Многомодальные тематические модели на гиперграфах

Жариков Илья Николаевич

zharikov.i.n@yandex.ru

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель: д.ф.-м.н. К. В. Воронцов

Июнь, 2018

Цель исследования

Транзакция — взаимодействие некоторого множества объектов.

Проблема

Большинство существующих тематических моделей описывают только попарные взаимодействия между объектами разных типов (модальностей).

Предлагается

Обобщить методы тематического моделирования на случай транзакционных данных, в которых взаимодействуют более двух объектов.

Гиперграф

Примеры транзакций

- **Данные социальной сети**
  (u, w, d) — пользователь u написал слово w в блоге d
- **Данные сети интернет-рекламы**
  (u, b, d) — 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 b купил товар g у продавца s

Транзакция $\leftrightarrow$ определенное ребро.

Данные $\leftrightarrow$ множество ребер **гиперграфа**.

Гиперграф — обобщение графа, в котором ребром могут соединяться не только две вершины, но и любое подмножество вершин.

Пример

Транзакционные данные — наблюдаемые ребра гиперграфа

Множество модальностей M:

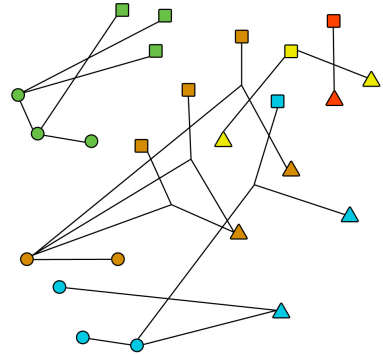
□ ○ △

Множество типов рёбер K:

d	□	□	○	△	□
x	○	△	○	○	△
n_{dx}	3	3	4	2	4

n_{dx} — число ребер (d, x) в гиперграфе

Множество тем T:



Задача: найти тематические вектора **всех** вершин

Постановка задачи

Для оптимизации параметров применим принцип максимума правдоподобия:

$$\sum_{k\in K}\tau_k \sum_{(d,x)\in E_k} n_{dx} \ln \sum_{t\in T} \theta_{td} \prod_{v\in x} \varphi_{vtk} + \underbrace{R(\Phi,\Theta)}_{\text{регуляризатор}} \to \max_{\Phi,\Theta},$$

где τ_k — вес ребёр типа k, $\theta_{td} = p(t \mid d)$, $\varphi_{vtk} = p_k(v \mid t)$.

Ограничения:

$$\sum_{v\in V_m} \varphi_{vtk} = 1,\ \varphi_{vtk} \geqslant 0$$

$$\sum_{t\in T} \theta_{td} = 1,\ \theta_{td} \geqslant 0$$

d — документ
v — вершина
m — модальность
k — тип ребра
t — тема

EM алгоритм

E step. Вычисление распределения тем для каждого (d, x):

$$p_{tdx} = \underset{t \in T}{\operatorname{norm}} \left(\theta_{td} \prod_{v \in x} \varphi_{vtk} \right), \text{ где } \underset{i \in I}{\operatorname{norm}} a_i = \frac{\max\{a_i, 0\}}{\sum_{j \in I} \max\{a_i, 0\}}.$$

M step. Оценивание параметров модели:

$$\varphi_{vtk} = \underset{v \in V_m}{\operatorname{norm}} \left(n_{vtk} + \varphi_{vtk} \frac{\partial R}{\partial \varphi_{vtk}} \right); \qquad n_{vtk} = \sum_{(d,x) \in E_k} [v \in x] \tau_k n_{dx} p_{tdx};$$

$$\theta_{td} = \underset{t \in T}{\operatorname{norm}} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); \qquad n_{td} = \sum_{k \in K} \sum_{(d,x) \in E_k} \tau_k n_{dx} p_{tdx};$$

Теорема

Если функция $R(\Phi, \Theta)$ непрерывно дифференцируема и (Φ, Θ) — точка локального максимума рассматриваемой задачи, то выполняется система уравнений, описанная выше, относительно параметров φ_{vtk}, θ_{td} и вспомогательных переменных p_{tdx}, n_{td} и n_{vtk}.

Эксперимент

Цели эксперимента:

1. Проверить, способен ли алгоритм восстановить параметры модели, с помощью которой были порождены данные.
2. Оценить устойчивость модели относительно инициализации и выбора числа тем.

Постановка эксперимента:

1. Вычисление матрицы Θ с использованием различных методов:
   - PLSA;
   - MultiARTM;
   - TransARTM.
2. Решение задачи классификации документов с использованием распределения тем $p(t \mid d)$ в качестве признаков.
3. Оценка качества восстановления матрицы Θ путем вычисления точности решения задачи классификации документов:

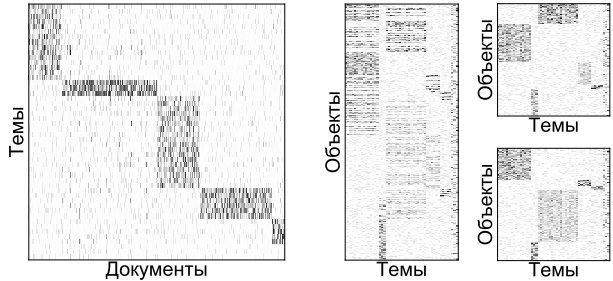
$$\text{Точность} = \frac{1}{N} \sum_{i=1}^{N} \left[y_i^{\text{pred}} = y_i^{\text{true}} \right]$$

Модельные данные:

1. Генерация матриц $\Theta = p(t \mid d)$ и $\Phi_k = p_k(v \mid t)$ для всех $k \in K$.

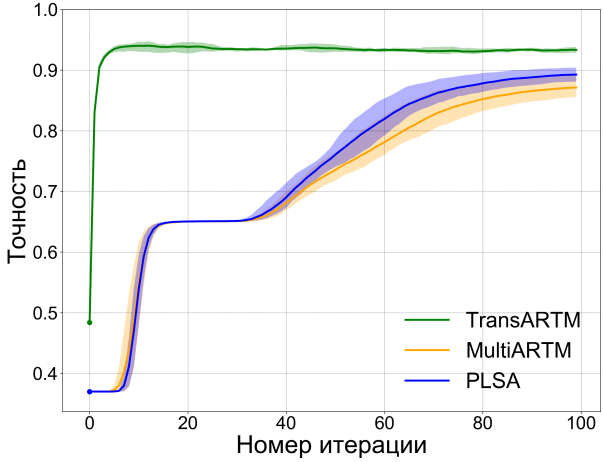
   Число тем: 50
   Число классов: 5
   Число документов: 5000
   Число объектов: 1000
   Число типов ребер: 9
   Число модальностей: 3

   $\Theta \in \mathbb{R}^{50 \times 5000}$
   $\Phi_k \in \mathbb{R}^{1000 \times 50}$



2. Генерация данных (транзакций) на основе матриц Θ, Φ_k.
   Общее число транзакций: $\sim 13\,500\,000$.

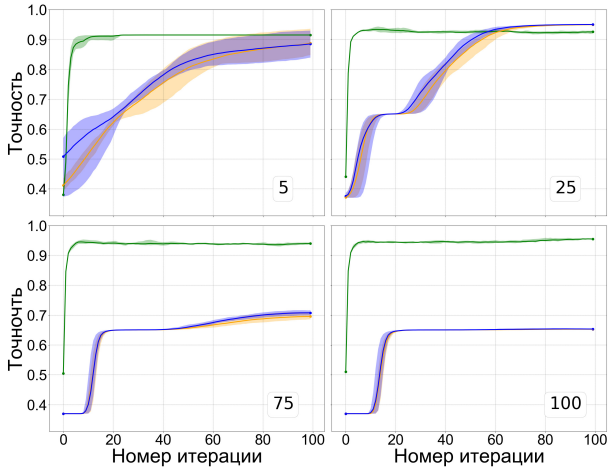
Результаты



Число тем совпадает с заданным при генерации

Вывод:
TransARTM достигает высокого качества быстрее других моделей на транзакционных данных.

Результаты

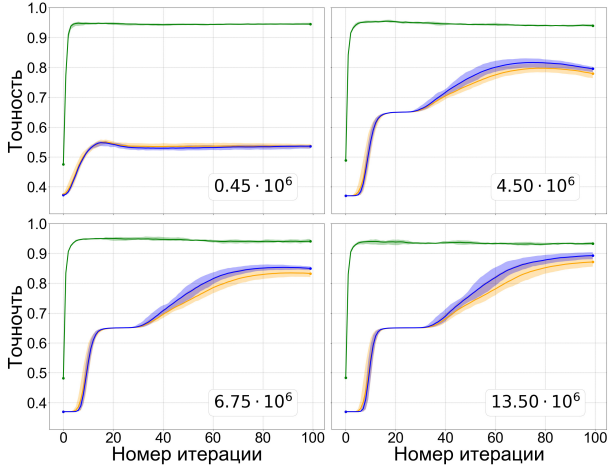


Меняем число тем от 5 до 100

Число тем в исходной матрице Θ равно 50.

Вывод:
TransARTM наиболее устойчива относительно инициализации и выбора числа тем.

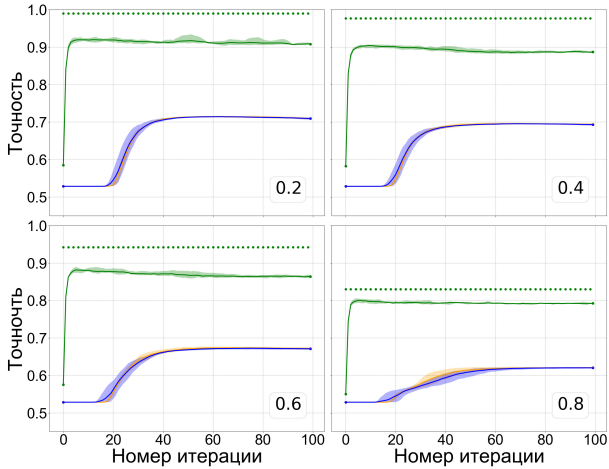
Результаты



Меняем число транзакций от 450 000 до 13 500 000

Вывод:
TransARTM восстанавливает изначальную структуры матрицы Θ с высоким качеством даже при небольшом числе данных.

Результаты



Меняем разреженность исходной матрицы Θ от 0.2 до 0.8

Вывод:
TransARTM показывает близкое к максимальному качество вне зависимости от разреженности матрицы Θ.

The Million Playlist Dataset (MPD)

$$\text{MPD} \begin{cases} \text{плейлист}_1 \begin{cases} \text{трек}_1, \text{альбом}_1, \text{исполнитель}_1 \\ \text{трек}_2, \text{альбом}_2, \text{исполнитель}_2 \\ \dots \end{cases} \\ \text{плейлист}_2 \\ \dots \\ \text{плейлист}_{1\,000\,000} \end{cases}$$

Цель:
Построить рекомендательную систему треков пользователям

	MPD	Train	Test	Valid
Число плейлистов:	1 000 000	100 000	1 000	1 000
Число треков:	66 346 428	9 875 306	232 613	232 808
Число уникальных треков:	2 262 292	296 882	39 368	38 641
Число уникальных альбомов:	734 684	140 983	20 690	20 483
Число уникальных артистов:	295 860	69 280	10 081	10 008

Ссылка на данные: https://recsys-challenge.spotify.com/dataset.

Эксперимент

Цели эксперимента:

1. Применить TransARTM к задаче построения рекомендаций.
2. Проанализировать различные гипотезы порождения плейлиста.
3. Сравнить результаты с другими моделями.

Постановка эксперимента:

1. Для нахождения параметров моделей используем обучающую выборку, состоящую из 100 000 плейлистов, каждый из которых содержит от 140 до 250 треков.
2. Настраиваем коэффициенты регуляризации по сетке на валидационной выборке.
3. Предсказываем ранжированный список из 500 треков для каждого плейлиста тестовой выборки (последние 70 каждого плейлиста используются для оценки качества).
4. Используем следующие метрики для оценки качества:
   - **precision**;
   - **recall**;
   - **fscore**;
   - **ndcg** — normalized discounted cumulative gain.

Результаты

Модель	Рассматриваемые взаимодействия	Метрики, @500 precision	recall	fscore	ndcg
TopTracks	-	0.0230	0.1646	0.0404	0.1152
PLSA	(Pl, Tr)	0.0592	0.4228	0.1038	0.3025
LDA	(Pl, Tr)	0.0583	0.4162	0.1022	0.2988
MultiARTM	(Pl, Al), (Pl, Tr)	0.0594	0.4245	0.1043	0.3029
	(Pl, Ar), (Pl, Tr)	**0.0608**	**0.4343**	**0.1067**	**0.3110**
	(Pl, Ar), (Pl, Al), (Pl, Tr)	0.0605	0.4321	0.1061	0.3098
TransARTM	(Pl, Al, Tr)	0.0490	0.3497	0.0859	0.2484
	(Pl, Ar, Tr)	**0.0504**	**0.3603**	**0.0885**	**0.2555**
	(Pl, Al, Tr), (Pl, Ar, Tr)	0.0502	0.3587	0.0879	0.2548
	(Pl, Ar, Al, Tr)	0.0476	0.3398	0.0835	0.2374

Playlist
Artist
Album
Track

Вывод: TransARTM показывает сравнимые результаты.

Интерпретируемость

Представление трех различных в виде 10 наиболее вероятных артистов в каждой теме (в убывающем порядке):

1	Nicki Minaj	Lil Jon	The Beatles
2	Beyonce	50 Cent	John Lennon
3	Rihanna	Snoop Dogg	George Harrison
4	Tinashe	J-Kwon	The Beach Boys
5	Omarion	Nelly	Elvis Presley
6	Jeremih	Usher	Paul McCartney
7	Trey Songz	Kanye West	David Bowie
8	Chris Brown	R. Kelly	Jim Sturgess
9	Big Sean	Youngbloodz	The Mamas & The Papas
10	Sage The Gemini	Bubba Sparxxx	The Turtles

Результаты

1. Предложено обобщение методов тематического моделирования на случай, когда исходные данные представимы в виде гиперграфа.
2. Проведены эксперименты на модельных данных, демонстрирующие корректность предложенного метода и преимущество его использования для сложноструктурированных данных.
3. Продемонстрировано применение гиперграфовой многомодальной тематической модели для построения рекомендательной системы.

Доклады:

1. Международная научная конференция «Ломоносов-2018», «Многомодальные тематические модели на гиперграфах».
2. Data Fest[5], «Гиперграфовые тематические модели для анализа транзакционных данных».