Открытые проблемы вероятностного тематического моделирования

Воронцов Константин Вячеславович
(МФТИ, ФИЦ ИУ РАН)

IDP-2020

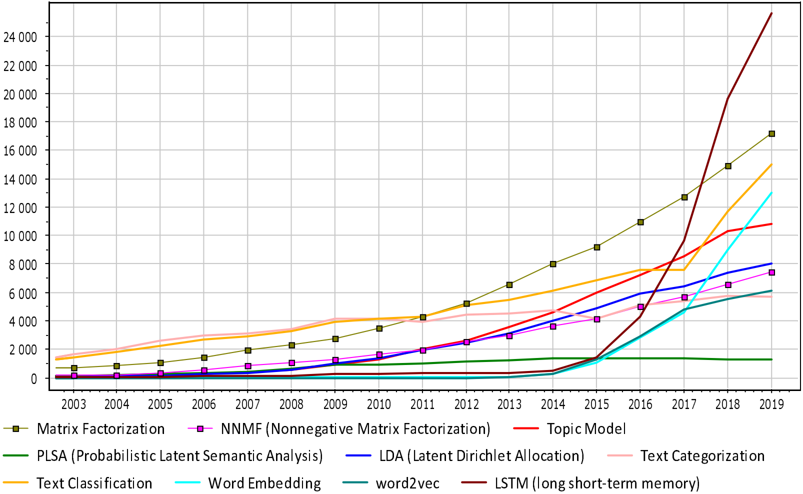
Интеллектуализация Обработки Информации
Москва • 8–12 декабря 2020

Содержание

1. **Некоторые решённые проблемы**
   - От регуляризации до анализа транзакционных данных
   - Лемма о максимизации на единичных симплексах
   - Однопроходная тематическая векторизация текстов
2. **Некоторые открытые проблемы**
   - Тематические модели внимания
   - Проблема несбалансированности тем
   - Обнаружение новых тем в текстовых потоках
3. **Некоторые перспективные проекты**
   - Мультиязыковой поиск документов по документам
   - Разведочный информационный поиск
   - Поиск потенциально опасного дискурса

Тематическое моделирование и смежные области исследований

Динамика цитирования, по данным Google Scholar:



Задача тематического моделирования (Topic Modeling)

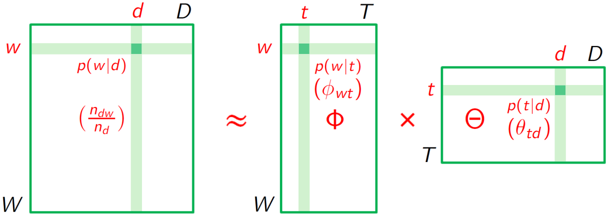
Дано: коллекция текстовых документов

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

Найти: тематическую модель $p(w|d) = \sum\limits_{t \in T} p(w|t)\, p(t|d)$

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

Это задача стохастического матричного разложения:



ARTM: аддитивная регуляризация тематических моделей

Максимизация log правдоподобия с регуляризатором R:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{array}{ll}
\text{E-шаг:} & \left\{\begin{array}{l}
p_{tdw} \equiv p(t|d,w) = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \\[2ex]
\phi_{wt} = \operatorname*{norm}_{w\in W}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right), \quad n_{wt} = \sum_{d\in D} n_{dw}p_{tdw} \\[2ex]
\theta_{td} = \operatorname*{norm}_{t\in T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \quad n_{td} = \sum_{w\in d} n_{dw}p_{tdw}
\end{array}\right. \\
\text{M-шаг:} &
\end{array}$$

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

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

Байесовский вывод — основной подход в Topic Modeling

$X = (d_i, w_i)_{i=1}^n$ — наблюдаемые переменные, коллекция длины n
$Z = (t_i)_{i=1}^n$ — скрытые переменные
$\Omega = (\Phi, \Theta)$ — искомые параметры модели
$\gamma = (\beta, \alpha)$ — гиперпараметры априорных распределений

Задача байесовского вывода — получить не Ω, а $p(\Omega|X,\gamma)$

Вариационный байесовский вывод:

вывести $p(Z,\Omega|X,\gamma) \propto p(X,Z|\Omega,\gamma)\, p(\Omega|\gamma)$

Сэмплирование Гиббса:

вывести $p(Z|X,\gamma)$
сэмплировать $Z \sim p(Z|X,\gamma)$
вывести $p(\Omega|X,Z,\gamma) \propto p(X,Z|\Omega,\gamma)\, p(\Omega|\gamma)$

Blei D., Ng A., Jordan M. Latent Dirichlet Allocation. JMLR, 2003.

