

Вероятностные тематические модели

Лекция 7. Мультиязычные, многоматричные и гиперграфовые тематические модели

Константин Вячеславович Воронцов
k.v.vorontsov@phystech.edu

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

1 Мультиязычные тематические модели

- Параллельные и сравнимые тексты
- Двужычные словари
- Кросс-язычный поиск

2 Трёх-матричные тематические модели

- Тематическая модель с порождающей модальностью
- Автор-тематическая модель
- Тематическая модель для анализа видеопотоков

3 Тематическая модель транзакционных данных

- Примеры транзакционных данных
- Гиперграфовая тематическая модель с регуляризацией
- Примеры транзакционных моделей на гиперграфах

Напоминание. Мультимодальная ARTM: постановка задачи

Модальности: слова, n -граммы, авторы, даты, категории, и т.д.

Дано: W^m — словарь термов m -й модальности, $m \in M$,
 D — коллекция текстовых документов $d \subset W = \bigsqcup_m W^m$,
 n_{dw} — сколько раз терм w встретился в документе d .

Найти: модель $p(w|d) = \sum_{t \in T} \phi_{wt} \theta_{td}$ с параметрами $\Phi^m_{W^m \times T}$ и $\Theta_{T \times D}$:
 $\phi_{wt} = p(w|t)$ — вероятности терма w в каждой теме t ,
 $\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d .

Критерий максимума регуляризованного правдоподобия:

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

$$\phi_{wt} \geq 0; \quad \sum_{w \in W^m} \phi_{wt} = 1; \quad \theta_{td} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1.$$

Максимизация \log правдоподобия с регуляризатором R :

$$\sum_{d \in D} \sum_{w \in \mathcal{W}} \tilde{n}_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta};$$

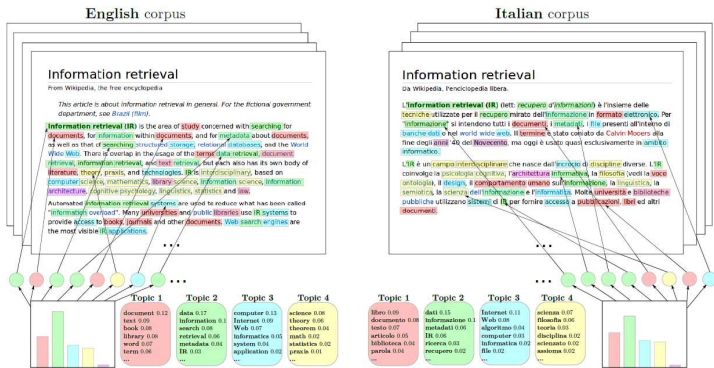
где $\tilde{n}_{dw} = \tau_{m(w)} n_{dw}$, $m(w)$ — модальность термина w .

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tdw} \equiv p(t|d, w) = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}); \\ \text{M-шаг:} & \begin{cases} \phi_{wt} = \operatorname{norm}_{w \in W^m} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right), & n_{wt} = \sum_{d \in D} \tilde{n}_{dw} p_{tdw}; \\ \theta_{td} = \operatorname{norm}_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right), & n_{td} = \sum_{w \in \mathcal{W}} \tilde{n}_{dw} p_{tdw}; \end{cases} \end{cases}$$

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

Многоязычные модели параллельных коллекций



Для построения мультиязычных тем достаточно иметь парные документы, без выравнивания, без двуязычных словарей!

I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. 2015

Параллельные и сравнимые корпуса текстов

Parallel — точный перевод (с выравниванием предложений),
пример: EuroParl, протоколы европарламента, 21 язык.

Comparable — не перевод, а пересказ на другом языке,
пример: Википедия.

W^ℓ — словарь языка ℓ из множества языков L .

Модель ML-P (MultiLingual Parallel)

- каждый язык — отдельная модальность
- $\theta_{td} = p(t|d)$ общее для всех связных документов $d = \bigsqcup_{\ell \in L} d^\ell$

Дополнительные данные — двуязычные словари:

- $P_k(w) \subset W^k$ — все переводы слова $w \in W^\ell$ в языке k

I. Vulić, W. De Smet, J. Tang, M.-F. Moens. Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications. 2015

Пример тем. Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
 Первые 10 слов и их вероятности $p(w|t)$ в %:

| 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 |

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример тем. Мультиязычная модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
 Первые 10 слов и их вероятности $p(w|t)$ в %:

| 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 |

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

Vorontsov, Frei, Apishev, Romov, Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Регуляризация по двуязычным словарям. Модель ML-TD

Гипотеза. Если $u \in \Pi_k(w)$, то тематика слов w и u близка:

$$\text{KL}(\hat{p}(t|u) \parallel p(t|w)) \rightarrow \min,$$

где $\hat{p}(t|u) = \frac{n_{ut}}{n_u}$, $p(t|w) = p(w|t) \frac{p(t)}{p(w)} = \phi_{wt} \frac{n_t}{n_w}$.

Модель ML-TD (MultiLingual Translation Dictionary)

$$R(\Phi) = \tau \sum_{\ell, k \in L} \sum_{w \in W^\ell} \sum_{u \in \Pi_k(w)} \sum_{t \in T} n_{ut} \ln \phi_{wt} \rightarrow \max_{\Phi}.$$

Недостатки. Модель ML-TD не учитывает два обстоятельства:

- тематику омонимов сближать не нужно,
- слово может иметь разные переводы в разных темах.

Дударенко М. А. Регуляризация многоязычных тематических моделей // Вычислительные методы и программирование. 2015. Т. 16. С. 26–36.

Матрица вероятностей переводов. Модель ML-TDP

Гипотеза. Переводы слов зависят от тем: $\pi_{uwt}^{kl} = p(u|w, t)$,
темы согласуются в разных языках через переводы слов:

$$\text{KL}(\hat{p}(u|t) \parallel p(u|t)) \rightarrow \min;$$

$\hat{p}(u|t) = \frac{n_{ut}}{n_t}$ — частотная оценка по модальности (языку) k ,
 $p(u|t)$ — модель темы t в языке k по языку ℓ :

$$p(u|t) = \sum_{w \in \Pi_\ell(u)} p(u|w, t)p(w|t) = \sum_{w \in \Pi_\ell(u)} \pi_{uwt}^{kl} \phi_{wt}.$$

Модель ML-TDP (MultiLingual Translation Dictionary Probability)

$$R(\Phi, \Pi) = \tau \sum_{\ell, k \in L} \sum_{u \in W^k} \sum_{t \in T} n_{ut} \ln \sum_{w \in \Pi_\ell(u)} \pi_{uwt}^{kl} \phi_{wt} \rightarrow \max_{\Phi, \Pi}.$$

Дударенко М. А. Регуляризация многоязычных тематических моделей // Вычислительные методы и программирование. 2015. Т. 16. С. 26–36.

