

# Подбор тематических моделей для построения порядка чтения на коллекции документов

Московский физико-технический институт

Мамонов Кирилл Романович

курс: Моя первая научная публикация, В.В. Стрижов

Консультант: Еремеев М. А.

Эксперт: Воронцов К. В.

30 апреля 2020

# Задача построения порядка чтения

## Цель

Разработать алгоритм для рекомендации порядка чтения коллекции документов.

## Решаемая проблема

Документы должны ранжироваться от простого к сложному, от общего к частному, то есть в том порядке, в котором пользователю будет легче разбираться в новой для него тематической области.

## Методы решения

Рассматриваются тематическая модель ARTM и её вариации (мультимодальная, иерархическая), предлагается новый подход к измерению общности документов.

# Постановка задачи построения порядка чтения

## Обозначения

Порядок чтения  $R(V, E)$  коллекции документов  $D$  — ориентированный ациклический граф.  $v_i \in V$  соответствует множеству эквивалентных документов  $D_i \neq \emptyset \subseteq D$ .

Ребро  $v_i \rightarrow v_j$  показывает, что документы, принадлежащие множеству  $D_i$ , предшествуют в порядке чтения документам  $D_j$ .

## Матрица смежности

Порядок чтения представим в виде матрицы смежности:

$$A_{ij} = \begin{cases} \frac{1}{\text{число прыжков}(d_i \rightarrow d_j)}, & \text{если есть путь } d_i \rightarrow d_j, \\ 0, & \text{иначе.} \end{cases}$$

## Метрика качества

Разность двух порядков чтения, представленных матрицами  $A$

$$\text{и } \hat{A}: MSE(A, \hat{A}) = \frac{1}{n} \sum_{i,j=1}^n (A_{ij} - \hat{A}_{ij})^2.$$

## Основная

- ❶ Georgia Koutrika, Lei Liu, and Steven J. Simske. *Generating reading orders over document collections*. 31st IEEE International Conference on Data Engineering, 2015, pages 507–518.
- ❷ Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. *Bigartm: open source library for regularized multimodal topic modeling of large collections*. In International Conference on Analysis of Images, Social Networks and Texts, pages 370–381. Springer, 2015.

# Решение задачи построения порядка чтения

## Первый этап

Строится тематическую модель документов из коллекции.

## Второй этап

Оценивается общность каждого документа.

## Третий этап

Самые общие документы объединяются в вершину графа  $N$ , остальные документы кластеризуются по взаимопересечению, алгоритм рекурсивно продолжается на каждом из кластеров, а потроенные деревья становятся детьми  $N$ .

Центральное предположение тематического моделирования, что вероятность появления слова  $w$  в документе  $d$ :

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

где матрица  $\Phi$  содержит распределение слов  $w$  в документ  $d$  ( $\phi_{wt}$ ), матрица  $\Theta$  — вероятности  $\theta_{td}$  появления темы  $t$  в документе  $d$ ,  $T$  — общее количество тем в модели.

## PLSA

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{wd} \log p(w \mid d) \rightarrow \max_{\Phi, \Theta},$$

ограничения на  $\Phi, \Theta$ :  $\sum_{w \in W} \phi_{wt} = 1, \phi_{wt} \geq 0, \sum_{t \in T} \theta_{td} = 1, \theta_{td} \geq 0$ .

## ARTM

$$L(\Phi, \Theta) + \sum R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}.$$

$$R(\Phi) = \sum_{t \in T} \sum_{w \in W} \beta_{wt} \log \phi_{wt}, \quad R(\Theta) = \sum_{d \in D} \sum_{t \in T} \alpha_{td} \log \theta_{td},$$

$$R = -\frac{\tau}{2} \sum_{t \in T} \sum_{s \in T} \sum_{w \in W} \phi_{wt} \phi_{ws}.$$

## Hierarchical ARTM (hARTM)

Слои — ARTM модели, связаны между собой регуляризатором

$$R = \sum_{t \in T} \sum_{w \in W} n_{wt} \log \sum_{s \in S} \phi_{ws} \psi_{st},$$

где  $T$  — темы родительского слоя,  $S$  — темы дочернего слоя,  $\Psi$  — вероятностная матрица смежности тем родительского и дочернего уровней.

Так как порядок чтения должен идти от общего к частному, то измерение общности каждого документа  $d$  — одна из главных проблем.

