

Тематическое моделирование для информационного поиска

Воронцов Константин Вячеславович
(МГУ, МФТИ, ФИЦ ИУ РАН)

Центр прикладного анализа больших данных
Томский государственный университет
17 марта 2022

1 Теория

- Вероятностное тематическое моделирование
- Теория аддитивной регуляризации
- Мультимодальные тематические модели

2 Проект BigARTM

- Библиотеки BigARTM и TopicNet
- Модульность аддитивной регуляризации
- Эксперименты: качество и скорость

3 Эксперименты с тематическим поиском

- Задача тематического информационного поиска
- Модели тематического поиска
- Результаты и выводы

Постановка задачи тематического моделирования

Дано: коллекция текстовых документов

- n_{dw} — частоты термов в документах, $\hat{p}(w|d) = \frac{n_{dw}}{n_d}$

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

- $\phi_{wt} = p(w|t)$ — вероятности термов w в каждой теме t
- $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Критерий: максимум правдоподобия

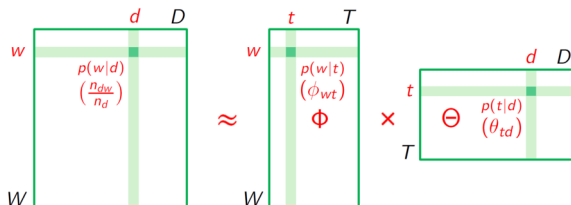
$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta}$$

при ограничениях неотрицательности и нормировки

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

Задача тематического моделирования некорректно поставлена

Тематическое моделирование — это задача низкорангового стохастического матричного разложения ($|T| \ll |D|, |W|$)



Задача *некорректно поставлена* по Адамару, так как множество её решений в общем случае бесконечно: если $\Phi\Theta$ — решение, то $(\Phi S)(S^{-1}\Theta)$ — тоже решение

Регуляризация — доопределение решения с помощью дополнительного критерия: $L(\Phi, \Theta) + \tau R(\Phi, \Theta) \rightarrow \max$

PLSA и LDA — наиболее известные тематические модели

PLSA, Probabilistic Latent Semantic Analysis:

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

LDA, Latent Dirichlet Allocation:

$$R(\Phi, \Theta) = \beta_0 \sum_t \sum_w \beta_w \ln \phi_{wt} + \alpha_0 \sum_d \sum_t \alpha_t \ln \theta_{td}$$

- распределения ϕ_t близки к заданному распределению β
- распределения θ_d близки к заданному распределению α

Можно также говорить об априорных распределениях Дирихле, но регуляризация проще и понятнее, а результат тот же

Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.

Blei D., Ng A., Jordan M. Latent Dirichlet allocation. 2003.

Воронцов К.В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM // www.MachineLearning.ru, 2022.

ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация \log правдоподобия с регуляризатором R :

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

EM-алгоритм: метод простой итерации для системы уравнений

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

где $\operatorname{norm}_{t \in T}(x_t) = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$ — операция нормировки вектора.

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

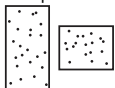
Регуляризаторы для улучшения интерпретируемости тем

background

Сглаживание фоновых тем $B \subset T$:

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

sparse

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

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

decorrelated



Декоррелирование для повышения различности тем:

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

interpretable



Сглаживание + разреживание + декоррелирование для улучшения интерпретируемости тем

Послойное построение уровней тематической иерархии

Шаг 1. Строим модель с небольшим числом тем.

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

Родительские темы приближаются смесями дочерних тем:

$$\sum_{t \in T} n_t \text{KL}_w \left(p(w|t) \parallel \sum_{s \in S} p(w|s)p(s|t) \right) \rightarrow \min_{\Phi, \Psi}$$