Общий взгляд на байесовское обучение, MAP и ARTM

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

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

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

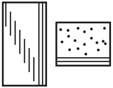
$$\Omega := \arg\max_{\Omega}\bigl(\ln p(X|\Omega) + \ln \mathrm{Prior}(\Omega|\gamma)\bigr)$$

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

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

Регуляризаторы, модальности, иерархии, графы, гиперграфы

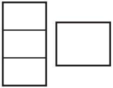
interpretable



Сглаживание, разреживание и декоррелирование тем

$$R(\Phi) = \beta_0 \sum_{t,w} \beta_{wt} \ln \phi_{wt} - \lambda \sum_{t,s} \sum_{w} \phi_{wt}\phi_{ws}$$

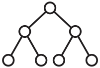
multilanguage



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

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

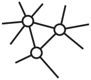
hierarchy



Связь родительских тем t с дочерними подтемами s:

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

graph



Модальность вершин графа v, содержащих $D_v \subset D$:

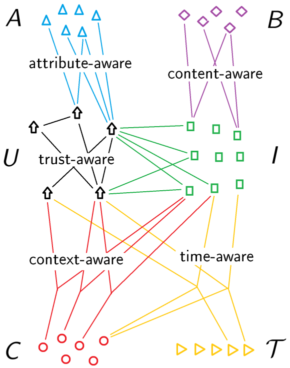
$$R(\Phi) = -\lambda \sum_{(u,v)\in E} S_{uv} \sum_{t\in T} n_t^2 \left(\frac{\phi_{vt}}{|D_v|} - \frac{\phi_{ut}}{|D_u|}\right)^2.$$

Разнородные данные в рекомендательных системах

A — словарь атрибутов клиентов (соцдем, регион, хобби...)
B — словарь свойств объектов (слова в текстовых объектах)
C — конечное множество (словарь) ситуативных контекстов
$\mathcal{T}$ — конечное множество (словарь) моментов времени

Виды данных:

$(i|u)$ — клиент u выбрал объект i
$(a|u)$ — клиент u имеет атрибут a
$(b|i)$ — объект i имеет свойство b
$(v|u)$ — клиент u доверяет клиенту v
$(i, b|u)$ — клиент u отметил i тэгом b
$(i|u, c)$ — клиент u выбрал объект i
в ситуативном контексте c
$(i|u, c, \tau)$ — клиент u выбрал объект i
в контексте c в момент времени τ



Мультимодальная (гипер)графовая тематическая модель

Ребро тем вероятнее, чем более схожи тематические *эмбединги* (векторные представления) инцидентных ему вершин:

$$
\begin{aligned}
&\sum_{u,i} r_{ui} \ln \sum_{t\in T} p(t|u)\, p(t|i)\, p^{-1}(t) + \\
&+ \lambda_1 \sum_{i,b} n_{ib} \ln \sum_{t\in T} p(t|i)\, p(t|b)\, p^{-1}(t) + \\
&+ \lambda_2 \sum_{u,a} n_{ua} \ln \sum_{t\in T} p(t|u)\, p(t|a)\, p^{-1}(t) + \\
&+ \lambda_3 \sum_{u,i,c} n_{uic} \ln \sum_{t\in T} p(t|i)\, p(t|u)\, p(t|c)\, p^{-2}(t) + \\
&+ \lambda_4 \sum_{u,i,c,t} n_{uic\tau} \ln \sum_{t\in T} p(t|i)\, p(t|u)\, p(t|c)\, p(t|\tau)\, p^{-3}(t) \;\to\; \max
\end{aligned}
$$

Оптимизация по всем эмбедингам — векторам вида $p(t|\bullet)$

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Самый быстрый онлайновый параллельный 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: Linux, MacOS, Windows (32/64 bit)
- Интерфейсы API: C++, Python, командная строка

Модульный подход ARTM: сравнение с байесовским подходом

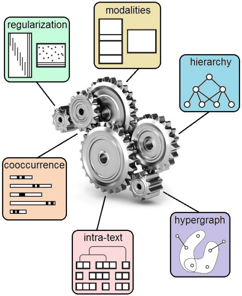
Для построения композитных моделей в ARTM не нужны ни математические выкладки, ни программирование «с нуля».

