Вероятностные тематические модели
Лекция 11. Разное.

К. В. Воронцов
vokov@forecsys.ru

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Вероятностные тематические модели (курс лекций, К.В.Воронцов)»

ВМК МГУ • 3 мая 2018

Содержание

Тематическая модель с порождающей модальностью

Основные предположения:

- Модальность C (категории, авторы) порождает темы
- $D \times W \times T \times C$ — дискретное вероятностное пространство
- коллекция — i.i.d. выборка $(d_i, w_i, t_i, c_i)_{i=1}^n \sim p(d, w, t, c)$
- d_i, w_i — наблюдаемые, темы t_i — скрытые
- два предположения об условной независимости:
  $p(w|d,t) = p(w|t), \;\; p(t|c,d) = p(t|c)$

Вероятностная модель порождения документа d:

$$p(w|d) = \sum_{t \in T} p(w|t) \sum_{c \in C} p(t|c)\, p(c|d) = \sum_{t \in T} \phi_{wt} \sum_{c \in C} \psi_{tc} \pi_{cd}$$

- $\phi_{wt} \equiv p(w|t)$ — распределение терминов в темах
- $\psi_{tc} \equiv p(t|c)$ — распределение тем в категориях
- $\pi_{cd} \equiv p(c|d)$ — распределение категорий в документах

ARTM для трёх-матричных разложений ΦΨΠ

Максимизация log правдоподобия с регуляризатором R:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T}\sum_{c\in C} \phi_{wt}\psi_{tc}\pi_{cd} + R(\Phi,\Psi,\Pi) \;\to\; \max_{\Phi,\Psi,\Pi};$$

EM-алгоритм: метод простой итерации для системы уравнений

E-шаг:
$$p_{tcdw} \equiv p(t,c|d,w) = \underset{(t,c)\in T\times C}{\operatorname{norm}}\big(\phi_{wt}\psi_{tc}\pi_{cd}\big);$$

M-шаг:
$$\begin{cases} \phi_{wt} = \underset{w\in W^m}{\operatorname{norm}}\left(n_{wt} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\right); & n_{wt} = \sum\limits_{d,c} n_{dw}p_{tcdw} \\ \psi_{tc} = \underset{t\in T}{\operatorname{norm}}\left(n_{tc} + \psi_{tc}\dfrac{\partial R}{\partial \psi_{tc}}\right); & n_{tc} = \sum\limits_{d,w} n_{dw}p_{tcdw} \\ \pi_{cd} = \underset{c\in C}{\operatorname{norm}}\left(n_{cd} + \pi_{cd}\dfrac{\partial R}{\partial \pi_{cd}}\right); & n_{cd} = \sum\limits_{w,t} n_{dw}p_{tcdw} \end{cases}$$

Автор-тематическая модель (Author-topic model)

$C_d \subset C$ — множество порождающих категорий документа d

- Если $\pi_{cd} = \frac{1}{|C_d|}\left[c \in C_d\right]$, вклады авторов равны,
  то матрица Π фиксирована, EM-алгоритм на Π отдыхает :)
- Если $\pi_{cd} = 0, \quad c \notin C_d$, вклады авторов определяет модель,
  фиксирована структура разреженности матрицы Π,
  EM-алгоритм определяет только ненулевые элементы.
- Если множество C_d задано неточно или частично:

$$R(\Pi) = \sum_{d \in D} \sum_{c \in C_d} \ln \pi_{cd} \to \max$$

- Если множества C_d неизвестны, но Π разрежена:

$$R(\Pi) = -\sum_{d \in D} \sum_{c \in C} \ln \pi_{cd} \to \max$$

M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth. The author-topic model for authors and documents. 2004.

Задача обработки видеопотоков

- Документ d — 1-секундный видеоклип
- Категория c — *поведение* (behaviour), сочетание действий
- Тема t — *действие* (action), сочетание событий
- Терм w — элементарное *визуальное событие* (event)

Задача: выделить в клипе одно основное поведение.

T. Hospedales, Shaogang Gong, Tao Xiang. Video behaviour mining using a dynamic topic model. 2011.

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 рекомендательной системы:**
  (u, f, s) — пользователь u оценил фильм f в ситуации s
- **Данные финансовых организаций:**
  (b, s, g) — покупатель u купил у продавца s товар g

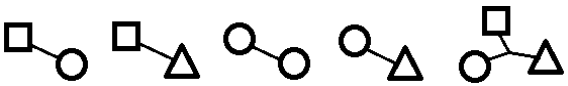
Задача: по наблюдаемой выборке рёбер гиперграфа выявить латентные темы его вершин.

