TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

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

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

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.

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

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.

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Пример

Транзакционные данные — наблюдаемые ребра гиперграфа

Множество модальностей M:

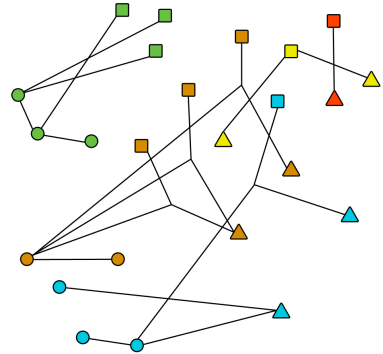
□ ○ △

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

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

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

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

EM алгоритм

E шаг. 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 шаг. Оценивание параметров модели:

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

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

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_{ktdx}, n_{td} и n_{vtk}.

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Эксперимент

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 Θ.

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:

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

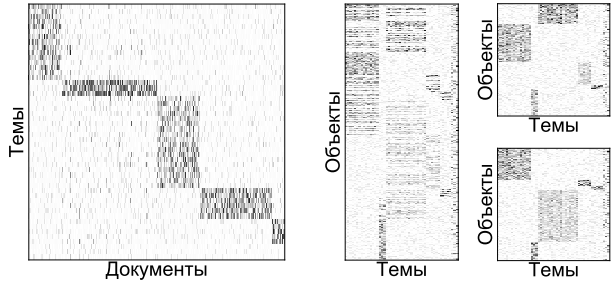
TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

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

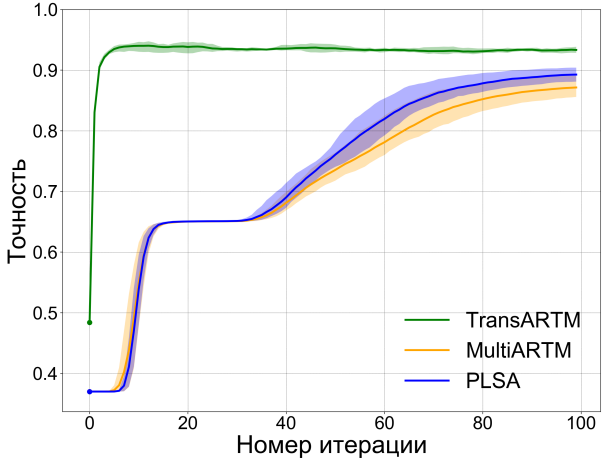
   $\Theta \in \mathbb{R}^{50 \times 5000}$
   $\Phi_k \in \mathbb{R}^{1000 \times 50}$



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

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Результаты

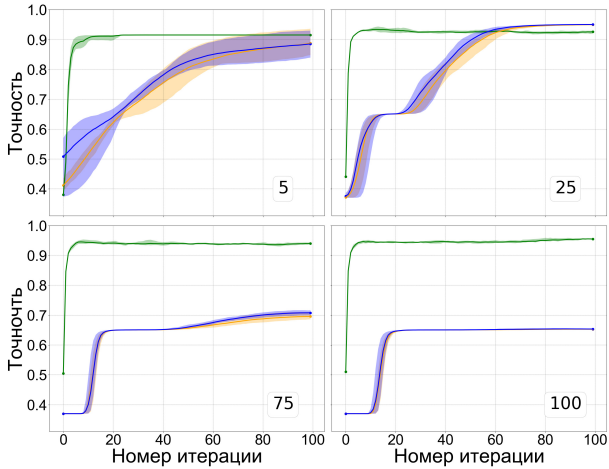


Число тем совпадает с заданным при генерации

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

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Результаты



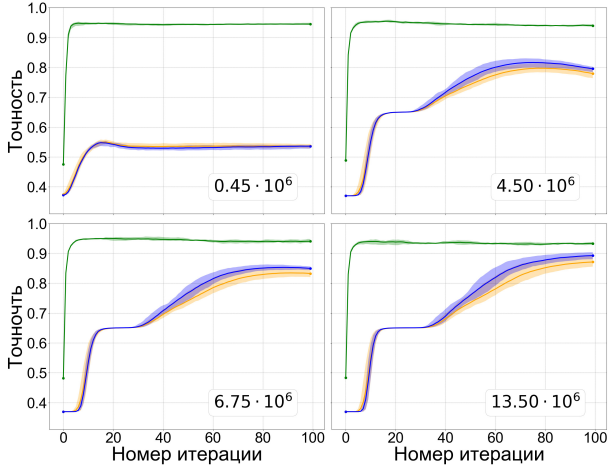
Меняем число тем от 5 до 100

Число тем в исходной матрице Θ равно 50.