Формулы M-шага для моделей ML-TD и ML-TDP

ML-TD (MultiLingual Translation Dictionary):

$$\phi_{wt} = \operatorname{norm}_{w \in W^\ell} \left(n_{wt} + \tau \sum_{k \in L \setminus \ell} \sum_{u \in \Pi_k(w)} n_{ut} \right)$$

ML-TDP (MultiLingual Translation Dictionary Probability):

$$\phi_{wt} = \operatorname{norm}_{w \in W^\ell} \left(n_{wt} + \tau \sum_{k \in L \setminus \ell} \sum_{u \in \Pi_k(w)} \pi_{wut}^{k\ell} n_{ut} \right)$$
$$\pi_{uwt}^{k\ell} = \operatorname{norm}_{u \in W^k} \left(\pi_{wut}^{k\ell} n_{ut} \right)$$

Смысл регуляризации:

условные вероятности $\phi_{wt} = p(w|t)$ согласуются
с их частотными оценками по словам других языков

Тематические переводы слов $\pi_{uwt}^{kl} = p(u|w, t)$ Темы, в которых $p(\langle \text{sum} \rangle | \langle \text{сумма} \rangle, t) > 0.9$

| Тема №6 | | Тема №12 | | Тема №20 | |
|--------------|----------|----------------|-------------|----------------|-------------|
| множество | set | математика | triangle | вектор | vector |
| пространство | space | треугольник | square | координата | coordinate |
| группа | point | теорема | number | пространство | field |
| точка | left | точка | point | преобразование | tensor |
| элемент | limit | математический | theorem | базис | transform |
| функция | symmetry | угол | angle | тензор | basis |
| предел | function | координата | mathematics | сила | space |
| отображение | open | экономика | real | векторный | force |
| симметрия | property | число | theory | точка | rotation |
| открытый | topology | квадрат | geometry | система | thermometer |

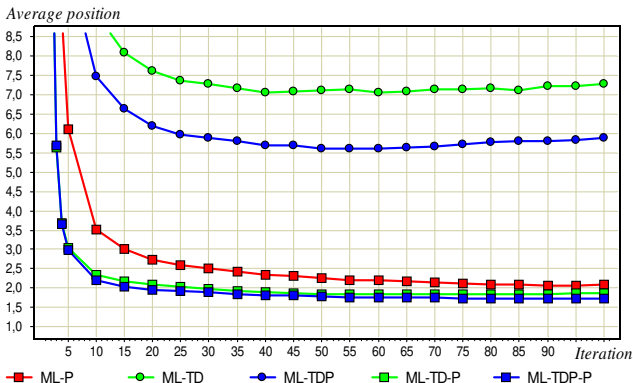
Темы, в которых $p(\langle \text{total} \rangle | \langle \text{сумма} \rangle, t) > 0.9$

| Тема №5 | | Тема №19 | | Тема №22 | |
|-------------|------------|-------------|------------|-------------|-----------|
| орбита | space | программный | software | игра | game |
| аппарат | nasum | версия | version | видеосигнал | character |
| космический | orbit | работа | news | игрок | video |
| земля | instrument | компания | company | фильм | player |
| поверхность | earth | анонимный | work | головоломка | series |
| солнечный | surface | примечание | note | серия | puzzle |
| станция | solar | терминатор | release | качество | movie |
| запуск | system | журнал | support | шахматы | jason |
| система | landing | рей | terminator | джейсон | world |
| атмосфера | camera | персонаж | anonymous | буква | chess |

Кросс-язычный поиск: ищем документ по его переводу

Wiki: $|D| = 586$, категория «Математика», $|T| = 100$,
 $|W_{\text{рус}}| = 19\,305$, $|W_{\text{eng}}| = 23\,413$, переводов 82 642 пар.

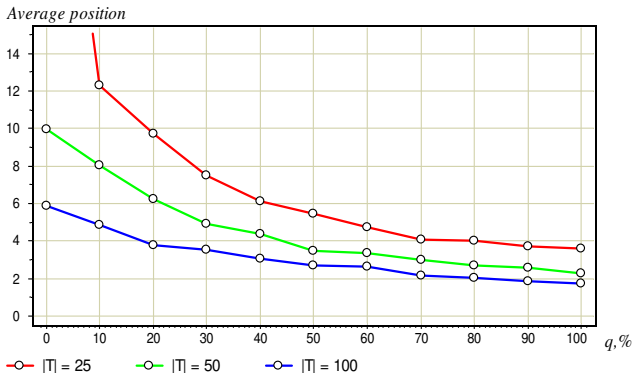
Качество поиска — средняя позиция перевода в выдаче:



Кросс-язычный поиск: ищем документ по его переводу

Wiki: $|D| = 586$, категория «Математика», $|T| = 25, 50, 100$,
 $|W_{\text{рус}}| = 19\,305$, $|W_{\text{eng}}| = 23\,413$, переводов 82 642 пар.

Зависимость средней позиции перевода в выдаче
от числа тем $|T|$ и доли q параллельных текстов в коллекции:



Поиск и рубрикация научных публикаций на 100 языках

Цель: мультиязыковой поиск и классификация научных публикаций по рубрикам УДК, ГРНТИ, ОЭСР, ВАК

| модель | ср.ч. УДК | ср.% УДК | ср.ч. ГРНТИ | ср.% ГРНТИ |
|-------------|--------------|--------------|--------------|--------------|
| Базовая TM | 0.558 | 0.165 | 0.536 | 0.220 |
| XLM-RoBERTa | 0.835 | 0.179 | 0.832 | 0.288 |
| ARTM | 0.995 | 0.225 | 0.852 | 0.366 |

Регуляризаторы:

$$\mathcal{L} \left(\begin{array}{c} \text{PLSA} \\ \left(\begin{array}{|c|} \hline \Phi \\ \hline \end{array} \begin{array}{|c|} \hline \Theta \\ \hline \end{array} \right) \end{array} \right) + R \left(\begin{array}{c} \text{interpretable} \\ \left(\begin{array}{|c|} \hline \text{matrix} \\ \hline \end{array} \begin{array}{|c|} \hline \text{matrix} \\ \hline \end{array} \right) \end{array} \right) + R \left(\begin{array}{c} \text{multimodal} \\ \left(\begin{array}{|c|} \hline \text{matrix} \\ \hline \end{array} \begin{array}{|c|} \hline \text{matrix} \\ \hline \end{array} \right) \end{array} \right) + R \left(\begin{array}{c} \text{multilanguage} \\ \left(\begin{array}{|c|} \hline \text{matrix} \\ \hline \end{array} \begin{array}{|c|} \hline \text{matrix} \\ \hline \end{array} \right) \end{array} \right) + R \left(\begin{array}{c} \text{supervised} \\ \left(\begin{array}{|c|} \hline \text{matrix} \\ \hline \end{array} \begin{array}{|c|} \hline \text{matrix} \\ \hline \end{array} \right) \end{array} \right) \rightarrow \max$$

Результаты:

- точность мультиязычного поиска 94%
- сокращение модели 128 Гб → 4.8 Гб при редукции словарей (ВРЕ-токенизация) до 11К токенов на каждый язык.