## Энтропия

$$g(d) = - \sum_{t \in T} \theta_{td} \log(\theta_{td})$$

## Иерархическая энтропия для двуслойной hARTM

$$g_h(d) = - \sum_{t \in T} \theta_{td}^1 \sum_{s \in S} \psi_{st} \theta_{sd}^2 \log \theta_{sd}^2$$



Еще одной важной характеристикой является мера пересечения документов по темам. Это определяет, какие документы могут быть прочитаны независимо друг от друга, а какие стоит читать в определённой последовательности.

## Пересечение по темам

$$o(d_i, d_j) = \frac{\theta_{td}^i \cdot \theta_{td}^j}{|\theta_{td}^i|^2 + |\theta_{td}^j|^2 - \theta_{td}^i \cdot \theta_{td}^j}$$

# Вычислительный эксперимент

## Данные

Два эталонных графа чтения русскоязычной Википедии: из категории Математика (глубина — 8, содержит 9503 документов) и из её подкатегории Машинное Обучение (глубина — 5, содержит 425 документов)

## Построенные тематические модели

Тип	Sparsity $\theta$	Sparsity $\Phi$ для слов	Sparsity $\Phi$ для биграмм
LDA	0.76	0.93	-
ARTM	0.77	0.93	-
PLSA	0.74	0.93	-
hARTM	0.87	0.95	-
ARTM с биграммами	0.82	0.90	0.75
hARTM с биграммами	0.90	0.93	0.81

# Результаты для каталога по Машинному Обучению

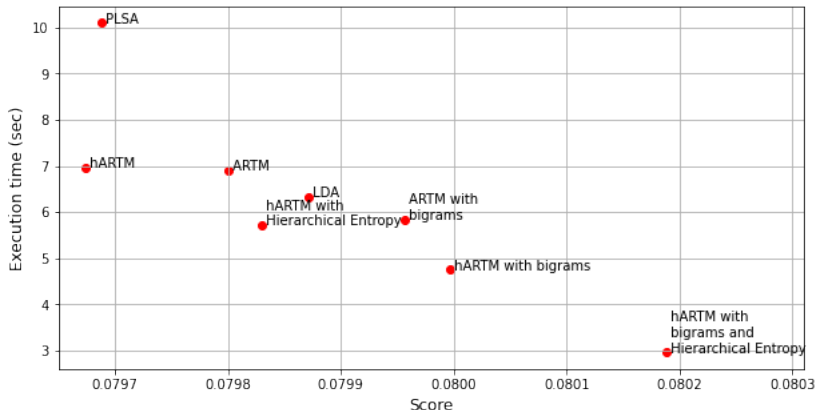


Рис.: Время построения и качество порядков чтения при разных Тематических моделях

# Результаты для каталога по Математике

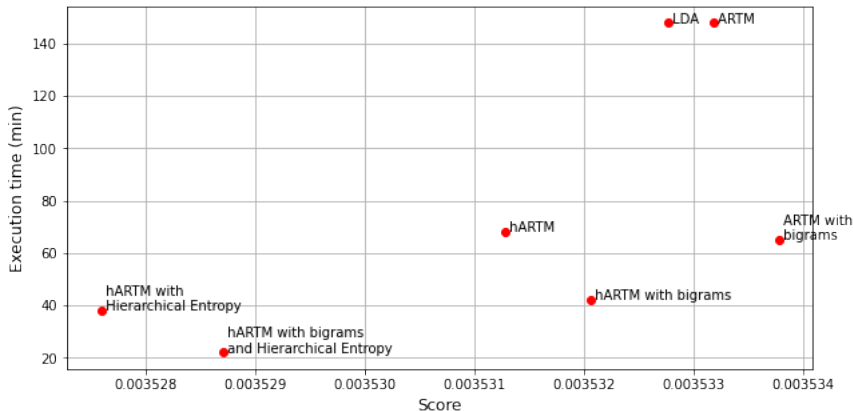


Рис.: Время построения и качество порядков чтения при разных Тематических моделях

- 1 Существуют оптимальные параметры, зависящие только от типа используемой энтропии, а не от тематических моделей, как можно было ожидать.
- 2 Рейтинг порядков чтения, построенных с использованием разных тематических моделей, по качеству не сохраняется при масштабировании.
- 3 Порядки чтения, построенные на основе тематической модели PLSA, дают одно из наилучших качеств, но применимы только к небольшим коллекциям документов из-за их большого времени построения в сравнении с другими моделями.
- 4 Иерархическая энтропия повышает качество порядков чтения на больших коллекциях документов.
- 5 Для произвольной коллекции лучше использовать hARTM и по качеству, и времени построения.
- 6 Использование биграмм повышет ошибку, но снижает время построения.