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

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

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

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

Построение второго уровня иерархии с подтемами S

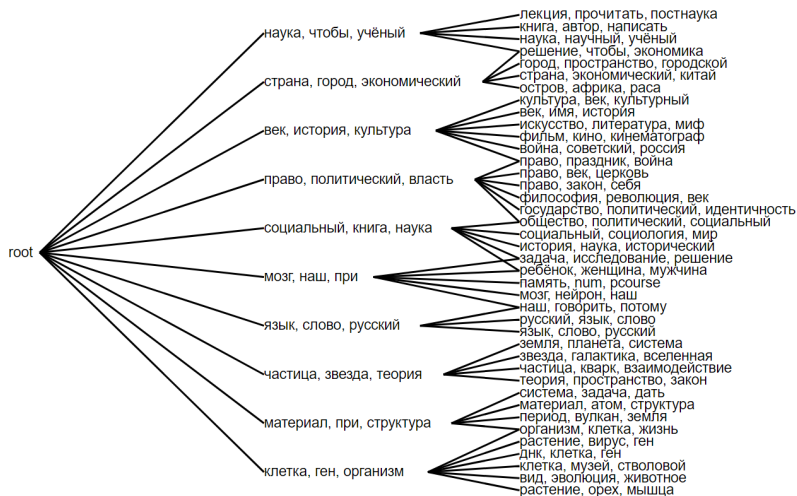
В коллекцию добавляются $|T|$ псевдодокументов родительских тем с частотами термов $n_{wt} = \tau n_t \phi_{wt}$, $t \in T$

$$\begin{array}{c}
 \begin{array}{ccc}
 & D & \\
 W & \begin{array}{|c|} \hline n_{dw} \\ \hline \end{array} & = \\
 & & \begin{array}{ccc}
 & T & \\
 W & \begin{array}{|c|} \hline \Phi^1 \\ \hline \end{array} & \times_T & \begin{array}{|c|} \hline \Theta^1 \\ \hline \end{array} \\
 & & & D
 \end{array} \\
 \\
 \begin{array}{ccc}
 & D & T \\
 W & \begin{array}{|c|c|} \hline n_{dw} & n_{wt} \\ \hline \end{array} & = \\
 & & \begin{array}{ccc}
 & S & \\
 W & \begin{array}{|c|} \hline \Phi^2 \\ \hline \end{array} & \times_S & \begin{array}{|c|c|} \hline \Theta^2 & \Psi \\ \hline \end{array} \\
 & & & D & T
 \end{array}
 \end{array}$$

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

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

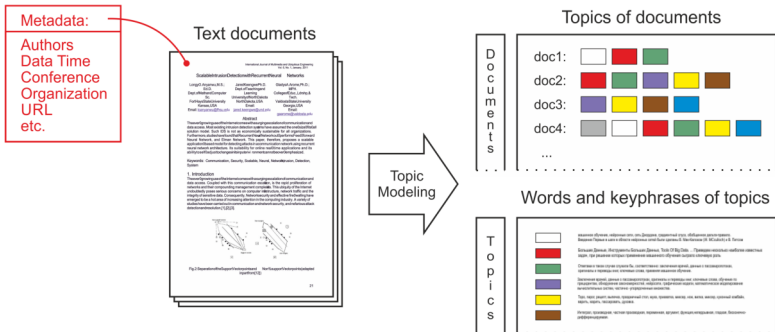
Пример тематической иерархии (коллекция postnauka.ru)



Дмитрий Федоряка. Технология интерактивной визуализации тематических моделей. Бакалаврская диссертация, МФТИ, 2017.

Мультимодальная тематическая модель

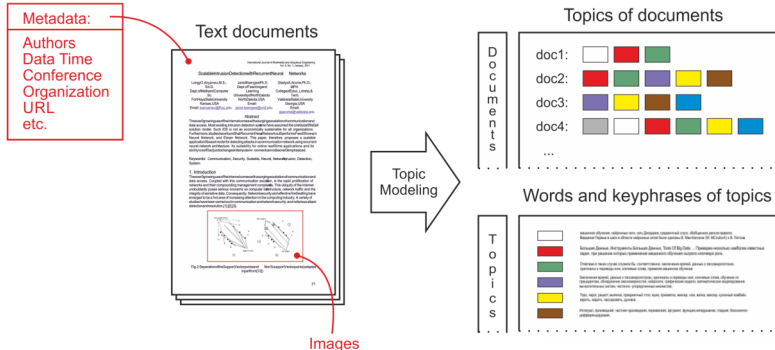
Тема может порождать термины различных *модальностей*:
 $p(\text{слово} | t)$, $p(n\text{-грамма} | t)$, $p(\text{автор} | t)$, $p(\text{время} | t)$, $p(\text{источник} | t)$,



Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

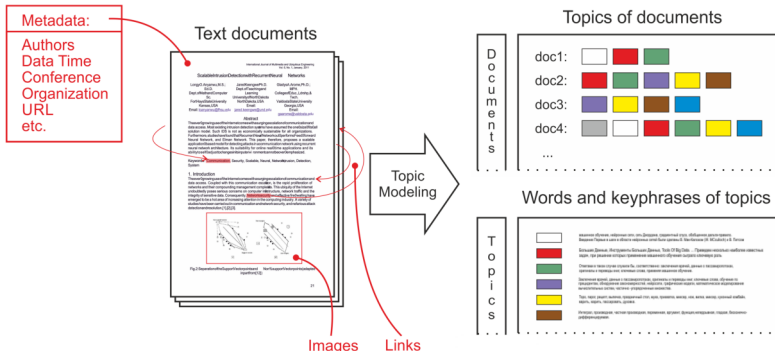
$p(\text{слово} | t)$, $p(n\text{-грамма} | t)$, $p(\text{автор} | t)$, $p(\text{время} | t)$, $p(\text{источник} | t)$,
 $p(\text{объект} | t)$,



Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

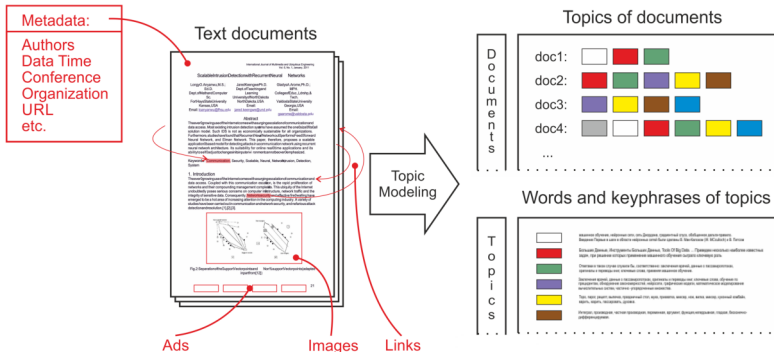
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
 $p(\text{объект}|t)$, $p(\text{ссылка}|t)$,



Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

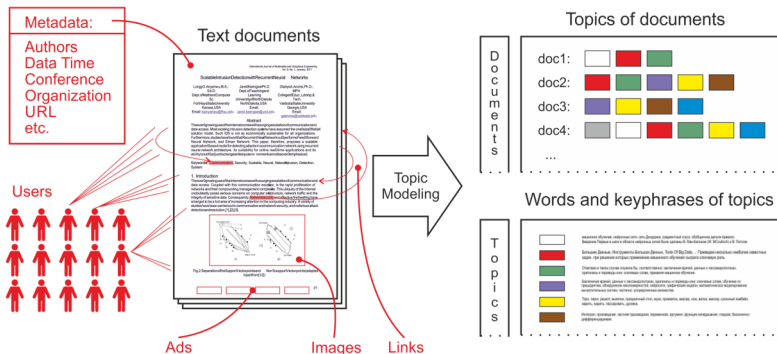
$p(\text{слово} | t)$, $p(n\text{-грамма} | t)$, $p(\text{автор} | t)$, $p(\text{время} | t)$, $p(\text{источник} | t)$,
 $p(\text{объект} | t)$, $p(\text{ссылка} | t)$, $p(\text{баннер} | t)$,



Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

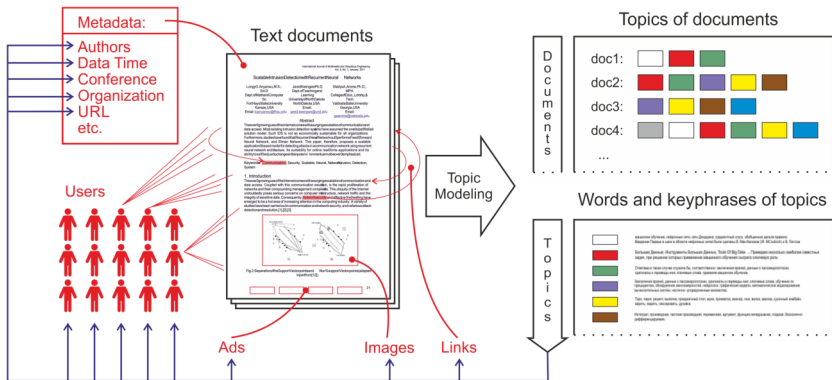
$p(\text{слово} | t)$, $p(n\text{-грамма} | t)$, $p(\text{автор} | t)$, $p(\text{время} | t)$, $p(\text{источник} | t)$,
 $p(\text{объект} | t)$, $p(\text{ссылка} | t)$, $p(\text{баннер} | t)$, $p(\text{пользователь} | t)$



Мультимодальная тематическая модель

Тема может порождать термины различных *модальностей*:

$p(\text{слово} | t)$, $p(n\text{-грамма} | t)$, $p(\text{автор} | t)$, $p(\text{время} | t)$, $p(\text{источник} | t)$,
 $p(\text{объект} | t)$, $p(\text{ссылка} | t)$, $p(\text{баннер} | t)$, $p(\text{пользователь} | t)$



Мультимодальная ARTM

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

Максимизация суммы \log правдоподобий с регуляризацией:

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

EM-алгоритм: метод простой итерации для системы уравнений

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

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

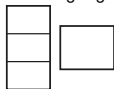
Регуляризаторы для мультимодальных тематических моделей

supervised



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

multilanguage

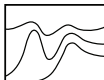


Модальность языков и регуляризация со словарём

$\pi_{uwt} = p(u|w, t)$ 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 :

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

geospatial



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

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

Пример 1. Мультиязычная 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

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

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

Пример 1. Мультиязычная 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

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Freij, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример 2. Биграммная модель научных конференций

Коллекция 1000 статей конференций ММРО, ИОИ на русском

распознавание образов в биоинформатике		теория вычислительной сложности	
униграммы	биграммы	униграммы	биграммы
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.

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Онлайн-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>



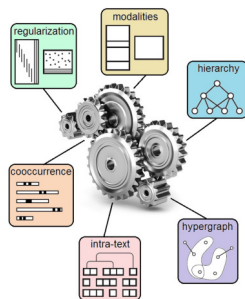
Лицензия и среда разработки:

- 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 связности текста



TopicNet

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

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

BigARTM упрощает разработку тематических моделей

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

Этапы моделирования

Bayesian TM

ARTM

	Анализ требований	Анализ требований	
<i>Формализация:</i>	Вероятностная модель порождения данных	Стандартные критерии	Свои критерии
<i>Алгоритмизация:</i>	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Единый регуляризованный EM-алгоритм для любых моделей и их композиций	
<i>Реализация:</i>	Исследовательский код (Matlab, Python, R)	Промышленный код BigARTM (C++, Python API)	
<i>Оценивание:</i>	Исследовательские метрики, исследовательский код	Стандартные метрики	Свои метрики
	Внедрение	Внедрение	

-- нестандартизуемые этапы, уникальная разработка для каждой задачи

-- стандартизуемые этапы

Модульность аддитивной регуляризации

Мешок регуляризаторов под каждую прикладную задачу

Выявления этнорелевантного дискурса в социальных сетях:

$$\mathcal{L} \left(\begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array} \right) + R \left(\begin{array}{c} \text{interpretable} \\ \text{[Bar chart]} \quad \text{[Scatter plot]} \end{array} \right) + R \left(\begin{array}{c} \text{n-gram} \\ \text{[Grid of boxes]} \end{array} \right) + R \left(\begin{array}{c} \text{seed words} \\ \text{[Bar chart]} \quad \text{[Box]} \end{array} \right) \rightarrow \max$$

