Вероятностное тематическое моделирование больших текстовых коллекций и транзакционных данных

Воронцов Константин Вячеславович
(МФТИ • ФИЦ ИУ РАН • ВМК МГУ)

ВМК МГУ • 26 февраля 2020

Содержание

1. **Вероятностное тематическое моделирование**
   - Постановка задачи
   - Теория аддитивной регуляризации
   - Примеры регуляризаторов
2. **Обобщения**
   - Мультимодальные тематические модели
   - Иерархические тематические модели
   - Гиперграфовые модели транзакционных данных
3. **Реализация**
   - Проект с открытым кодом BigARTM
   - Ключевые механизмы BigARTM
   - Некоторые эксперименты с BigARTM

Что такое «тема» в коллекции текстовых документов?

Выделение тематики — первый шаг к пониманию смысла текста

- *тема* — специальная терминология предметной области
- *тема* — набор часто совместно встречающихся терминов
- *тема* — семантически однородный кластер текстов

Имея коллекцию текстовых документов, хотим узнать:

- из каких тем состоит коллекция;
- $p(t|d)$ — из каких тем состоит каждый документ;
- $p(w|t)$ — из каких слов или терминов состоит каждая тема.

Тематическая модель выявляет латентные темы по наблюдаемым распределениям слов $p(w|d)$ в документах.

Пример. Мультиязычная модель Википедии. Интерпретируемость тем.

216K русско-английских пар статей. Первые 10 слов и их вероятности в теме, %:

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. Мультиязычная модель Википедии. Интерпретируемость тем.

216K русско-английских пар статей. Первые 10 слов и их вероятности в теме, %:

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.

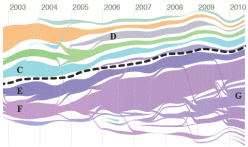
Приложения тематического моделирования

Тематическое моделирование — «мягкая кластеризация» коллекции текстов

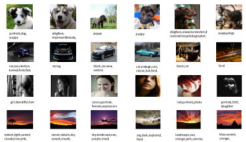
разведочный поиск в электронных библиотеках

поиск тематического контента в соцсетях

детектирование и трекинг новостных сюжетов



мультимодальный поиск текстов и изображений

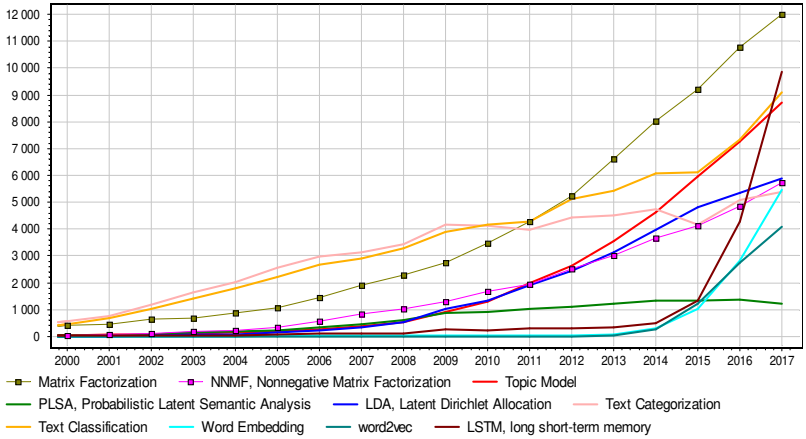


анализ банковских транзакционных данных

управлением диалогом в разговорном интеллекте

Тематическое моделирование и смежные области исследований

Динамика цитирования в академических публикациях, по данным Google Scholar:



Математическая постановка задачи тематического моделирования

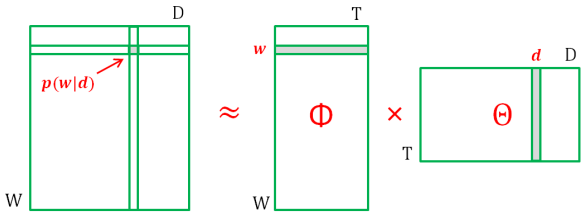
Дано: коллекция текстовых документов D, словарь слов или *термов* W