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 возможности библиотек BigARTM и TopicNet



BigARTM (с 2014 г.)

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

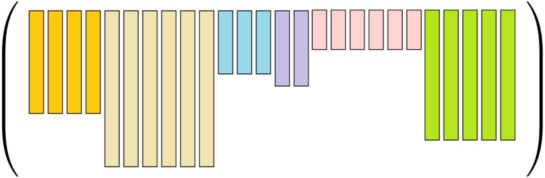
TopicNet (с 2020 г.)

- 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 regularization for topic modelling accessible. LREC-2020

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 $|I_j|$:

$\Omega =$ 

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

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

Лемма о максимизации функции на единичных симплексах

Операция нормировки вектора: $p_i = \operatorname*{norm}_{i\in I}(x_i) = \dfrac{\max\{x_i, 0\}}{\sum\limits_{k\in I}\max\{x_k, 0\}}$

Лемма. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω. Тогда векторы ω_j локального экстремума задачи $f(\Omega) \to \max$ удовлетворяют системе уравнений

$$\begin{cases} \omega_{ij} = \operatorname*{norm}_{i\in I_j}\left(\omega_{ij}\dfrac{\partial f}{\partial \omega_{ij}}\right), & \text{если } \exists i\colon \omega_{ij}\dfrac{\partial f}{\partial \omega_{ij}} > 0 \\ \omega_{ij} = \operatorname*{norm}_{i\in I_j}\left(-\omega_{ij}\dfrac{\partial f}{\partial \omega_{ij}}\right), & \text{иначе, если } \exists i\colon \omega_{ij}\dfrac{\partial f}{\partial \omega_{ij}} < 0 \\ \omega_{ij}\dfrac{\partial f}{\partial \omega_{ij}} = 0, & \text{иначе} \end{cases}$$

Замечания к Лемме о максимизации на единичных симплексах

- Лемма применима для построения широкого класса моделей, параметрами которых являются дискретные распределения вероятности (нормированные неотрицательные векторы)
- Численное решение системы — методом простых итераций
- Существование стационарной точки Ω гарантировано
- Первый из трёх случаев является основным:

$$\omega_{ij} := \operatorname*{norm}_{i \in I_j} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right)$$

- В остальных случаях нормирующий знаменатель нулевой; такие векторы будем удалять из модели как вырожденные
- Итерации похожи на градиентную оптимизацию, но учитывают ограничения и не требуют подбора шага η:

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

Доказательство Леммы

Запишем условия Каруша–Куна–Таккера для $\omega_j = (\omega_{ij} \colon i \in I_j)$:

$$\frac{\partial f}{\partial \omega_{ij}} = \lambda_j - \mu_{ij}; \quad \mu_{ij}\omega_{ij} = 0.$$

Предполагая $\omega_{ij} > 0$, умножим обе части равенства на ω_{ij}:

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

Возможны три случая:

1. Если $\lambda_j > 0$, то либо $A_{ij} > 0$, либо $\omega_{ij} = 0$. Тогда $\omega_{ij}\lambda_j = (A_{ij})_+$; $\lambda_j = \sum\limits_i (A_{ij})_+ \;\Rightarrow\; \omega_{ij} = \operatorname*{norm}\limits_i(A_{ij})$.
2. Если $\lambda_j < 0$ и $(\exists i)\ A_{ij} < 0$, то $(\forall i)\ A_{ij} \leqslant 0$. Тогда $\omega_{ij}\lambda_j = -(-A_{ij})_+$; $\lambda_j = -\sum\limits_i (-A_{ij})_+ \;\Rightarrow\; \omega_{ij} = \operatorname*{norm}\limits_i(-A_{ij})$.
3. Иначе $\lambda_j = 0$ и ω_j находится из уравнений $\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} = 0$. ■

Доказательство основной теоремы ARTM

Применим лемму к log-правдоподобию с регуляризатором:

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

Дифференцируя, выделим вспомогательную переменную p_{tdw}:

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

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

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

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

Исключение матрицы Θ из модели

Мотивации:

- Ограничение-равенство $\theta_{td} = \theta_{td}(\Phi)$ играет роль регуляризатора и повышает устойчивость модели
- Вычисление $\theta_{td}(\Phi)$ за один линейный проход по документу
- Сокращение размерности модели, уменьшение переобучения