Тематический поиск научных и научно-популярных статей:

$$\mathcal{L} \left(\begin{array}{c} \text{multimodal} \\ \text{[Stacked boxes]} \quad \text{[Box]} \end{array} \right) + R \left(\begin{array}{c} \text{interpretable} \\ \text{[Bar chart]} \quad \text{[Scatter plot]} \end{array} \right) + R \left(\begin{array}{c} \text{n-gram} \\ \text{[Grid of boxes]} \end{array} \right) + R \left(\begin{array}{c} \text{hierarchy} \\ \text{[Tree diagram]} \end{array} \right) \rightarrow \max$$

Выявление и прослеживание событий в новостном потоке:

$$\mathcal{L} \left(\begin{array}{c} \text{multimodal} \\ \text{[Stacked boxes]} \quad \text{[Box]} \end{array} \right) + R \left(\begin{array}{c} \text{interpretable} \\ \text{[Bar chart]} \quad \text{[Scatter plot]} \end{array} \right) + R \left(\begin{array}{c} \text{temporal} \\ \text{[Line graph]} \end{array} \right) + R \left(\begin{array}{c} \text{sentiment} \\ \text{[Sentiment diagram]} \end{array} \right) \rightarrow \max$$

Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

3.7М статей Википедии, 100К 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.

Две коллекции новостей про технологии

Nabrahbr.ru

175 143 статей на русском
10 552 слов (униграмм)
742 000 биграмм
524 авторов статей
10 000 авторов комментариев
2546 тегов
123 хаба (категории)

TechCrunch.com

759 324 статей на английском
11 523 слов (униграмм)
1.2 млн. биграмм
605 авторов
184 категорий

Предобработка текстов

- отброшены 5% наиболее частотных слов (общая лексика)
- удаление пунктуации
- нижний регистр, ё→е
- лемматизация rymorphy2

Методика оценивания качества разведочного поиска

Поисковый запрос

набор ключевых слов или фрагментов текста, около одной страницы A4

Поисковая выдача

документы d с распределением $p(t|d)$, близким к распределению $p(t|q)$ запроса

Два задания ассессорам

- 1 найти как можно больше статей, пользуясь любыми средствами поиска (и засечь время)
- 2 оценить релевантность поисковой выдачи на том же запросе

Поисковик MapReduce

Поисковик MapReduce – программа поиска (поисковик) написанная распределенным вычислением для больших объемов данных и работающая параллельно шардებს, представляющая собой набор Java-классов и исполняемых утилит для создания и обработки данных на параллельной обработке.

Основные возможности Поисковика MapReduce можно сформулировать как:

- обработка написанных больших объемов данных;
- масштабируемость;
- автоматическое распределение данных;
- работа на невидимых оборудовании;
- автоматическая обработка отказов написанных данных.

Поисковик – популярная программная платформа (язык Java, библиотека) построения распределенных приложений для массово-параллельной обработки (задачи: анализ, обработка, поиск, МРТ) данных.

Поисковик включает в себе следующие компоненты:

1. HDFS – распределенная файловая система;
2. **Поисковик MapReduce** – программная платформа (язык Java) написанная распределенным вычислением для больших объемов данных и работающая параллельно шардებს.

Клиентские приложения в архитектуре **Поисковика MapReduce** и структура HDFS, стали привычной частью жизни в среде вычисления, в том числе и в отношении точки отказа. Что, в конечном итоге, определило ограничение платформ **Поисковик** в целом. К сожалению можно отметить:

Ограничение масштабируемости кластера **Поисковик** –4K вычислительных узлов, –40K параллельных заданий.

Сильная зависимость **Поисковика** распределенного вычисления и клиентских приложений, реализованных распределенной программой. Как следствие:

Отсутствие поддержки альтернативной программной модели написанных распределенных вычислений в **Поисковике** v1.0 поддерживается только модель написанных шардებს.

Наличие единой точки отказа и как следствие, необходимость написываться в среде с высоким требованием к надежности;

Проблема совместности требований по единовременному обслуживанию всех вычислительных узлов кластера при обслуживании платформ **Поисковик** (установка новых версий или пакета обновлений).