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

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

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

Это задача стохастического матричного разложения, T — заданное число тем:



PLSA — Probabilistic Latent Semantic Analysis [T.Hofmann, 1999]

Максимизация log-правдоподобия при $\phi_{wt} \geqslant 0$, $\theta_{td} \geqslant 0$, $\sum_w \phi_{wt} = 1$, $\sum_t \theta_{td} = 1$:

$$\mathscr{L}(\Phi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}$$

EM-алгоритм: метод простой итерации для системы уравнений

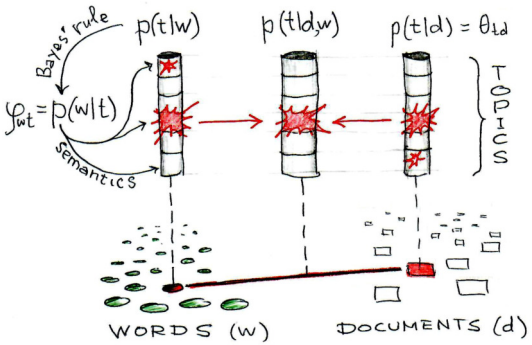
$$\text{E-шаг:} \quad p_{tdw} = p(t|d,w) = \underset{t\in T}{\operatorname{norm}}\big(\phi_{wt}\theta_{td}\big)$$

$$\text{M-шаг:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\operatorname{norm}}\Big(\sum\limits_{d\in D} n_{dw} p_{tdw}\Big) \\ \theta_{td} = \underset{t\in T}{\operatorname{norm}}\Big(\sum\limits_{w\in d} n_{dw} p_{tdw}\Big) \end{cases}$$

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

Тематические векторные представления слов и документов

- Коллекция текстов — это двудольный граф с рёбрами (d, w)
- Слово w встречается в документе d потому, что у них есть общие темы t
- Темы интерпретируются благодаря распределению слов $p(w|t) = p(t|w)\frac{p(w)}{p(t)}$



Задачи, некорректно поставленные по Адамару

Задача *корректно поставлена*, если её решение

- существует,
- единственно,
- устойчиво.

Жак Саломон Адамар
(1865–1963)

Задача матричного разложения *некорректно поставлена*:
если Φ, Θ — решение, то стохастические Φ', Θ' — тоже решения

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

Регуляризация — доопределения решения с помощью дополнительных критериев

ARTM — Аддитивная Регуляризация Тематических Моделей

Максимизация log-правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для решения системы уравнений

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

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

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

Два частных случая: классические модели PLSA и LDA

PLSA (Probabilistic Latent Semantic Analysis) — никакой регуляризации нет:

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

LDA (Latent Dirichlet Allocation, латентное размещение Дирихле)
— столбцы ϕ_t похожи на вектор β ($\beta_w \geqslant -1$);
— столбцы θ_d похожи на вектор α ($\alpha_t \geqslant -1$):

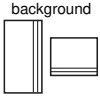
$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} + \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td}.$$

M-шаг — сглаженные частотные оценки с параметрами β_w, α_t:

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

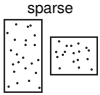
Hofmann T. Probabilistic latent semantic indexing. SIGIR 1999.
Blei D., Ng A., Jordan M. Latent Dirichlet allocation. 2003.

Регуляризаторы для улучшения интерпретируемости тем



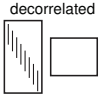
Сглаживание фоновых тем $t \in B \subset T$ (аналогично модели LDA):

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



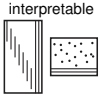
Разреживание предметных тем $t \in S \subset T$ (обобщение LDA):

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



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

Иерархические, темпоральные, регрессионные модели

hierarchy

Связь родительских тем t с дочерними подтемами s:

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

temporal

Темпоральные модели с модальностью времени i:

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

regression

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

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

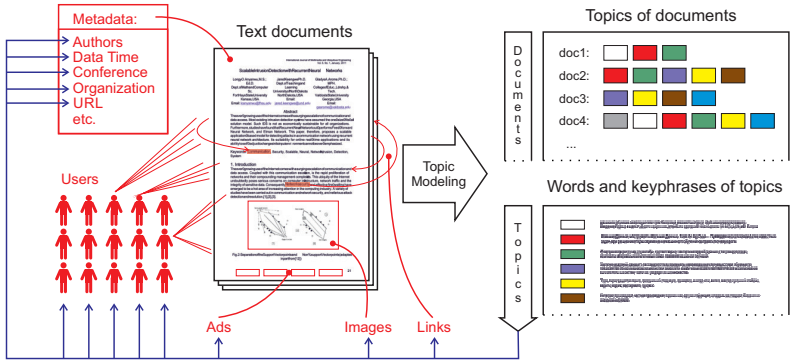
n of topics

Разреживание $p(t)$ для отбора тем:

$$R(\Theta) = -\tau \sum_{t\in T} \frac{1}{|T|} \ln p(t), \quad p(t) = \sum_{d\in D} p(d)\theta_{td}.$$

Задачи мультимодального тематического моделирования

Темы определяют распределения термов различных *модальностей* $p(w|t)$:
$p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{категория}|t)$, $p(\text{класс}|t)$, $p(\text{тег}|t)$, $p(\text{ссылка}|t)$,
$p(\text{баннер}|t)$, $p(\text{элемент\_изображения}|t)$, $p(\text{пользователь}|t), \ldots$



Мультимодальная ARTM

Максимизация log-правдоподобий модальностей со словарями термов W^m, $m \in M$:

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

EM-алгоритм: метод простой итерации для решения системы уравнений

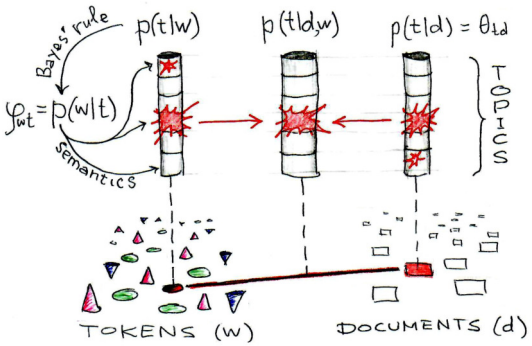
$$\text{E-шаг:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt} \theta_{td}\right)$$

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

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

Мультимодальные тематические векторные представления

- Документы содержат слова и термы других *модальностей*
- Примеры модальностей: авторы, время, теги, пользователи,...
- Через темы смыслы слов передаются другим модальностям



Пример. Модальность n-грамм улучшает качество тем

Коллекция 1000 статей конференций ММРО, ИОИ на русском языке

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

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

Построение тематической иерархии: по уровням, сверху вниз

Шаг 1. Строим модель верхнего уровня с небольшим числом тем.

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

Родительские темы t — *псевдо-документы* с частотами слов $n_{wt} = \phi_{wt} n_t$:

$$\sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \phi_{ws} \theta_{st} \;\to\; \max,$$

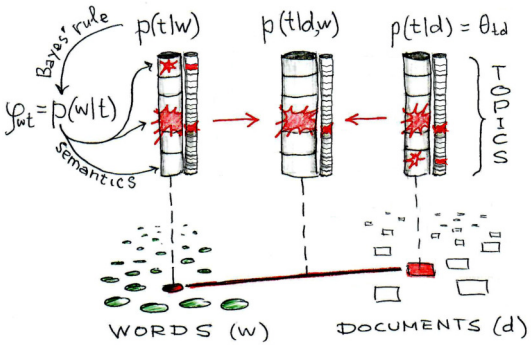
где $\theta_{st} = p(s|t)$ — вероятность подтемы s в родительской теме t.

N.A.Chirkova, K.V.Vorontsov. Additive Regularization for Hierarchical Multimodal Topic Modeling. JMLDA, 2016.