П.Потапова, А.Грабовой, О.Бахтеев, Е.Егоров, Н.Зиновкин, Ю.Чехович, К.Воронцов и др. Мультиязыковая автоматическая рубрикация научных документов. 2023.

Резюме по мультиязычным моделям

- Главное чудо: для построения мультиязычных тем достаточно иметь сравнимые корпуса
- Сравнимая коллекция является более сильным источником многоязычной информации, чем словарь переводов (!)
- Модель с вероятностями переводов — самая сильная
- Не обязательно, чтобы все документы имели параллельные
- Главное применение — по запросу на одном языке ищем:
 - тексты на другом языке — *кросс-язычный поиск*
 - тексты на всех языках — *мульти-язычный поиск*
- Каждая тема получает представление в каждом языке
- Объёмы темы в языках могут различаться существенно
- Устаревшее применение в машинном переводе:
выбор варианта перевода согласно тематике контекста

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 термов 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 $\Phi\Psi\Pi$

Максимизация \log правдоподобия с регуляризатором R :

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

EM-алгоритм: метод простой итерации для системы уравнений

$$\begin{cases} \text{E-шаг:} & p_{tcdw} \equiv p(t, c | d, w) = \mathop{\text{norm}}_{(t,c) \in T \times C} (\phi_{wt} \psi_{tc} \pi_{cd}); \\ \text{M-шаг:} & \left\{ \begin{array}{l} \phi_{wt} = \mathop{\text{norm}}_{w \in W^m} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right); \quad n_{wt} = \sum_{d,c} n_{dw} p_{tcdw} \\ \psi_{tc} = \mathop{\text{norm}}_{t \in T} \left(n_{tc} + \psi_{tc} \frac{\partial R}{\partial \psi_{tc}} \right); \quad n_{tc} = \sum_{d,w} n_{dw} p_{tcdw} \\ \pi_{cd} = \mathop{\text{norm}}_{c \in C} \left(n_{cd} + \pi_{cd} \frac{\partial R}{\partial \pi_{cd}} \right); \quad n_{cd} = \sum_{w,t} n_{dw} p_{tcdw} \end{array} \right. \end{cases}$$

Автор-тематическая модель (Author-topic model)

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

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

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

- Если множества C_d неизвестны, но Π разрежена:

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

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

Тематическая модель для анализа 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, \dots , n -ки термов разных модальностей.

- **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

- **Данные о пассажирских авиаперелётах:**

(u, a, b, c) — перелёт клиента u из a в b авиакомпанией c

Задача: по наблюдаемой выборке рёбер гиперграфа найти латентные тематические векторные представления его вершин.

Тематическая модель гиперграфа: определения и обозначения

$\Gamma = \langle V, E \rangle$ — ориентированный гиперграф.

$V = V^1 \sqcup \dots \sqcup V^M$ — разбиение вершин по модальностям

M — множество модальностей:

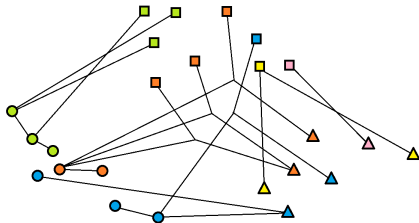
□ ○ △

K — множество типов рёбер:

□○ ○△ ○○△ ○△

T — множество тем:

● ● ● ● ●



E_k — исходные данные: выборка рёбер-транзакций типа k

(d, x) — ребро: вершина-контейнер $d \in V$ и вершины $x \subset V$

n_{kdx} — число вхождений ребра (d, x) в выборку E_k

K. V. Vorontsov. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization // Data Analysis and Optimization, Springer, 2023.

Тематическая модель гиперграфа: основные предположения

- в ребре (d, x) подмножество $x \subset V$ может быть любым, независимо от типа ребра k
- первая *гипотеза условной независимости*: тематика контейнера $p(t|d)$ не зависит от типа ребра k
- вторая *гипотеза условной независимости*: распределение $p(v|t)$ термов v модальности V^m в теме t не зависит ни от контейнера d , ни от типа ребра k
- третья *гипотеза условной независимости*: термы $v \in x$ в ребре (d, x) не зависят друг от друга
- *гипотеза «мешка транзакций»*: выборка транзакций типа k порождается случайно и независимо из

$$p_k(d, x) = p(d) \sum_{t \in T} p(t|d) \prod_{v \in x} p(v|t)$$

Тематическая модель гиперграфа

Вероятностная тематическая модель рёбер типа k

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

Задача максимизации взвешенной суммы log-правдоподобий по всем типам рёбер:

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

при ограничениях неотрицательности и нормировки:

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

где $\tau_k > 0$ — веса типов рёбер.

EM-алгоритм для гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} n_{kdx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in X} \phi_{vt} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

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

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

Доказательство (по лемме о максимизации на симплексах)

Применим Лемму к log-правдоподобию с регуляризатором R :

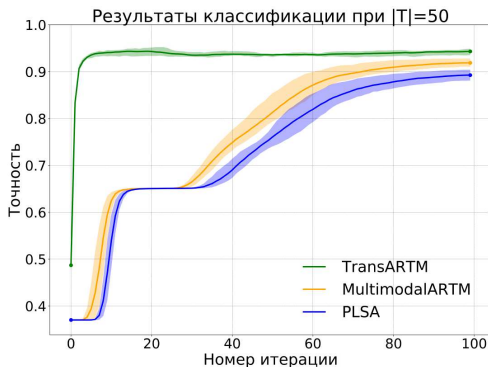
$$\begin{aligned}\phi_{vt} &= \operatorname{norm}_{v \in V_m} \left(\phi_{vt} \sum_{k \in K} \tau_k \sum_{dx \in E_k} n_{kdx} \frac{\theta_{td}}{p(x|d)} \frac{\partial}{\partial \phi_{vt}} \prod_{u \in X} \phi_{ut} + \phi_{vt} \frac{\partial R}{\partial \phi_{vt}} \right) = \\ &= \operatorname{norm}_{v \in V_m} \left(\sum_{k \in K} \sum_{dx \in E_k} \tau_k n_{kdx} [v \in x] p_{tdx} + \phi_{vt} \frac{\partial R}{\partial \phi_{vt}} \right)\end{aligned}$$

$$\begin{aligned}\theta_{td} &= \operatorname{norm}_{t \in T} \left(\theta_{td} \sum_{k \in K} \tau_k \sum_{x \in d} n_{kdx} \frac{1}{p(x|d)} \prod_{v \in X} \phi_{vt} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) = \\ &= \operatorname{norm}_{t \in T} \left(\sum_{k \in K} \sum_{x \in d} \tau_k n_{kdx} p_{tdx} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)\end{aligned}$$

■

Эксперименты на модельных данных

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



Вывод: обычные ТМ восстанавливают гиперграф плохо и долго

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

Транзакционные данные в рекомендательных системах

U — конечное множество (словарь) клиентов (users)

I — конечное множество (словарь) объектов (items)

A — словарь атрибутов клиентов (соцдем, регион, хобби...)