Пример запроса для разведочного поиска

Векторный поиск тематически близких документов

$\theta_{tq} = p(t|q)$ — тематический вектор запроса q

$\theta_{td} = p(t|d)$ — тематические векторы документов $d \in D$

Косинусная мера близости документа d и запроса q :

$$\text{sim}(q, d) = \frac{\sum_t \theta_{tq} \theta_{td}}{(\sum_t \theta_{tq}^2)^{1/2} (\sum_t \theta_{td}^2)^{1/2}}.$$

Ранжируем документы коллекции $d \in D$ по убыванию $\text{sim}(q, d)$

Выдача тематического поиска — k первых документов.

Реализация: *векторный индекс* для быстрого поиска документов d по каждой из тем t запроса

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

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

Какие модели поиска сравнивались

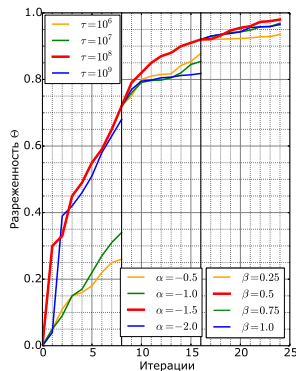
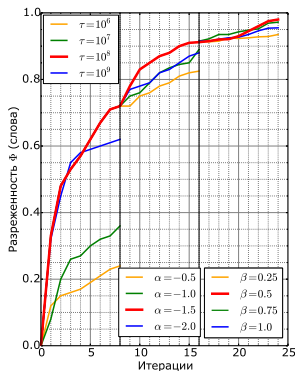
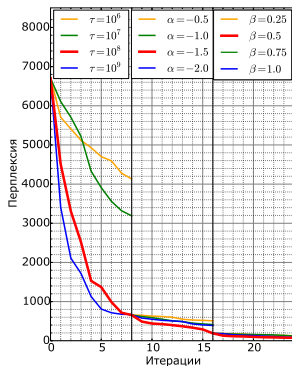
- **assessors**: результаты поиска, выполненного ассессорами
- **TF-IDF, BM25**: сравнение документов по частотам слов
- **word2vec**: нетематические векторные представления слов
- **PLSA**: Probabilistic Latent Semantic Analysis (1999)
- **LDA**: Latent Dirichlet Allocation (2003)
- **ARTM**: тематическая модель с тремя регуляризаторами
- **hARTM**: двухуровневая иерархическая модель ARTM

Задачи регуляризаторов в ARTM и hARTM:

- сделать темы как можно более различными
- сделать векторы $p(t|d)$ как можно более разреженными
- не допустить вырожденности распределений $p(w|t)$

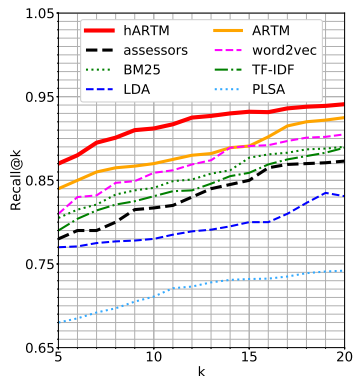
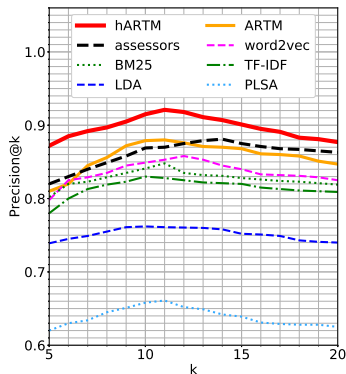
Последовательный подбор коэффициентов регуляризации

- декоррелирование распределений термов в темах (τ),
- разреживание распределений тем в документах (α),
- сглаживание распределений термов в темах (β).