A.V.Belyy, M.S.Seleznova, A.K.Sholokhov, K.V.Vorontsov. Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue 2018.

Иерархического тематического моделирования

- Разбиение родительских тем на дочерние подтемы: по уровням, сверху вниз
- На дочернем уровне родительские темы превращаются в *псевдо-документы*
- Связь «много-ко-многим»: дочерняя тема может иметь много родительских



Транзакционные данные

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

Примеры:

- **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 финансовых организаций:**
  (b, s, g) — покупатель u купил у продавца s товар g
- **Данные о пассажирских авиаперевозках:**
  (u, x, y, a) — клиент u вылетел из аэропорта x в аэропорт y авиакомпанией a

Задача: по выборке рёбер гиперграфа выявить латентные темы его вершин.

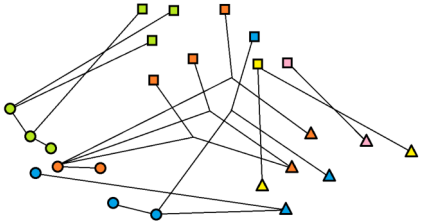
Тематическая модель гиперграфа: определения и обозначения

$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.
$V = V^1 \sqcup \cdots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

K — множество типов рёбер:

T — множество тем:



X^k — наблюдаемая выборка транзакций — рёбер типа k
ребро (d, x) состоит из вершины-контейнера $d \in V$ и множества вершин $x \subset V$,
n_{dx} — число вхождений ребра (d, x) в выборку X^k
$p(d, x)$ — неизвестное распределение на рёбрах типа k

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k:

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

$\theta_{td} = p(t|d)$ — тематика контейнера не зависит от типа ребра k
$\phi_{vt} = p(v|t)$ — распределение термов модальности v в теме t

Задача максимизации log-правдоподобия:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in X^k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \phi_{vt} \to \max_{\Phi, \Theta},$$

$$\phi_{vt} \geqslant 0, \quad \sum_{v \in V^m} \phi_{vt} = 1; \qquad \theta_{td} \geqslant 0, \quad \sum_{t \in T} \theta_{td} = 1;$$

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного log-правдоподобия:

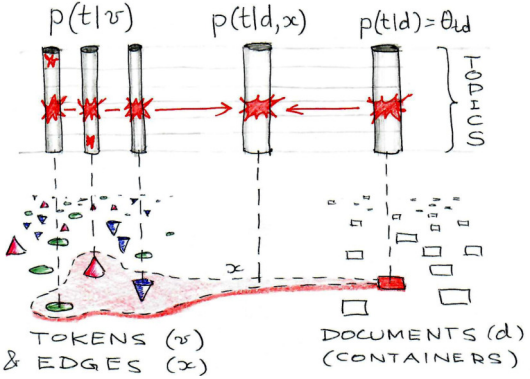
$$\sum_{k\in K}\tau_k \sum_{(d,x)\in X^k} n_{dx} \ln \sum_{t\in T}\theta_{td}\prod_{v\in x}\phi_{vt} \;+\; R(\Phi,\Theta) \to \max_{\Phi,\Theta}.$$

EM-алгоритм: метод простой итерации для решения системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdx} = \underset{t\in T}{\text{norm}}\Big(\theta_{td}\prod_{v\in x}\phi_{vt}\Big) \\ \text{M-шаг:} & \phi_{vt} = \underset{v\in V^m}{\text{norm}}\Big(\sum_{k\in K}\tau_k \sum_{(d,x)} \big[v\in x\big] n_{dx}p_{tdx} + \phi_{vt}\frac{\partial R}{\partial \phi_{vt}}\Big) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\Big(\sum_{k\in K}\tau_k \sum_{(d,x)} n_{dx}p_{tdx} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

Интерпретируемые эмбедингги транзакционных данных

- *Гиперграф* — это система подмножеств вершин-термов
- Транзакция = подмножество термов = ребро гиперграфа
- Транзакция тем более вероятна, чем больше общих тем имеют её термы



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.

