

# 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**.

## Транзакционные данные — наблюдаемые ребра гиперграфа

Множество модальностей  $M$ :

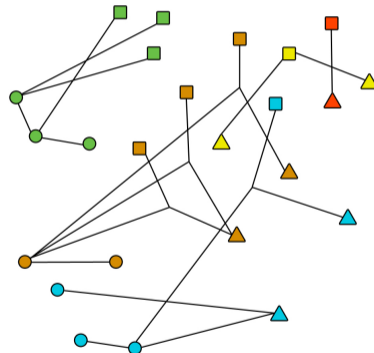
□ ○ △

Множество типов рёбер  $K$ :

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

 $n_{dx}$  — число рёбер  $(d, x)$  в гиперграфеМножество тем  $T$ :

■ ■ ■ ■ ■



Задача: найти тематические вектора всех вершин

## Постановка задачи

Для оптимизации параметров применим принцип максимума правдоподобия:

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

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

Ограничения:

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

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

$d$  — документ

$v$  — вершина

$m$  — модальность

$k$  — тип ребра

$t$  — тема

## EM алгоритм

**E шаг.** Вычисление распределения тем для каждого  $(d, x)$ :

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

**M шаг.** Оценивание параметров модели:

$$\varphi_{kvtk} = \mathop{\text{norm}}_{v \in V_m} \left( n_{vtk} + \varphi_{vtk} \frac{\partial R}{\partial \varphi_{vtk}} \right); \quad n_{vtk} = \sum_{(d,x) \in E_k} [v \in x] \tau_k n_{dx} p_{ktdx};$$

$$\theta_{td} = \mathop{\text{norm}}_{t \in T} \left( n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); \quad n_{td} = \sum_{k \in K} \sum_{(d,x) \in E_k} \tau_k n_{dx} p_{ktdx};$$

## Теорема

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

## Цели эксперимента:

- 1 Проверить, способен ли алгоритм восстановить параметры модели, с помощью которой были порождены данные.
- 2 Оценить устойчивость модели относительно инициализации, выбора числа тем, размера данных и разреженности исходной матрицы  $\Theta$ .

## Постановка эксперимента:

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

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

## Модельные данные:

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

Число тем: 50

Число классов: 5

Число документов: 5000

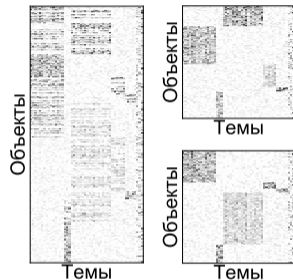
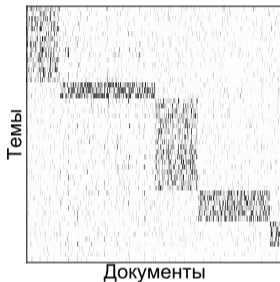
Число объектов: 1000

Число типов ребер: 9

Число модальностей: 3

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

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

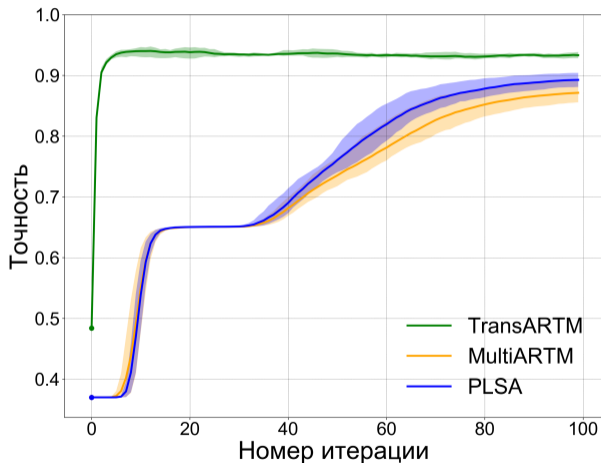


- 2 Генерация данных (транзакций) на основе матриц  $\Theta$ ,  $\Phi_k$ .  
Общее число транзакций:  $\sim 13\,500\,000$ .



## Результаты

Число тем совпадает с заданным при генерации



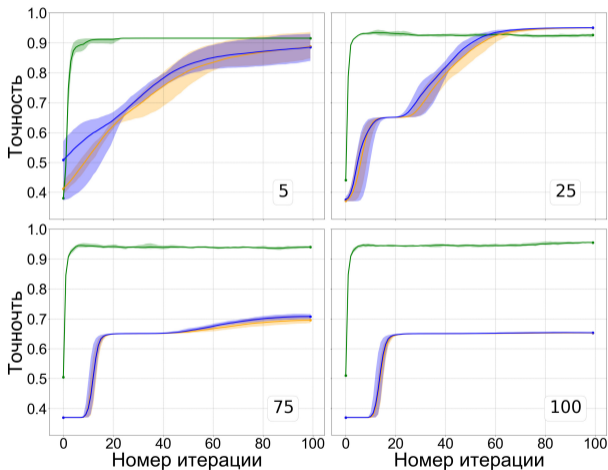
### Вывод:

TransARTM достигает высокого качества быстрее других моделей на транзакционных данных.