Тематическая модель гиперграфа: определения и обозначения

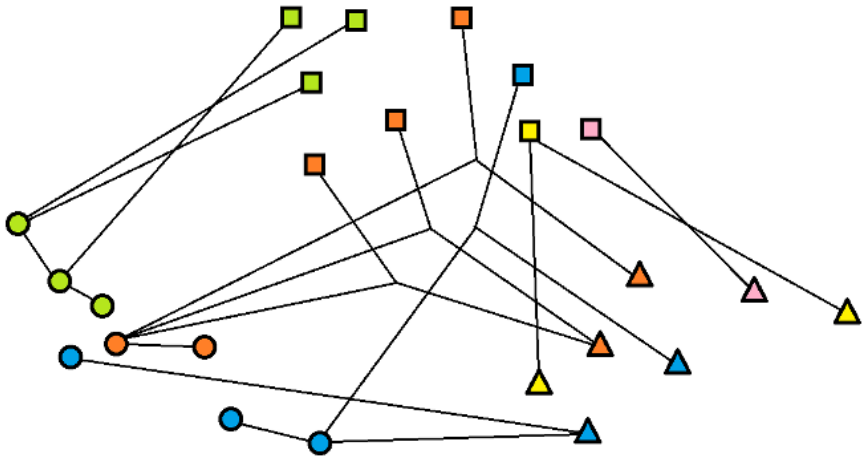
$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.
$V = V^1 \sqcup \cdots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

K — множество типов рёбер:



T — множество тем:



X^k — наблюдаемая выборка транзакций — рёбер типа k
ребро (d, x): вершина-контейнер $d \in V$ и вершины $x \subset V$,
n_{dx} — число вхождений ребра (d, x) в выборку X^k
$p_k(d, x)$ — неизвестное распределение на рёбрах типа k

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k:

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

$\theta_{td} = p(t|d)$ — тематика контейнера не зависит от типа ребра k
$\phi_{kvt} = p_k(v|t)$ — для модальности v в теме t на рёбрах типа k

Задача максимизации log правдоподобия:

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

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

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

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

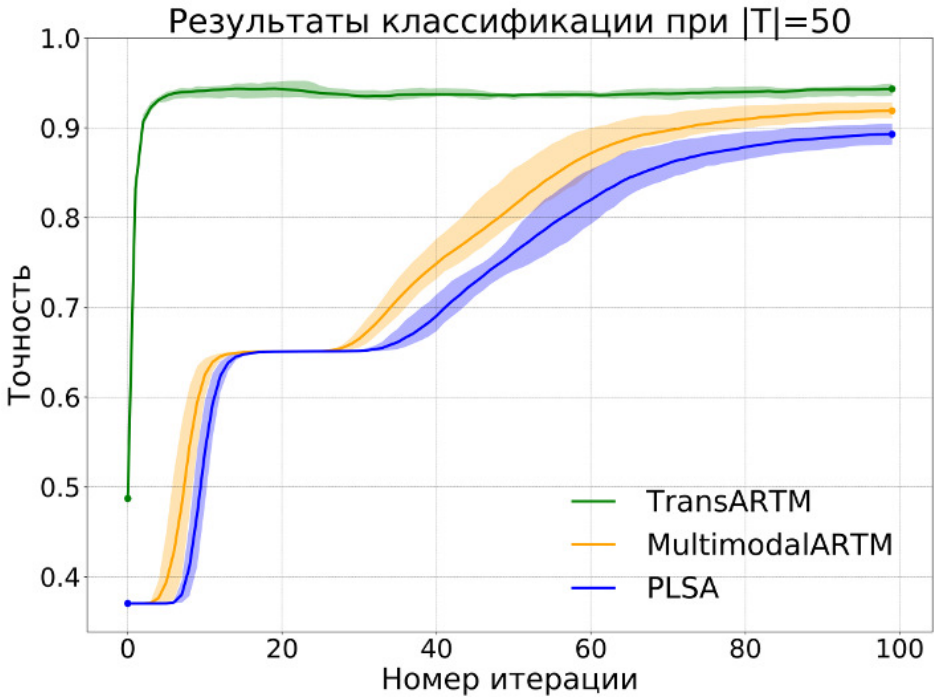
EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{ktdx} = p_k(t|d,x)$:

E-шаг:
$$p_{ktdx} = \underset{t\in T}{\operatorname{norm}}\Big(\theta_{td} \prod_{v\in x} \phi_{kvt}\Big)$$

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

Эксперименты на модельных данных

13M транзакций, 3 модальности, 5 классов, 9 типов рёбер



Вывод: обычные модели не могут восстановить гиперграф.

Илья Кариков. Гиперграфовые тематические модели транзакционных данных. Магистерская диссертация, МФТИ, 2018.