Первая итерация EM-алгоритма без регуляризации
при равномерном начальном приближении $\theta_{td}^0 = \frac{1}{|T|}$:

$$\theta_{td}(\Phi) = \operatorname*{norm}_{t\in T}\Bigl(\sum_w n_{dw} p_{tdw}\Bigr) = \sum_w \frac{n_{dw}}{n_d} \frac{\phi_{wt}\theta_{td}^0}{\sum_s \phi_{ws}\theta_{sd}^0} = \sum_w \frac{p_{dw}\phi_{wt}}{\sum_s \phi_{ws}},$$

где $p_{dw} = \frac{n_{dw}}{n_d}$ — частотная оценка условной вероятности $p(w|d)$

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

EM-алгоритм для ARTM с исключённой матрицей Θ

Максимизация логарифма правдоподобия:

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

EM-алгоритм: метод простой итерации для системы уравнений

$$p_{tdw} = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr);$$

$$n_{wt} = \sum_{d\in D} n_{dw} p_{tdw}; \qquad n_{td} = \sum_{w\in d} n_{dw} p_{tdw};$$

$$\phi_{wt} = \operatorname*{norm}_{w\in W}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \phi_{wt}\sum_{d\in D}\sum_{s\in T}\left(\frac{n_{sd}}{\theta_{sd}} + \frac{\partial R}{\partial \theta_{sd}}\right)\frac{\partial \theta_{sd}}{\partial \phi_{wt}}\right)$$

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 М-шага относительно Φ и $\Theta(\Phi)$:

$$Q(\Phi) = \sum_{d\in D}\sum_{u\in W}\sum_{s\in T} n_{du} p_{sdu}\bigl(\ln\phi_{us} + \ln\theta_{sd}(\Phi)\bigr) + R\bigl(\Phi, \Theta(\Phi)\bigr) \to \max_{\Phi}$$

Применим Лемму к регуляризованному log-правдоподобию Q:

$$\phi_{wt}\frac{\partial Q}{\partial \phi_{wt}} = \sum_{d\in D} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} +$$

$$+ \sum_{d,s,u} n_{du} p_{sdu} \frac{\phi_{wt}}{\theta_{sd}}\frac{\partial \theta_{sd}}{\partial \phi_{wt}} + \phi_{wt}\sum_{d,s}\frac{\partial R}{\partial \theta_{sd}}\frac{\partial \theta_{sd}}{\partial \phi_{wt}} =$$

$$= n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}} + \phi_{wt}\sum_{d,s}\left(\frac{n_{sd}}{\theta_{sd}} + \frac{\partial R}{\partial \theta_{sd}}\right)\frac{\partial \theta_{sd}}{\partial \phi_{wt}} \qquad \blacksquare$$

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

Частные производные: $\frac{\partial \theta_{sd}}{\partial \phi_{wt}} = p_{wd} h_w \bigl(\delta_{st} - \phi_{ws} h_w\bigr)$

ЕМ-алгоритм: метод простой итерации для системы уравнений

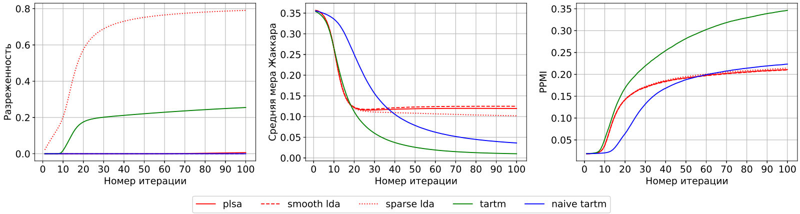
$$
\begin{aligned}
\theta_{td} &= \sum_{w\in d} p_{dw}\phi_{wt}h_w; &\qquad h_w &= \Bigl(\textstyle\sum_t \phi_{wt}\Bigr)^{-1};\\
p_{tdw} &= \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr); & c_{td} &= \frac{n_{td}}{\theta_{td}} + \frac{\partial R}{\partial \theta_{td}};\\
n_{td} &= \sum_{w\in d} n_{dw}p_{tdw}; & \gamma_{dw} &= \sum_{t\in T} \phi_{wt}c_{td};\\
p'_{tdw} &= p_{tdw} + n_d^{-1}\phi_{wt}h_w\bigl(c_{td} - h_w\gamma_{dw}\bigr);\\
\phi_{wt} &= \operatorname*{norm}_{w\in W}\Biggl(\sum_{d\in D} n_{dw}p'_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Biggr).
\end{aligned}
$$

Е-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.