Гиперграфовые тематические модели языка

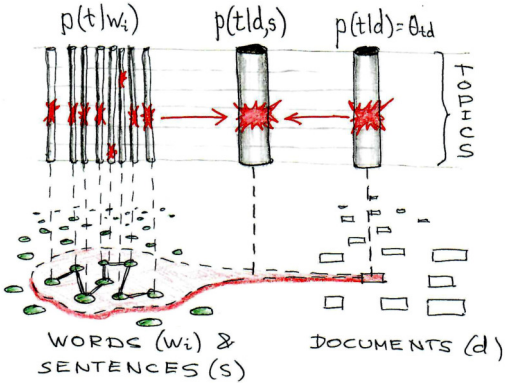
Что ещё может быть ребром гиперграфа?

Любое подмножество связанных по смыслу термов, порождаемых общей темой.

- предложение
- синтагма, ветка синтаксического дерева
- именная группа
- факт «объект, субъект, действие»
- пары термов в одном или соседних предложениях, связанных тезаурусными отношениями: синонимы, гипоним–гипероним, мероним–холоним
- лексическая цепочка
- текст сообщения и его автор
- финансовая транзакция с текстом платёжного поручения

Интерпретируемые эмбединги предложений

- Предложение — это наиболее семантически однородная единица языка
- Предложение = подмножество слов = ребро гиперграфа
- Предложение тем более вероятно, чем больше общих тем имеют его слова



BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновый параллельный мультимодальный ARTM
- Большие данные: коллекция не хранится в памяти
- 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

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	
Формализация:	Вероятностная порождающая модель данных	Стандартные критерии	Свои критерии
Алгоритмизация:	Байесовский вывод для данной порождающей модели (VI, GS, EP)	Общий регуляризованный EM-алгоритм для любых моделей	
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

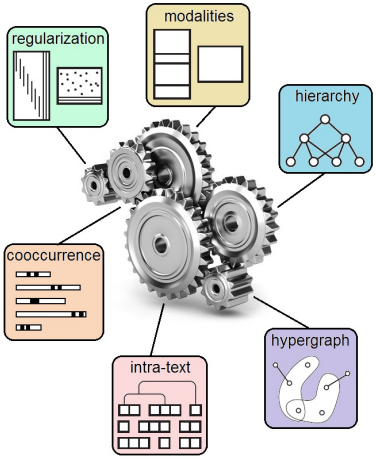
-- стандартизуемые этапы

Ключевые механизмы BigARTM

Благодаря ARTM, эти механизмы можно комбинировать в любых сочетаниях:

1. регуляризация
2. модальности
3. иерархия тем
4. гиперграфы транзакций
5. парная встречаемость термов
6. обработка последовательного текста

Механизмы 4–6 позволяют учитывать порядок слов в обход гипотезы «мешка слов»



Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

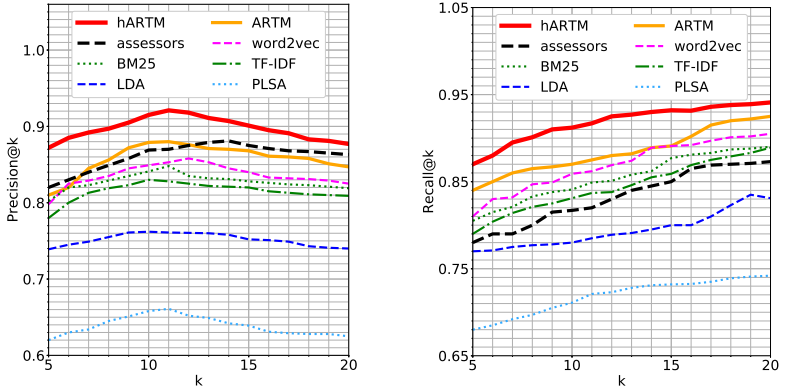
3.7M статей Википедии, 100K слов. В ячейках: «время в минутах (перплексия)»

процессоров	$\|T\|$	Gensim	Vowpal Wabbit	BigARTM синхр	BigARTM асинхр
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.