B — словарь свойств объектов (слова в текстовых объектах)

C — словарь ситуативных контекстов

J — словарь интервалов времени

Возможные виды данных:

n_{ui} — клиент u выбрал объект i

n_{ua} — клиент u имеет атрибут a

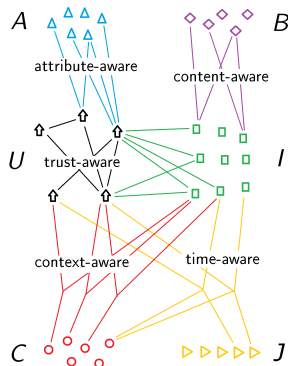
n_{ib} — объект i имеет свойство b

n_{uv} — клиент u доверяет клиенту v

n_{uib} — клиент u отметил i тэгом b

n_{uic} — клиент u выбрал i в контексте c

n_{uicj} — u выбрал i в c в интервале j



Симметризованная модель транзакционных данных

Симметризованные модели подходят для задач, в которых нет «естественных контейнеров» с неизменным содержимым

$x \subset V$ — рёбра гиперграфа

$\phi_{vt} = p(v|t)$ — распределение термов $v \in V^m$ в теме t

$\pi_t = p(t)$ — распределение тем во всей коллекции

E_k — наблюдаемая выборка рёбер-транзакций $x \subset V$ типа k

n_{kx} — число наблюдений ребра x в выборке E_k

Вероятностная тематическая модель рёбер гиперграфа:

$$p(x) = \sum_{t \in T} p(t) \prod_{v \in x} p(v|t) = \sum_{t \in T} \pi_t \prod_{v \in x} \phi_{vt}$$

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k \in K} \tau_k \sum_{x \in E_k} n_{kx} \ln \left(\sum_{t \in T} \pi_t \prod_{v \in x} \phi_{vt} \right) + R(\Phi, \pi) \rightarrow \max_{\Phi, \pi}$$

EM-алгоритм для симметризованной гиперграфовой ARTM

Задача максимизации регуляризованного правдоподобия:

$$\sum_{k \in K} \tau_k \sum_{x \in E_k} n_{kx} \ln \left(\sum_{t \in T} \pi_t \prod_{v \in X} \phi_{vt} \right) + R(\Phi, \pi) \rightarrow \max_{\Phi, \pi}$$

EM-алгоритм: метод простой итерации для системы уравнений со вспомогательными переменными $p_{tx} = p(t|x)$:

$$\begin{cases} \text{E-шаг:} & p_{tx} = \operatorname{norm}_{t \in T} \left(\pi_t \prod_{v \in X} \phi_{vt} \right) \\ \text{M-шаг:} & \begin{cases} \phi_{vt} = \operatorname{norm}_{v \in V^m} \left(\sum_{k \in K} \tau_k \sum_{x \in E_k} [v \in X] n_{kx} p_{tx} + \phi_{vt} \frac{\partial R}{\partial \phi_{vt}} \right) \\ \pi_t = \operatorname{norm}_{t \in T} \left(\sum_{k \in K} \tau_k \sum_{x \in E_k} n_{kx} p_{tx} + \pi_t \frac{\partial R}{\partial \pi_t} \right) \end{cases} \end{cases}$$

Симметризованная модель рекомендательной системы

Сумма log-правдоподобий для четырёх типов транзакций (content-attribute-context-time-aware model):

$$\begin{aligned}
 & \sum_{u,i} n_{ui} \ln \sum_{t \in T} \pi_t \phi_{ut} \phi_{it} \\
 & + \tau_1 \sum_{i,b} n_{ib} \ln \sum_{t \in T} \pi_t \phi_{it} \phi_{bt} \\
 & + \tau_2 \sum_{u,a} n_{ua} \ln \sum_{t \in T} \pi_t \phi_{ut} \phi_{at} \\
 & + \tau_3 \sum_{u,i,c,j} n_{uicj} \ln \sum_{t \in T} \pi_t \phi_{it} \phi_{ut} \phi_{ct} \phi_{jt} \rightarrow \max_{\Phi, \pi}
 \end{aligned}$$

Как построить симметризованную модель в BigARTM:

- 1 документы становятся модальностью
- 2 коллекция разбивается на документы d по времени
- 3 столбцы θ_{td} сглаживаются по n_t или по $\theta_{t,d-1}$

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 :

$$p(s|d) = \sum_{t \in T} p(t|d) \prod_{w \in \mathcal{S}} p(w|t)^{n_{sw}} = \sum_{t \in T} \theta_{td} \prod_{w \in \mathcal{S}} \phi_{wt}^{n_{sw}}$$

Максимизация регуляризованного log-правдоподобия

$$\sum_{d \in D} \sum_{s \in S_d} \ln \sum_{t \in T} \theta_{td} \prod_{w \in \mathcal{S}} \phi_{wt}^{n_{sw}} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

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

Гипер-рёбрами могут быть *сегментоиды* — подмножества термов, связанные по смыслу и порождаемые общей темой:

- предложение / фраза / синтагма
- ветка синтаксического дерева / именная группа
- факт «объект, субъект, действие»
- пары синонимов, гипоним–гипероним, мероним–холоним
- лексическая цепочка
- текст комментария, дата–время, автор

Модель даёт интерпретируемые тематические эмбединги:

- $p(t|d)$ — каждого контейнера, в частности, документа
- $p(t|w) = \phi_{wt} \frac{p(t)}{p(w)}$ — каждого терма, в частности, слова
- $p(t|d, x)$ — каждой отдельной транзакции (фразы, факта)

Анализ транзакций розничных клиентов банка

Дано (Sberbank Data Science Contest):

D — множество клиентов (15 000)

W — категории = MCC-коды (Merchant Category Code) (328)

n_{dw} — сумма транзакций клиента d по категории w

Найти: темы — типы экономического поведения (потребления)

$\phi_{wt} = p(w|t)$ — структура потребления для темы t

$\theta_{td} = p(t|d)$ — типы потребления клиента d

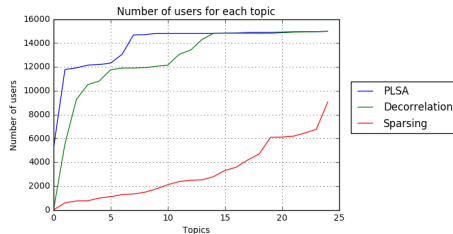
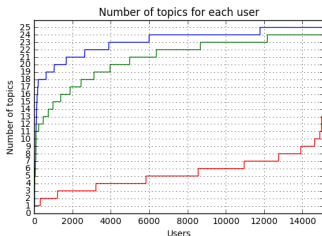
Регуляризаторы:

- повышение различности тем
- разреживание $p(t|d)$
- учёт модальностей времени, типа транзакции, терминала

Egorov E., Nikitin F., Goncharov A., Alekseev V., Vorontsov K. Topic Modelling for Extracting Behavioral Patterns from Transactions Data // IC-AIAI 2019.

