итерация EM-алгоритма с инициализацией $\theta_{td}^0 = \frac{1}{|T|}$:

$$\theta_{td}(\Phi) = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in W} n_{dw} p_{tdw}\Big) = \sum_{w\in d}\frac{n_{dw}}{n_d}\operatorname*{norm}_{t\in T}(\phi_{wt}\theta_{td}^0)$$

- формула полной вероятности:

$$\theta_{td}(\Phi) = \sum_{w\in d} p(w|d)\,p(t|w) = \sum_{w\in d}\frac{n_{dw}}{n_d}\operatorname*{norm}_{t\in T}\big(\phi_{wt}p(t)\big)$$

Однопроходный по документу EM-алгоритм для ARTM

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

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

EM-алгоритм: метод простой итерации для системы уравнений

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

$$p'_{tdw} = p_{tdw} + \frac{\phi_{wt}}{n_{dw}} \sum_{z \in T} \frac{n_{zd}}{\theta_{zd}(\Phi)} \frac{\partial \theta_{zd}}{\partial \phi_{wt}}$$

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

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 и $\Theta(\Phi)$:

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

Применим лемму к регуляризованному log-правдоподобию Q:

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

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

$$= \sum_{d\in D} n_{dw}\underbrace{\left(p_{tdw} + \frac{\phi_{wt}}{n_{dw}}\sum_{z\in T}\frac{n_{zd}}{\theta_{zd}}\frac{\partial\theta_{zd}}{\partial\phi_{wt}}\right)}_{p'_{tdw}} + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}}$$

■

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

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\phi'_{tw} = \operatorname*{norm}_{t \in T}\left(\phi_{wt} p_t\right); \qquad \theta_{td} = \sum_{w \in d} p_{wd} \phi'_{tw}$$

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

$$p'_{tdw} = p_{tdw} + \frac{\phi'_{tw}}{n_d} \left(\frac{n_{td}}{\theta_{td}} - \sum_{z \in T} \phi'_{zw} \frac{n_{zd}}{\theta_{zd}} \right)$$

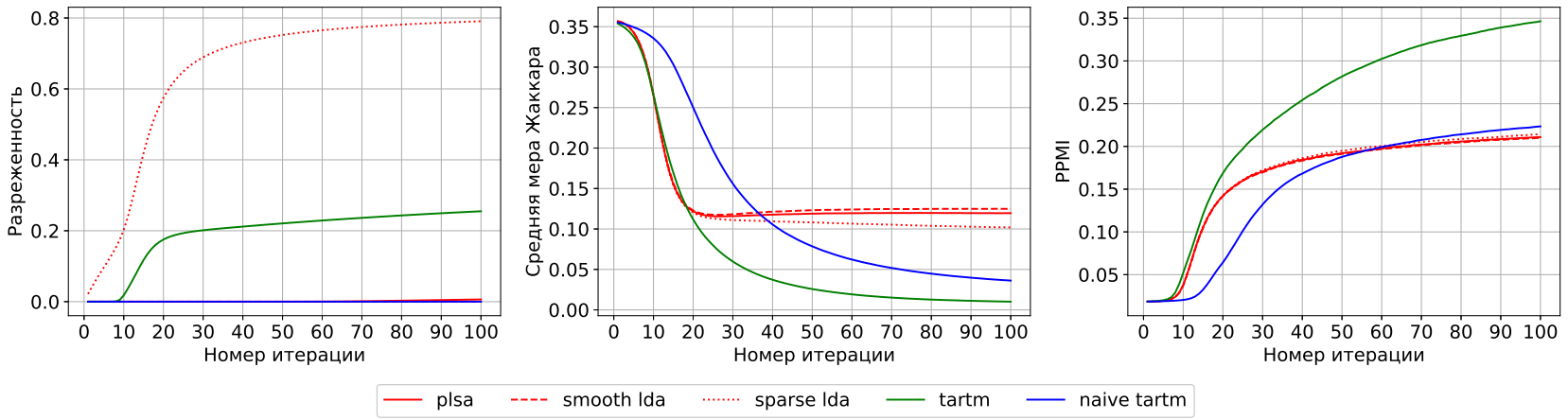
$$\phi_{wt} = \operatorname*{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p'_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$$

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.