Сравнение качества поиска с асессорами и простыми моделями

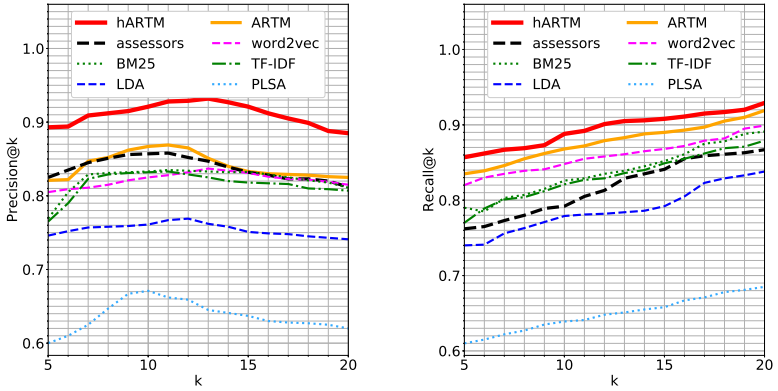
Точность и полнота по первым k позициям поисковой выдачи (Habrahabr.ru)



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

Сравнение качества поиска с асессорами и простыми моделями

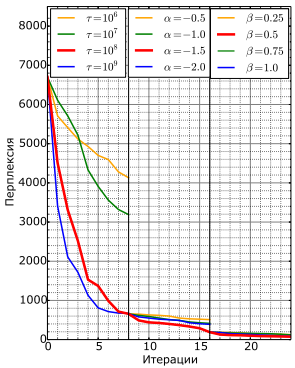
Точность и полнота по первым k позициям поисковой выдачи (TechCrunch.com)

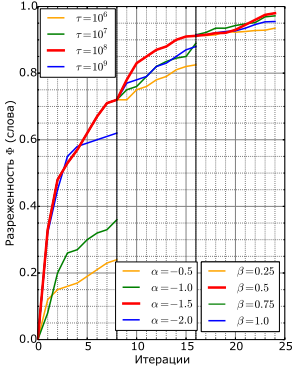


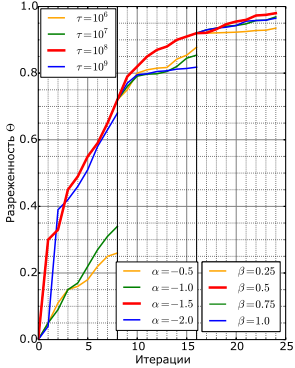
A.Ianina, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Последовательное включение регуляризаторов с подбором коэффициентов

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







Резюме

- Тематическое моделирование — способ векторизации токенов
- Численный метод — регуляризация матричного разложения, EM-алгоритм
- ARTM — аддитивная регуляризация для комбинирования тематических моделей и построения моделей с заданными свойствами
- BigARTM — эффективная реализация этого подхода с открытым кодом
- Наиболее «сильные» обобщения — модальности, иерархии, гиперграфы

Литература

- *К.В.Воронцов.* Обзор вероятностных тематических моделей. 2020. – NEW!
  http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf
- *К.В.Воронцов.* Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.
- *K.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.
- *N.Chirkova, K.Vorontsov.* Additive regularization for hierarchical multimodal topic modeling. JMLDA, 2016.
- *Ianina A., Golitsyn L., Vorontsov K.* Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.
- *A.Potapenko, A.Popov, K.Vorontsov.* Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.
- *A.Belyy, M.Seleznova, A.Sholokhov, K.Vorontsov.* Quality Evaluation and Improvement for Hierarchical Topic Modeling. Dialogue, 2018.
- *Ianina A., Vorontsov K.* Regularized Multimodal Hierarchical Topic Model for Document-by-Document Exploratory Search. FRUCT-ISMW, 2019.
- *Egorov E., Nikitin F., Goncharov A., Alekseev V., Vorontsov K.* Topic Modelling for Extracting Behavioral Patterns from Transactions Data. IC-AIAI, 2019.