Тематическая модель банковских транзакционных данных

F — конечное множество фирм одной отрасли
n_{bs} — объём транзакций *покупателя* $b \in F$ и *продавца* $s \in F$
V — конечное множество *видов деятельности* (аналог тем)
$\Omega = F \times V \times F \times V$ — дискретное вероятностное пространство

Вероятность транзакции между покупателем b и продавцом s:

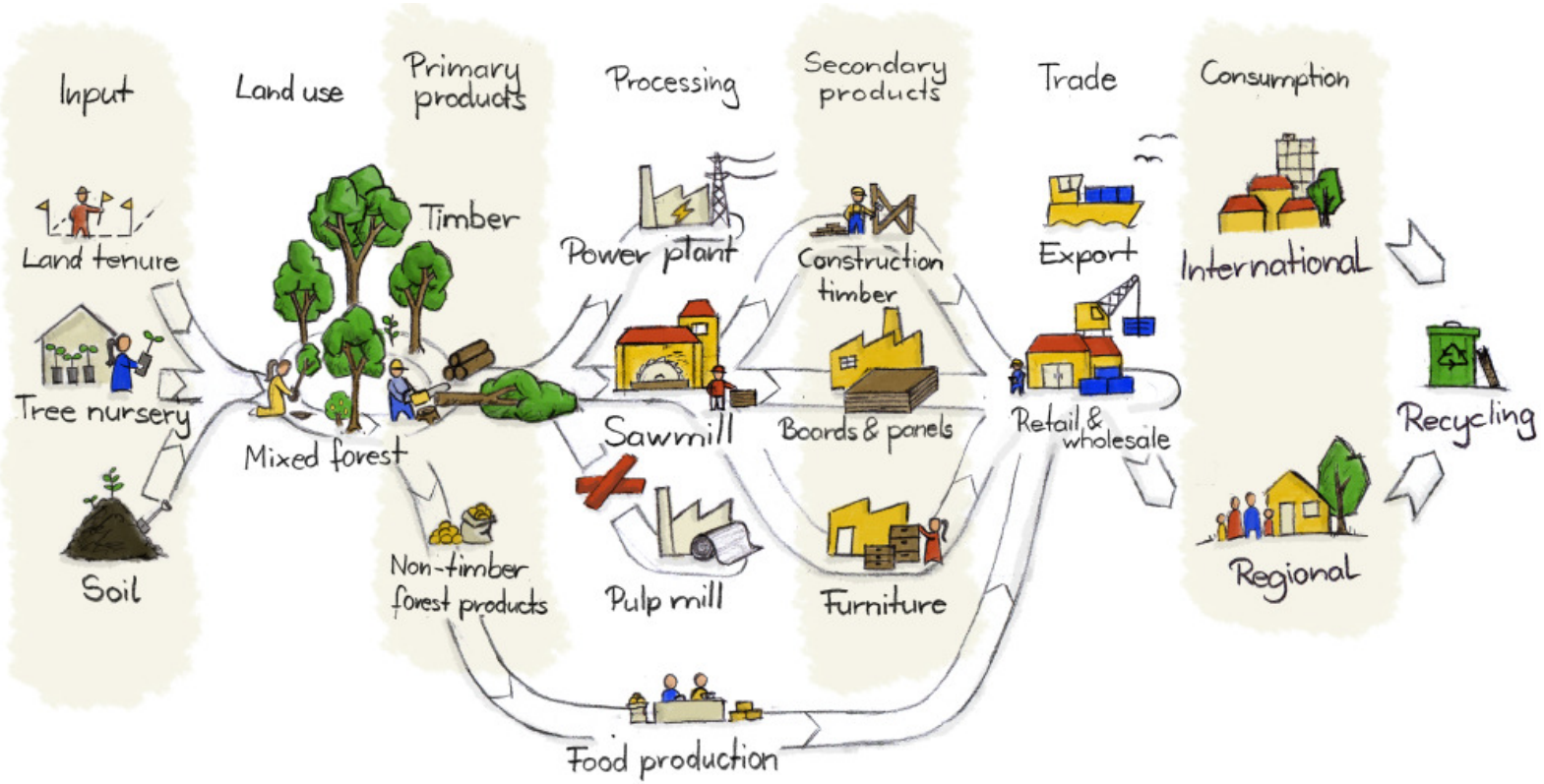
$$p(b,s) = \sum_{u \in V} \sum_{v \in V} p(b|u)p(s|v)p(u,v) = \sum_{u \in V} \sum_{v \in V} p_{bu} p_{sv} \lambda_{uv},$$

$p_{bu} = p(b|u)$ — вероятность, что b осуществляет деятельность u
$p_{sv} = p(s|v)$ — вероятность, что s осуществляет деятельность v
$\lambda_{uv} = p(u,v)$ — вероятность, что продукт деятельности v необходим для осуществления деятельности u

$p(b,s|u,v) = p(b|u)p(s|v)$ — гипотеза условной независимости

Цели тематического моделирования транзакционных данных

- Выявление структуры товарно-денежных потоков отрасли
- Выявление семантики видов деятельности компаний
- Выявление моделей бизнеса компаний



Максимизация правдоподобия и EM-алгоритм

Принцип максимума правдоподобия:

$$\sum_{b,s\in F} n_{bs} \log \sum_{u,v\in V} p_{bu} p_{sv} \lambda_{uv} + R(P,\Lambda) \to \max_{P,\Lambda}$$