Резюме по тематическому моделированию

- 20 лет развития BTM, сотни моделей, тысячи публикаций — переусложнённость теории оставалась незамеченной, был пропущен этап развития теории
- ARTM — математический аппарат тематического моделирования, альтернативный байесовскому обучению
- Тематическое моделирование — «теория одной леммы»
- Эта же лемма применима для обучения нейросетей с неотрицательными нормированными векторами
- Нейросетевые тематические модели — основной тренд TM
- Неотрицательность и нормировка векторов — путь к интерпретируемости нейросетевых моделей?

К.Воронцов. Вероятностное тематическое моделирование: теория регуляризации и проект BigARTM. 2023. (Готовится для издательства URSS)
http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

Концепция «мастерской знаний»

«Огромное и все возрастающее богатство знаний разбросано сегодня по всему миру. Этих знаний, вероятно, было бы достаточно для решения всего громадного количества трудностей наших дней, но они рассеяны и неорганизованы. Нам необходима очистка мышления в *своеобразной мастерской*, где можно получать, сортировать, суммировать, усваивать, разъяснять и сравнивать знания и идеи»
— *Герберт Уэллс, 1940*

"An immense and ever-increasing wealth of knowledge is scattered about the world today; knowledge that would probably suffice to solve all the mighty difficulties of our age, but it is dispersed and unorganized. We need a sort of mental clearing house for the mind: a depot where knowledge and ideas are received, sorted, summarized, digested, clarified and compared"
— *Herbert Wells, 1940*

Концепция сервиса тематического разведочного поиска

Подборка — долгосрочный поисковый интерес пользователя

Поисково-рекомендательные функции:

- поиск тематически близких документов по подборке
- мониторинг новых документов для подборки
- контекстные рекомендации по документу из подборки

Аналитические функции:

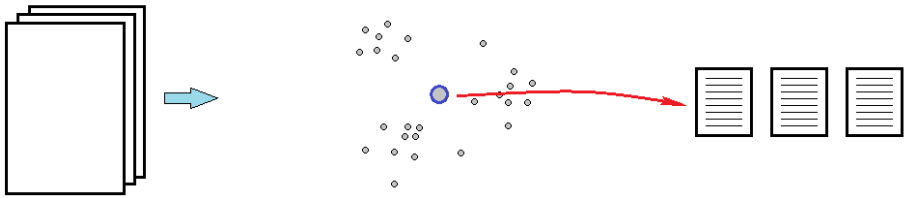
- автоматизация реферирования подборки
- кластеризация трендов, методов, мнений в подборке
- рекомендация порядка чтения внутри подборки
- выделение «важных мест» в документе из подборки

Коммуникативные функции:

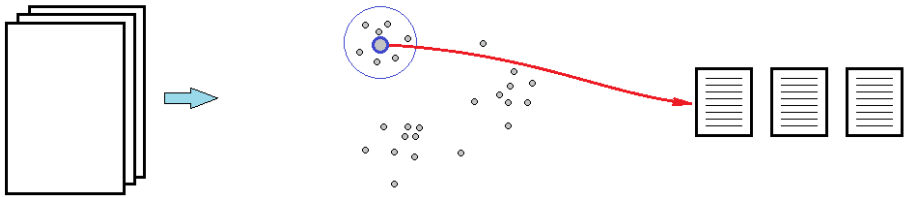
- совместное составление и использование подборок
- интерактивная визуализация и инфографика по подборке

Стратегии поиска документов по тематическим векторам

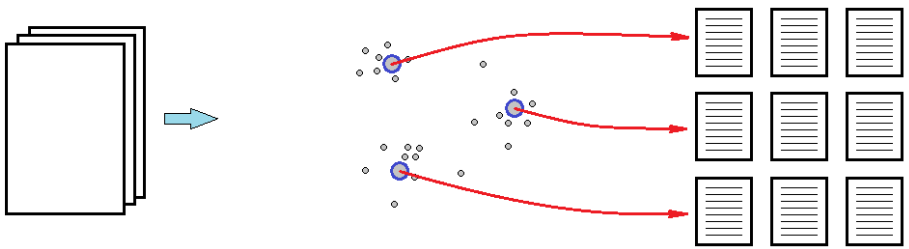
Поиск по среднему вектору подборки (неудачная стратегия):