Быстрая векторизация текста за линейное время

Тематический вектор текста $p(t|d)$ вычисляется за один проход усреднением тематических векторов $p(t|w)$ всех слов текста:

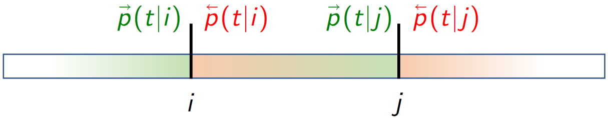
$$\theta_{td}(\Phi) \equiv p(t|d) = \frac{1}{n_d} \sum_{i=1}^{n_d} p(t|w_i)$$

Тематические векторы локального контекста $p(t|i)$ вычисляются для всех $i = 1, \dots, n_d$ экспоненциальным скользящим средним за два прохода «слева направо» и «справа налево»:

$$\vec{p}(t|i) = \alpha_i \cdot p(t|w) + (1 - \alpha_i) \cdot \vec{p}(t|i-1)$$
$$\overleftarrow{p}(t|i) = \alpha_i \cdot p(t|w) + (1 - \alpha_i) \cdot \overleftarrow{p}(t|i+1)$$

α_i — коэффициент сглаживания в позиции i;
$\alpha_i \approx \frac{1}{m}$, где m — число усредняемых позиций;
α_i можно умножать на вес (важность, TF-IDF) слова в тексте,
α_i можно увеличивать до 1, если надо забыть контекст.

Использование тематических векторов локального контекста



Двунаправленные тематические векторы определяют:

- $\vec{p}(t|i)$ — тематику левого контекста слова w_i
- $\overleftarrow{p}(t|i)$ — тематику правого контекста слова w_i
- $\frac{1}{2}\bigl(\vec{p}(t|i) + \overleftarrow{p}(t|i)\bigr)$ — тематику двустороннего контекста w_i
- $p(t|i\ldots j) = \frac{1}{2}\bigl(\overleftarrow{p}(t|i) + \vec{p}(t|j)\bigr)$ — тематику сегмента $[i\ldots j]$
- тематическую однородность сегмента $[i\ldots j]$: насколько распределения $\overleftarrow{p}(t|i)$ и $\vec{p}(t|j)$ схожи
- позиции i границ между сегментами: насколько распределения $\vec{p}(t|i)$ и $\overleftarrow{p}(t|i)$ не схожи
- короткие и длинные контексты при различных α_i

Тематические модели внимания (self-attention)

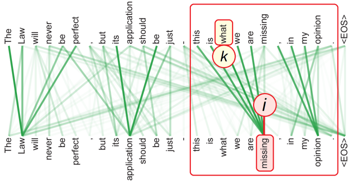
Внимание в нейросетевых моделях языка:

x_i — эмбединги (размерности T) термов w_i, $i = 1, \dots, n$

$\alpha_{ik} = \mathrm{norm}_k \langle x_i, x_k \rangle$ — важность терма w_k в контексте терма w_i

$c_i = \sum\limits_k V x_k \alpha_{ik}$ — эмбединг контекста терма w_i с обучаемой $\underset{T\times T}{V}$

α_{ik}



Аналогичная конструкция в тематической модели:

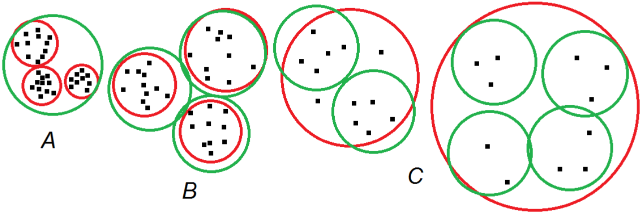
$$\underbrace{p(t|i)}_{c_i} = \sum_k \sum_{t'\in T} \underbrace{p(t|t')}_{v_{tt'}} \underbrace{p(t'|w_k)}_{x_k} \mathop{\mathrm{norm}}_k \Big\langle \underbrace{p(t''|w_k)}_{x_k}, \underbrace{p(t''|w_i)}_{x_i} \Big\rangle$$

Vaswani et al. Attention is all you need. 2017.