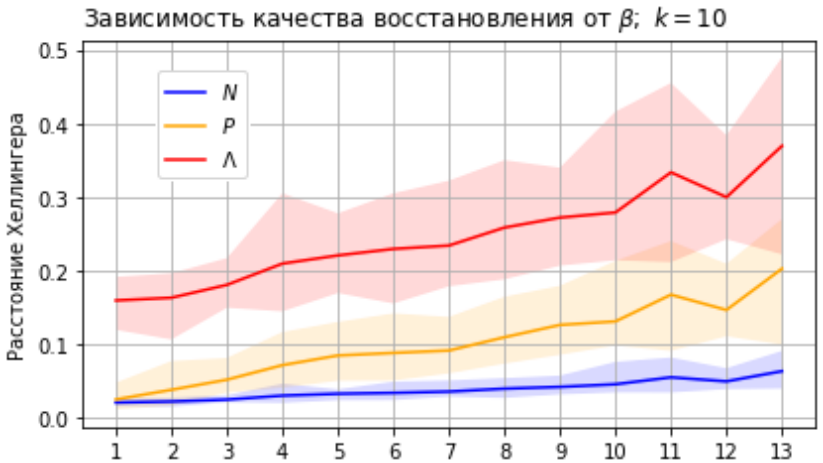
EM-алгоритм: метод простой итерации для системы уравнений

$$\text{E-шаг:} \quad p_{uvbs} = \underset{(u,v)\in V^2}{\text{norm}} p_{bu} p_{sv} \lambda_{uv}$$

$$\text{M-шаг:} \quad \begin{cases} p_{fz} = \underset{f\in F}{\text{norm}} \left(\sum_{s,v} n_{fs} p_{zvfs} + \sum_{b,u} n_{bf} p_{uzbf} + p_{fz} \frac{\partial R}{\partial p_{fz}} \right) \\ \lambda_{uv} = \underset{(u,v)\in V^2}{\text{norm}} \left(\sum_{b,s} n_{bs} p_{uvbs} + \lambda_{uv} \frac{\partial R}{\partial \lambda_{uv}} \right) \end{cases}$$

Эксперименты на модельных данных

$|F| = 40$, $|V| = 15$, $\alpha = 0.9$ — разреженность Λ,
k — среднее число действительных видов деятельности фирм
$k + \beta$ — число предполагаемых видов деятельности фирм



Выводы: чем разреженнее модель, и чем больше мы знаем о видах деятельности фирм, тем лучше восстановление.

Роза Айсина. Тематическое моделирование финансовых потоков корпоративных клиентов банка по транзакционным данным. Бакалаврская диссертация, ВМК МГУ, 2017.

Выводы

- Тематическое моделирование легко обобщается на мультимодальные транзакционные данные
- Скрытых переменных также может быть несколько
- Исследования новых типов моделей начинаются с экспериментов по восстановлению модельных данных
- Много нетекстовых приложений: изображения, видео, реклама, логи, игры, банки, ...

Регуляризатор для сокращения числа тем

Цель: избавиться от «мелких» незначимых тем.

Разреживаем распределение $p(t) = \sum_d p(d)\theta_{td}$, максимизируя кросс-энтропию между $p(t)$ и равномерным распределением:

$$R(\Theta) = -\tau \sum_{t \in S} \ln \sum_{d \in D} p(d)\theta_{td} \to \max.$$

Подставляем, получаем:

$$\theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td} - \tau \frac{n_d}{n_t}\theta_{td}\right), \text{ вариант: } \theta_{td} = \underset{t \in T}{\text{norm}}\left(n_{td}\left(1 - \frac{\tau}{n_t}\right)\right).$$

Эффект: обнуляются строки матрицы Θ с малыми n_t, заодно (неожиданно) удаляются зависимые и расщеплённые темы.

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization. SLDS 2015.

Эксперименты с регуляризатором отбора тем

Коллекция статей NIPS (Neural Information Processing System)

- $|D| = 1566$ обучающих документов; $|D'| = 174$ тестовых
- $|W| = 13\,\text{K}$ — мощность словаря

Синтетическая коллекция:

- строим PLSA за 500 итераций, $|T_0| = 50$ тем на NIPS
- генерируем (n^0_{dw}) из полученных Φ и Θ:

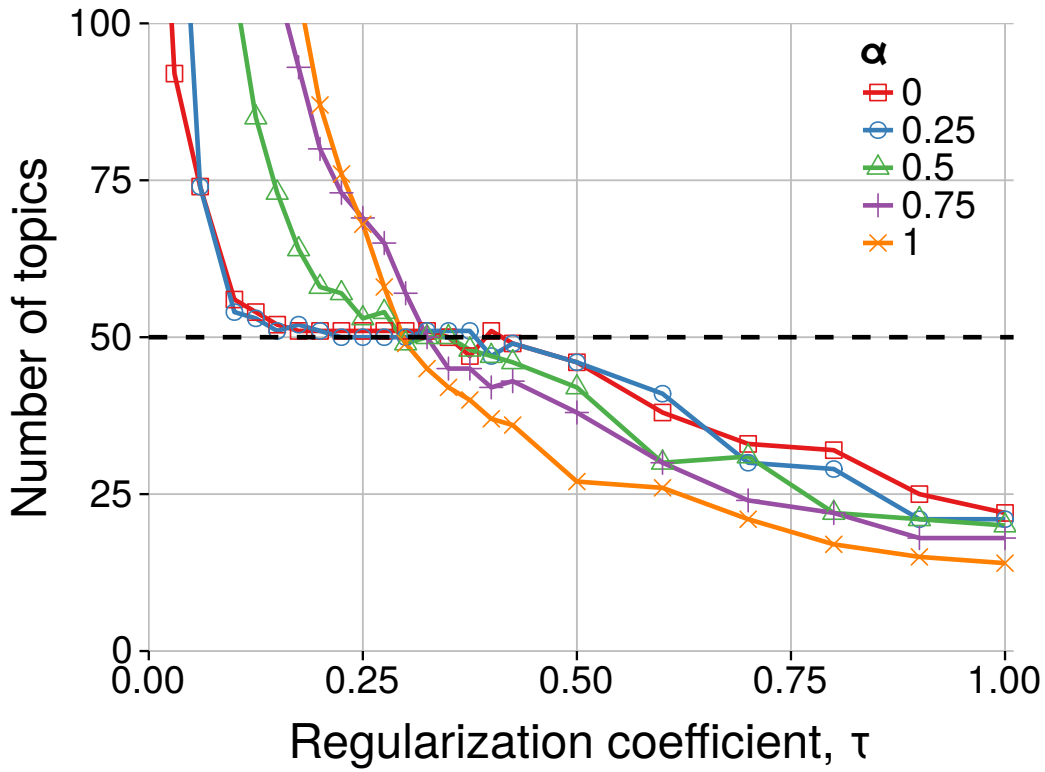
$$n^0_{dw} = n_d \sum_{t \in T} \phi_{wt} \theta_{td}$$

Параметрическое семейство полусинтетических данных:

- n^α_{dw} — смесь синтетических данных n^0_{dw} и реальных n_{dw}:

$$n^\alpha_{dw} = \alpha n_{dw} + (1 - \alpha) n^0_{dw}$$

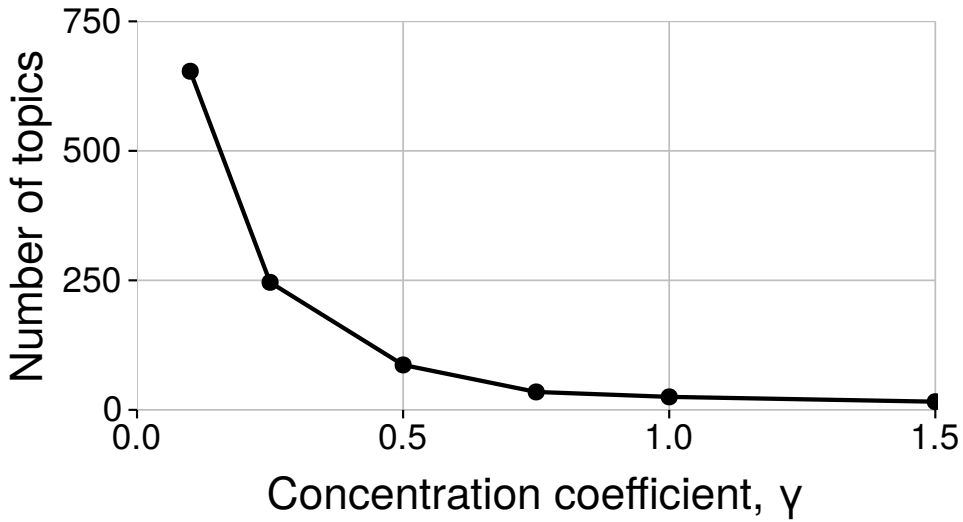
Попытка определения числа тем



- На синтетических данных надёжно находим $|T| = 50$,
- в широком интервале значений коэффициента τ;
- однако на реальных данных нет столь чёткого интервала.

Сравнение с байесовской тематической моделью HDP

HDP, Hierarchical Dirichlet Process [Tech et.al, 2006] — «state-of-the-art» байесовский подход к определению числа тем