Построение модели ARTM, 25 тем

- 30 итераций PLSA — без регуляризаторов
- 10 итераций — повышение различности тем
- 10 итераций — разреживание $p(t|d)$



Декоррелирование Φ и разреживание Θ определяют минимальное число типов экономического поведения каждого клиента, достаточное для описания его расходов.

Пользуюсь картой только чтобы снять наличные

$\phi_{wt}, \%$ МСС-код (категория расходов)

72 Финансовые институты — снятие наличности вручную

27 Финансовые институты — снятие наличности автоматически

0.23 Денежные переводы MasterCard MoneySend

0.1 Денежные переводы

0.012 Финансовые институты — снятие наличности вручную

0.0055 Легковой и грузовой транспорт: продажа, сервис, ремонт, лизинг

0.0027 Магазины игрушек

Наличные + авто, спорт, компьютеры

$\phi_{wt}, \%$ МСС-код (категория расходов)

- 55 Финансовые институты — снятие наличности автоматически
- 44 Денежные переводы
- 0.111 Станции техобслуживания
- 0.105 Автозапчасти и аксессуары
- 0.094 Компьютерная сеть/информационные услуги
- 0.043 Спортивная одежда, одежда для верховой езды и езды на мотоцикле
- 0.024 Финансовые институты — снятие наличности вручную
- 0.020 СТО общего назначения
- 0.018 Горючее топливо — уголь, нефть, разжиженный бензин, дрова
- 0.015 Магазины мужской и женской одежды
- 0.015 Финансовые институты — снятие наличности вручную
- 0.013 Магазины спорттоваров
- 0.012 Садовые принадлежности (в том числе для ухода за газонами) в розницу
- 0.011 Паркинги и гаражи
- 0.011 Бакалейные магазины, супермаркеты
- 0.010 Различные магазины одежды и аксессуаров

Цивилизованный потребитель: разные магазины, связь, авто

$\phi_{wt}, \%$ МСС-код (категория расходов)

- 27 Станции техобслуживания
- 20 Различные продовольственные магазины, рынки, полуфабрикаты
- 15 Звонки с использованием телефонов, считывающих магнитную ленту
- 12 Финансовые институты — снятие наличности автоматически
- 4.7 Горючее топливо — уголь, нефть, разжиженный бензин, дрова
- 4.1 Универсальные магазины
- 3.4 Автозапчасти и аксессуары
- 1.4 Аптеки
- 1.2 Магазины с продажей спиртных напитков на вынос
- 1.1 Бакалейные магазины, супермаркеты
- 0.57 Автошины
- 0.37 Прямой маркетинг — торговля через каталог
- 0.35 Товары для дома
- 0.33 Универмаги
- 0.32 Плавательные бассейны — распродажа
- 0.21 Места общественного питания, рестораны

Всего 24 категории с $\phi_{wt} > 0.1\%$; 61 категория с $\phi_{wt} > 0.01\%$

Продвинутые мамки

$\phi_{wt}, \%$ МСС-код (категория расходов)

- 56 Бакалейные магазины, супермаркеты
- 8.6 Финансовые институты — снятие наличности автоматически
- 5.4 Аптеки
- 4.0 Звонки с использованием телефонов, считывающих магнитную ленту
- 2.2 Рестораны, закусочные
- 1.8 Обувные магазины
- 1.5 Различные продовольственные магазины — рынки, полуфабрикаты
- 1.4 Магазины спорттоваров
- 1.4 Детская одежда, включая одежду для самых маленьких
- 1.3 Магазины игрушек
- 1.3 Места общественного питания, рестораны
- 1.1 Магазины мужской и женской одежды
- 1.1 Магазины с продажей спиртных напитков на вынос
- 1.1 Магазины косметики
- 1.0 Садовые принадлежности в розницу
- 0.73 Одежда для всей семьи

Всего 41 категория с $\phi_{wt} > 0.1\%$; 95 категорий с $\phi_{wt} > 0.01\%$

Бизнес-леди: забыла про наличку — всё по карте

$\phi_{wt}, \%$ МСС-код (категория расходов)

- 12 Магазины мужской и женской одежды
- 7.3 Оборудование, мебель и бытовые принадлежности
- 7.0 Места общественного питания, рестораны
- 5.6 Магазины по продаже часов, ювелирных изделий и изделий из серебра
- 5.3 Обувные магазины
- 4.7 Магазины косметики
- 4.6 Одежда для всей семьи
- 3.8 Универмаги
- 3.2 Готовая женская одежда
- 2.8 Практикующие врачи, медицинские услуги
- 1.8 Прямой маркетинг — торговля через каталог
- 1.5 Салоны красоты и парикмахерские
- 1.3 Детская одежда, включая одежду для самых маленьких
- 1.3 Аптеки
- 1.0 Изготовление и продажа меховых изделий
- 1.0 Центры здоровья

Всего 70 категорий с $\phi_{wt} > 0.1\%$; 134 категории с $\phi_{wt} > 0.01\%$

Продвинутый активный потребитель всего, и по карте

$\phi_{wt}, \%$ МСС-код (категория расходов)

20 Финансовые институты — снятие наличности вручную

15 Универсальные магазины

13 Туристические агентства и организаторы экскурсий

11 Автозапчасти и аксессуары

8.8 Коммунальные услуги — электричество, газ, санитария, вода

4.2 Веломагазины — продажа и обслуживание

3.7 СТО общего назначения

0.9 Услуги курьера — по воздуху и на земле, агентство по отправке грузов

0.8 Рекламные услуги

0.7 Компьютеры, периферия, программное обеспечение

0.5 Образовательные услуги

0.4 Бакалейные магазины, супермаркеты

0.4 Практикующие врачи, медицинские услуги

0.3 Продажа мотоциклов

0.3 Оборудование, мебель и бытовые принадлежности

0.2 Автошины

Всего 35 категорий с $\phi_{wt} > 0.1\%$; 93 категории с $\phi_{wt} > 0.01\%$

Бизнес-класс: авиа, отели, казино, рестораны, ценные бумаги

$\phi_{wt}, \%$ МСС-код (категория расходов)

- 28 Авиа линии, авиакомпании
- 19 Финансовые институты — торговля и услуги
- 9.5 Отели, мотели, базы отдыха, сервисы бронирования
- 8.6 Транзакции по азартным играм (плюс)
- 5.2 Финансовые институты — торговля и услуги
- 3.2 Места общественного питания, рестораны
- 3.1 Не-финансовые институты: ин.валюта, переводы, дорожн.чеки, квази-кэш
- 2.2 Пассажирские железнодорожные перевозки
- 1.7 Бизнес-сервис
- 1.4 Жилье — отели, мотели, курорты
- 1.3 Галереи/учреждения видеоигр
- 1.3 Транзакции по азартным играм (минус)
- 0.6 Ценные бумаги: брокеры/дилеры
- 0.5 Туристические агентства и организаторы экскурсий
- 0.3 Лимузины и такси
- 0.3 Беспшлинные магазины Duty Free

Всего 50 категорий с $\phi_{wt} > 0.1\%$; 103 категории с $\phi_{wt} > 0.01\%$

Провинциальный малый бизнес

$\phi_{wt}, \%$ МСС-код (категория расходов)