Поиск по части подборки или по отдельному документу:

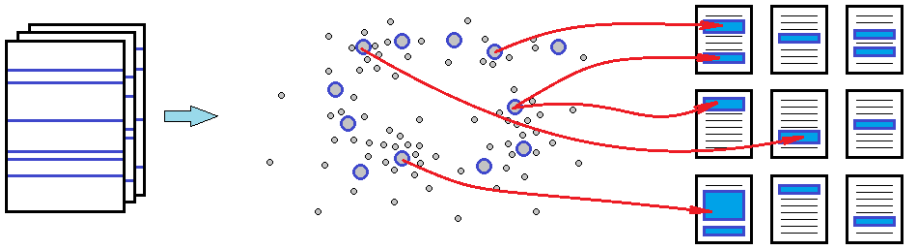


Поиск по тематике кластеров, на которые делится подборка:

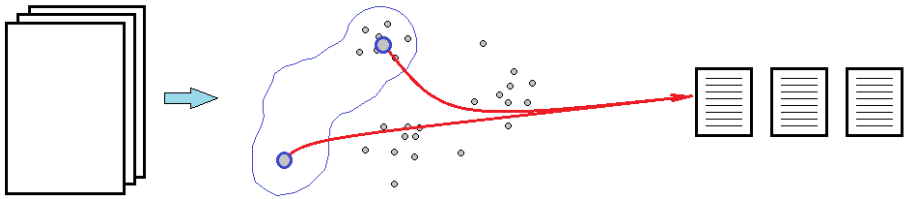


Стратегии поиска документов по тематическим векторам

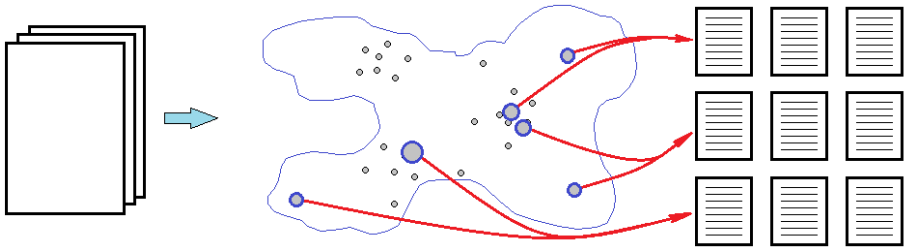
Поиск по тематике сегментов документов:



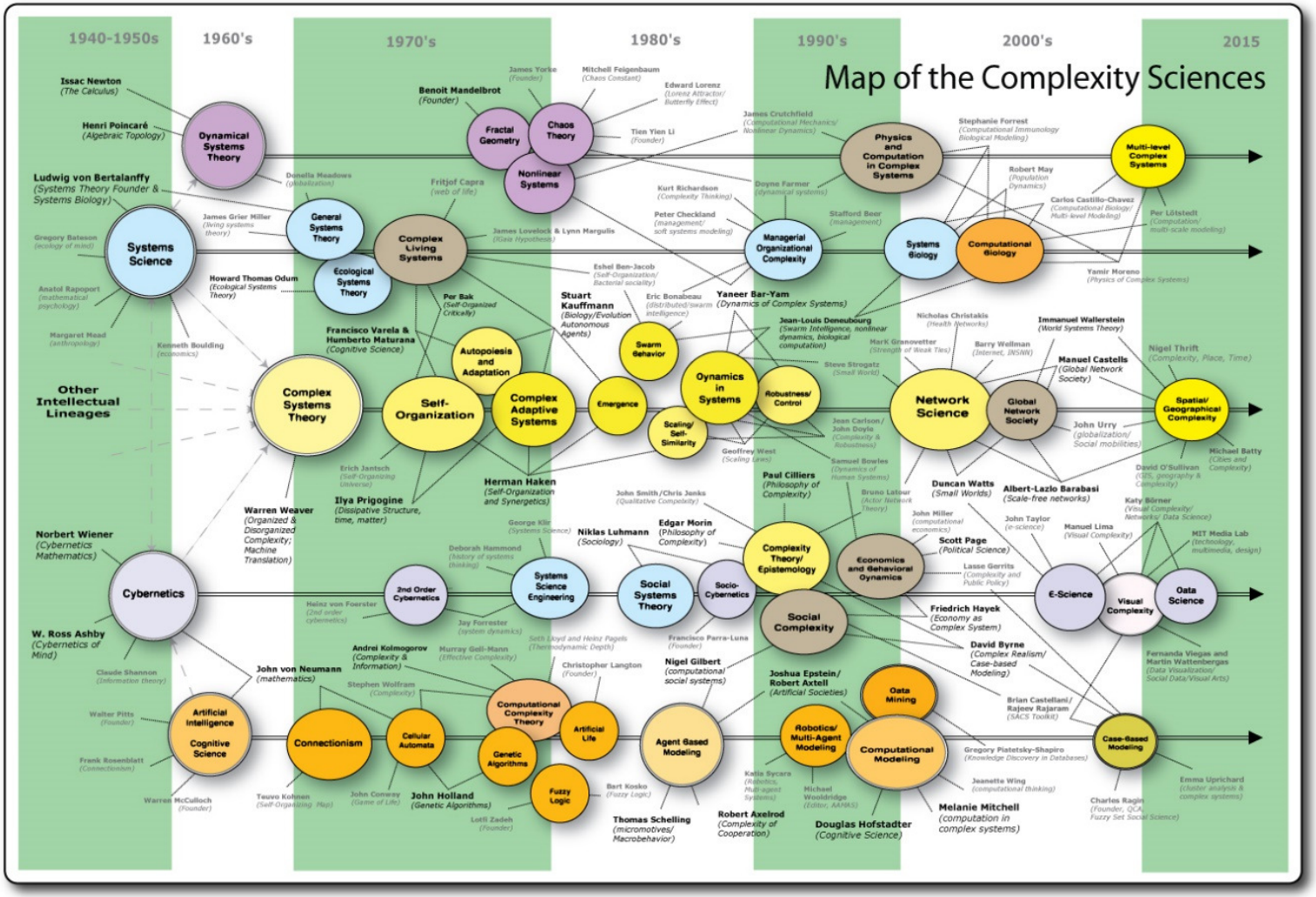
Поиск по тематике, смежной для части подборки:



Поиск по тематике, смежной для всей подборки:



Пример карты предметной области (построено вручную)



http://www.theoryculturesociety.org/brian-castellani-on-the-complexity-sciences

Концепция «новостного коллайдера»

Цель создания адронного коллайдера — сталкивая потоки частиц, узнать больше о строении материи

Цель создания новостного коллайдера — сталкивая потоки новостей, защитить общество от угроз эпохи постправды и информационных войн

Типология угроз и задачи их автоматической детекции

воздействия → фейки → пропаганда → инф.война

1. детекция приёмов манипулирования
2. детекция замалчивания
3. детекция обмана (deception detection), слухов (rumors d.), мистификаций (hoaxes d.)
4. детекция кликбэйта (clickbait detection)
5. автоматическая проверка фактов (auto fact-checking)
6. детекция позиции (stance d.), противоречий (controversy d.), поляризации (polarization d.)
7. выявление конструктов картины мира: идеологем, мифологем
8. оценивание возможных психо-эмоциональных реакций
9. выявление целевых аудиторий воздействия
10. оценивание и предсказание скорости распространения (virality prediction)
11. оценивание достоверности источников (credibility scores)
12. детекция прямой агрессии (угрозы, призывы, провокации, вербовка, экстремизм)