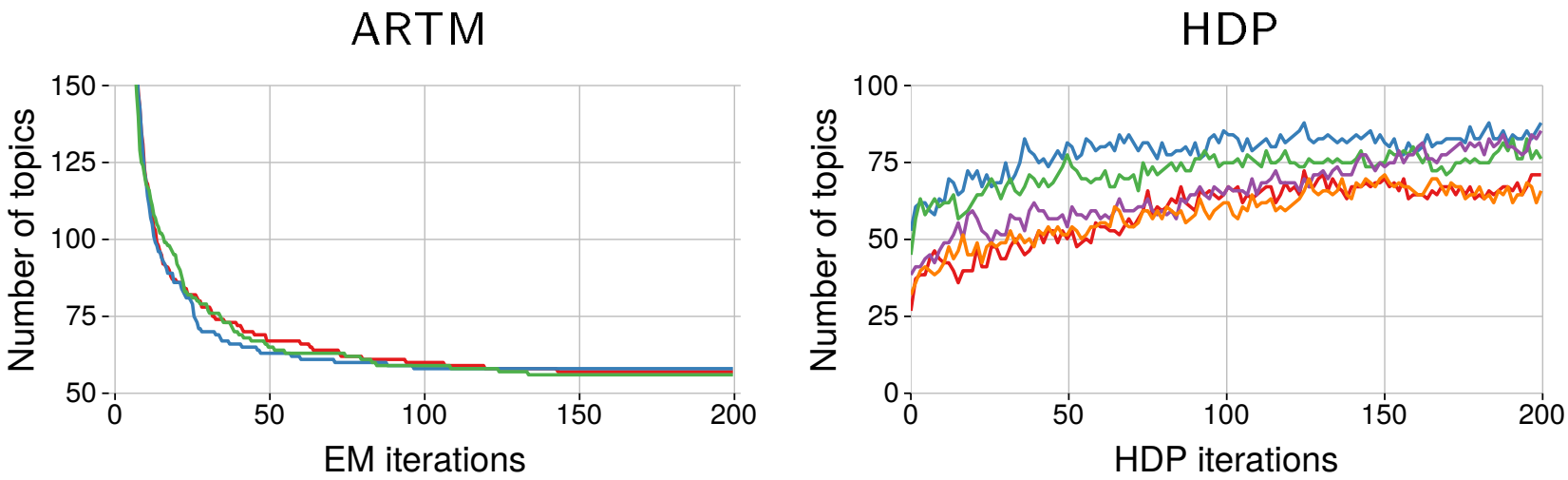


- *Коэффициент концентрации* γ в HDP влияет на $|T|$ так же сильно, как выбор коэффициента τ в ARTM.

Y.W.Teh, M.Jordan, M.Beal, D.Blei. Hierarchical Dirichlet Processes. 2005.

Сравнение ARTM и HDP по устойчивости

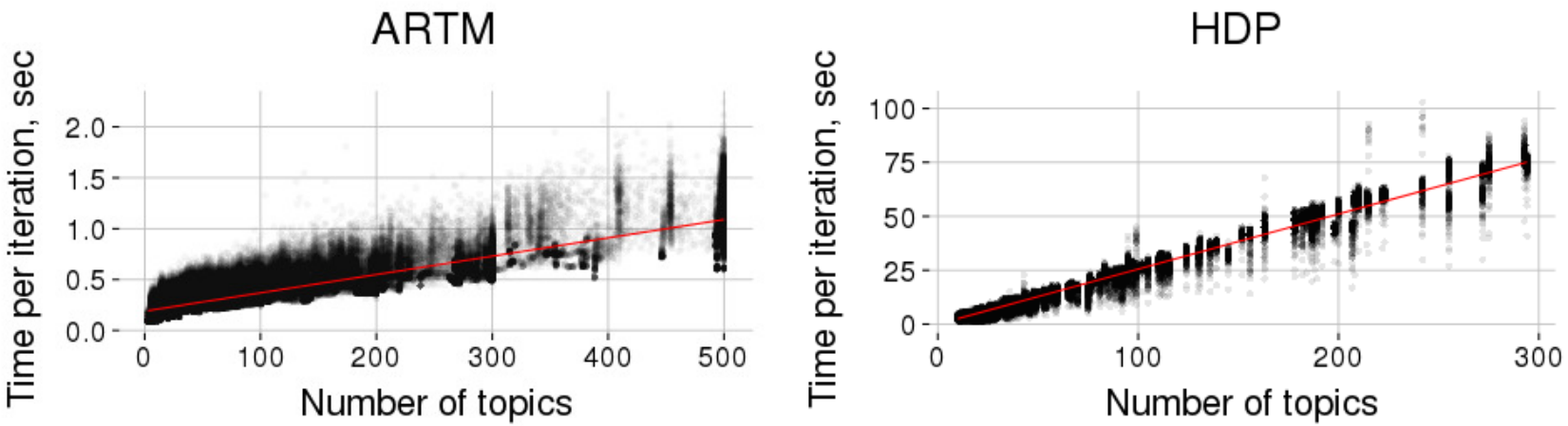
Запуск ARTM и HDP много раз из случайных инициализаций:



- HDP менее устойчив, причём в двух смыслах:
  - число тем сильнее флуктуирует от итерации к итерации;
  - результаты нескольких запусков различаются сильнее.
- «Рекомендуемые» значения параметров γ в HDP и τ в ARTM дают примерно равное число тем $|T| \approx 60$

Сравнение ARTM и HDP по времени вычислений

Сравнение времени одного прохода коллекции (sec)