Сравнение с ассессорами по качеству поиска

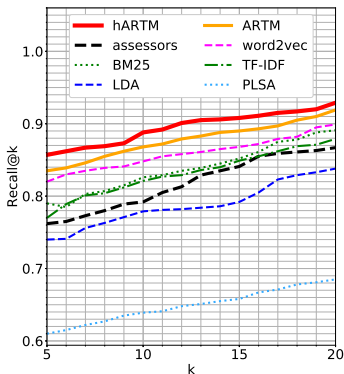
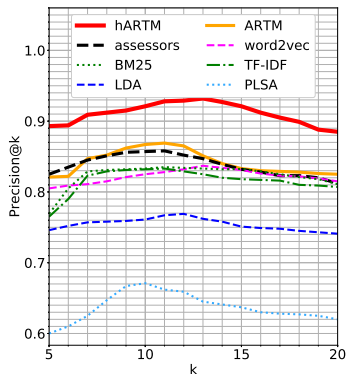
Точность и полнота по первым k позициям поисковой выдачи (коллекция Habrhabr.ru)



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

Сравнение с ассессорами по качеству поиска

Точность и полнота по первым k позициям поисковой выдачи (коллекция TechCrunch.com)



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

Влияние числа тем на качество поиска

Nabrahabr. Все регуляризаторы и модальности, **три уровня**

$ T_1 $	20		25					30			
$ T_2 $	150	200	250		275			300		400	450
$ T_3 $	750	800	1200	1300	1300	1400	1500	1500	1600	3000	3500
Pr@5	0.625	0.743	0.840	0.852	0.869	0.872	0.870	0.805	0.771	0.705	0.672
Pr@10	0.648	0.754	0.851	0.867	0.882	0.915	0.901	0.811	0.799	0.722	0.694
Pr@15	0.632	0.752	0.850	0.872	0.878	0.895	0.889	0.809	0.785	0.729	0.703
Pr@20	0.629	0.745	0.845	0.861	0.871	0.877	0.882	0.803	0.778	0.710	0.681
R@5	0.632	0.780	0.845	0.869	0.883	0.889	0.872	0.851	0.841	0.721	0.695
R@10	0.654	0.792	0.859	0.873	0.905	0.922	0.881	0.873	0.850	0.749	0.703
R@15	0.675	0.805	0.874	0.892	0.932	0.942	0.905	0.889	0.863	0.787	0.725
R@20	0.684	0.824	0.889	0.901	0.958	0.961	0.912	0.904	0.878	0.805	0.734

- существует оптимальное число тем на каждом уровне
- три уровня лучше, чем один или два
- увеличивается оптимальное число тем на нижнем уровне

Влияние числа тем на качество поиска

TechCrunch. Все регуляризаторы и модальности, три уровня

$ T_1 $	80		100						120		
$ T_2 $	300	350	500		550			600		700	750
$ T_3 $	1500	1700	2500	2600	2600	2800	3000	3000	3200	4500	4700
Pr@5	0.655	0.707	0.751	0.792	0.887	0.893	0.890	0.789	0.722	0.703	0.678
Pr@10	0.678	0.712	0.773	0.823	0.895	0.922	0.905	0.805	0.741	0.722	0.692
Pr@15	0.692	0.715	0.775	0.831	0.902	0.921	0.907	0.821	0.743	0.725	0.703
Pr@20	0.687	0.709	0.761	0.819	0.889	0.885	0.898	0.809	0.736	0.719	0.683
R@5	0.751	0.795	0.802	0.856	0.871	0.877	0.863	0.852	0.831	0.738	0.705
R@10	0.767	0.812	0.825	0.875	0.892	0.908	0.879	0.871	0.842	0.751	0.711
R@15	0.772	0.824	0.841	0.887	0.912	0.927	0.901	0.893	0.854	0.772	0.721
R@20	0.783	0.830	0.854	0.892	0.931	0.949	0.935	0.905	0.871	0.790	0.732

- существует оптимальное число тем на каждом уровне
- три уровня лучше, чем один или два
- увеличивается оптимальное число тем на нижнем уровне