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

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Результаты

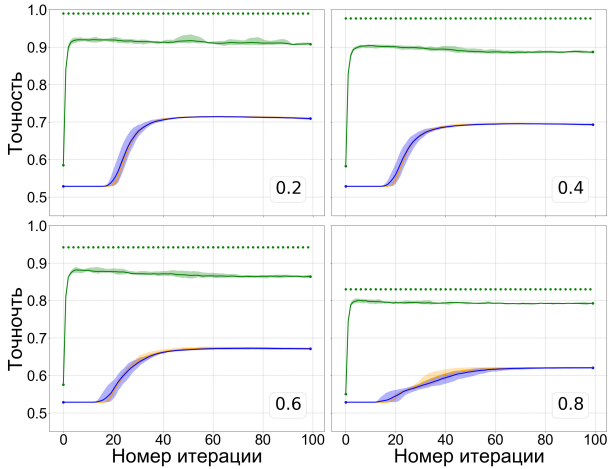


Меняем число транзакций от 450 000 до 13 500 000

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

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

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.

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

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.

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

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.
2. Настраиваем коэффициенты регуляризации по сетке на валидационной выборке.
3. Предсказываем ранжированный список 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.

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

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

Pl aylist
Ar tist
Al bum
Tr ack

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

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

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

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

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».

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Литература

1. Daud A. et al. Knowledge discovery through directed probabilistic topic models: a survey // Frontiers of computer science in China. – 2010. – Т. 4. – №. 2. – С. 280-301.
2. Vorontsov K. et al. Non-Bayesian additive regularization for multimodal topic modeling of large collections // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. – ACM, 2015. – С. 29-37.
3. Li L., Li T. News recommendation via hypergraph learning: encapsulation of user behavior and news content // Proceedings of the sixth ACM international conference on Web search and data mining. – ACM, 2013. – С. 305-314.
4. Zhou D., Huang J., Schölkopf B. Learning with hypergraphs: Clustering, classification, and embedding // Advances in neural information processing systems. – 2007. – С. 1601-1608.

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Обозначения

Каждая вершина $v \in V$ имеет **модальность** $m = \mu(v) \in M$:

$$V = \bigsqcup_{m \in M} V_m, \text{ where } V_m = \{v \in V \colon \mu(v) = m\}.$$

Каждое транзакция $e \in E$ имеет **тип** $k = \varkappa(e) \in K$:

$$E = \bigsqcup_{k \in K} E_k, \text{ where } E_k = \{e \in E \colon \varkappa(e) = k\}.$$

Каждому типу рёбер k соответствует дискретное вероятностное пространство:

$$\Omega_k \subseteq 2^V \times T$$

с функцией вероятности $p_k \colon \Omega_k \to [0, 1]$.

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Обозначения

Будем полагать, что ребра гиперграфа $e \in E_k$ это выборка независимых наблюдений $(e, t) \in \Omega_k$, и каждое ребро входит в выборку n_e раз, и с каждым вхождением ребра связана своя латентная тема $t \in T$.

Каждая вершина v связана с латентными темами:

$$p_k(v, t) = p_k(v \mid t)p_k(t) = p_k(t \mid v)p_k(v).$$

Вероятностные распределения нормируются внутри каждой модальности:

$$\sum_{v \in V_m} p_k(v) = 1; \qquad \sum_{v \in V_m} p_k(v \mid t) = 1.$$

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Предположения

1. Для каждого типа рёбер k первая модальность является *контейнером*.

   D — множество всех вершин-контейнеров в гиперграфе.
   $(d, x) \in E_k$ произвольное ребро типа k, где x — множество всех остальных вершин ребра.