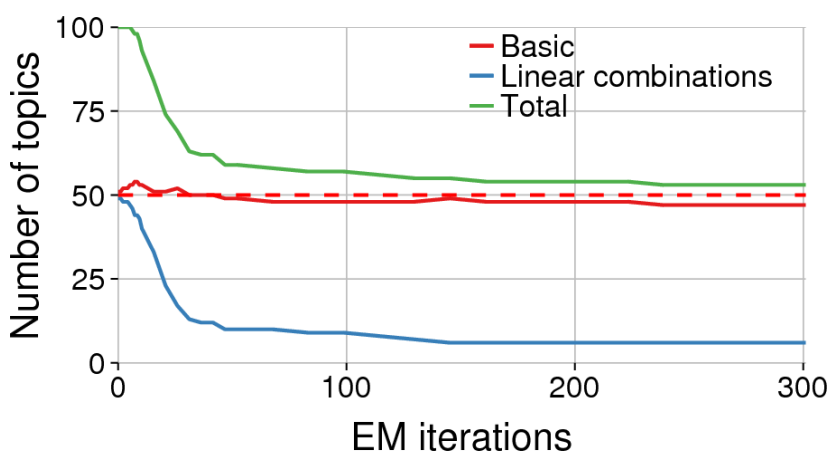


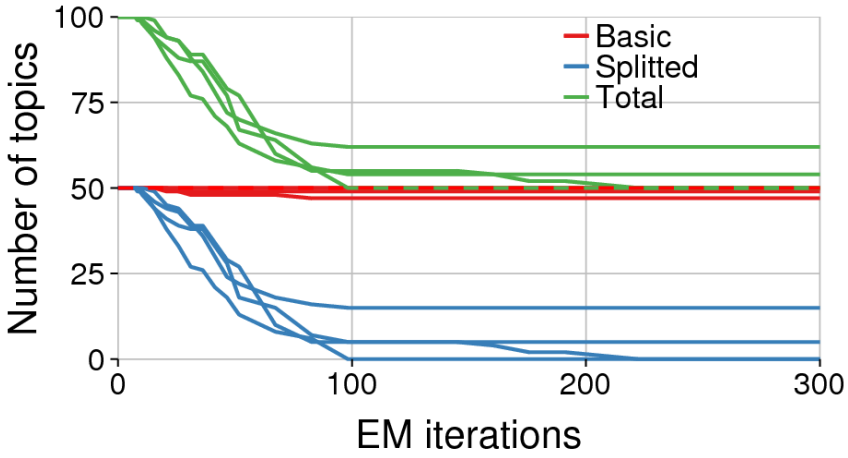
- ARTM в 100 раз быстрее!

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.

Удаление линейно зависимых и расщеплённых тем

Добавили 50 линейных комбинаций тем в модельную Φ.
Расщепили 50 тем, каждую на две подтемы в модельной Φ.





- Удаляются линейно зависимые и расщеплённые темы
- Остаются более различные темы исходной модели.

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.

Выводы

- Регуляризатор отбора тем — для удаления незначимых, зависимых, расщеплённых тем.
- Оптимального числа тем вообще не существует! Оно задаётся исходя из целей моделирования.
- Есть простой метод для удаления лишних тем, но пока в ARTM нет простых критериев добавления тем.
- **Открытая проблема:** почему этот регуляризатор удаляет линейно зависимые и расщеплённые темы?

Задача автоматического именования тем (topic labeling)

Требования к *названию темы* (topic label):

- релевантность названия теме
- интерпретируемость и грамматическая корректность
- непохожесть на названия похожих тем

Гипотеза 1: тройка топ-слов — плохое название.
Гипотеза 2: все названия уже придуманы, осталось их найти.

Подзадачи

- формирование названий-кандидатов $\ell_1, \ldots, \ell_m$
- построение (обучение) функции релевантности $s(\ell, t)$
- выбор названия с учётом названий похожих тем

Qiaozhu Mei, Xuehua Shen, Chengxiang Zhai. Automatic labeling of multinomial topic models. KDD 2007.

Способы формирования названий-кандидатов

Специфичные для данной темы:

- топовые n-граммы данной темы
- синтаксические ветки наиболее тематичных предложений
- тематичные именные группы (вырезанные OpenNLP chunker)
- тематичные фразы «объект, субъект, действие»
- заголовки тематичных документов или их фрагменты
- метаданные (теги, категории) тематичных документов

Общие для всех тем:

- n-граммы из внешней коллекции, например, Википедии
- заголовки статей или категорий Википедии
- термины из внешних тезаурусов:
  WordNet, PyTez, Викисловарь, и др.