Влияние модальностей на качество поиска

Все регуляризаторы и модальности, 3 уровня, оптимальное $|T|$

Модальности: Words, Bigrams, Authors, Comments, Tags, Hubs, Categories

	Habrahabr						TechCrunch					
	асесс	W	Com	WB	WBTH	All	асесс	W	C	WB	WBC	All
Pr@5	0.821	0.621	0.558	0.673	0.871	0.872	0.822	0.718	0.569	0.795	0.891	0.893
Pr@10	0.869	0.645	0.567	0.712	0.911	0.915	0.851	0.729	0.592	0.807	0.919	0.922
Pr@15	0.875	0.631	0.532	0.693	0.894	0.895	0.835	0.737	0.603	0.803	0.920	0.921
Pr@20	0.863	0.628	0.531	0.688	0.877	0.877	0.813	0.729	0.594	0.792	0.883	0.885
R@5	0.780	0.725	0.645	0.797	0.888	0.889	0.762	0.754	0.659	0.775	0.874	0.877
R@10	0.817	0.748	0.652	0.812	0.921	0.922	0.792	0.778	0.671	0.808	0.908	0.908
R@15	0.850	0.782	0.679	0.842	0.941	0.942	0.835	0.783	0.679	0.825	0.927	0.927
R@20	0.873	0.789	0.672	0.852	0.960	0.961	0.867	0.785	0.711	0.837	0.949	0.949

- лучше использовать все модальности
- биграммы и категории выигрывают у ассессоров
- авторы и комментаторы наименее важны

Влияние регуляризаторов на качество поиска

Все регуляризаторы и модальности, 3 уровня, оптимальное |T|

Регуляризаторы: Decorrelation, Θ-sparsing, Φ-smoothing, Hierarchy









	Habrahabr					TechCrunch				
	нет	D	DΘ	DΦ	DΘΦ	нет	D	DΘ	DΦ	DΘΦ
Pr@5	0.628	0.772	0.771	0.865	0.872	0.652	0.777	0.779	0.879	0.893
Pr@10	0.653	0.781	0.812	0.883	0.915	0.679	0.788	0.819	0.895	0.922
Pr@15	0.642	0.785	0.792	0.891	0.895	0.669	0.791	0.798	0.901	0.921
Pr@20	0.643	0.771	0.783	0.875	0.877	0.673	0.775	0.792	0.892	0.885
R@5	0.692	0.820	0.805	0.875	0.889	0.673	0.825	0.812	0.869	0.877
R@10	0.714	0.831	0.834	0.905	0.922	0.685	0.856	0.845	0.881	0.908
R@15	0.725	0.847	0.867	0.921	0.942	0.712	0.877	0.869	0.912	0.927
R@20	0.735	0.873	0.891	0.943	0.961	0.723	0.892	0.895	0.934	0.949

- Лучше использовать все регуляризаторы
- Модели со слабой регуляризацией (PLSA, LDA) слабы

Выводы по результатам экспериментов

- Ассессорские данные относятся не к темам, а к коллекции; поэтому с их помощью можно оценивать новые модели
- Небольших ассессорских данных хватает для оценивания тематических моделей, т.к. они обучаются *без учителя*
- Регуляризаторы, улучшающие интерпретируемость модели, повышают также и качество поиска
- Иерархия улучшает качество поиска (в основном точность) благодаря постепенному сужению области поиска
- Подбор траектории регуляризации и оптимизация коэффициентов регуляризации влияет на качество поиска
- При тщательной оптимизации тематический поиск превосходит как ассессоров, так и конкурирующие модели

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

-  *К.Воронцов.* Вероятностное тематическое моделирование: теория регуляризации и проект BigARTM. 2022.
<http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf>
-  *К.Воронцов.* Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.
-  *К.Vorontsov, A.Potapenko.* Additive regularization of topic models. Machine Learning, 2015.
-  *O.Frei, M.Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016.
-  *A.Ianina, K.Vorontsov.* Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.
-  *A.Ianina, K.Vorontsov.* Regularized Multimodal Hierarchical Topic Model for Document-by-Document Exploratory Search. FRUCT-ISMW, 2019.
-  *И.А.Ирхин, В.Г.Булатов, К.В.Воронцов.* Аддитивная регуляризация тематических моделей с быстрой векторизацией текста. КИМ, 2020.
-  *V.Bulatov, E.Egorov, E.Veselova, D.Polyudova, V.Alekseev, A.Goncharov, K.Vorontsov.* TopicNet: making additive regularisation for topic modelling accessible. LREC-2020