- 27 Финансовые институты — снятие наличности автоматически
- 8.5 Лесо- и строительный материал
- 8.4 Бытовое оборудование
- 6.6 Плавательные бассейны — распродажа
- 5.5 Продажа электронного оборудования
- 4.1 Бакалейные магазины, супермаркеты
- 3.3 Универсальные магазины
- 3.0 Садовые принадлежности в розницу
- 2.6 Телекоммуникационное оборудование, включая продажу телефонов
- 2.4 Легковой и грузовой транспорт: продажа, сервис, ремонт, лизинг
- 2.2 Товары для дома
- 2.1 Пассажирские железнодорожные перевозки
- 1.5 Оборудование, мебель и бытовые принадлежности
- 1.3 Скобяные товары в розницу
- 1.2 Магазины спорттоваров
- 1.1 Аптеки

Всего 54 категории с $\phi_{wt} > 0.1\%$; 104 категории с $\phi_{wt} > 0.01\%$

Анализ транзакций корпоративных клиентов банка

Данные:

лесная отрасль, 2016 г., 10.7М транзакций, 1М компаний.

Транзакция — это тройка (покупатель, продавец, текст).

Некоторые *тексты* платёжных поручений (далеко не все!) содержат названия товаров и услуг.

Документ — это история транзакций одной компании

Семь модальностей:

- компании: поставщики / покупатели
- слова в платёжных поручениях: поставщики / покупатели
- ОКВЭДы данной компании
- ОКВЭДы контрагентов: поставщики / покупатели

Примеры тем — видов деятельности компаний

| покупка | продажа |
|--------------------|--------------------|
| 0.11: услуга | 0.12: лдсп |
| 0.07: классик | 0.08: дсп |
| 0.05: дрова | 0.03: мдф |
| 0.05: пиловочник | 0.03: поставка |
| 0.05: материал | 0.02: услуга |
| 0.03: порода | 0.02: охранный |
| 0.03: лесоматериал | 0.02: ламинировать |
| 0.03: сертум | 0.02: хдф |
| 0.02: хвойный | 0.02: материал |
| 0.01: дерево | 0.01: накл |
| 0.01: транспортный | 0.01: товар |

| покупка | продажа |
|---------------------|---------------------|
| 0.19: право | 0.16: арендный |
| 0.17: сбис | 0.10: часть |
| 0.16: использование | 0.08: плата |
| 0.03: аккаунт | 0.04: минимальный |
| 0.02: электронный | 0.04: участок |
| 0.02: лицевой | 0.04: использование |
| 0.02: устный | 0.02: земля |
| 0.01: устройство | 0.02: лесничество |
| 0.01: генерация | 0.02: земельный |
| 0.01: хранение | 0.01: фонд |
| 0.01: ключевой | 0.01: федеральный |

Примеры тем — видов деятельности компаний

| покупка | продажа |
|--------------------|---------------------|
| 0.09: ткань | 0.16: мебель |
| 0.09: поставка | 0.05: плёнка |
| 0.02: мебельный | 0.04: стул |
| 0.02: деревянный | 0.03: кресло |
| 0.02: транспортный | 0.03: изделие |
| 0.02: фанера | 0.02: краска |
| 0.02: поролон | 0.02: фанера |
| 0.01: механизм | 0.01: лкм |
| 0.01: плата | 0.01: лакокрасочный |
| 0.01: частичный | 0.01: лак |
| | 0.01: материал |
| | 0.01: клеить |

| покупка | продажа |
|--------------------|-----------------|
| 0.06: лдсп | 0.37: товар |
| 0.05: фурнитура | 0.15: мебель |
| 0.02: плёнка | 0.04: поставка |
| 0.02: материал | 0.04: накладный |
| 0.02: мебельный | 0.03: накл |
| 0.02: стекло | 0.03: рубль |
| 0.02: мдф | |
| 0.02: кромка | |
| 0.01: транспортный | |
| 0.01: клеить | |
| 0.01: профиль | |
| 0.01: пвх | |

Примеры тем — видов деятельности компаний

| покупка | продажа |
|-------------|------------------|
| 0.52: гсм | 0.14: вывоз |
| 0.43: далее | 0.09: тбо |
| | 0.04: мусор |
| | 0.03: отход |
| | 0.02: утилизация |
| | 0.01: тко |

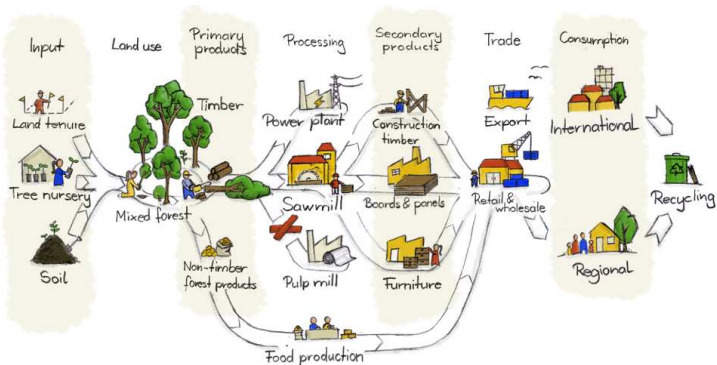
| покупка | продажа |
|---------------------|----------------------|
| 0.19: налог | 0.11: бумага |
| 0.06: услуга | 0.08: гофроящик |
| 0.04: макулатура | 0.04: гофрокартон |
| 0.03: поставка | 0.03: гофрокороб |
| 0.03: транспортный | 0.03: поставка |
| 0.02: лесопродукция | 0.03: фактура |
| 0.02: автоуслуга | 0.02: гофропродукция |
| 0.01: перевозка | 0.02: гофротару |
| 0.01: плата | 0.02: гофрирование |
| | 0.02: гофролоток |
| | 0.02: товар |
| | 0.01: лоток |

Примеры тем — видов деятельности компаний

| | | | |
|---------------------|------------------|------------------|------------------|
| покупка | продажа | продажа | продажа |
| 0.15: программа | 0.13: фурнитура | 0.14: рекламный | 0.21: тмц |
| 0.11: право | 0.09: материал | 0.13: размещение | 0.06: накл |
| 0.09: сертификат | 0.08: лдсп | 0.09: материал | 0.04: инструмент |
| 0.07: эвм | 0.04: кромка | 0.05: проект | 0.03: пила |
| 0.07: использование | 0.04: мебельный | 0.05: яндекс | 0.02: заточка |
| 0.07: лицензия | 0.04: фрз | 0.04: директ | 0.02: нож |
| 0.04: криптопро | 0.04: мдф | 0.04: реклама | 0.02: материал |
| 0.03: абонентский | 0.03: клеить | 0.02: рубль | 0.02: фреза |
| 0.02: обслужа | 0.03: пвх | 0.01: стек | 0.02: клеить |
| 0.02: пользование | 0.02: тмц | | 0.01: товар |
| 0.02: контур | 0.02: комплект | | 0.01: перчатка |
| 0.01: проверка | 0.02: профиль | | |
| | 0.02: столешница | | |