Проблема расщепления и слияния тем

Тема — кластер на единичном симплексе размерности $|W| - 1$ с центром $p(w|t)$ и точками $p(w|t,d)$, $d \in D$: $\theta_{td} > 0$

- Тематические модели стремятся выравнивать темы по их мощности (красные кластеры).
- Это приводит к появлению тем-дубликатов (A) и семантически разнородных тем (C).
- Выравнивание тем по *радиусу семантической однородности* (зелёные кластеры) должно решать обе проблемы.



Гипотеза условной независимости

$$\left.\begin{aligned} p(w,d|t) &= p(w|t)\,p(d|t) \\ p(w|d,t) &= p(w|t) \\ p(d|w,t) &= p(d|t) \end{aligned}\right\}\ \text{три эквивалентных представления}$$

Гипотеза семантической однородности темы t
— в теме t термы и документы порождаются независимо:

$$H_0(t): \quad \hat{p}(w,d|t) \sim p(w|t)\,p(d|t)$$

Гипотеза согласованности документа d **с темой** t
— термы темы t порождаются независимо от документов:

$$H_0(t,d): \quad \hat{p}(w|d,t) \sim p(w|t)$$

Гипотеза согласованности терма w **с темой** t
— тема t распределена по документам независимо от термов:

$$H_0(t,w): \quad \hat{p}(d|w,t) \sim p(d|t)$$

Статистики для оценивания семантической однородности

Статистики для проверки гипотез $H_0(t)$, $H_0(d,t)$, $H_0(w,t)$, основанные на KL-дивергенции, устроены единообразно:

$$S_t = \mathrm{KL}\bigl(\hat{p}(w,d|t) \,\|\, p(w|t)p(d|t)\bigr) \qquad = \operatorname*{avg}_{d,w}\bigl(n_{tdw}, \ell(d,w)\bigr)$$

$$S_{td} = \mathrm{KL}\bigl(\hat{p}(w|d,t) \,\|\, p(w|t)\bigr) \qquad = \operatorname*{avg}_{w\in d}\bigl(n_{tdw}, \ell(d,w)\bigr)$$

$$S_{wt} = \mathrm{KL}\bigl(\hat{p}(d|w,t) \,\|\, p(d|t)\bigr) \qquad = \operatorname*{avg}_{d\in D}\bigl(n_{tdw}, \ell(d,w)\bigr)$$

где $\operatorname*{avg}_{i\in I}(\gamma_i, x_i) = \dfrac{\sum_{i\in I}\gamma_i x_i}{\sum_{i\in I}\gamma_i}$ — средневзвешенное x_i с весами γ_i;

$\ell(d,w) = \dfrac{\hat{p}(w|d)}{p(w|d)}$ — функция потерь, общая для всех статистик.

Veselova E., Vorontsov K. Topic balancing with additive regularization of topic models. 2020

Идея. Регуляризатор семантической однородности

Минимизация суммарной семантической неоднородности тем:

$$\sum_{t\in T} S_t = \sum_{d\in D}\sum_{w\in d}\left(\sum_{t\in T}\frac{n_{tdw}}{n_t}\right)\ln\frac{\hat{p}(w|d)}{p(w|d)} \to \min_{\Phi,\Theta}$$

Регуляризатор в сумме с log-правдоподобием, $\beta_{dw} = \sum_{t\in T}\frac{p_{tdw}}{p_t}$:

$$\sum_{d\in D}\sum_{w\in d} n_{dw}\bigl(1+\tau\beta_{dw}\bigr)\ln\sum_{t\in T}\phi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

Модифицированный EM-алгоритм

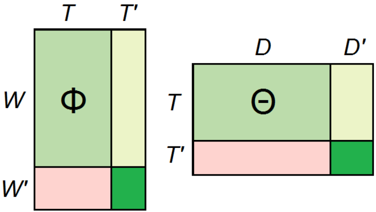
$$p_{tdw} = \operatorname*{norm}_{t\in T}\bigl(\phi_{wt}\theta_{td}\bigr) \qquad \beta_{dw} = \sum_{t\in T}\frac{p_{tdw}}{p_t}$$

$$\phi_{wt} = \operatorname*{norm}_{w\in W}\Bigl(\sum_{d\in D}\tilde{n}_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}}\Bigr) \qquad \tilde{n}_{dw} = n_{dw}(1+\tau\beta_{dw})$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\Bigl(\sum_{w\in W}\tilde{n}_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\Bigr) \qquad p_t = \frac{1}{n}\sum_{d,w}n_{dw}p_{tdw}$$

Обнаружение новых тем в текстовых потоках

Добавление пакета документов D' к коллекции D.
В словарь W добавляются новые слова W'.
К темам T добавляются новые темы T'.