Функция релевантности (relevance score)

Релевантность нулевого порядка:

$$s(\ell, t) = \sum_{w \in \ell} \log \frac{p(w|t)}{p(w)} \to \max$$

Релевантность первого порядка: слова темы t неслучайно часто появляются рядом (в одном контексте C) с названием ℓ:

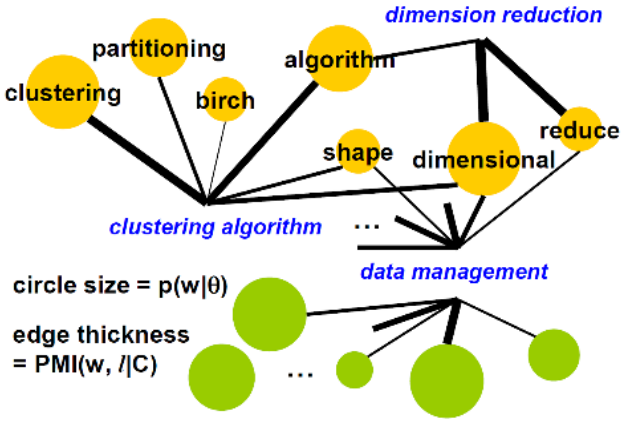
$$s(\ell, t) = \sum_{w \in \ell} p(w|t) \underbrace{\log \frac{p(w, \ell|C)}{p(w|C)p(\ell|C)}}_{\mathrm{PMI}(w,\ell|C)} \to \max$$

где C — релевантный теме контекст, в котором ожидается появление как слов темы t, так и названия ℓ целиком.

Qiaozhu Mei, Xuehua Shen, Chengxiang Zhai. Automatic labeling of multinomial topic models. KDD 2007.

Выбор нескольких названий для темы

Пример: оранжевая тема покрывается двумя названиями:
— *clustering algorithm*
— *dimension reduction*
но название *data management* неудачно, конкурирует с другой темой



Выбирать каждое следующее название, чтобы оно было

- максимально релевантно, $s(\ell, t) \to \max$,
- максимальное не похоже на уже выбранные названия ℓ':

$$s(\ell, t) + \lambda \max_{\ell'} \mathrm{KL}(\ell' \| \ell) \to \max$$

где параметр λ подбирается эмпирически.

Qiaozhu Mei, Xuehua Shen, Chengxiang Zhai. Automatic labeling of multinomial topic models. KDD 2007.

Максимизация различности названий различных тем

Модифицированная функция релевантности $s'(\ell, t)$:

- максимизирует релевантность своей темы, $s(\ell, t) \to \max$
- минимизирует релевантность других тем, $s(\ell, t') \to \min$

$$s'(\ell, t) = s(\ell, t) - \mu \sum_{t' \in T \setminus t} s(\ell, t') \to \max$$

где параметр μ подбирается эмпирически.

Методика оценивания качества именования тем:

- 3 асессора, каждый асессор видит для каждой темы:
  — список топ-слов темы, список топ-документов темы
  — варианты названия, сгенерированные разными методами
- асессор ранжирует методы $0, 1, 2, \ldots$ (чем выше, тем лучше)

Qiaozhu Mei, Xuehua Shen, Chengxiang Zhai. Automatic labeling of multinomial topic models. KDD 2007.

Оценивание качества именования тем

Две коллекции: научная (SIGMOD), новостная (Assoc.Press)
Автоматические и асессорские названия тем, SIGMOD:

	SIGMOD			
Auto Label	**clustering algorithm**	**r tree**	**data streams**	**concurrency control**
Man. Label	**clustering algorithms**	**indexing methods**	**Stream data management**	**transaction management**
θ	clustering clusters video dimensional cluster partitioning quality birch	tree trees spatial b r disk array cache	stream streams continuous monitoring multimedia network over ip	transaction concurrency transactions recovery control protocols locking log

Победил выбор n-грамм по релевантности 1-го порядка,
но он всё ещё заметно хуже человеческого именования тем:

Baseline v.s. Zero-order v.s. First-order				
Dataset	#Label	Baseline	Ngram-0-B	Ngram-1
SIGMOD	1	0.76	0.75	**1.49**
SIGMOD	5	0.36	1.15	**1.51**
AP	1	0.97	0.99	**1.02**
AP	5	0.85	0.66	**1.48**

System v.s. Human			
Dataset	#Label	Ngram-1	Human
SIGMOD	1	0.35	**0.65**
SIGMOD	5	0.25	**0.75**
AP	1	0.24	**0.76**
AP	5	0.21	**0.79**

Резюме

- *Automatic Topic Labeling* — очень узкое направление, всего 20–30 статей за 10 лет
- Важно для автоматизации создания приложений
- Близко к задаче суммаризации темы
- Для иерархических моделей добавляется требование *полноты*: названия дочерних тем должны адекватно описывать разделение родительской темы

Из последнего:

Wanqiu Kou, Fang Li, T.Baldwin. Automatic Labelling of Topic Models using Word Vectors and Letter Trigram Vectors. 2015.

S.Bhatia, Jey Han Lau, T.Baldwin. Automatic Labelling of Topics with Neural Embeddings. COLING-2016

Xiaojun Wan, Tianming Wang. Automatic Labeling of Topic Models Using Text Summaries. 2016.

M.Allahyari, S.Pouriyeh, K.Kochut, H.R.Arabnia. A Knowledge-based Topic Modeling Approach for Automatic Topic Labeling. 2017.