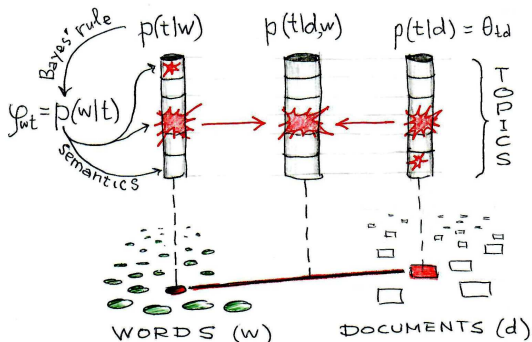
Конечные цели моделирования транзакционных данных

- Получение векторных представлений компаний
- Поиск схожих и конкурирующих компаний
- Восстановление структуры товарных потоков отрасли



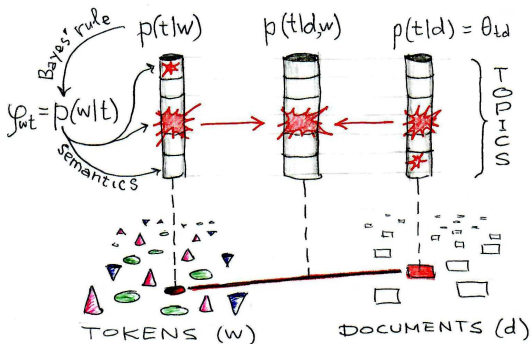
Интерпретируемые тематические эмбединги слов и документов

- Коллекция текстов — двудольный граф с рёбрами (d, w)
- Тематические эмбединги: $p(t|d)$, $p(t|w)$, $p(t|d, w)$
- Интерпретируемость тем через частоты термов $p(w|t)$
- Слово w встречается в d , когда у них есть общие темы



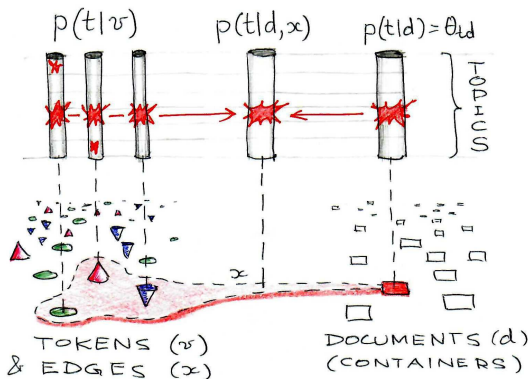
Интерпретируемые эмбединги мультимодальных документов

- Документы содержат термины различных *модальностей*
- Примеры: слова, n -граммы, теги, категории, авторы, ...
- Через темы смыслы слов передаются другим модальностям
- Терм w встречается в d , когда у них есть общие темы



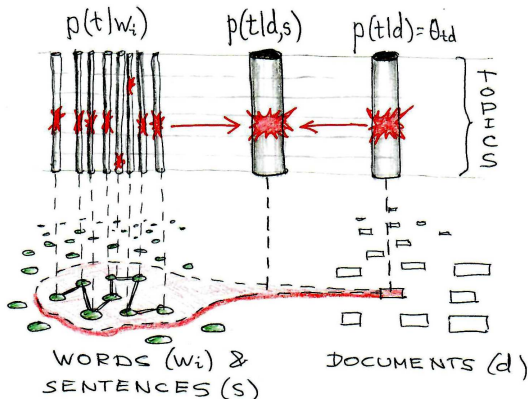
Интерпретируемые эмбединги транзакционных данных

- Транзакция — взаимодействие двух или более термов
- Примеры: $\langle \text{buyer, seller, item} \rangle$, $\langle \text{user, site, banner} \rangle$
- Транзакционные данные — это выборка рёбер гиперграфа
- Транзакция происходит, когда её термы имеют общие темы



Интерпретируемые эмбединги предложений

- Ребро гиперграфа — семантически связанные слова
- Примеры: предложение, синтагма, лексическая цепочка, именная группа, факт «объект, субъект, действие»
- Слова связываются, когда они имеют общие темы



- Контексты, модальности, транзакции — существенно расширяют спектр приложений тематических моделей
- Особенности данных, определяющие структуру модели:
 - данная модальность имеет локальные контексты или относится к документу (контейнеру) целиком?
 - имеет ли документ (контейнер) неизменное содержимое или пополняется динамически?
 - что порождается одной и той же темой — контекст? предложение? абзац? сообщение? транзакция?
 - чем образуется контекст — последовательностью токенов во времени? источником? автором?
- Не лучше ли перейти от $p(w|t)$ к эмбедингам $p(t|w)$? (в модели локальных контекстов это тоже напрашивается) Теоретически эти варианты эквивалентны... а вычислительно?
- Не попробовать ли разреженные тематические эмбединги на выходе свёрточных слоёв при обработке изображений?

Задача-минимум: научиться решать задачи NLP с использованием тематического моделирования в BigARTM

Задача-максимум: сделать полезное мини-исследование

| виды деятельности | оценка |
|-------------------------------|--------------|
| теоретические задания | $\sum_i X_i$ |
| решение прикладной задачи | 5X |
| обзор по последним NeuralTM | 5X |
| интеграция ARTM в pyTorch | 5X |
| участие в одном из проектов | 10X |
| работа над открытой проблемой | 10X |

где X — оценка за вид деятельности по 5-балльной шкале.
score — суммарная оценка по всем видам деятельности.

Итоговая оценка: $\min(10, \lfloor \text{score}/5 \rfloor)$ по 10-балльной шкале.

Теоретическое задание к лекции 1

Упражнения на принцип максимума правдоподобия:

1. Униграммная модель документов: $p(w|d) = \xi_{dw}$

Найти параметры модели ξ_{dw} .

2. Униграммная модель коллекции: $p(w|d) = \xi_w$ для всех d

Найти параметры модели ξ_w .

Подсказка: применить условия ККТ или основную лемму.

3. Творческое задание (возможны разные решения)

Предложите модель, определяющую роли слов в текстах:

— тематические слова

— специфичные слова документа (шум)

— слова общей лексики (фон)

Подсказка 1: искать распределение ролей слов $p(r|w)$, $r \in \{\text{т, ш, ф}\}$.

Подсказка 2: можно разреживать $p(r|w)$ для жёсткого определения ролей.

Подсказка 3: можно использовать документную частоту слов.

Теоретическое задание к лекции 2

4. Запишите критерий логарифма правдоподобия с регуляризацией для тематической модели $p(w|d) = \sum_t \phi_{wt} \theta_{td}$, используя исходные данные $(d_i, w_i)_{i=1}^n$ вместо счётчиков n_{dw} . Выведете из него EM-алгоритм, докажите его эквивалентность обычному EM-алгоритму для ARTM.

5. Запишите критерий логарифма правдоподобия для локализованной тематической модели $p(w|C_i) = \sum_t \phi_{wt} p(t|C_i)$. Выведете из него EM-алгоритм с локализованным E-шагом.