## Результаты

Меняем число тем  
от 5 до 100

Число тем в исходной  
матрице  $\Theta$  равно 50.

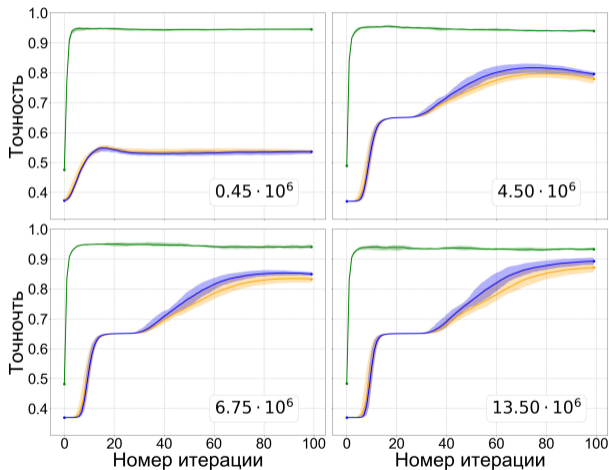


### Вывод:

TransARTM наиболее  
устойчива относительно  
инициализации и выбора  
числа тем.

## Результаты

Меняем число  
транзакций  
от 450 000 до 13 500 000



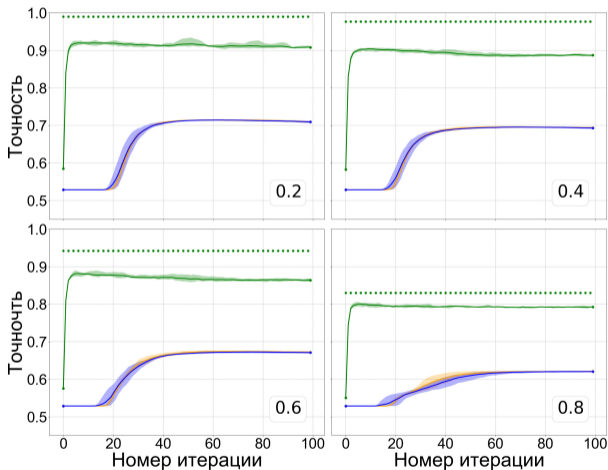
### Вывод:

TransARTM

восстанавливает  
изначальную структуру  
матрицы  $\Theta$  с высоким  
качеством даже при  
небольшом числе данных.

## Результаты

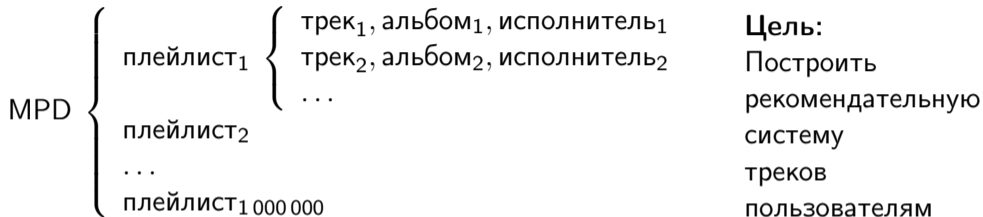
Меняем разреженность  
исходной матрицы  $\Theta$   
от 0.2 до 0.8



### Вывод:

TransARTM показывает  
близкое к максимальному  
качество вне зависимости  
от разреженности  
матрицы  $\Theta$ .

# The Million Playlist Dataset (MPD)



	<b>MPD</b>	<b>Train</b>	<b>Test</b>	<b>Valid</b>
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

## 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.

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
	<b>(Pl, Ar), (Pl, Tr)</b>	<b>0.0608</b>	<b>0.4343</b>	<b>0.1067</b>	<b>0.3110</b>
	(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
	<b>(Pl, Ar, Tr)</b>	<b>0.0504</b>	<b>0.3603</b>	<b>0.0885</b>	<b>0.2555</b>
	(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

Pl aylist  
Ar tist  
Al bum  
Tr ack

**Вывод:** TransARTM не выигрывает рассматриваемые тематические модели на данных, которые могут быть адекватно описаны парными взаимодействиями.

## Интерпретируемость

Представление трех различных тем 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<sup>5</sup>, «Гиперграфовые тематические модели для анализа транзакционных данных».