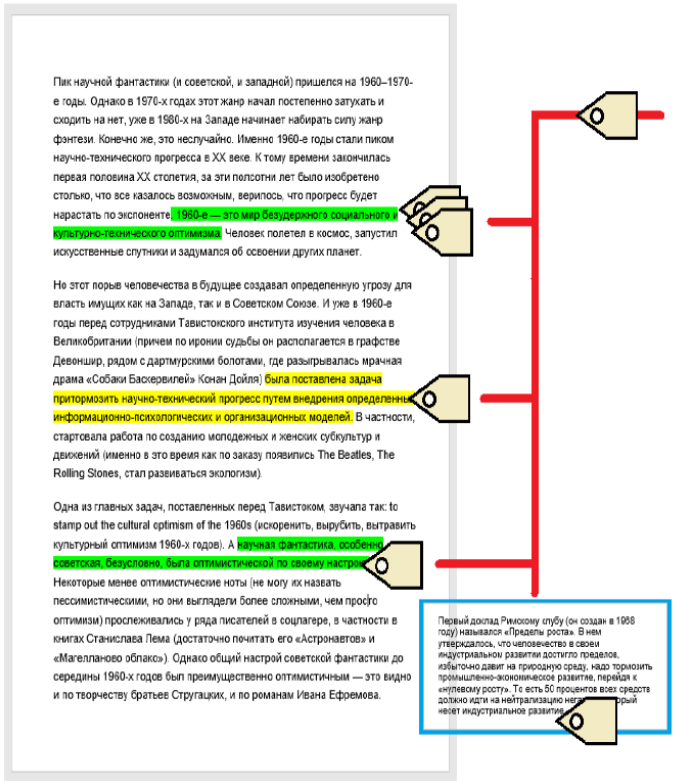
E.Saquete, D.Tomas, P.Moreda, P.Martinez-Barco, M.Palomar. Fighting post-truth using natural language processing: A review and open challenges // Expert Systems With Applications, Elsevier, 2020.

Типы задач ML/NLU для мониторинга медиа-пространства

1. **Классификация текста (сообщения/предложения) целиком**
   - deception detection, fact-checking, text credibility
2. **Классификация пары текстов**
   - stance, controversy, polarization, clickbait detection
   - выявление противоречий, разногласий, замалчивания
3. **Разметка текста (выделение и классификация фрагментов)**
   - поиск лингвистических маркеров (linguistic-based cues) в тексте
   - детекция приёмов манипулирования
   - выявление конструктов картины мира: мифологем, идеологем
   - выявление психо-эмоциональных реакций и целевых аудиторий
4. **Кластеризация или тематическое моделирование**
   - кластеризация мнений по заданной теме (controversy detection)
   - выявление поляризованных мнений (polarization detection)
   - выявление мнений как сочетаний слов, семантических ролей и тональностей
   - выявление «картин мира» – устойчивых сочетаний суждений и идеологем

Профессиональная разметка: путь к стандартизации

Обобщение классических задач компьютерной лингвистики NER, SemAn, SemRL, SyntPars и др.; выявления манипуляций, поляризации, смысловых ошибок в академических эссе



Разметка состоит из элементов

Элемент разметки может содержать любое число фрагментов, затекстов и тегов

Теги (классы) выбираются из словаря тегов

Фрагмент задаётся началом и концом, может иметь один или несколько тегов:

BEGIN была поставлена задача притормозить научно-технический прогресс путем внедрения определен информационно-психологических и организационных моделей. END TAG

Затекст может выбираться из словаря фраз или свободно генерироваться по контексту, может иметь один или несколько тегов

Технический регламент конкурса ПРО//ЧТЕНИЕ (http://ai.upgreat.one)

Оценивание моделей разметки: путь к стандартизации

- В основе методики — сравнение пар разметок текста: «модель – эксперт», «эксперт-1 – эксперт-2», путём оптимального сопоставления их элементов
- Вводится мера согласованности пары разметок (A,B); если она измеряется несколькими критериями, то берётся их средневзвешенная согласованность Con(A,B)
- СТАР (Средняя Точность Алгоритмической Разметки) — средняя по размеченной выборке согласованность Con(A,E) разметки модели A и разметки эксперта E
- СТЭР (Средняя Точность Экспертной Разметки) — средняя по размеченной выборке согласованность Con(E1,E2) разметок двух экспертов, E1 и E2
- ОТАР = СТАР / СТЭР, если больше 100%, то модель лучше экспертов

Технический регламент конкурса ПРО//ЧТЕНИЕ (http://ai.upgreat.one)

Проект «Тематизатор»: общие требования

Переход от библиотек (BigARTM, VisARTM, TopicNet)
к приложению «Тематизатор» для конечного пользователя —
аналитика в области цифровых гуманитарных исследований