Задачи:

- как определить число новых тем $|T'|$
- как определить наличие новой темы в документе
- как инициализировать новые темы
- как исключать из модели старые темы

Десять открытых проблем тематического моделирования

1. Гарантирование качества (интерпретируемости) всех тем
2. Надёжное разделение лексики на тематическую и общую
3. Моделирование тематики связного текста
4. Динамическое создание новых тем в текстовых потоках
5. Обеспечение устойчивости тематических моделей
6. Оптимизация гиперпараметров в потоковом режиме
7. Бережное слияние моделей нескольких коллекций
8. Автоматическое именование и реферирование тем
9. Создание больших предобученных тематических моделей
10. Применение гиперграфовых тематических моделей

Эти задачи должны решаться на любых данных «из коробки», т.е. без экспериментов по подбору гиперпараметров.

Мультиязыковой поиск научных публикаций

- 100 языков, использованы данные Википедии
- 300 тем, из них 1 фоновая
- BPE токенизация
- словарь редуцирован: 120K $\to$ 2K на язык
- обучение занимает 4 часа
- модель занимает 4.8 Гб

Средняя позиция текста перевода в поисковой выдаче

	cs	de	es	fr	it	ja
cs	1.00	2.33	1.15	2.11	5.44	3.24
de	1.54	1.00	7.18	4.58	2.33	2.80
es	1.28	7.40	1.00	3.43	7.38	6.76
fr	2.96	6.30	6.97	1.00	3.69	4.49
it	5.07	4.45	5.19	1.34	1.00	1.65
ja	4.03	5.49	5.68	3.90	1.75	1.00

Работа выполняется лабораторией МИ МФТИ по заказу компании Антиплагиат в рамках проекта «Пан-языковой анализ больших текстовых коллекций на естественных языках» (федеральный проект «Цифровые технологии» национальной программы «Цифровая экономика Российской Федерации»)

Поисково-рекомендательная система arXiv-search.mipt.ru

Подборка — долгосрочный поисковый интерес пользователя

Поисково-рекомендательные функции:

- поиск тематически близких документов по **подборке**
- мониторинг новых документов для подборки

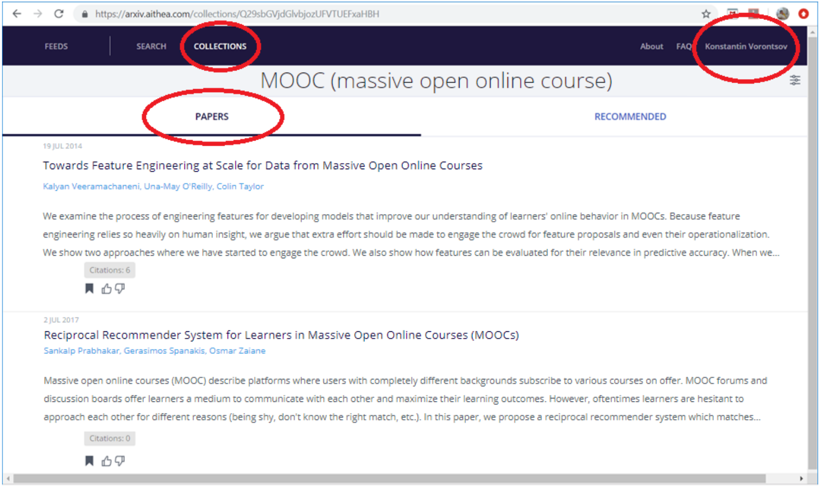
Аналитические функции:

- автоматизация реферирования **подборки**
- рекомендация порядка чтения внутри подборки
- кластеризация тем, идей, мнений во всей подборке
- выделение ключевых понятий, фактов, идей из документа

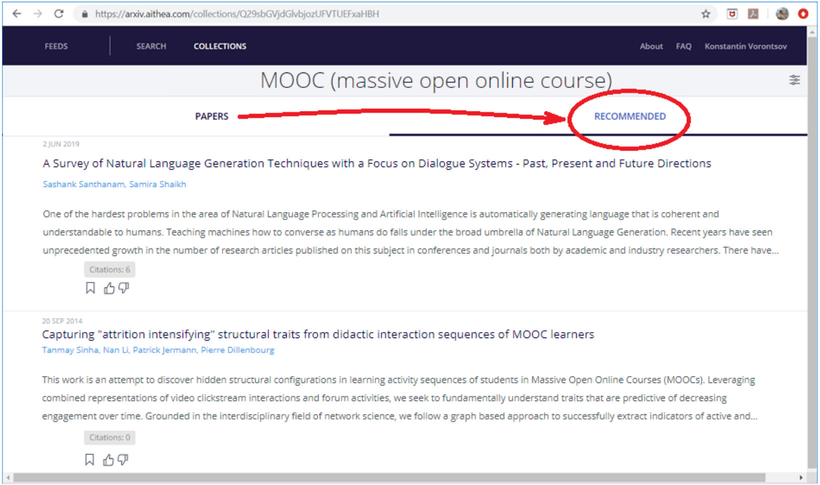
Коммуникативные функции:

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

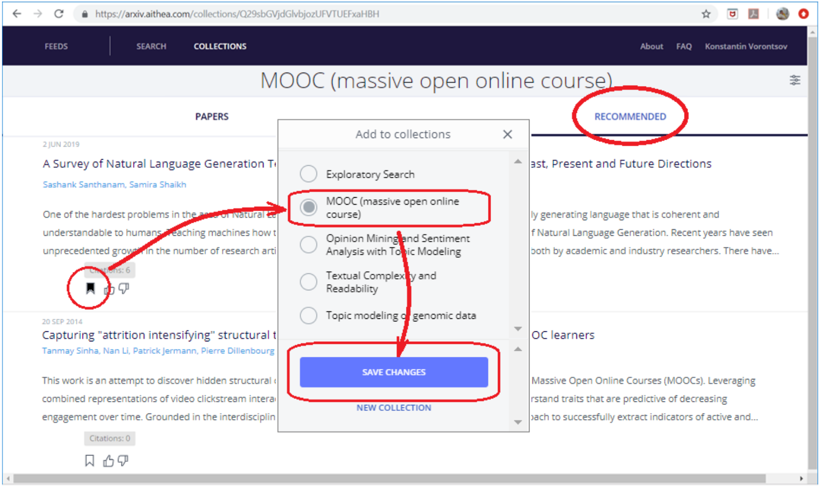
arXiv-search.mipt.ru: Тематическая подборка пользователя



arXiv-search.mipt.ru: Список рекомендуемых статей



arXiv-search.mipt.ru: Добавление статьи в подборку



Задача поиска потенциально опасного дискурса в СМИ и СМ

Явление *потенциально опасного дискурса* не ограничивается фейками, призывами к насилию, разжиганию розни и т.п.

В зависимости от источника и целевой аудитории п.о.дискурс характеризуется устойчивым сочетанием

- тональностей по отношению к определённым субъектам
- упоминаемых и неупоминаемых фактов
- приёмов манипулирования общественным мнением: обесценивание, гиперболизация, умалчивание, демагогия,...
- конструктов мифо(идео)логизированной картины мира

Новое приложение тематического моделирования:

- «тема» — тип потенциально опасного дискурса
- «термы» — атомарные элементы дискурса, *конструкты*, распознавание которых в тексте — отдельная задача

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

Резюме

- Байесовский вывод избыточен для Topic Modeling
- ARTM убирает этап оценивания Posterior$(\Phi, \Theta | X, \gamma)$
- Тематическое моделирование — это векторизация графов, а не раздел анализа текстов или байесовского обучения
- Теперь это «теория одной леммы»
- Интерпретируемость возникает, когда есть вершины-слова
- Полшага до решения проблемы слитых и дублирующих тем
- Полшага до тематических моделей внимания
- Главный тренд тематического моделирования — поиск вариантов симбиоза с нейросетевыми моделями языка

- Открытые библиотеки: BigARTM, TopicNet
- Полезный сервис: https://arxiv-search.mipt.ru/

К.В.Воронцов. Обзор вероятностных тематических моделей. 2020. – NEW!
http://www.MachineLearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf