Генерация порядков чтения из неструктурированного корпуса текстов

RAAI Summer School

Мамонов Кирилл Романович
Московский физико-технический институт

18 июля 2021

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 ацикличный граф. $v_i \in V$ соответствует множеству эквивалентных документов $D_i \neq \emptyset \subseteq D$.
Ребро $v_i \to v_j$ показывает, что документы, принадлежащие множеству D_i, предшевствуют в порядке чтения документам D_j.

Матрица смежности

Порядок чтения представим в виде матрицы смежности:

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

Метрика качества

Разность двух порядков чтения, представленных матрицами A и $\hat{A}$: $MSE(A, \hat{A}) = \frac{1}{n} \sum_{i,j=1}^{n} (A_{ij} - \hat{A}_{ij})^2$.

Литература

Основная

1. 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.
2. 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 Φ содержит распределение слов w в теме t (ϕ_{wt}), матрица Θ — вероятности θ_{td} появления темы t в документе d, T — общее количество тем в модели.

PLSA

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

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

ARTM

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

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

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

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, Ψ — вероятностная матрица смежности тем родительского и дочернего уровней.

Метрики общности

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

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

Метрика пересечения документов

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

$$o(d_i, d_j) = \frac{\theta^i_{td} \cdot \theta^j_{td}}{|\theta^i_{td}|^2 + |\theta^j_{td}|^2 - \theta^i_{td} \cdot \theta^j_{td}}$$

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)

Построеннные тематические модели

Тип	Sparsity θ	Sparsity Φ
LDA	0.76	0.93
ARTM	0.77	0.93
PLSA	0.74	0.93
hARTM	0.87	0.95

Результаты для каталога по Машинному Обучению

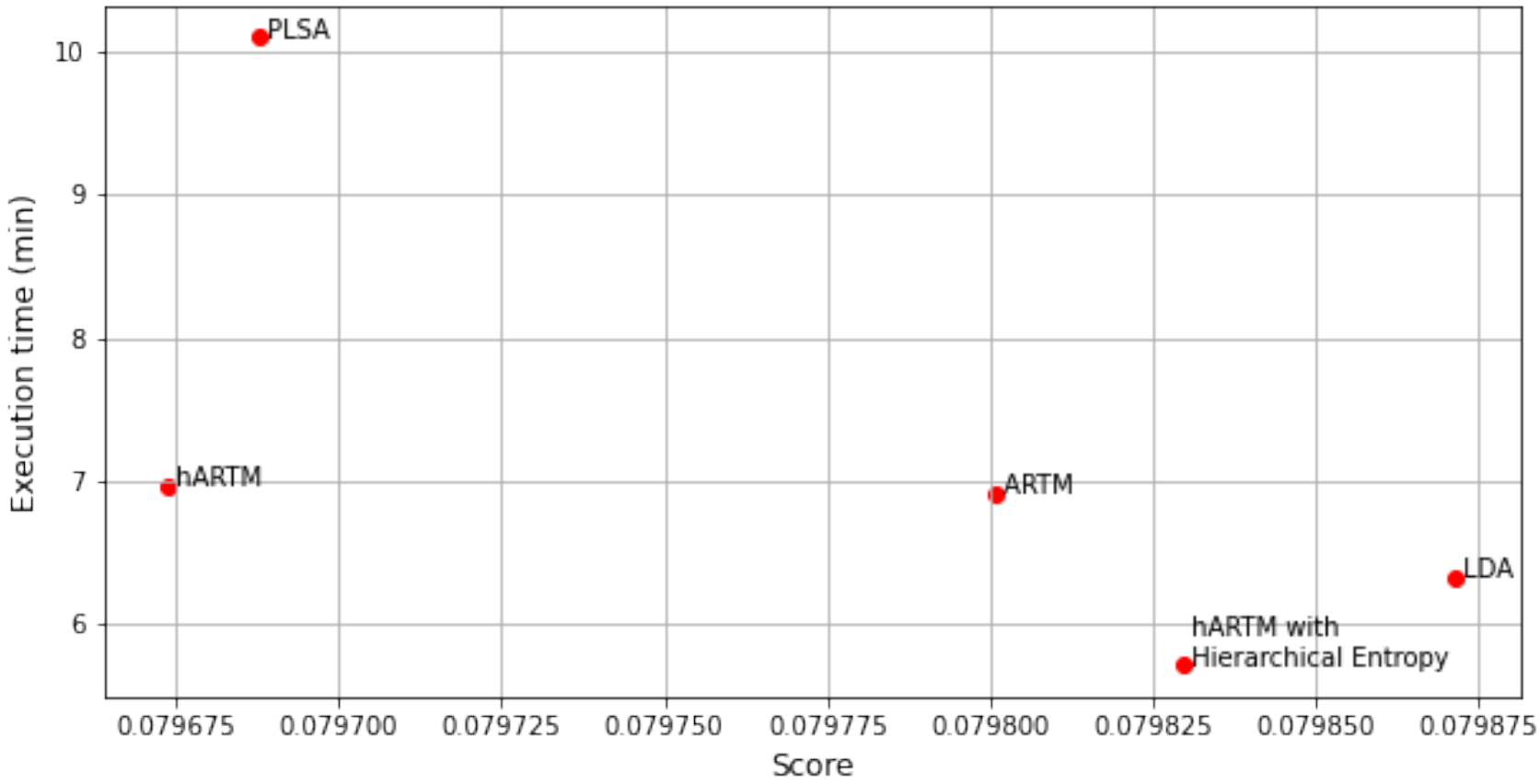

Рис.: Время построения и качество порядков чтения при разных Тематических моделях