1. Цели пользователя — разведочный анализ, понимание тематической структуры данных, «о чём эта коллекция»
2. Пользователь не обязан знать
   — форматы исходных данных и способы их предобработки
   — теорию TM и ARTM, виды регуляризаторов
   — методики подбора гиперпараметров
   — критерии качества моделей
   — библиотеку BigARTM
3. Интуитивная визуальная среда, веб-интерфейс
4. Пользователю должны быть доступны настройки
5. Дефолтные настройки должны работать на любых данных

Приложения и исследования, взятые для анализа требований

1. Выявление этно-релевантных тем в социальных медиа
2. Анализ программ развития российских вузов
3. Проекты Школы Прикладного Анализа Данных
4. Тематизация twitter о российско-украинских отношениях
5. Тематизация научно-просветительского онлайн-журнала
6. Тематизация пресс-релизов внешнеполитических ведомств
7. Выявление трендов в коллекции научных публикаций
8. Выявление событийных тем в новостных потоках
9. Разведочный поиск в технологических блогах
10. Поиск и рубрикация научных статей на 100 языках
11. Поиск похожих дел в актах арбитражных судов
12. Построение «банков тем» в проекте TopicNet

Функциональные требования (по приоритетности)

1. Визуализация множества всех тем и их характеристик
2. Визуализация каждой темы с её «рассказом о себе»
3. Возможность задавать словари затравок для (групп) тем
4. Определение динамики тем во времени
5. Выявление трендов — тем с динамикой быстрого роста
6. Разбиение тем на подтемы иерархически
7. Выявление связей тем по сочетаемости в документах
8. Возможность отбора «банка тем» вручную
9. Тематическая фильтрация коллекции
10. Возможность группировки тем вручную
11. Тематический поиск по документу или фрагменту
12. Рекомендательный поиск и построение подборок

Требования к интерпретируемости (по приоритетности)

1. Доля интерпретируемых тем близка к 100%
2. Темы строятся более на терминах, чем на словах
3. Общая лексика выводится в отдельные фоновые темы
4. Нет мусорных тем, нет тем-дубликатов (декорреляция)
5. Решена проблема несбалансированности тем
6. Темы способны рассказать о себе словами и фразами
7. Темы именуются автоматически
8. Нетекстовые термы способны рассказать о себе словами
9. Темы тестируются на однородность по сочетаемости слов
10. В иерархии имена дочерних тем уточняют родительские
11. Для коротких текстов строится дистрибутивная модель
12. Длинные тексты разбиваются на тематические сегменты

Основной пользовательский сценарий (без детализации)

1. **Загрузка**
   - данные в различных «сырых» форматах
   - возможна дозагрузка данных порциями
2. **Предобработка**
   - автоматический выбор обработчиков на основании данных
   - выделение модальностей: языков, времени, терминов и т.д.
3. **Моделирование**
   - визуализация метрик качества в процессе обучения модели
   - возможность перехода к анализу, не прерывая обучения
4. **Визуализация**
   - каждая тема должна уметь «рассказать о себе»
   - много разных графиков (distant reading)
5. **Коррекция**
   - перебор моделей и накопление «банка тем»
   - пользовательские темы как подборки с рекомендациями

Требования к функциям Загрузки

1. Загрузка коллекций из различных сырых форматов
2. — txt, json, docx, odt, pdf и др.
3. — СМИ, соцмедиа, Википедия, статьи, патенты и др.
4. Представление метаданных и модальностей
5. Возможность загрузки как локально, так и из облака
6. Возможность дозагрузки данных из источника порциями
7. Текст как последовательность или как «мешок слов»
8. В одном файле один документ или много документов

Требования к функциям Предобработки

1. Автоматическая токенизация и лемматизация
2. Автоматическое исправление опечаток (соцсети)
3. Автоматическое выделение терминов n-грамм
4. Метаданные: авторы, время, категории, заголовки и др.
5. Модальности: онимы, теги, ссылки, пользователи и др.
6. Настройка шаблонов для выделения модальностей
7. Сортировка по времени и нарезка по пакетам
8. Автоматическое определение коротких текстов
9. Автоматическая редукция словарей (по необходимости)
10. Автоматическое определение языков
11. Машинный перевод для получения параллельных текстов
12. Предобработка не должна идти дольше тематизации

Требования к функциям Моделирования

1. Визуализация процесса обучения модели
2. Вывод метрик на графиках от #итерации, #пакета
3. Метрики перплексии, разреженности, вырожденности и др.
4. Автоматическая подстройка под короткие тексты
5. Автоматическая подстройка под длинные тексты
6. Темпоральная модель, если есть модальность времени
7. Подбор числа тем или построение иерархии тем
8. Автоматический подбор гиперпараметров, AutoML
9. Логирование информации о найденных аномалиях
10. Логирование данных о моделях, журнал экспериментов
11. Возможность перехода к анализу, не прерывая обучения
12. Возможность замены BigARTM на альтернативы

Требования к функциям Визуализации

1. Визуальная навигация по темам, документам, терминам
2. XY-график тем в осях характеристик тем
3. XY-график объектов в осях объёмов тем или групп тем
4. Построение спектра тем по семантической близости
5. XY-график документов в осях «время–спектр тем»
6. XY-график документов темы с осью тематичности
7. Визуализация динамики тем в осях «время–объём темы»
8. Визуализация иерархии тем
9. Визуализация связей тем по их сочетаемости в документах
10. Визуализация тематической структуры документа
11. Выбор характеристик тем для осей XY-графиков
12. Выбор характеристик объектов и документов для осей

Требования к функциям Коррекции

1. Разметка тем на релевантные, нерелевантные, мусорные
2. Разметка релевантных термов, документов в темах
3. Термы-затравки для «классификации иголок в стоге сена»
4. Обнаружение и расщепление неоднородных тем
5. Автоматический переход к тематической иерархии
6. Детекция новых событийных тем в темпоральных моделях
7. Накопление «банка тем» по множеству моделей
8. Многокритериальное оценивание качества моделей
9. Планирование экспериментов по улучшению моделей
10. Тематическая фильтрация коллекции и потока
11. Создание пользовательских тем — подборок документов
12. Ранжирование рекомендаций для пользовательских тем

Требования к рабочему пространству проекта пользователя

1. Настройки входных данных — коллекций и потоков
2. Настройки модулей предобработки
3. Структура и гиперпараметры сравниваемых моделей
4. Структура и гиперпараметры финальной модели
5. Визуализации процесса обучения модели
6. Визуализации количественных результатов моделирования
7. Визуализации качественных результатов (аннотации тем)
8. Банк тем — множество тем, отобранных из моделей
9. Пользовательские темы — подборки документов
10. Настройка подробности отчёта по проекту
11. Настройка комментариев к пунктам отчёта по проекту
12. Сгенерированный отчёт по проекту

Открытые (частично) задачи, где требуются исследования

1. Доля интерпретируемых тем близка к 100%
2. Проблема несбалансированности тем
3. Проблема внутритекстовой когерентности тем
4. Модель Θ-less для коротких и длинных текстов
5. Обнаружение и расщепление неоднородных тем
6. Обнаружение новых тем в пакетах
7. Накопление «банка тем» по множеству моделей
8. Автоматический переход к тематической иерархии
9. Автоматический подбор гиперпараметров, AutoML
10. Оптимизация гиперпараметров в потоковом режиме
11. Именование и аннотирование темы
12. Применение гиперграфовых тематических моделей

Вместо резюме

Основная тенденция ТМ — нейросетевые тематические модели

- Открытая проблема — «объединить лучшее от двух миров»:
  — покоординатную интерпретируемость BTM
  — глубину и выразительность нейросетевых моделей языка
- Что для этого уже есть:
  — лемма о максимизации на симплексах
  — тематические векторы слов-в-контексте $p(t|d, w_i)$
  — одно(двух)проходные алгоритмы тематизации текста
  — реализация ARTM в библиотеке BigARTM —
  возможность вычислять градиенты методом BackProp
- Чего не хватает:
  — уверенности, что смысл определяется тематикой
  — реализации и экспериментов

К.Воронцов. Вероятностное тематическое моделирование: теория регуляризации и проект BigARTM. 2023. (Готовится для издательства URSS)
http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf