Multimodal Topic Models on Hypergraphs

Ilya Zharikov

ABSTRACT

Topic modeling that is one of modern trends of statistical analysis of texts actively develops over the past 20 years. Probabilistic topic modeling is aimed to identify topics of documents. The main purpose of topic modeling is to understand large text collection and systematize its content. Topic models also can be applied to non-textual and heterogeneous data such as images, audio or video signals.

Modern literature on topic modeling contains hundreds of models adapted to real word problems. But the most popular choice are models of Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA) that is a Bayesian version of PLSA with Dirichlet priors. PLSA and LDA describe interactions between pairs of objects of two types (modalities). To deal with higher number of modalities multimodal topic modeling approach is developed.

The problem to be solved by the current research is that existing latent topic models still describe only pairwise interactions between objects. However, in real word problems relationships between objects are more complex, and considering them as a set of pairwise interactions leads to loss of valuable information. The main contribution of this research is a generalization of existing topic models to the case when original data has complex structure, and can be represented as a hypergraph. The proposed algorithm for the considered extension is called TransARTM and implemented as a part of BigARTM open source project.

The experiments have been carried out both on simulated transaction data and real data. Stability of the proposed method has been investigated on simulated data. Also it has been shown that TransARTM achieves high quality faster than other conventional models and gives the best solution even with relatively small amount of data. On the real data the application of multimodal and hypergraphic multimodal models has been applied for construction of recommendation systems. TransARTM gives significantly better results than considered baseline model that is a choice of the most popular tracks among unheard ones.

Table of Contents

1. INTRODUCTION

1.1. Background

At present human society generates massive amounts of complicated structured information including digital collections of text documents, social media data, ad networks, recommendation systems, etc. These informational environments generate transaction data between objects of multiple types (modalities). The objects can be represented as text documents, word or key phrases, users, advertisements, products or services, etc. Examples of transactions are of the relationship or interaction between objects: a user created (read, rated, “liked”) a document clicked on an ad, a word was found in a document, in an ad, in a user’s request, etc.

A common feature for many approaches is an assumption that these objects have vector descriptions that correspond to topic interests of people, i.e. they describe semantics of these objects. Semantics of objects is latent but indirectly manifested in texts related to objects or in data about the joint use of objects by people. The identification of latent semantic descriptions of objects is the essence of *topic modeling.* Knowledge of these descriptions allows to solve a lot of data analysis tasks, to create qualitatively new Internet services. In particular, topic modeling methods allow to carry out a semantic information search, to build taxonomies or ”road maps” of subject areas, to make recommendations, targeting advertising, etc. Topic models are also used for non-textual and heterogeneous data that can be images, audio or video signals.

Conventional latent topic models describe only pairwise interaction between objects. But in real word problems relationships between objects are more complex, and considering them as a set of pairwise interactions will lead to loss of valuable information. The goal of this research is to develop a topic model that considers all complex interactions among objects of different modalities.

1.2. Literature Review

Methods of topic modeling have been developing over the past 20 years. Topic modeling is applied to determine trends in news streams or scientific papers [1, 2], for multilingual information retrieval [3], in analysis of social network structures [4, 5], in problems of classification, clustering and categorization of documents [6], for topic segmentation of texts [7]. Numerous ideas, models and applications of topic modeling are described in the survey [8].

Classical problems use only two modalities: documents and words. Probabilistic topic modeling [9, 10] methods allow to construct for each word and each document a *topic profile* — a discrete probability distribution on a set of latent topics. Different computational methods of low-rank matrix expansions can be used to solve this problem. Similar approaches are applied in recommendation systems and collaborative filtering task [11,12] with the only difference that other modalities — users and items — are used instead of documents and words. A notable trend is consolidation of data about content and use of objects [13].

Large collections of complex structured and heterogeneous data come from web-sources. Documents usually contain not only words but also links, images, a lot of metadata containing authors, date-time stamps, etc. Social networks provide an example of complex data structure [12, 14–16]. Important information is not only a text of a message but also its metadata including time count, an author of the message, sender's and recipient's geolocation, socio-demographic data, opinions of other users about this message, etc. In the examined case, there are relationships not only between pairs of objects of different modalities but also between triples or any number of objects. For example, (u, w, d) — user u wrote word w in message d. At the same time for topic modeling pairwise interrelations between elements of various modalities remain important. For example, a document is associated with its creation time, an author is associated with geolocation, an advertising banner is associated with words of an advertising text, etc. Therefore, the actual problem is a generalization of multimodal topic modeling [17] methods for analysis of transaction data that includes pairs, triples or more complex interactions.

An adequate mathematical model of transaction data representation is a hypergraph. A *hypergraph* is a generalization of an ordinary graph which edges can connect not only two vertices but any number of vertices. So, vertices of a hypergraph are objects of different modalities, and an unknown hidden topic profile is associated with each vertex. There are transactions between objects that are described by hyperedges. Representation of data as hypergraph improves results of recommendation systems [18, 19], classification and clustering [20].

In the current research the problem of restoring topic profiles of objects from transaction data is stated. It is assumed that probability of a transaction is determined by a similarity degree of included topic profiles of objects. Mathematical models that solve this problem differ by the way of this assumption's formalization. In this research *the hypergraph extension of ARTM approach* [21] is developed, it allows describe more complex interactions of objects than pairwise. The algorithm for the proposed extension is called TransARTM and implemented as a part of the BigARTM open source project.

The experiments were carried out both on simulated transaction data and real data. On simulated data stability of the proposed method was investigated. It was shown that the considered extension takes into account more complex relationships among objects that leads to a significant increase in results. The million playlist dataset (MPD[1]) is used as real data for the problem of playlists extension.

1.3. Potential impacts and novelty

Many real world problems appeal to complex structured data with non-pairwise interactions between objects. Considering such complex relationships as a set of pairwise interactions leads to loss of valuable information. For such data conventional topic models are not suitable. The main contribution of this paper is generalization of topic models to the case of complex structured data. The proposed model takes into account the relationships among any number of objects and finds topic profiles of all objects regardless of its modality.

The developed in this paper topic model is supposed to be used for transaction data in financial organizations. Transaction data analysis is currently considered by some major banks as an important step towards targeting financial services and providing new services in the field of industry consulting. Therefore, it becomes relevant to create tools for analysis of transaction data such that these instruments can give a general understanding of financial flows structure within the industry. The hypergraphic topic model allows to restore latent information about company activities types on observed transaction data.

The rest of this paper is organized as follows. In section 2 general problem statement of topic modeling, ARTM approach, basic models PLSA and LDA, multimodal topic model are described. In section 3 multimodal topic model on hypergraph with generalized EM-algorithm is introduced. Section 4 is devoted to experiments on both simulated (see subsection 4.1) and real data (see subsection 4.2). In this section results of conducted experiments are presented. Section 5 concludes the results and contribution of this research.

[1] Million Playlist Dataset, official website hosted at `https://recsys-challenge.spotify.com/`

2. TOPIC MODELING

In this section a general problem statement of probabilistic topic modeling is described. Classic topic models PLSA and LDA are introduced as well as multimodal topic models. The latter two models are considered as an extension of PLSA. In addition, an approach to solve a problem of a stochastic matrix factorization using additive regularization of topic models (ARTM) is discussed.

2.1. Problem Statement

Consider collections of documents D and let T be a finite set of topics and W be a dictionary or a finite set of terms. Each document $d \in D$ is represented as a sequence of terms $w_1,\ w_2,\ \dots,\ w_{n_d}\ \in\ W$ where n_d is a length of a document d. Assume that each occurrence of a term w in a document d is associated with some topic $t\ \in\ T$. Taking into account the bag-of-words hypothesis suppose that term order in a document is not important and does not affect topic of a document. Therefore, consider only a number of occurrences n_{dw} of each term w in a document d. Define a *probabilistic topic model of text generation* using the law of total probability and the hypothesis of conditional independence:

$$p(w \,|\, d) = \sum_{t\in T} p(w \,|\, t, d)\, p(t \,|\, d) = \sum_{t\in T} p(w \,|\, t)\, p(t \,|\, d) = \sum_{t\in T} \varphi_{wt}\theta_{td}, \tag{1}$$

where θ_{td} is a distribution of topics in a document d and φ_{wt} is a distribution of terms in topic t. Matrices $\Theta = (\theta_{td})_{T\times D}$ and $\Phi = (\varphi_{wt})_{W\times T}$ are used to denote model parameters.

The *topic modeling problem goal* is to find model parameters for which the model (1) gives a close approximation for frequency estimations of conditional probabilities $\hat{p}(w \,|\, d) = n_{dw}/n_d$ for the given collection of documents.

The equation (1) can be rewritten in matrix form if the following way. The left part of the equation contains known matrix of term frequencies $F = (\hat{p}(w \,|\, d))_{W\times D}$. The right part is product of two unknown matrices Φ and Θ. Therefore, the topic modeling problem is equivalent to the stochastic *matrix factorization problem.*

2.2. Additive regularization for topic models

The stochastic matrix factorization problem is an ill-posed since it has an infinite number of solutions in a general case. In fact, if a pair Φ and Θ is a solution, then a pair (ΦS) and $(S^{-1}\Theta)$ is also a solution for all non-singular matrices S for which the matrices ΦS and $S^{-1}\Theta$ are stochastic.

There is a general approach for solving ill-posed inverse problems called *regularization*. When an optimization problem is under defined, an additional criterion (regularizer) is added to the main criterion taking into account the specifics of the problem and knowledge of the subject area.

Additive regularization for topic models (ARTM) [21] is based on maximizing a linear combination of a main objective L and *regularizers* $R_i(\Phi, \Theta)$ with non-negative *coefficients* τ_i:

$$L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi,\Theta}, \quad \text{where} \quad R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta). \tag{2}$$

In the paper [21] authors showed that ARTM allows to improve topic interpretability along with model sparsity and common parlance words allocation [22]. It also makes possible to discard dependent and uninformative topics [23], use specific dictionaries to highlight highly specialized topics, in particular, for study of inter-ethnic relations using social networks data [24].

2.3. Topic Models PLSA and LDA

Probabilistic Latent Semantic analysis (PLSA)

In *probabilistic latent semantic analysis* [9] estimation of topic model parameters is done by maximizing the likelihood of documents collection:

$$p(D, \Phi, \Theta) = \prod_{i=1}^{n} p(d_i, w_i) = \prod_{d \in D} \prod_{w \in d} p(d, w) = \prod_{d \in D} \prod_{w \in d} p(w \mid d)^{n_{dw}} p(d)^{n_{dw}} \to \max_{\Phi,\Theta} \tag{3}$$

After taking the logarithm:

$$\ln p(D, \Phi, \Theta) = \ln \prod_{i=1}^{n} p(d_i, w_i) = \sum_{d \in D} \sum_{w \in d} \ln p(w \mid d) + \sum_{d \in D} n_d \ln p(d) \to \max_{\Phi,\Theta} \tag{4}$$

Taking into account (1) and dropping the last term the maximization problem is obtained:

$$L\left(D, \Phi, \Theta\right) = \sum_{d \in D} \sum_{w \in d} \ln p\left(w \,|\, d\right) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} \to \max_{\Phi, \Theta}, \tag{5}$$

with the constraints of non-negativity and normalization:

$$\theta_{td} \geqslant 0, \ \sum_{t \in T} \theta_{td} = 1 \quad \text{and} \quad \varphi_{wt} \geqslant 0, \ \sum_{w \in W} \varphi_{wt} = 1. \tag{6}$$

The log-likelihood maximization problem (5), (6) can be solved using *expectation-maximization algorithm* (EM-algorithm). It consists of random model parameters initialization and two steps that are repeated in a loop.

At the E-step conditional distributions for latent topics $p(t \,|\, d, w)$ are calculated for each term w in each document d according to the Bayes rule for current values of model parameters φ_{wt}, θ_{td}. At the M-step, on the contrary, a new approximation of model parameters is calculated based on conditional probabilities for topics $p(t \,|\, d, w)$. The formulas for the E and M steps can be found in [25].

Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) topic model was proposed to deal with over-fitting in PLSA. LDA model is based on an assumption that θ_d and φ_t columns are random vectors from Dirichlet distribution with parameters $\alpha \in \mathbb{R}^{|T|}$ and $\beta \in \mathbb{R}^{|W|}$ respectively.

For LDA model the maximization problem also can be written:

$$\begin{aligned} L\left(D, \Phi, \Theta\right) &= \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + \ln \prod_{t \in T} \operatorname{Dir}(\varphi_t; \beta) \prod_{d \in D} \operatorname{Dir}(\theta_d; \alpha) = \\ &= \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + \underbrace{\sum_{t \in T} \sum_{w \in W} (\beta_w - 1) \ln \varphi_{wt} + \sum_{d \in D} \sum_{t \in T} (\alpha_t - 1) \ln \theta_{td}}_{R(\Phi, \Theta)} \to \max_{\Phi, \Theta}. \end{aligned} \tag{7}$$

Now it is clear that LDA model is the PLSA model with constrained parameters Φ, Θ. Moreover, if $\beta_w = 1$ and $\alpha_t = 1$, a prior Dirichlet distribution coincides with uniform distribution and LDA model corresponds to PLSA [26].

2.4. Multimodal Topic Models

PLSA and LDA models use only one modality of terms (usually, words). *Multimodal topic model* describes documents that contain not only text but some additional metadata that helps to identify document topic.

Each type of metadata forms a separate *modality* with its own dictionary. Examples of non-textual modalities are authors, time stamps, geodata, genres, categories, classes, etc. Each document is considered as a universal container that comprises tokens of various modalities.

Let M be a set of modalities. As noted, each modality has its own dictionary of tokens W_m, $m \in M$. These sets are disjoint. Their union is denoted by W as previously. Modality of a particular token $w \in W$ is denoted by $m(w)$.

Topic model of modality m is similar to the model (1):

$$p(w \,|\, d) = \sum_{t \in T} p(w \,|\, t)\, p(t \,|\, d) = \sum_{t \in T} \varphi_{wt}\theta_{td}, \qquad w \in W_m, \quad d \in D. \tag{8}$$

Each modality m responds to a stochastic matrix $\Phi_m = (\varphi_{wt})_{W_m \times T}$. Set of matrices Φ_m forms $W \times T$ matrix Φ. Assume that topic distribution for each document is common to all modalities.

Multimodal model is constructed by maximizing a linear combination of the modalities log-likelihood and regularizers with weights:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W_m} n_{dw} \ln \sum_{t \in T} \varphi_{wt}\theta_{td} + R(\Phi, \Theta) \;\to\; \max_{\Phi,\Theta}; \tag{9}$$

$$\sum_{w \in W_m} \varphi_{wt} = 1, \quad \varphi_{wt} \geqslant 0, \quad m \in M; \quad \sum_{t \in T} \theta_{td} = 1; \quad \theta_{td} \geqslant 0, \tag{10}$$

where weights τ_m help to balance modalities according to their importance and frequency of occurrences in documents. Optimization problem (9), (10) can be solved using a regularized EM-algorithm [17, 27].

3. MULTIMODAL TOPIC MODELS ON HYPERGRAPHS

3.1. Problem Statement

Topic models on hypergraphs are further generalization of multimodal models.

Consider the case when observed data can be represented as a hypergraph. Define *hypergraph* $\Gamma = \langle V, E \rangle$ which is determined by a set of vertices V and set of edges E where each edge $e \in E$ is a subset of vertices $e \subset V$ that corresponds to a transaction (interaction or relation). An example of hypergraph one can find in Figure 1.

For example, in the Internet advertising transaction "user u clicked on an ad b on a page d" is an edge of three vertices $e = (u, d, b)$. In social network transaction "user u wrote a word w in a blog d" is an edge of three vertices $e = (d, u, w)$. In recommendation system transaction "a user u rated a movie f in situational context s" is an edge of three vertices $e = (u, f, s)$. Moreover, in these examples an interaction of three vertices can not be reduced to pairwise interactions. On the contrast, in music recommendation system transaction "a track r by an artist a refers to an album d that was published in year y" is an edge of four vertices $e = (r, a, d, y)$, however, it still may be represented by a set of pairwise relationships (d, r), (d, a), (d, y).

Each vertex $v \in V$ has *modality* $m = \mu(v)$ that belongs to a given finite set of modalities M. A set of all vertices consists of disjoint subsets of nodes of different modalities:

$$V = \bigsqcup_{m \in M} V_m, \qquad V_m = \{v \in V \colon \mu(v) = m\}. \tag{11}$$

Each edge $e \in E$ has a *transaction type* $k = \kappa(e)$ from a given finite set K. A set of all transactions (edges of a hypergraph) is represented by disjoint subsets in the following way:

$$E = \bigsqcup_{k \in K} E_k, \qquad E_k = \{e \in E \colon \kappa(e) = k\}. \tag{12}$$

For example, in conventional topic models only two modalities are considered: documents D and terms W, see 2.3. There is also only one transaction type: an occurrence of the term w in a document d that is associated with an edge $e = (d, w)$. In this case the graph is bipartite.

Each type of edges k corresponds to a discrete probability space $\Omega_k \subseteq 2^V \times T$ with a prob-

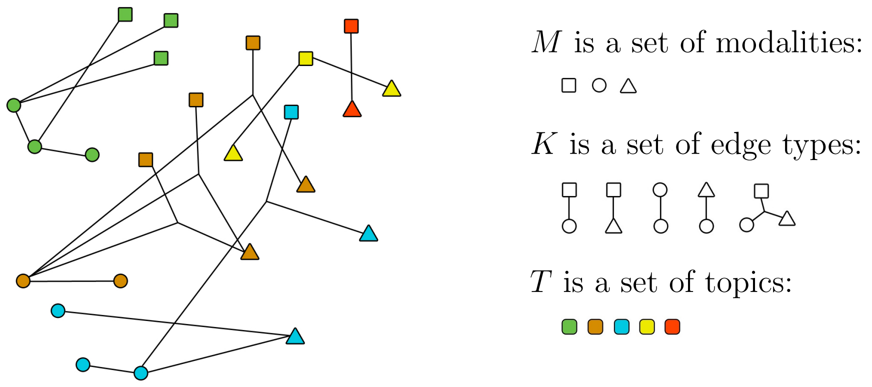


Figure 1: Example of the hypergraph with vertices of 3 different modalities and 5 edge types.

ability function $p_k\colon \Omega_k \to [0,1]$. It is assumed that edges of the hypergraph $e \in E_k$ are independent observations $(e,t) \in \Omega_k$ and each edge is sampled n_e times, and every entry of edge is associated with an unobserved (latent) topic $t \in T$. Probability distribution is normalized within each modality: $\sum_{v \in V_m} p_k(v) = 1$ and $\sum_{v \in V_m} p_k(v \,|\, t) = 1$.

Usually topic models are asymmetrical. Documents are associated with conditional distributions $p(t \,|\, d)$, other modalities with conditional distributions $p(v \,|\, t)$. Therefore, documents are allocated in particular modality which is known as a *container*. Asymmetry makes building the model easier.

Turning on to generalization of the asymmetric model for the case of hypergraph assume that for each type of edges k first modality is a container (for example, document or user). Denote by D a set of all container vertices in a hypergraph and by $(d,x) \in E_k$ an arbitrary edge of type k where x is a set of all other vertices except d.

Probabilistic model of transaction data generation is based on two basic assumptions.

Firstly, suppose that the distribution of the topics in container vertex d does not depend on edge type $p_k(t \,|\, d) = p(t \,|\, d)$ for all $k \in K$. This is a generalization of conventional multimodal topic modeling assumption that topics distribution for the document is equally valid for all modalities. Simultaneously, vertices distribution for the topic $p_k(v \,|\, t)$ is not assumed to be the same for all edge types. For example, the distribution of words used in texts on web-pages, custom queries and ad banners can significantly vary for one topic. An additional requirement that these distributions are similar can be posted with a help of regularization.

Secondly, introduce a hypothesis of conditional independence of vertices that are a part of edge (d, x):

$$p_k(x \,|\, t) = \prod_{v \in x} p_k(v \,|\, t). \qquad (13)$$

Under the made assumptions generation process of each edge $(d, x) \in E_k$ consists of two steps. Firstly a topic t from a distribution $p(t \,|\, d)$ is generated. Then a set of vertices $x \subset V$ is generated, and each vertex $v \in x$ of modality m is generated according to its distribution $p_k(v \,|\, t)$ over a set V_m.

Mathematical model expresses the probability of hypergraph edges occurrence using a distribution associated with vertices:

$$\begin{aligned} p_k(d, x) = p_k(d)\, p_k(x \,|\, d) = p_k(d) \sum_{t \in T} p_k(x \,|\, d, t)\, p_k(t \,|\, d) = p_k(d) \sum_{t \in T} p_k(x \,|\, t)\, p_k(t \,|\, d) = \\ = p_k(d) \sum_{t \in T} p(t \,|\, d) \prod_{v \in x} p_k(v \,|\, t) = p_k(d) \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vtk}. \quad (14) \end{aligned}$$

The parameters of this model are a conditional probability of vertices in topics $\varphi_{vtk} = p_k(v \,|\, t)$ that is normalized for each modality $v \in V_m$, and a conditional probability of topics in the containers $\theta_{td} = p(t \,|\, d)$. Probability $p_k(d)$ is estimated from observed data and does not depend on parameters of the model:

$$p_k(d) = \sum_{(d,x) \in E_k} n_{dx} \Big/ \sum_{e \in E_k} n_e. \qquad (15)$$

Therefore, considered hypergraphic topic model is defined by:

- the oriented hypergraph $\Gamma = \langle V, E \rangle$,
- the set of modalities M,
- the decomposition of the vertices set into subsets of different modalities $\mu \colon V \to M$,
- the set of edge types K,
- the decomposition of the edges set into subsets of different edge types $\kappa \colon E \to K$,
- the set of topics T,
- the probability space Ω_k with the distribution p_k for all $k \in K$,
- the model parameters $\varphi_{vtk} = p_k(v \,|\, t)$ and $\theta_{td} = p(t \,|\, d)$.

Hypergraphic topic model describes a wide class of topic models mentioned in 3.4. The subcases of the proposed model are described in 3.4.

To optimize model parameters the *principle of maximum likelihood* for each edge type k is applied. Therefore, the weighted sum of the log-likelihood with weights τ_k is maximized discarding terms of the form $\tau_k n_{dx} \ln p_k(d)$:

$$L(\Phi, \Theta) = \sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vtk} \to \max_{\Phi, \Theta}. \tag{16}$$

Regularizer $R(\Phi, \Theta)$ is aimed to improve stability of solution. The problem of building a topic model with constraints of normalization and non-negativity can be formulate as follows:

$$\sum_{k \in K} \tau_k \sum_{(d,x) \in E_k} n_{dx} \ln \sum_{t \in T} \theta_{td} \prod_{v \in x} \varphi_{vtk} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}; \tag{17}$$

$$\sum_{v \in V_m} \varphi_{vtk} \in \{0, 1\}, \qquad \varphi_{vtk} \geqslant 0, \qquad k \in K,\ m \in M,\ t \in T. \tag{18}$$

$$\sum_{t \in T} \theta_{td} \in \{0, 1\}, \qquad \theta_{td} \geqslant 0, \qquad t \in T,\ d \in D; \tag{19}$$

Constraints (18) and (19) provide an opportunity for distributions to be equal to 0. If $\varphi_{vtk} = 0$ for each $v \in V_m$, topic t is not involved in process of generation edges of type k with vertices of modality m. If $\theta_{td} = 0$ for all $t \in T$, consider that a topic model is not able to describe the content of container vertex d.

3.2. EM-algorithm

Denote non-negative valuation operator which converts an arbitrary vector $(a_i)_{i \in I}$ to a vector of probabilities of a discrete distribution is introduced in the following way:

$$\underset{i \in I}{\operatorname{norm}}\, a_i = \frac{\max\{a_i, 0\}}{\sum\limits_{j \in I} \max\{a_j, 0\}} = \frac{(a_i)_+}{\sum\limits_{j \in I} (a_j)_+}, \text{ for all } i \in I, \tag{20}$$

and if $a_i \leqslant 0$ for all $i \in I$, suppose that $\underset{i \in I}{\operatorname{norm}}\, a_i = 0$.

Regularized EM-algorithm is used to solve the problem (17), (18), (19), E and M steps are performed on each iteration.

On E-step for each of the observed edges (d, x) of hypergraph distribution $p_{ktdx} = p_k(t \mid d, x)$

is calculated using Bayes rule:

$$p_{ktdx} = \operatorname*{norm}_{t \in T}\left(\theta_{td} \prod_{v \in x} \varphi_{vtk}\right). \tag{21}$$

On M-step the obtained values of auxiliary variables p_{tdx} are used to estimate model parameters:

$$\varphi_{vtk} = \operatorname*{norm}_{v \in V_m}\left(n_{vtk} + \varphi_{vtk}\frac{\partial R}{\partial \varphi_{vtk}}\right), \qquad n_{vtk} = \sum_{(d,x) \in E_k} [v \in x]\, \tau_k n_{dx} p_{tdx}, \tag{22}$$

$$\theta_{td} = \operatorname*{norm}_{t \in T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \qquad n_{td} = \sum_{k \in K} \sum_{(d,x) \in E_k} \tau_k n_{dx} p_{tdx}, \tag{23}$$

where n_{vtk} is interpreted as a total weight of edges of type k containing a vertex v and relevant to a topic t, n_{td} — as a total weight of all edges of all types with container vertex d relevant to a topic t.

The considered EM-algorithm adapted to large collections is described in Algorithms 1 and 2.

Algorithm 1 Fast online EM-algorithm for TransARTM.

```
    Input: collection $\bigcup_{k \in K} D_k$ split into batches $D_b, b = 1, \dots, B$;
    Output: $\varphi_{vtk}$ for all $v \in V, t \in T, k \in K$;

initialize $\varphi_{vtk}$ for all $v \in V, t \in T, k \in K$;
$n_{vtk} := 0$, $\widetilde{n}_{vtk} := 0$ for all $v \in V, t \in T, k \in K$;
for all batches $D_b, b = 1, \dots, B$ do
    iterate each document $d \in D_b$ at a constant matrix $\Phi$:
    $(\widetilde{n}_{vtk}) := (\widetilde{n}_{vtk}) +$ ProcessBatch $(D_b, \Phi)$;        ▷ see Algorithm 2
    if synchronize then
        $n_{vtk} := n_{vtk} + \widetilde{n}_{vtk}$ for all $v \in V, t \in T, k \in K$;
        $\varphi_{vtk} = \operatorname*{norm}_{v \in V_m}\left(n_{vtk} + \varphi_{vtk}\frac{\partial R}{\partial \varphi_{vtk}}\right)$ for all $v \in V_m, m \in M, t \in T, k \in K$;
        $\widetilde{n}_{vtk} := 0$ for all $v \in V, t \in T, k \in K$;
```

Algorithm 2 ProcessBatch iterates $d \in D_b$ at a constant Φ.

```
    Input: set of vertices-containers $D_b$, matrix $\Phi$;
    Output: matrix $(\widetilde{n}_{vtk})$;

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

3.3. Theoretical justification

Topic $t \in T$ is called *regular in a modality* $m \in M$ for edge type $k \in K$ if the following inequality is held at least for one vertex $v \in V_m$:

$$n_{vtk} + \varphi_{vtk} \frac{\partial R}{\partial \varphi_{vtk}} > 0. \tag{24}$$

Container vertex d is called *regular*, if the following inequality is held at least for one topic $t \in T$:

$$n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} > 0. \tag{25}$$

The regularity condition is not overloaded and means that the regularizer R slightly effects on model when its partial derivative is negative. In PLSA and LDA models the regularity conditions are held always.

If a topic t is not regular, assume that $\varphi_{vtk} = 0$ for all $v \in V_m$. This means that the topic t is not involved in generation process of type k edges.

If a container vertex d is not regular, assume that $\theta_{td} = 0$ for all $t \in T$. This means that the model is not able to describe the content of vertex-container d.

The following theorem shows that formulas of iteration process (21), (22) and (23) represent a system of equations that is equivalent to the Karush–Kuhn–Tucker conditions for the problem (17), (18), (19).

Theorem 1. If function $R(\Phi, \Theta)$ is continuously differentiable and (Φ, Θ) is a local maximum of the problem (17), (18), (19), then the following system of equations for model parameters φ_{vtk}, θ_{td} and auxiliary variables p_{tdx}, n_{td} and n_{vtk} is held:

$$p_{ktdx} = \underset{t\in T}{\mathrm{norm}}\left(\theta_{td}\prod_{v\in x}\varphi_{vtk}\right), \tag{26}$$

$$\varphi_{vtk} = \underset{v\in V_m}{\mathrm{norm}}\left(n_{vtk} + \varphi_{vtk}\frac{\partial R}{\partial \varphi_{vtk}}\right), \qquad n_{vtk} = \sum_{(d,x)\in E_k} [v \in x]\, \tau_k n_{dx} p_{tdx}, \tag{27}$$

$$\theta_{td} = \underset{t\in T}{\mathrm{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \qquad n_{td} = \sum_{k\in K}\sum_{(d,x)\in E_k} \tau_k n_{dx} p_{tdx}, \tag{28}$$

Proof. First of all, let's prove the equation (21) using Bayes rule:

$$p_{ktdx} = p_k(t\,|\,d,x) = \frac{p_k(t,d,x)}{p_k(d,x)} = \frac{p_k(x\,|\,d,t)\,p_k(t\,|\,d)}{p_k(x\,|\,d)} = \frac{p_k(x\,|\,d,t)\,p_k(t\,|\,d)}{\sum\limits_{t\in T} p_k(x\,|\,d,t)\,p_k(t\,|\,d)} = \tag{29}$$

$$= \underset{t\in T}{\mathrm{norm}}\left(p_k(x\,|\,d,t)\,p_k(t\,|\,d)\right) = \underset{t\in T}{\mathrm{norm}}\left(p_k(x\,|\,t)\,p(t\,|\,d)\right) = \underset{t\in T}{\mathrm{norm}}\left(\theta_{td}\prod_{v\in x}\varphi_{vtk}\right). \tag{30}$$

Using Karush–Kuhn–Tucker conditions the Lagrangian of the optimization problem (17) can be written as follows:

$$\mathscr{L}(\Phi,\Theta) = \sum_{k\in K}\tau_k \sum_{(d,x)\in E_k} n_{dx}\ln\sum_{t\in T}\theta_{td}\prod_{v\in x}\varphi_{vtk} + R(\Phi,\Theta) - \tag{31}$$

$$-\sum_{k\in K}\sum_{m\in M}\sum_{t\in T}\lambda_{kmt}\left(\sum_{v\in V_m}\varphi_{vtk} - 1\right) - \sum_{k\in K}\sum_{m\in M}\sum_{v\in V_m}\sum_{t\in T}\lambda_{kmvt}\varphi_{vtk} - \tag{32}$$

$$-\sum_{d\in D}\mu_d\left(\sum_{t\in T}\theta_{td} - 1\right) - \sum_{d\in D}\sum_{t\in T}\mu_{td}\theta_{td}. \tag{33}$$

Set the derivatives of the Lagrangian for model parameters to be zero:

$$\frac{\partial\mathscr{L}}{\partial\varphi_{vtk}} = \sum_{(d,x)\in E_k}[v\in x]\,\tau_k n_{dx}\frac{\theta_{td}\prod_{u\in x\setminus v}\varphi_{utk}}{p_k(x\,|\,d)} + \frac{\partial R}{\partial\varphi_{vtk}} - \lambda_{k\mu(v)t} - \lambda_{k\mu(v)vt} = 0; \tag{34}$$

$$\frac{\partial\mathscr{L}}{\partial\theta_{td}} = \sum_{k\in K}\sum_{(d,x)\in E_k}\tau_k n_{dx}\frac{\prod_{v\in x}\varphi_{vtk}}{p_k(x\,|\,d)} + \frac{\partial R}{\partial\theta_{td}} - \mu_d - \mu_{td} = 0. \tag{35}$$

Multiply left and right parts of the first equality by φ_{vtk}, left and right parts of the second equality by θ_{td}:

$$\sum_{(d,x)\in E_k} [v \in x]\, \tau_k n_{dx} \underbrace{\frac{\theta_{td} \prod_{u\in x} \varphi_{utk}}{p_k(x \,|\, d)}}_{p_{ktdx} = p_k(t \,|\, d,x)} + \varphi_{vtk} \frac{\partial R}{\partial \varphi_{vtk}} = \lambda_{k\mu(v)t} \varphi_{vtk}; \tag{36}$$

$$\sum_{k\in K} \sum_{(d,x)\in E_k} \tau_k n_{dx} \underbrace{\frac{\theta_{td} \prod_{v\in x} \varphi_{vtk}}{p_k(x \,|\, d)}}_{p_{ktdx} = p_k(t \,|\, d,x)} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} = \mu_d \theta_{td}. \tag{37}$$

Rewrite these equations within variables n_{vtk} from (27) and n_{td} from (28):

$$\varphi_{vtk} \lambda_{k\mu(v)t} = n_{vtk} + \varphi_{vtk} \tfrac{\partial R}{\partial \varphi_{vtk}}. \tag{38}$$

$$\theta_{td} \mu_d = n_{td} + \theta_{td} \tfrac{\partial R}{\partial \theta_{td}}; \tag{39}$$

Suppose that $\lambda_{kmt} \leqslant 0$, then the regularity condition (24) is not held, and in this case, according to the agreement, $\varphi_{vtk} = 0$ is held for each $v \in V_m$. If the dual variable λ_{kmt} is positive, then both parts of the equation (38) are non-negative. Combining these two cases into one formula the following expression is obtained:

$$\varphi_{vtk} \lambda_{k\mu(v)t} = \left(n_{vtk} + \varphi_{vtk} \tfrac{\partial R}{\partial \varphi_{vtk}} \right)_+. \tag{40}$$

Similarly, if $\mu_d \leqslant 0$, then the regularity condition (25) is not held, and according to the agreement, $\theta_{td} = 0$ for all $t \in T$. If $\mu_d > 0$, then both parts of the equation (39) are non-negative. Combining these two cases into one formula the following expression is obtained:

$$\theta_{td} \mu_d = \left(n_{td} + \theta_{td} \tfrac{\partial R}{\partial \theta_{td}} \right)_+. \tag{41}$$

After summing left and right parts of the equation (40) by $v \in V_m$, left and right parts of the equation (41) by $t \in T$, and applying normalization conditions, it is possible to express the dual variables:

$$\lambda_{kmt} = \sum_{v\in V_m} \left(n_{vtk} + \varphi_{vtk} \tfrac{\partial R}{\partial \varphi_{vtk}} \right)_+; \tag{42}$$

$$\mu_d = \sum_{t\in T} \left(n_{td} + \theta_{td} \tfrac{\partial R}{\partial \theta_{td}} \right)_+. \tag{43}$$

After substituting obtained expressions (40) and (41) of dual variables, (22) and (23) are derived. The theorem is proved. ■

3.4. Special cases

In this subsection PLSA, LDA, MultiARTM topic models described in 2.3, 2.4 are represented in the proposed notations as special cases of hypergraphic multimodal topic model. The optimization problem of the considered model is the following:

$$\sum_{k\in K} \tau_k \sum_{(d,x)\in E_k} n_{dx} \ln \sum_{t\in T} \theta_{td} \prod_{v\in x} \varphi_{vtk} + R(\Phi, \Theta) \to \max_{\Phi,\Theta} \tag{44}$$

with constraints on matrices Φ, Θ defined in (18) and (19).

3.4.1. PLSA and LDA

Consider hypergraphic multimodal topic model with only one edge type — occurrence the word w in the document d. In this case there are two modalities: document and word, and each edge consists of two vertices (d, w). Documents are used as container vertices. According to these characteristics the optimization problem (44) can be rewritten:

$$\tau \sum_{(d,x)\in E} n_{dx} \ln \sum_{t\in T} \theta_{td} \varphi_{xt} + R(\Phi, \Theta) \to \max_{\Phi,\Theta}. \tag{45}$$

Substituting $\tau = 1$ in (45) and renaming x by w one can see that this problem is exactly the same as (7) and (5) without regularization. Therefore, LDA and PLSA models are special cases of the proposed topic model.

3.4.2. Multimodal Topic Models

Consider multimodal topic model with a set of modalities $M = \{\mu_1, \mu_2, \ldots, \mu_l\}$, and add one more modality μ_d for container vertices: $M' = M \cup \{\mu_d\}$. Consider only pairwise interactions between documents and objects of different modalities that leads to $|M'| - 1 = |M|$ edge types of degree 2. Suppose that documents are container vertices and have μ_d modality. In this case D is equal to the set of all documents and edge type is defined by the modality of the second vertex. Therefore, $K = M$ and $(d, x) \in E_m$ if and only if $x \in W_m$, where W_m is a set of objects with modality m. Putting it all together one can obtain the following optimization problem:

$$\sum_{m \in M} \tau_k \sum_{(d,x) \in E_m} n_{dx} \ln \sum_{t \in T} \theta_{td} \varphi_{xtk} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}. \tag{46}$$

Renaming k by m, x by w and considering φ_{vtk} separately by modality one can see that the problem (46) coincides with the optimization problem (9), (10). Therefore, multimodal topic model is also a special case of the considered hypergraphic generalization.

3.5. Regularizers

This section describes regularizers used in the current research. Each of the considered regularizers is written for objects that are elements of one particular edge type and of the same modality. Overall regularization is a weighted linear combination of all using regularizers.

Smoothing regularizer introduces a requirement for distributions φ_{wt} and θ_{td} to be from the given distributions of β_w and α_t as for LDA model:

$$R\left(\Phi, \Theta\right) = R(\Phi) + R(\Theta) = \beta \sum_{t \in T} \sum_{w \in W} \beta_w \ln \varphi_{wt} + \alpha \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max. \tag{47}$$

Sparsing regularizer has the same form but regularization coefficients β and α are negative that leads to appearance of zero elements in distributions φ_{wt} and θ_{td}.

4. COMPUTATIONAL EXPERIMENTS

This section is devoted to experiments carried out on both simulated transaction data and real data. The aim of a series of experiments is to study the behavior of TransARTM in comparison with other topic models described above.

4.1. Simulated Data

Simulated data is an example of transaction data in which the probability of object occurrence depends on the type of transaction. The goal is to investigate the quality of restoring the structure of the matrix of topics distribution for documents denoted by Θ. Move on to the description of the generation procedure for simulated data.

4.1.1. Generation procedure

A generation procedure consists of three main steps:

1. Determination of the following sets: a set of container vertices or documents D, a set of modalities M, a set of edge types or transaction types K, a set of vertices or objects $V = \bigsqcup\limits_{m \in M} V_m$, a set of topics T;
2. Generation of matrices Θ and Φ_k for all $k \in K$;
3. Generation of transaction data according to obtained Θ and Φ_k matrices.

The last two steps should be described in more detail.

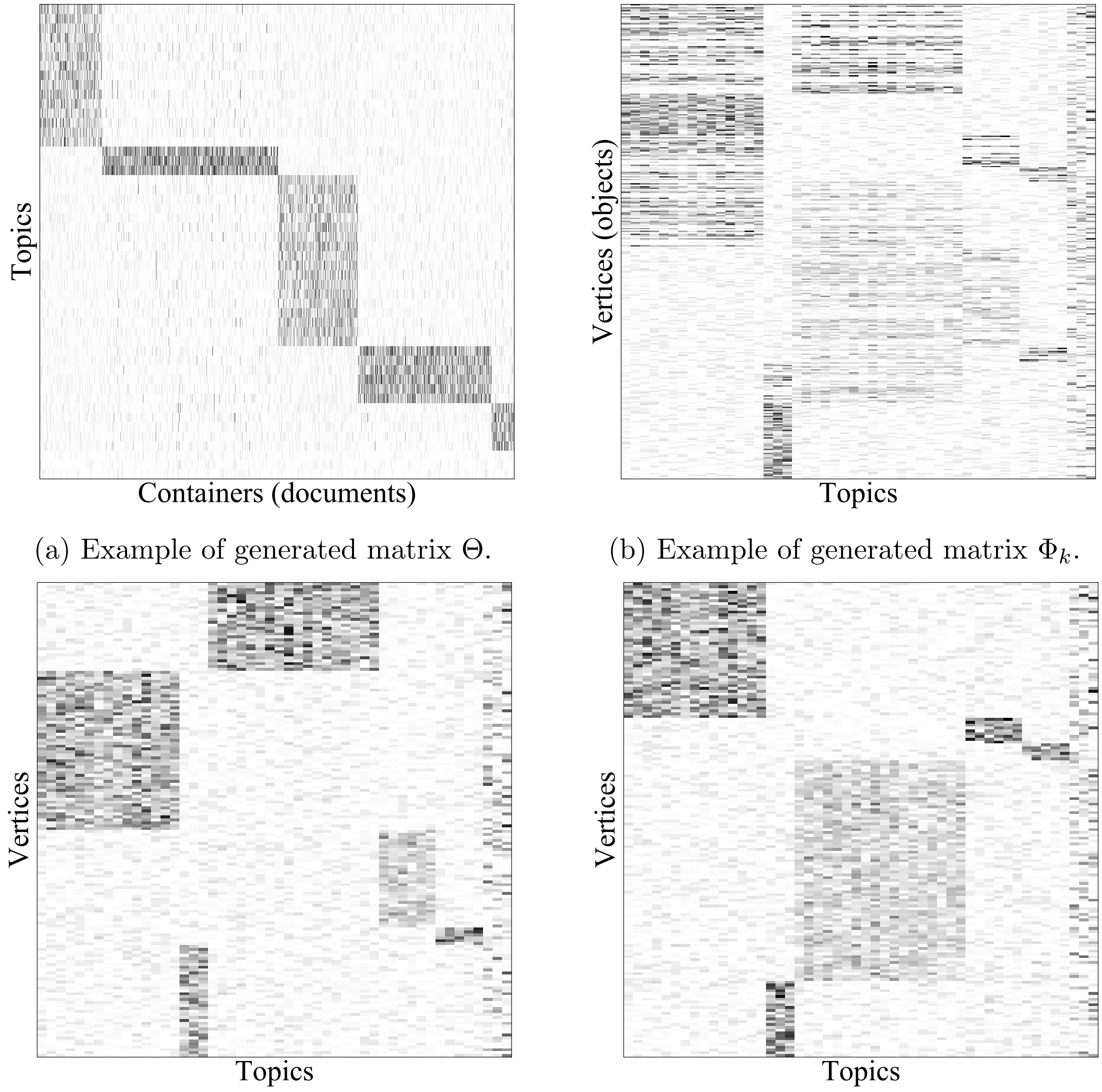
Step 2. Since the goal of the experiment is to learn how the structure of Θ matrix is restored it is necessary to set the structure during generation procedure. To comply with it one can specify the assignment of documents into several classes. According to this partition the dominant topics for each class and the dominant objects among the topics of one class are randomly determined depending on the edge type. Dominant objects / topics are those ones with the probability much higher compared to others. All elements of Θ and Φ_k matrices are generated from normal distribution (only positive elements are considered). Θ matrix also contains background topics with elements generated from uniform distribution. For Θ and Φ_k matrices introduce sparsity parameter that regulates a fraction of non-zero elements. When this parameter is of a high value there are a lot of zero elements in the matrix. Further, Θ matrix is normalized so that the sum of elements in each column is equal to one:

$$\sum_{t \in T} \theta_{td} = 1 \text{ for all } d \in D. \tag{48}$$

The elements of Φ_k matrices are normalized within the objects of the same modality separately:

$$\sum_{v \in V_m} \varphi_{vtk} = 1 \text{ for all } k \in K, \, m \in M, \, t \in T. \tag{49}$$

The examples of Θ and Φ_k are shown in Figure 2.



(a) Example of generated matrix Θ.

(b) Example of generated matrix Φ_k.

(c) Matrix Φ_k slices on objects of the same modality.

Figure 2: Examples of generated matrices Θ and Φ_k for particular k.

It is important to note that according to this generation procedure depending on the edge type dominant objects can differ for the same class.

Step 3. After generation of matrices Θ and Φ_k for all $k \in K$ simulated transaction data can be generated according to Algorithm 3.

Algorithm 3 Probabilistic process of simulated transaction data generation.

```
 Input: $K$, distributions $p(t\,|\,d)$, $p_k(v\,|\,t)$, for all $k \in K$;
Output: edges of the hypergraph (transactions);

for all $d \in D$ do                                   ▷ $D$ is a set of container vertices
    define $K' \subset K$;                              ▷ $K$ is a set of edge types
    for all $k \in K'$ do
        define the number of hyperedges — $n_{dk}$;
        for all $i = 1, \dots, n_{dk}$ do
            $d_i := d$;
            choose random topic $t_i$ from $p(t\,|\,d_i)$;
            for all $j = 2, \dots, h(k)$ do             ▷ $h(k) = |e|$, $e \in E_k$
                choose random object $v_i$ from $p_k(v|t_i)$;
```

Transaction data adaptation. Since transaction data is not suitable for conventional topic models it is necessary to transform it. For multimodal topic models it is proposed to consider each transaction as set of pairwise interactions between container vertex and other vertices in this transaction. For PLSA and LDA topic models consider the same transformation but in addition combine all modalities into one.

4.1.2. Experimental Setup

The main pipeline of all experiments on the simulated data is described below.

First of all, using PLSA, Multimodal and TransARTM topic models Θ is restored. This experiment does not imply any regularization so LDA model is not considered.

The next step is the following. According to the generation procedure the assignment of documents into several classes is already known. Denote this splitting by y. The goal is to understand how the considered models restore structure of matrix Θ. It can be managed through solving document classification problem using restored topics distribution $p(t\,|\,d)$ in a document as features of this document. The prediction $\hat{y}$ is constructed by Logistic Regression without tuning its parameters using 5-fold cross-validation. The quality of Θ matrix structure reconstruction is measured as accuracy of the document classification problem solution:

$$\text{Accuracy} = \frac{1}{|D|} \sum_{i=1}^{|D|} [y_i = \hat{y}_i] \,. \tag{50}$$

4.1.3. Experiment 1: Restoring matrix Θ

The goal of this experiment is to confirm that the proposed TransARTM model can restore the initial complex structure of transaction data.

In this experiment generation parameters are the following: number of topics $|T|$=50 where 3 out of them are background topics, $|D| = 5000$, $|M| = 3$, $|K| = 9$ where $h(k) \leqslant 4$ for all $k \in K$, number of classes is equal to 5, sparsity equals to 0.65 and number of other vertices is equal to 1000. It means that matrix $\Theta \in \mathbb{R}^{50\times5000}$ and matrices $\Phi_k \in \mathbb{R}^{1000\times50}$ for all $k \in K$. Using the generation procedure 3 about 13.5 million transactions were synthesized.

According to the experimental setup (see 4.1.2) on each iteration of EM-algorithm the quality of Θ matrix structure reconstruction is measured for all considered topic models. The total number of the iterations equals to 100. For each model 5 different random initializations are used. The number of topics for restored matrix Θ was the same as for generation. The results are demonstrated in Figure 3, the main curve is the mean among all initializations.

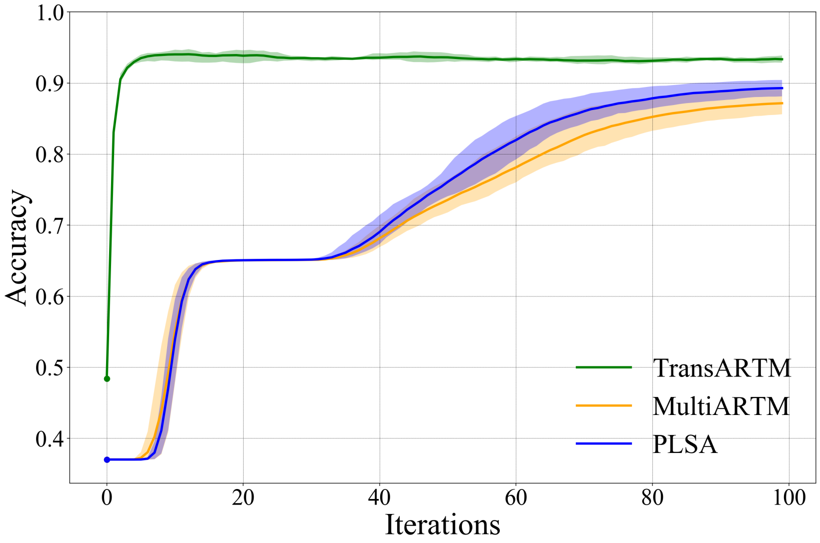


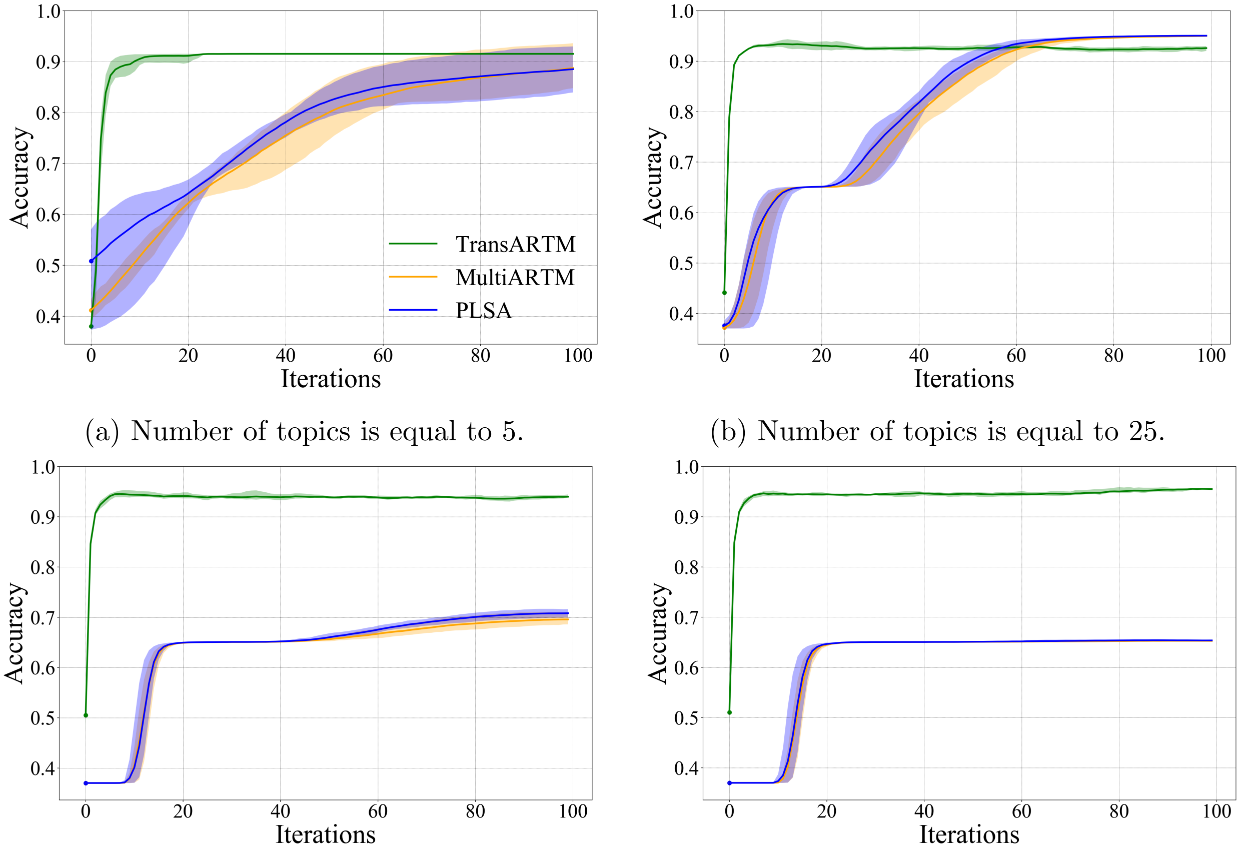
Figure 3: The number of topics is the same as specified during generation.

Conclusion. From this experiment it can be concluded that the proposed hypergraphic multimodal topic model (TransARTM) achieves high quality faster than other compared models on the simulated transaction data.

4.1.4. Experiment 2: Varying number of topics

The goal of this experiment is to evaluate stability of the proposed model with respect to initialization and the number of topics of matrix Θ being reconstructed.

This experiment is conducted on the same data as the experiment 4.1.3. All experimental setups are also the same except the number of topics of matrix Θ being reconstructed. It varies from 5 to 100. The results for different number of topics are represented in Figure 4.



(a) Number of topics is equal to 5.

(b) Number of topics is equal to 25.

(c) Number of topics is equal to 75.

(d) Number of topics is equal to 100.

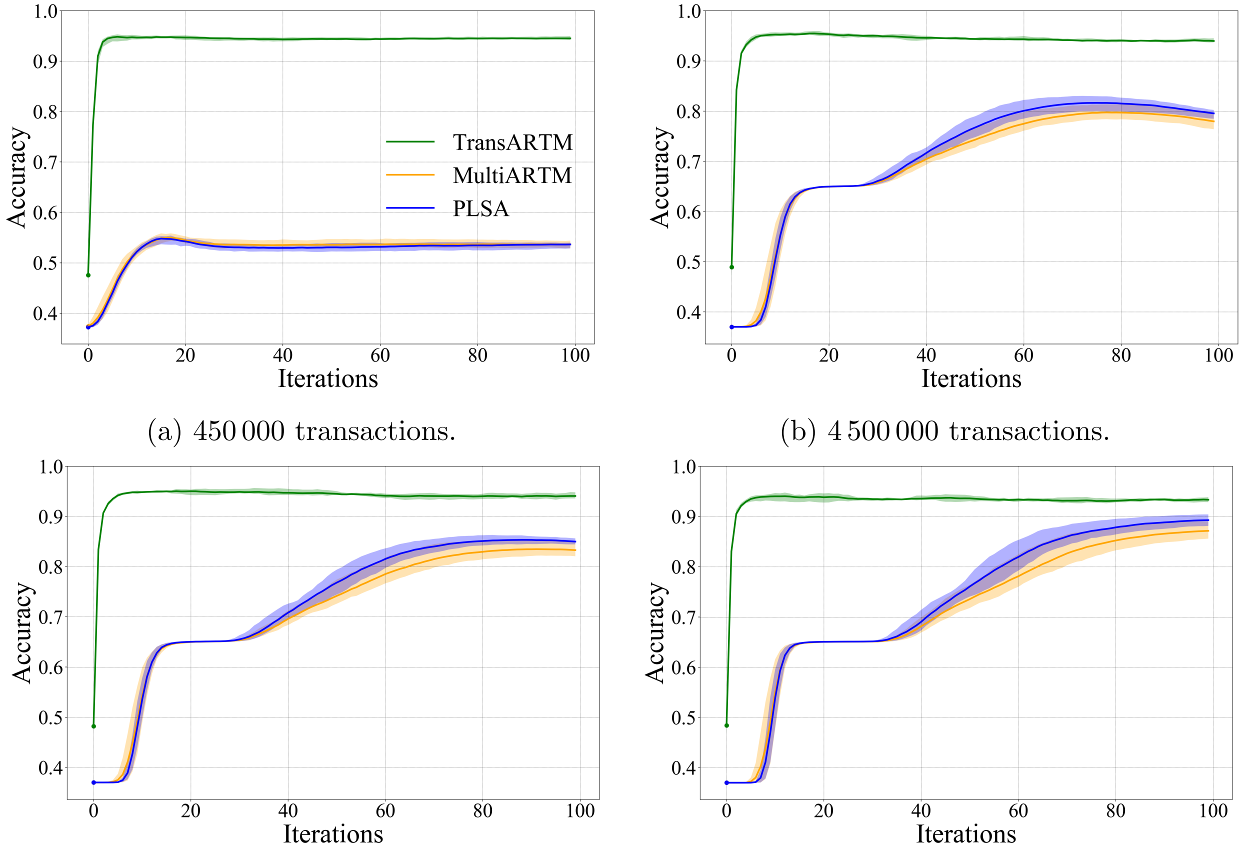
Figure 4: The number of topics varies from 5 to 100.

Conclusion. TransARTM is the most stable model with respect to initialization and selection of the number of topics.

4.1.5. Experiment 3: Varying data size

The goal of this experiment is to analyze how quality of reconstruction of matrix Θ depends on the size of transaction data.

This experiment also uses the same generated matrices Θ and Φ_k as for the previous experiments. The experiment setup is also the same. Only the size of data varies from 450 thousand to 13.5 million transactions. The number of topics $|T| = 50$ is the same as specified during generation. The results are illustrated as a series of graphics in Figure 5.



(a) 450 000 transactions.

(b) 4 500 000 transactions.

(c) 6 750 000 transactions.

(d) 13 500 000 transactions.

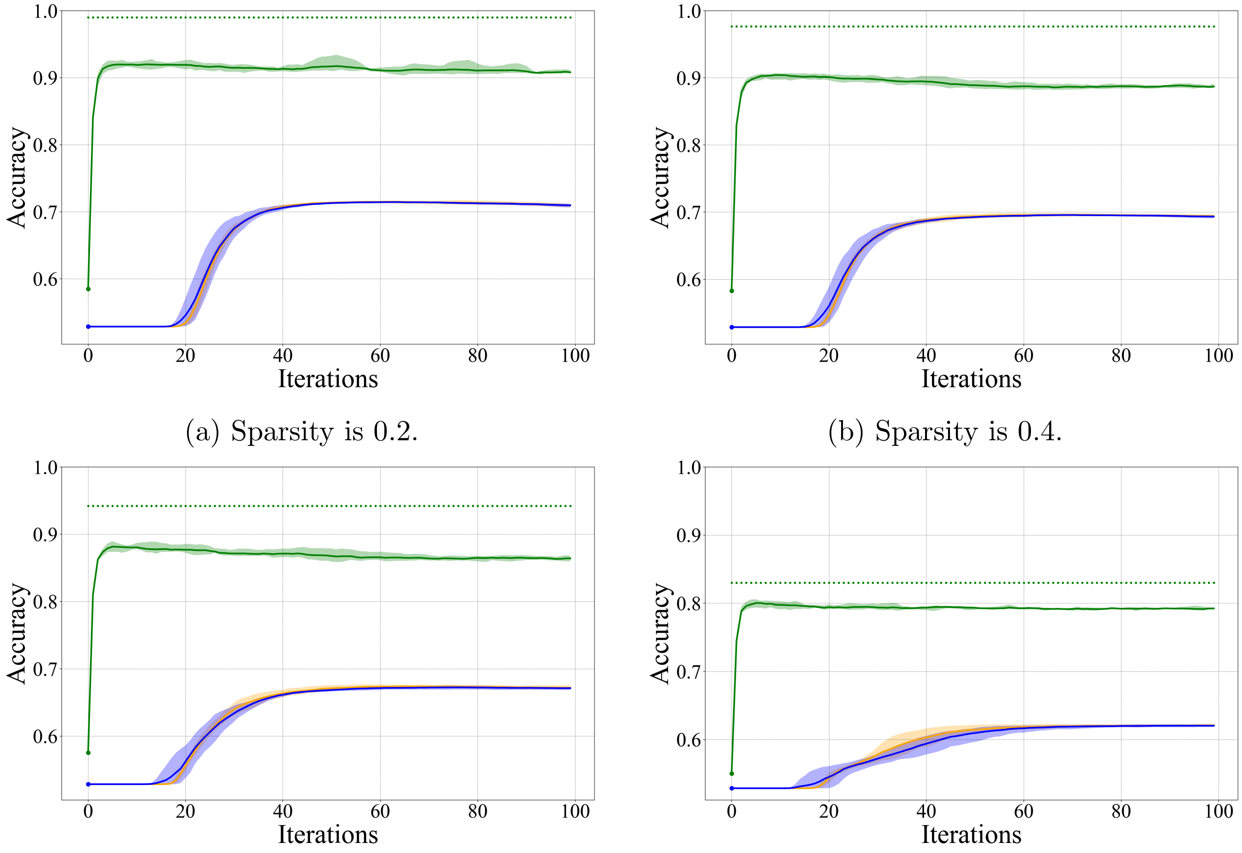
Figure 5: The number of transactions varies from 450 000 to 13 500 000.

Conclusion. From these experiments it can be concluded that the proposed model TransARTM comprehends the initial structure of Θ matrix even with the small amount of data. On the contrast, quality of reconstruction for other compared models depends on data size. As expected because of larger data size more accurate frequency estimations and differences between classes are achieved.

4.1.6. Experiment 4: Varying sparsity

The goal of this experiment is to analyze how quality of the reconstruction of matrix Θ depends on sparsity of matrices Θ and Φ_k for all $k \in K$.

This experiment uses the same generation parameters as for experiment 4.1.3 except sparsity parameter that varies from 0.2 to 0.8. The experiment setup is also the same. The number of transactions is equal to 6.75 million. Number of topics $|T| = 50$. The results one can see in Figure 6. The green dotted graph represents classification accuracy on the ground truth matrix Θ.



(a) Sparsity is 0.2.

(b) Sparsity is 0.4.

(c) Sparsity is 0.6.

(d) Sparsity is 0.8.

Figure 6: The sparsity of Θ matrix varies from 0.2 to 0.8.

Conclusion. Hypergraphic multimodal topic model TransARTM shows quality close to ground truth both at high and low sparsity. The overall quality of reconstruction decreases with an increase of the number of zero elements in Θ matrix.

4.2. Real Data

This section is devoted to the experiments on real data. For these experiments Million Playlist Dataset (MPD) is used. The aim is to apply the TransARTM model to music track recommendation task.

4.2.1. Description

MPD is an array of 1 000 000 playlists. Each playlist is an ordered list each element of which contains tracks, albums and artists names. There are also some metadata for each playlist such as number of editing sessions or time stamp when the playlist was previously update, etc. All the experiments in this subsection use only track, album and artist representation of playlists.

For the experiments three different datasets are created: train, valid and test sets with the same distribution of the number of tracks. Each playlist contains at least 100 but no more than 200 tracks. For the playlists from valid and test datasets the last 70 tracks are removed to evaluate the quality of topic models used in these experiments. All tracks in test and valid datasets as well as all holdout tracks appear in the train dataset. The main characteristics of the whole dataset and divided sets one can see in Table 1.

Table 1: Different characteristics of MPD and divided sets.

Number of	MPD	Train	Test	Valid
Playlists	1 000 000	100 000	1 000	1 000
Tracks	66 346 428	9 875 306	232 613	232 808
Unique tracks	2 262 292	296 882	39 368	38 641
Unique albums	734 684	140 983	20 690	20 483
Unique artists	295 860	69 280	10 081	10 008

The goal is to recommend tracks for each of playlists from test dataset.

4.2.2. Metrics

To evaluate solutions the following metrics are used. Denote the ground truth set of objects G and the ordered predicted list of objects R of size k. Consider the function $r^{rel} : R \leftarrow \{0, 1\}$ that equals to 1 if some object from R are in the set G. Used metrics can be defined according to these notations.

Precision is the number of relevant objects divided by the number of predicted objects:

$$\text{precision@k} = \frac{|G \cap R|}{|R|}. \tag{51}$$

Recall is the number of relevant objects divided by the number of known relevant objects:

$$\text{recall@k} = \frac{|G \cap R|}{|G|}. \tag{52}$$

F-measure is the harmonic mean of precision and recall:

$$\text{fscore@k} = 2 \cdot \frac{\text{recall@k} \cdot \text{precision@k}}{\text{recall@k} + \text{precision@k}}. \tag{53}$$

These metrics reflect total number of retrieved relevant tracks regardless of order.

Normalized discounted cumulative gain (NDCG) measures ranking quality of predicted objects. It increases when relevant objects are placed higher in list R.

NDCG is equal to DCG divided by the ideal DCG:

$$\text{ndcg@k} = \frac{\text{dcg@k}}{\text{idcg}} = \sum_{i=1}^{|R|} \frac{r^{rel}(R_i)}{\log_2(i+1)} \Bigg/ \sum_{i=1}^{|G|} \frac{1}{\log_2(i+1)}. \tag{54}$$

All described above metrics are averaged across all playlists in test set.

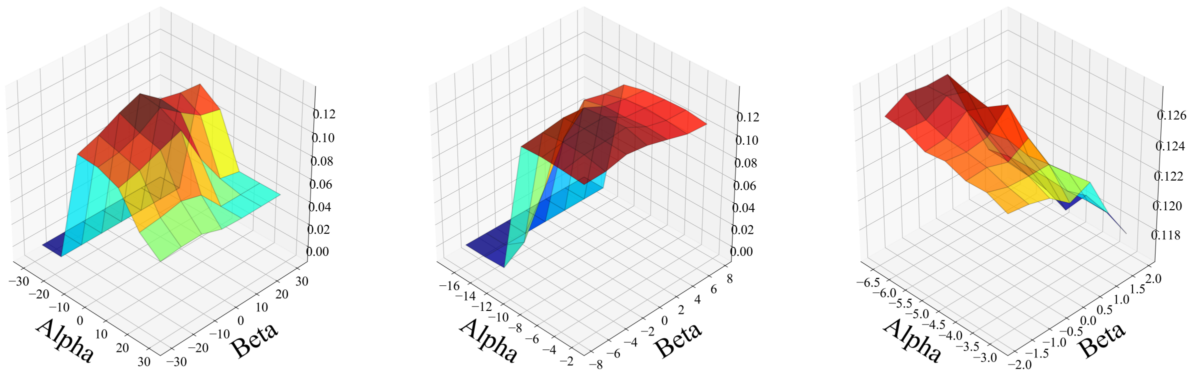
4.2.3. Experimental Setup

In experiments the PLSA, LDA, multimodal and the proposed hypergraphic multimodal topic models with smoothing and sparsing regularizers from 3.5 are used. This regularizers applied to the whole matrix Θ with coefficient α and to track modality of Φ matrix with coefficient β. In LDA and PLSA models only one modality is considered, so the regularizers are applied to the whole matrices Θ and Φ. For multimodal and hypergraphic multimodal topic models different interactions between artist, album and tracks modalities are considered.

For each topic model parameters α and β are tuned using valid dataset. This procedure is repeated for different number of topics. The metrics are calculated for different number of predicted tracks that is from 70 to 500. The ground truth sets of tracks is obtained by removing the last 70 tracks of each playlist in test and valid datasets.

4.2.4. Parameters tuning

Optimal coefficients of the regularizers α and β are selected for each topic model and number of topics on the valid dataset. Parameters are tuned by grid search during 5 steps over 5 × × 5 parameters. On each step current optimal parameters are determined as a center of grid for the next step. Size of a new grid is also decreased seven times. One can find an example of the first three steps of tuning procedure in Figure 7.



(a) Step 1.

(b) Step 2.

(c) Step 3.

Figure 7: The first three steps of tuning procedure.

4.2.5. Results

For all considered models experiment setup is the same as described in 4.2.3. The regularization coefficients are tuned in accordance with 4.2.4 for different number of topics separately. Number of topics varies from 50 to 2 000 while the number of predicted tracks for each playlist — from 70 to 500. All metrics are averaged across the holdout test dataset.

Baseline. The following baseline model is supposed to clarify the overall quality of topic modeling. The ordered list of tracks is calculated using the train dataset according to their popularity. An order of each track is determined by number of its occurrences in a dataset. Therefore, a track with high number of occurrences is on the top of the list. Then for each playlist in the test dataset according to the list of tracks popularity a list of recommended tracks is predicted. The tracks that are already in the playlist are ignored during prediction. The number of the retrieved tracks varies from 70 to 500. Finally, metrics are calculated and averaged across all playlists. The results of baseline model are presented in Table2.

Table 2: The results of baseline model (TopTracks) for different number of predicted tracks.

Number of	**Metrics**			
predicted tracks	**precision**	**recall**	**fscore**	**ndcg**
70	0.0425	0.0425	0.0425	0.0479
100	0.0387	0.0553	0.0455	0.0565
300	0.0268	0.1149	0.0435	0.0905
500	0.0230	0.1646	0.0404	0.1152

PLSA and LDA. These topic models describe the interaction between document and terms. In the examined case documents are presented by playlists, and terms correspond to tracks. A set of playlists that contain tracks is supposed as a dataset. The results are introduced in Table 3 and Table 4 for PLSA and LDA models respectively. The best values of each metric are highlighted by bold for varying number of topics.

Table 3: The results of PLSA topic model for different number of predicted tracks and number of topics.

	Number of topics								
	50	**100**	**150**	**300**	**500**	**750**	**1 000**	**1 500**	**2 000**
precision@70	0.1208	0.1226	0.1230	0.1211	**0.1247**	0.1221	0.1235	0.1203	0.1247
precision@100	0.1090	0.1109	0.1118	0.1105	0.1133	0.1114	**0.1134**	0.1097	0.1126
precision@300	0.0735	0.0747	0.0753	0.0751	0.0750	0.0746	**0.0762**	0.0748	0.0761
precision@500	0.0572	0.0585	0.0586	0.0583	0.0583	0.0583	**0.0592**	0.0583	0.0591
recall@70	0.1208	0.1226	0.1230	0.1211	**0.1247**	0.1221	0.1235	0.1203	0.1247
recall@100	0.1558	0.1584	0.1598	0.1578	0.1619	0.1591	**0.1620**	0.1567	0.1609
recall@300	0.3150	0.3203	0.3226	0.3219	0.3216	0.3199	**0.3264**	0.3204	0.3261
recall@500	0.4084	0.4178	0.4188	0.4163	0.4164	0.4162	**0.4228**	0.4166	0.4218
fscore@70	0.1208	0.1226	0.1230	0.1211	**0.1247**	0.1221	0.1235	0.1203	0.1247
fscore@100	0.1283	0.1304	0.1316	0.1300	0.1333	0.1310	**0.1334**	0.1290	0.1325
fscore@300	0.1192	0.1212	0.1221	0.1218	0.1217	0.1210	**0.1235**	0.1212	0.1234
fscore@500	0.1003	0.1026	0.1029	0.1022	0.1023	0.1022	**0.1038**	0.1023	0.1036
ndcg@70	0.1319	0.1346	0.1334	0.1324	**0.1354**	0.1333	0.1344	0.1314	0.1324
ndcg@100	0.1553	0.1585	0.1579	0.1569	0.1602	0.1580	**0.1602**	0.1557	0.1593
ndcg@300	0.2465	0.2514	0.2512	0.2510	0.2518	0.2502	**0.2545**	0.2496	0.2531
ndcg@500	0.2931	0.3001	0.2992	0.2981	0.2991	0.2982	**0.3025**	0.2975	0.3024

Multimodal topic model. This model describes pairwise interactions between documents and objects of different modalities. For the given task documents also are presented by playlists. Artist, album and track are considered as modalities. These experiments are aimed to analyze three models that are defined by used modalities. The results of two conducted experiments are presented in Table 5.

Table 4: The results of LDA topic model for different number of predicted tracks and number of topics.

	Number of topics								
	50	**100**	**150**	**300**	**500**	**750**	**1 000**	**1 500**	**2 000**
precision@70	0.1205	0.1214	**0.1223**	0.1223	0.1131	0.1212	0.1198	0.0613	0.0426
precision@100	0.1090	0.1101	0.1108	**0.1116**	0.1031	0.1103	0.1094	0.0554	0.0387
precision@300	0.0735	**0.0747**	0.0746	0.0746	0.0704	0.0738	0.0729	0.0376	0.0268
precision@500	0.0573	**0.0583**	0.0579	0.0583	0.0554	0.0576	0.0571	0.0305	0.0231
recall@70	0.1205	0.1214	**0.1223**	0.1223	0.1131	0.1212	0.1198	0.0613	0.0426
recall@100	0.1557	0.1572	0.1582	**0.1594**	0.1473	0.1576	0.1563	0.0792	0.0553
recall@300	0.3151	**0.3202**	0.3198	0.3199	0.3017	0.3163	0.3126	0.1612	0.1151
recall@500	0.4095	**0.4163**	0.4137	0.4162	0.3958	0.4118	0.4082	0.2177	0.1648
fscore@70	0.1205	0.1214	**0.1223**	0.1223	0.1131	0.1212	0.1198	0.0613	0.0426
fscore@100	0.1282	0.1295	0.1303	**0.1313**	0.1213	0.1298	0.1287	0.0652	0.0456
fscore@300	0.1192	**0.1211**	0.1210	0.1210	0.1142	0.1197	0.1183	0.0610	0.0435
fscore@500	0.1006	**0.1022**	0.1016	0.1022	0.0972	0.1011	0.1003	0.0535	0.0405
ndcg@70	0.1320	0.1320	0.1333	**0.1338**	0.1238	0.1314	0.1301	0.0685	0.0480
ndcg@100	0.1555	0.1560	0.1573	**0.1587**	0.1467	0.1558	0.1545	0.0805	0.0565
ndcg@300	0.2469	0.2495	0.2499	**0.2508**	0.2352	0.2468	0.2442	0.1274	0.0906
ndcg@500	0.2939	0.2974	0.2968	**0.2988**	0.2821	0.2944	0.2919	0.1556	0.1153

In the first experiment different multimodal topic models with various combinations of modalities are compared. Considered combinations are the following: track and album, track and artist, track and album and artist. Number of topics is equal to 750. The model that uses track and artist modalities shows the best results.

The second experiment uses the best model from the first experiment. Number of topics varies from 500 to 2 000. A further increase of the number of topics does not improve results.

Hypergraphic multimodal topic model. TransARTM proposed in this research describes interactions between any number of objects. For the examined dataset containers (documents) are also represented by playlists. Artist, album and track are considered as modalities. These experiments are aimed to analyze four models defined by interacting objects (transactions). The results of two conducted experiments are presented in Table 6.

In the first experiment different hypergraphic multimodal topic models with various types of transactions are compared. Considered transactions are following: playlist – album – track, playlist – artist – track, playlist – album – track, playlist – artist – track, playlist – track – album – artist. Number of topics is equal to 750. The model considering playlist – artist – track interaction shows the best results.

The second experiment uses the best model from the first experiment. Number of topics varies from 500 to 2 000. A further increase of number of topics does not improve results.

Table 5: The resulting scores of multimodal topic models for different combinations of modalities (Track, Album, Artist) for 750 topics, and scores of the best model with track and artist modalities for different number of topics.

	Modalities			**Number of topics**				
	Al Tr	**Ar Tr**	**Ar Al Tr**	**500**	**750**	**1 000**	**1 500**	**2 000**
precision@70	0.1243	**0.1290**	0.1260	0.1273	**0.1290**	0.1264	0.1264	0.1260
precision@100	0.1139	**0.1171**	0.1150	0.1159	**0.1171**	0.1152	0.1155	0.1151
precision@300	0.0766	**0.0782**	0.0774	0.0775	**0.0782**	0.0781	0.0782	0.0774
precision@500	0.0594	**0.0608**	0.0601	0.0603	**0.0608**	0.0608	0.0606	0.0605
recall@70	0.1243	**0.1290**	0.1260	0.1273	**0.1290**	0.1264	0.1264	0.1260
recall@100	0.1627	**0.1673**	0.1643	0.1656	**0.1673**	0.1646	0.1649	0.1644
recall@300	0.3281	**0.3353**	0.3317	0.3319	**0.3353**	0.3348	0.3352	0.3317
recall@500	0.4245	**0.4343**	0.4292	0.4308	**0.4343**	0.4343	0.4328	0.4323
fscore@70	0.1243	**0.1290**	0.1260	0.1273	**0.1290**	0.1264	0.1264	0.1260
fscore@100	0.1340	**0.1378**	0.1353	0.1364	**0.1378**	0.1356	0.1358	0.1354
fscore@300	0.1242	**0.1269**	0.1255	0.1256	**0.1269**	0.1267	0.1268	0.1255
fscore@500	0.1043	**0.1067**	0.1054	0.1058	**0.1067**	0.1067	0.1063	0.1062
ndcg@70	0.1343	**0.1394**	0.1364	0.1375	**0.1394**	0.1367	0.1365	0.1358
ndcg@100	0.1600	**0.1651**	0.1620	0.1631	**0.1651**	0.1622	0.1622	0.1615
ndcg@300	0.2548	**0.2616**	0.2580	0.2585	**0.2616**	0.2599	0.2599	0.2574
ndcg@500	0.3029	**0.3110**	0.3066	0.3078	**0.3110**	0.3095	0.3085	0.3076

Table 6: The resulting scores of TransARTM models for different types of transactions (Track, Album, Artist, the playlist is omitted) for 750 topics, and scores of the model considered playlist – artist – track interaction for different number of topics.

	Transactions				**Number of topics**				
	Al Tr	**Ar Tr**	**Ar Tr** **Al Tr**	**Ar Al Tr**	**500**	**750**	**1 000**	**1 500**	**2 000**
precision@70	0.0791	**0.1014**	0.1005	0.0973	0.1020	0.1014	**0.1044**	0.1023	0.0884
precision@100	0.0729	**0.0922**	0.0913	0.0883	0.0924	0.0922	**0.0949**	0.0935	0.0814
precision@300	0.0495	**0.0633**	0.0629	0.0593	0.0623	0.0633	0.0641	**0.0641**	0.0571
precision@500	0.0396	0.0495	**0.0499**	0.0465	0.0490	0.0495	0.0500	**0.0504**	0.0458
recall@70	0.0791	**0.1014**	0.1005	0.0973	0.1020	0.1014	**0.1044**	0.1023	0.0884
recall@100	0.1041	**0.1317**	0.1304	0.1262	0.1320	0.1317	**0.1356**	0.1335	0.1162
recall@300	0.2120	**0.2712**	0.2697	0.2539	0.2672	0.2712	0.2746	**0.2746**	0.2445
recall@500	0.2826	0.3534	**0.3567**	0.3322	0.3501	0.3534	0.3571	**0.3603**	0.3272
fscore@70	0.0791	**0.1014**	0.1005	0.0973	0.1020	0.1014	**0.1044**	0.1023	0.0884
fscore@100	0.0857	**0.1084**	0.1074	0.1039	0.1087	0.1084	**0.1117**	0.1099	0.0957
fscore@300	0.0802	**0.1026**	0.1021	0.0961	0.1011	0.1026	0.1039	**0.1039**	0.0925
fscore@500	0.0694	0.0868	**0.0876**	0.0816	0.0860	0.0868	0.0877	**0.0885**	0.0804
ndcg@70	0.0862	**0.1120**	0.1095	0.1063	0.1129	0.1120	**0.1152**	0.1110	0.0972
ndcg@100	0.1029	**0.1322**	0.1295	0.1257	0.1329	0.1322	**0.1361**	0.1318	0.1158
ndcg@300	0.1647	**0.2122**	0.2094	0.1989	0.2104	0.2122	**0.2158**	0.2128	0.1893
ndcg@500	0.1999	**0.2532**	0.2528	0.2379	0.2517	0.2532	**0.2570**	0.2555	0.2305

Putting it all together. PLSA and LDA models consider pairwise interactions between playlists and tracks. Multimodal topic model allows to describe several pairwise interactions separately within one model. TransARTM model takes into account interactions between more than two object. It is important to note the all compared models are special cases of TransARTM that was proved in 3.4. Gathering the best results of all considered models for 500 predicted tracks oen can find the summary in Table 7.

Table 7: The best results of all considered models for 500 predicted tracks.

Model	**Considered iteractions**	**Metrics, @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

To make sure that predicted topics make sence one can pay attention to the top-10 artists constituents of several topics that are presented in Table 8. This representations are obtained using TransARTM models with number of topics equal to 750 and considered interaction between playlist, artist and track.

Table 8: Representation of five different topics by its top-10 artists (descending order).

Linkin Park	Nicki Minaj	Lil Jon	The Beatles	Guns N' Roses
3 Doors Down	Beyonce	50 Cent	John Lennon	Bon Jovi
Evanescence	Rihanna	Snoop Dogg	George Harrison	AC/DC
Nickelback	Tinashe	J-Kwon	The Beach Boys	Def Leppard
Hinder	Omarion	Nelly	Elvis Presley	Ozzy Osbourne
Papa Roach	Jeremih	Usher	Paul McCartney	Journey
Hoobastank	Trey Songz	Kanye West	David Bowie	Aerosmith
Creed	Chris Brown	R. Kelly	Jim Sturgess	Scorpions
Daughtry	Big Sean	Youngbloodz	The Mamas & The Papas	Metallica
Finger Eleven	Sage The Gemini	Bubba Sparxxx	The Turtles	Survivor

It can be concluded that topic modeling approach for the problem of playlists extension improves overall results. The best multimodal topic model uses combination of track and artist modalities. It implies that users tend to listen tracks by several artists they like but not necessary from particular albums. The proposed TransARTM model shows comparable results that are still slightly lower. It can be explained by the fact that artist, album and track are linked hierarchically that means they are not truly independent.

5. CONCLUSION

In this research hypergraphic multimodal topic model called TransARTM has been proposed. This model generalizes currently existing topic models of matrix factorization to the case when original data can be represented as a hypergraph. TransARTM allows to describe more complex relationships between objects than pairwise interactions. It has been shown that conventional topic modeling approaches PLSA, LDA and multimodal topic models are actually subcases of the developed topic model. The proposed extension has been implemented as a part of BigARTM open source project.

The experiments have been carried out both on simulated transaction data and real data. The results on simulated transaction data have shown that the proposed model which takes into account relationships of any number of objects tends to converge faster than other methods to the best solution even with a relatively small number of data. Also the stability with respect to the number of topics has been investigated comparing with other models in case of sparse ground truth matrix Θ. Application of multimodal and hypergraphic multimodal models for the construction of recommendation systems has been demonstrated on real data.

Further experiments are supposed to use transaction data from financial organizations that is not convenient for the current research due to small sizes of freely distributed financial datasets. The proposed model is supposed to give a general understanding of structure of financial flows within the industry.

REFERENCES

[1] Jianwen Zhang, Yangqiu Song, Changshui Zhang, and Shixia Liu. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1079–1088. ACM, 2010.

[2] Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE transactions on visualization and computer graphics*, 17(12):2412–2421, 2011.

[3] Ivan Vulić, Wim De Smet, and Marie-Francine Moens. Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval*, 16(3):331–368, 2013.

[4] Wayne Xin Zhao, Jinpeng Wang, Yulan He, Jian-Yun Nie, Ji-Rong Wen, and Xiaoming Li. Incorporating social role theory into topic models for social media content analysis. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):1032–1044, 2015.

[5] Devesh Varshney, Sandeep Kumar, and Vineet Gupta. Modeling information diffusion in social networks using latent topic information. In *International Conference on Intelligent Computing*, pages 137–148. Springer, 2014.

[6] Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine learning*, 88(1-2):157–208, 2012.

[7] Hongning Wang, Duo Zhang, and ChengXiang Zhai. Structural topic model for latent topical structure analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1526–1535. Association for Computational Linguistics, 2011.

[8] Ali Daud, Juanzi Li, Lizhu Zhou, and Faqir Muhammad. Knowledge discovery through directed probabilistic topic models: a survey. *Frontiers of computer science in China*, 4(2):280–301, 2010.

[9] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[11] Hongzhi Yin, Bin Cui, Yizhou Sun, Zhiting Hu, and Ling Chen. Lcars: A spatial item recommender system. *ACM Transactions on Information Systems (TOIS)*, 32(3):11, 2014.

[12] Hongzhi Yin, Bin Cui, Ling Chen, Zhiting Hu, and Chengqi Zhang. Modeling location-based user rating profiles for personalized recommendation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 9(3):19, 2015.

[13] Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer, 2011.

[14] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *Proceedings of the 17th international conference on World Wide Web*, pages 101–110. ACM, 2008.

[15] Sang Su Lee, Tagyoung Chung, and Dennis McLeod. Dynamic item recommendation by topic modeling for social networks. In *Information Technology: New Generations (ITNG), 2011 Eighth International Conference on*, pages 884–889. IEEE, 2011.

[16] Hongzhi Yin, Bin Cui, Yizhou Sun, Zhiting Hu, and Ling Chen. Lcars: A spatial item recommender system. *ACM Transactions on Information Systems (TOIS)*, 32(3):11, 2014.

[17] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, Marina Suvorova, and Anastasia Yanina. Non-bayesian additive regularization for multimodal topic modeling of large collections. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, pages 29–37. ACM, 2015.

[18] Lei Li and Tao Li. News recommendation via hypergraph learning: encapsulation of user behavior and news content. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 305–314. ACM, 2013.

[19] Jiajun Bu, Shulong Tan, Chun Chen, Can Wang, Hao Wu, Lijun Zhang, and Xiaofei He. Music recommendation by unified hypergraph: combining social media information and music content. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 391–400. ACM, 2010.

[20] Denny Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in neural information processing systems*, pages 1601–1608, 2007.

[21] Konstantin Vorontsov. Additive regularization for topic models of text collections. In *Doklady Mathematics*, volume 89, pages 301–304. Springer, 2014.

[22] Konstantin Vorontsov and Anna Potapenko. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. In *International Conference on Analysis of Images, Social Networks and Texts_x000D_*, pages 29–46. Springer, 2014.

[23] Konstantin Vorontsov, Anna Potapenko, and Alexander Plavin. Additive regularization of topic models for topic selection and sparse factorization. In *International Symposium on Statistical Learning and Data Sciences*, pages 193–202. Springer, 2015.

[24] Murat Apishev, Sergei Koltcov, Olessia Koltsova, Sergey Nikolenko, and Konstantin Vorontsov. Additive regularization for topic modeling in sociological studies of user-generated texts. In *Mexican International Conference on Artificial Intelligence*, pages 169–184. Springer, 2016.

[25] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1-2):177–196, 2001.

[26] Mark Girolami and Ata Kabán. On an equivalence between plsi and lda. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 433–434. ACM, 2003.

[27] Konstantin Vorontsov and Anna Potapenko. Additive regularization of topic models. *Machine Learning*, 101(1-3):303–323, 2015.