Результаты для каталога по Математике

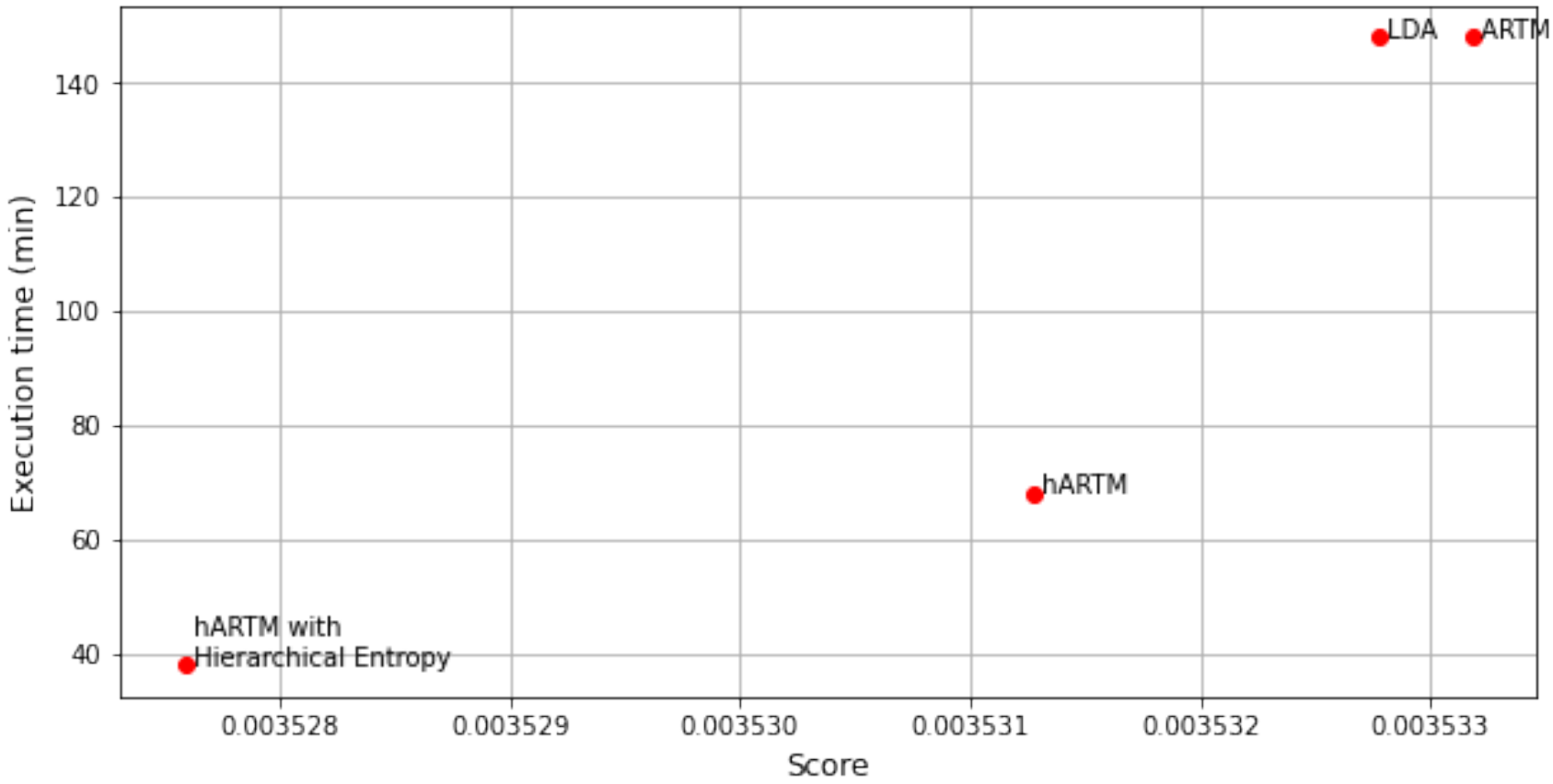

Рис.: Время построения и качество порядков чтения при разных Тематических моделях

Сделано

1. Собран корпус данных
2. Данные декомпозированы
3. Корпус визуализирован
4. Данные очищены
5. Для русского языка успешно воспроизведена статья [1]
6. Построены более точные тематические модели, в том числе иерархическая hARTM, что дало прирост к качеству порядков чтения
7. Разработан новый подход к оценке общности текста — иерархическая энтропия

Литература

1. 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.

Литература

1. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. *Latent dirichlet allocation.* J. Mach. Learn. Res., 3:993–1022, 2003.
2. Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, Marina Suvorova, and Anastasia Yanina. *Non-bayesian additive regularization for multimodal topic modeling of large collections.* 10 2015.
3. N.A. Chirkova. *Additive regularization for hierarchical multimodal topic modeling.* Machine Learning and Data Analysis, 2:187–200, 01 2016.
4. Anton Belyy. *Construction and quality evaluation of heterogeneous hierarchical topic models.* CoRR, abs/1811.02820, 2018

Возможные пути дальнейшего развития

1. Использование биграмм
2. Построение бимодальных тематических моделей
3. Исследование оптимальных гиперпараметров