Какие приближения пришлось сделать в процессе вывода?

Какие переменные удобнее оставить в модели, ϕ_{wt} или ϕ'_{tw} ?

6. Творческое задание (возможны разные решения)

Предложите «какую-нибудь разумную» параметризацию для тематической модели внимания. Используя «основную лемму», получите уравнения для новых параметров модели.

Открытая проблема. Продолжить исследование Ильи Ирхина:

- Освоить код: https://github.com/ilirhin/python_artm
- Реализовать локализованный E-шаг

Исследовать зависимость метрик качества от параметров (перплексия, разреженность, различность, когерентность):

- L — число проходов
- $\vec{\gamma}_i, \overleftarrow{\gamma}_i$ — длина скользящего среднего
- $\vec{\gamma}_i, \overleftarrow{\gamma}_i$ — асимметричность левого и правого контекста
- $\vec{\gamma}_i, \overleftarrow{\gamma}_i$ — учёт границ предложений, абзацев, глав
- β — баланса левого и правого контекста
- α, δ — параметры онлайн-алгоритма EM
- опция «подставлять p_{ti}/n_t вместо $\phi_{w_i t}$ на E-шаге»
- опция «исключать p_{ti} позиции i из контекстов $\vec{\theta}_{ti}, \overleftarrow{\theta}_{ti}$ »

Теоретическое задание к лекции 3

7. Выведите формулы EM-алгоритма в случае, когда логарифм в функции потерь заменяется гладкой монотонно возрастающей функцией ℓ :

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ell \left(\sum_{t \in T} \phi_{wt} \theta_{td} \right) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

Подумайте, какие замены логарифма полезны, и почему.

8. Замените \ln гладкой монотонно возрастающей функцией μ в регуляризаторе сглаживания–разреживания (модель LDA):

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

Как изменится M-шаг и воздействие регуляризатора на модель?

9. Какому регуляризатору соответствует формула M-шага

$$\phi_{wt} = \underset{w}{\text{norm}}(n_{wt} [n_{wt} > \gamma n_t])$$

Теоретическое задание к лекции 3

Аналитик построил тематическую модель Φ^0, Θ^0 и отметил среди столбцов матрицы Φ^0 темы двух типов: удачные $T_+ \subset T$ и неудачные $T_- \subset T$.

Теперь он хочет построить модель ещё раз так, чтобы

- удачные темы остались в матрице Φ ;
- остальные темы построились по-другому и были не похожи на каждую из неудачных тем $t \in T_-$.

10. Предложите регуляризаторы для этого.

11. Не получится ли так, что новые темы будут отдаляться от суммы неудачных тем $\sum_{t \in T_-} \phi_{wt}^0$ вместо того, чтобы отдаляться от каждой из неудачных тем по отдельности? Почему это плохо и как этого избежать?

12. Предложите способ инициализации Φ для новой модели.

- Проблема несбалансированности тем
 - генераторы синтетических несбалансированных коллекций
 - модели локального контекста лишены этой проблемы?
 - регуляризаторы декоррелирования + семантической однородности
- Семейство средневзвешенных статистик
 - генераторы синтетических коллекций, удовлетворяющих гипотезе условной независимости
 - как (и нужно ли) определять пороги для построения статистических тестов условной независимости?
 - как ослабить проверку гипотезы условной независимости в модели локального контекста?
 - как перестраивать несогласованные темы?
- Критерий внутритекстовой когерентности
 - найти лучший вариант критерия с помощью калибровки по размеченным тематическим цепочкам
 - вычисление критерия должно естественным образом встраиваться в модель локального контекста

13. Для иерархической тематической модели с рег. $R(\Phi, \Psi)$ предложите способ разреживания матрицы связей $\Psi = (p(s|t))$, гарантирующий, что

- 1) у каждой родительской темы будет хотя бы одна дочерняя;
- 2) у каждой дочерней темы будет хотя бы одна родительская.

Подсказка: можно придумывать критерий регуляризации, а можно — формулу М-шага для матрицы Ψ .

14. Предложите способ гарантировать, что если родительская тема t получает только одну дочернюю s , то она переходит в неё целиком и как распределение: $p(w|s) = p(w|t)$, то есть тема t на данном уровне не расщепляется на подтемы.

15. Предложите способ согласования вероятностных смесей $p(w|t) \approx \sum_{s \in S} p(w|s)p(s|t)$ и $p(t|d) \approx \sum_{s \in S} p(t|s)p(s|d)$ с учётом тождества $p(s|t)p(t) = p(t|s)p(s)$.

Участие в проекте «Мастерская знаний»

Дано:

- подборки, сгенерированные SciRus по одной статье
- ассессорская разметка статей подборки по релевантности
- несколько вариантов токенизации
 - в том числе с автоматическим выделением терминов

Найти:

- тематическую модель
- модель ранжирования подборки по релевантности
- оптимальные: токенизацию, число тем, регуляризаторы
- распределение терминов по тематичности

Критерий:

- качество ранжирования
- (визуально) интерпретируемость тем
 - в том числе автоматического именованя тем

- Открытые датасеты (английский): 20 newsgroups, NIPS, KOS
- Научные статьи: eLibrary, Semantic Scholar, arXiv, PubMed
- Научно-популярные статьи: ПостНаука, Элементы, Хабр,...
- TechCrunch (английский)
- Данные социальных сетей: VK, Twitter, Telegram,...
- Википедия
- Новостной поток (20 источников на русском языке)
- Данные кадровых агентств: резюме + вакансии
- Транзакции клиентов Sberbank DSD 2016
- Акты арбитражных судов РФ

- «Мастерская знаний» для научного поиска
 - пользователь строит тематические подборки статей,
 - поисковая выдача формируется моделью SciRus.
 - задача: показать пользователю тематику подборки
 - понадобится автоматическое выделение терминов,
 - выделение тематических фраз из документов,
 - автоматическое именование и суммаризация тем
 - конечная цель: ускорить понимание предметной области
- «Тематизатор» для социо-гуманитарных исследований
 - пользователь задаёт грубый фильтр текстового потока
 - задача: «классифицировать иголки в стоге сена»,
 - разделив темы на информативные и мусорные,
 - выделив аспекты и тональности в каждой теме
 - конечная цель: q&q аналитика проблемной среды

- 1 Проблема несбалансированности тем в коллекции
- 2 Обеспечение 100%-й интерпретируемости тем
- 3 Тематические модели внимания последовательного текста
- 4 Обнаружение новых тем или трендов в потоке текстов
- 5 Автоматическое именованное и аннотирование тем
- 6 Обзор подходов в нейросетевых тематических моделях
- 7 Обеспечение полноты и устойчивости множества тем
- 8 Автоматический подбор гиперпараметров, AutoML
- 9 Оптимизация гиперпараметров в потоковом режиме
- 10 Проблема несбалансированности текстов по длине
- 11 Бережное слияние моделей нескольких коллекций
- 12 Гиперграфовые тематические модели в RecSys