2. Распределение тем в вершине-контейнере d не зависит от типа ребра:

$$p_k(t \mid d) = p(t \mid d) \text{ for all } k \in K.$$

3. Гипотеза условной независимости вершин в ребрах:

$$p_k(x \mid t, d) = p_k(x \mid t) = \prod_{v \in x} p_k(v \mid t).$$

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Гиперграфовая тематическая модель

Вероятности появления рёбер гиперграфа:

$$p_k(d,x) = p_k(d)\, p_k(x \mid d) = p_k(d) \sum_{t \in T} p_k(x \mid d,t)\, p_k(t \mid d) =$$
$$= p_k(d) \sum_{t \in T} p(t \mid d) \prod_{v \in x} p_k(v \mid t) = p_k(d) \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vtk}.$$

Параметры модели:

1. Распределения вершин в темах: $\varphi_{vtk} = p_k(v \mid t)$;
2. Распределение тем в вершинах-контейнерах $\theta_{td} = p(t \mid d)$;
3. Вероятности $p_k(d)$ оцениваются по наблюдаемым данным:

$$p_k(d) = \sum_{(d,x) \in E_k} n_{dx} \Big/ \sum_{e \in E_k} n_e.$$

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Гиперграфовая тематическая модель (TransARTM)

Гиперграфовая тематическая модель определяется:

1. ориентированным гиперграфом $\Gamma = \langle V, E \rangle$,
2. множеством модальностей M,
3. разбиением множества вершин по модальностям $\mu\colon V \to M$,
4. множеством типов рёбер K,
5. разбиением множества рёбер по типам $\varkappa\colon E \to K$,
6. множеством тем T,
7. вероятностным пространством Ω_k с распределением p_k для каждого $k \in K$,
8. параметрами модели $\varphi_{vtk} = p_k(v \mid t)$ и $\theta_{td} = p(t \mid d)$.

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Постановка задачи

Для оптимизации параметров модели применим *принцип максимума правдоподобия* для каждого типа рёбер k:

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

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

Ограничения:

$$\sum_{v\in V_m} \varphi_{vtk} \in \{0,1\}, \qquad \varphi_{vtk} \geqslant 0, \qquad k \in K,\ m \in M,\ t \in T.$$

$$\sum_{t\in T} \theta_{td} \in \{0,1\}, \qquad \theta_{td} \geqslant 0, \qquad t \in T,\ d \in D;$$

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Теоретическое обоснование

Теорема 1. Если функция $R(\Phi, \Theta)$ непрерывно дифференцируема и (Φ, Θ) — точка локального максимума рассматриваемой задачи, то выполняется система уравнений, приведенная ниже, относительно параметров φ_{vtk}, θ_{td} и вспомогательных переменных p_{ktdx}, n_{td} и n_{vtk}:

$$p_{ktdx} = \underset{t \in T}{\operatorname{norm}} \left(\theta_{td} \prod_{v \in x} \varphi_{vtk} \right),$$

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

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

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Доказательство

Сначала выведем уравнение для p_{ktdx} используя формулу Байеса:

$$p_{ktdx} = p_k(t \mid d, x) = \frac{p_k(t, d, x)}{p_k(d, x)} = \frac{p_k(x \mid d, t)\, p_k(t \mid d)}{p_k(x \mid d)} = \frac{p_k(x \mid d, t)\, p_k(t \mid d)}{\sum\limits_{t \in T} p_k(x \mid d, t)\, p_k(t \mid d)} =$$

$$= \underset{t \in T}{\operatorname{norm}} \left(p_k(x \mid d, t)\, p_k(t \mid d)\right) = \underset{t \in T}{\operatorname{norm}} \left(p_k(x \mid t)\, p(t \mid d)\right) = \underset{t \in T}{\operatorname{norm}} \left(\theta_{td} \prod_{v \in x} \varphi_{vtk}\right).$$

Воспользуемся условиями Каруша–Куна–Таккера и запишем лагранжиан оптимизационной задачи:

$$\mathscr{L}(\Phi, \Theta) = \sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vtk} + R(\Phi, \Theta) - \sum_{k \in K} \sum_{m \in M} \sum_{t \in T} \lambda_{kmt} \left(\sum_{v \in V_m} \varphi_{vtk} - 1\right) -$$

$$- \sum_{k \in K} \sum_{m \in M} \sum_{v \in V_m} \sum_{t \in T} \lambda_{kmvt} \varphi_{vtk} - \sum_{d \in D} \mu_d \left(\sum_{t \in T} \theta_{td} - 1\right) - \sum_{d \in D} \sum_{t \in T} \mu_{td} \theta_{td}.$$

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Доказательство

Приравняем нулю производные лагранжиана по параметрам модели:

$$\frac{\partial \mathscr{L}}{\partial \varphi_{vtk}} = \sum_{(d,x)\in E_k} [v \in x]\, \tau_k n_{dx} \frac{\theta_{td} \prod_{u\in x\setminus v} \varphi_{utk}}{p_k(x \mid d)} + \frac{\partial R}{\partial \varphi_{vtk}} - \lambda_{k\mu(v)t} - \lambda_{k\mu(v)vt} = 0;$$

$$\frac{\partial \mathscr{L}}{\partial \theta_{td}} = \sum_{k\in K} \sum_{(d,x)\in E_k} \tau_k n_{dx} \frac{\prod_{v\in x} \varphi_{vtk}}{p_k(x \mid d)} + \frac{\partial R}{\partial \theta_{td}} - \mu_d - \mu_{td} = 0.$$

Умножим левую и правую части первого равенства на φ_{vtk}, второго — на θ_{td}:

$$\sum_{(d,x)\in E_k} [v \in x]\, \tau_k n_{dx} \underbrace{\frac{\theta_{td} \prod_{u\in x} \varphi_{utk}}{p_k(x \mid d)}}_{p_{ktdx} = p_k(t \mid d,x)} + \varphi_{vtk} \frac{\partial R}{\partial \varphi_{vtk}} = \lambda_{k\mu(v)t} \varphi_{vtk};$$

$$\sum_{k\in K} \sum_{(d,x)\in E_k} \tau_k n_{dx} \underbrace{\frac{\theta_{td} \prod_{v\in x} \varphi_{vtk}}{p_k(x \mid d)}}_{p_{ktdx} = p_k(t \mid d,x)} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} = \mu_d \theta_{td}.$$

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Доказательство

Запишем эти уравнения короче через переменные n_{vtk} и n_{td}:

$$\varphi_{vtk}\lambda_{k\mu(v)t} = n_{vtk} + \varphi_{vtk}\frac{\partial R}{\partial \varphi_{vtk}}.$$
$$\theta_{td}\mu_d = n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}};$$

Если $\lambda_{kmt} \leqslant 0$, то тема t вырождена: $\varphi_{vtk} = 0$ для всех $v \in V_m$. Если $\lambda_{kmt} > 0$, то либо обе части равенства выше положительны, либо $\varphi_{vtk} = 0$. Объединяя оба случая в одну формулу, получим:

$$\varphi_{vtk}\lambda_{k\mu(v)t} = \left(n_{vtk} + \varphi_{vtk}\frac{\partial R}{\partial \varphi_{vtk}}\right)_+.$$

Если $\mu_d \leqslant 0$, то вершина-контейнер d вырождена: $\theta_{td} = 0$ для всех $t \in T$. Если $\mu_d > 0$, то либо обе части равенства выше положительны, либо $\theta_{td} = 0$. Объединяя оба случая в одну формулу, получим:

$$\theta_{td}\mu_d = \left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+.$$

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Доказательство

Просуммируем равенства, полученные на предыдущем слайде, по $v \in V_m$ и по $t \in T$, применим условия нормировки и выразим двойственные переменные:

$$\lambda_{kmt} = \sum_{v \in V_m} \left(n_{vtk} + \varphi_{vtk} \frac{\partial R}{\partial \varphi_{vtk}} \right)_+ ;$$

$$\mu_d = \sum_{t \in T} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)_+ .$$

Подставив λ_{kmt} и μ_d, получим искомые выражения. Теорема доказана. ■

Онлайновый EM-алгоритм для TransARTM

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Вход: коллекция $\bigcup\limits_{k\in K} D_k$, разделенная на части $D_b, b = 1, \ldots, B$;

Выход: φ_{vtk} для всех $v \in V, t \in T, k \in K$;

1: инициализировать φ_{vtk} для всех $v \in V, t \in T, k \in K$;
2: $n_{vtk} := 0$, $\widetilde{n}_{vtk} := 0$ для всех $v \in V, t \in T, k \in K$;
3: **for all** $D_b, b = 1, \ldots, B$ **do**
4: обработать каждый $d \in D_b$ при неизменной матрице Φ:
$(\widetilde{n}_{vtk}) := (\widetilde{n}_{vtk}) + \mathbf{ProcessBatch}\,(D_b, \Phi)$;
5: **if** synchronize **then**
6: $n_{vtk} := n_{vtk} + \widetilde{n}_{vtk}$ для всех $v \in V, t \in T, k \in K$;
7: $\varphi_{vtk} = \underset{v\in V_m}{\operatorname{norm}}\left(n_{vtk} + \varphi_{vtk}\frac{\partial R}{\partial \varphi_{vtk}}\right)$ для всех $v \in V_m, m \in M, t \in T, k \in K$;
8: $\widetilde{n}_{vtk} := 0$ для всех $v \in V, t \in T, k \in K$;

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Онлайновый EM-алгоритм для TransARTM

ProcessBatch обрабатывает $d \in D_b$ не изменяя матрицу Φ.

Вход: множество вершин-контейнеров D_b, матрица Φ;
Выход: матрица $(\widetilde{n}_{vtk})$;

$\widetilde{n}_{vtk} := 0$ для всех $v \in V, t \in T, k \in K$;
for all $d \in D_b$ **do**
    инициализировать $\theta_{td} := \frac{1}{|T|}$ для всех $t \in T$;
    repeat
        $p_{tdx} = \underset{t \in T}{\operatorname{norm}} \left(\theta_{td} \prod_{v \in x} \varphi_{vtk} \right)$ для всех $t \in T, k \in K, (d,x) \in E_k$;
        $n_{td} = \sum_{k \in K} \sum_{(d,x) \in E_k} \tau_k n_{dx} p_{tdx}$ для всех $t \in T$;
        $\theta_{td} = \underset{t \in T}{\operatorname{norm}} \left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$ для всех $t \in T$;
    until θ_{td} не сойдутся;
    $\widetilde{n}_{vtk} := \widetilde{n}_{vtk} + \sum_{(d,x) \in E_k} [v \in x] \tau_k n_{dx} p_{tdx}$ для всех $v \in V, t \in T, k \in K$;

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

Модельные данные

Вероятностный процесс порождения данных:

Вход: K, распределения $p(t \mid d)$, $p_k(v \mid t)$, для всех $k \in K$;
Выход: ребра гиперграфа (транзакции);

for all $d \in D$ **do** ▷ D — множество вершин-контейнеров
  определить $K' \subset K$; ▷ K — множество типов ребер
  for all $k \in K'$ **do**
    определить число транзакций — n_{dk};
    for all $i = 1, \ldots, n_{dk}$ **do**
      $d_i := d$;
      выбрать случайную тему t_i из $p(t \mid d_i)$;
      for all $j = 2, \ldots, h(k)$ **do** ▷ $h(k) = |e|$, $e \in E_k$
        выбрать случайный объект v_i из $p_k(v|t_i)$;

TransARTM
Гиперграф
Постановка задачи
Алгоритм
Эксперимент
Модельные данные
Постановка
Результаты
Реальные данные
Постановка
Результаты
Заключение

The Million Playlist Dataset (MPD)

Датасет содержит 1 000 000 плейлистов:

pid – playlist id;
name – the name of the playlist;
description – the description given to the playlist;
modified_at – timestamp when this playlist was last updated;
num_artists – the total number of unique artists;
num_albums – the number of unique albums;
num_tracks – the number of tracks in the playlist;
num_followers – the number of followers this playlist;
num_edits – the number of separate editing sessions;
duration_ms – the total duration of all the tracks in the playlist;
collaborative – if true, the playlist is a collaborative playlist;
tracks – an array of information about each track in the playlist:
- **track_name** – the name of the track;
- **track_uri** – the Spotify URI of the track;
- **album_name** – the name of the track's album;
- **album_uri** – the Spotify URI of the album;
- **artist_name** – the name of the track's primary artist;
- **artist_uri** – the Spotify URI of track's primary artist;
- **duration_ms** – the duration of the track in milliseconds;
- **pos** – the position of the track in the playlist (zero-based).

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