Topic Modelling for Extracting Behavioral Patterns from Transactions Data

Evgeny Egorov*, Filipp Nikitin†, Vasiliy Alekseev‡, Alexey Goncharov§, Konstantin Vorontsov¶,
Moscow Institute of Physics and Technology
Moscow, Russian Federation
Email: *egorov.eo@mipt.ru, †filipp.nikitin@phystech.edu, ‡vasiliy.alekseyev@phystech.edu, §alex.goncharov@phystech.edu
and ¶k.v.vorontsov@phystech.edu

***Abstract*—With the increasing popularity of cashless payment methods for everyday, seasonal and special expenses popular banks accumulate huge amount of data about customer operations. In the article, we report a successful application of topic modelling to extract behaviour patterns from the data. The resulting models are built with BigARTM framework: flexible and efficient tool for topic modelling. The framework allows us to experiment with various models including PLSA, LDA and beyond.**

Results demonstrate ability of the approach to aggregate information about behaviour patterns of different customer groups. The results analysis allows to see the topics of such people clusters varying from travellers to mortgage holders. Moreover, low-dementional embeddings of the customers, which was given with topic model, were studied. We display that the client vector representations store demographic information as well as source data. We also test for a best way of preparing data for the model with metric above in mind.

***Index Terms*—Topic Modelling, Transactions, BigARTM, Additive Regularization**

I. Introduction

In our days' people make a huge amount of transactions in their day-to-day life. Therefore, banks collecting loads of information about their clients' transaction history daily. Application of such information could include but not limited to improving their services, attracting new clients, increasing income, developing of pricing strategies, predicting loan return probabilities and enabling the better understanding of the client's needs. In the last few years, machine learning algorithms succeeded in solving different tasks and have the potential to help in others [12]. The transaction analysis is among those problems because the transaction data is big and unstructured, therefore ML helps to draw some insight from it.

Many solutions in the field have been developed using ML techniques [6], [8], [13]. Mostly, applications from the field concentrate on the areas like consumer profiling [3], [15], [20], assessment of buying patterns and purchase predictions [23], or discovering client communities in the data for better product campaign clustering [19]. While customer profiling can be performed by various clustering algorithms as tried [5], the result of such endeavours is usually quite coarse and poorly interpretable for such sensitive field as Banking. Buying patterns and purchase prediction is a widely discussed topic because of its direct applications. However, several attempts were made trying to tie together psychological user-profiles and their transaction data with the help of LDA mode [7] or helping marketing campaigns with predicting client consumption with neural networks and random forests [14].

The approach solving all these problems at once involves the construction of a vector space of client embeddings. Some authors ventured that path [3] using autoencoders. In the article, we focus on constructing a low-dimensional vector space with topic modelling. Topic modelling is a powerful tool which was successfully applied to various ML problems [18]. Originally, the approach was invented as a method of natural language processing [4], [9]. Nowadays it is applied to a non-trivial data such as analysis of weblogs [15], [16], mining a behaviour pattern from video [10]. We reporting similar attempt on an individual transaction data from a bank.

In our experiments, we use a technique called Additive Regularization of Topic Models (ARTM) [22]. We use the technique because it includes many popular topic models such as PLSA [9], LDA [4] and their probabilistic and non-probabilistic generalisation. However, the technique is implemented in fast and efficient library BigARTM [21]. The library was successfully applied to different problems [2]. Another feature of ARTM is the ability to use multimodal data [11]. The fact is really important in our case because transaction information consists of different data about the client: the amount of money, merchant category code (MCC), gender, age.

Overall, we adopt the latter approach for a number of reasons: topic model is not a black box solution and can be easily interpreted, our approach to topic modelling allows us to satisfy multiple criteria posed by a business to an ML solution, topic model itself is fast in inference and training solution suitable for practical applications.

The rest of this paper is organised as follows. In the following section, we consider related theory and display its mathematical part. In the third section, we give information about the used dataset and data reprocessing. The third section provides experimental results. The last section concludes the results and discusses the contribution of this paper.

II. Theory

A. Problem Formulation

In natural language processing the researcher deals with a collection of *documents* composed of sequences of words or *tokens* as we would relate to them later. When dealing with large document collections it becomes useful to cluster the documents by their meaning. This type of clustering is called *topic modelling* and it constructs a hidden dimension of *topics* that provide a **short** description for each document in the collection. In bank transactions we are dealing with a collection of *clients transactions history* composed of sequences of transactions described by their date, MCC code and sum spent on that code. Thus, we could read the clients like a book, by applying a topic model to their transactions history. The result of our efforts would be a latent *embedding space* which represents the types of consumption derived from the statistics of the transactions data. The *topics* provide the representation of any client through breaking them into a selection of the consumption types describing them in an **interpretable** way.

B. Topic Modelling

Topic modelling was initially designed to work on the large collections of documents.

Since then, it was implemented to any sequential data: being it purchase baskets at a grocery store, web behaviour, bank transactions. However distant these fields seem to be, they all fall in line with hypotheses imposed by Topic modelling.

After agreeing on which entities in the collection should be treated as tokens which compose documents we can apply the topic modelling formalism to the data.

If we treat a transaction MCC code as a token with token frequency as sum spent on that code in the document which is represented by a client's transaction history.

Or more formally, by denoting a collection of clients transactions as D with individual client transaction history as $d \in D$ and a collection of all possible transactions in D as W with every transaction being denoted as $w \in W$ we can formulate the hypotheses as:

- **Themes existence hypothesis.** Every token w entry in the document d is determined by existence of some topic t from some finite set T. These topics t are latent (hidden) variables of our approach.
- **Bag of words hypothesis.** This hypothesis states that an order of tokens in the document is not essential for topic retrieval. Restating that for the actual use cases: the order of products on the grocery store bill won't change the way the bill is categorized by the model. As an extension of that hypothesis, we can assume "the bag of documents" that states that our topics shouldn't change from the order in which we feed the documents into the model.
- **Probabilistic generative model hypothesis.** This hypothesis states that the observed collection is generated from unknown distributions $p(t\,|\,d)$ and $p(w\,|\,t)$. Finding these distributions is the goal of the topic modelling.
- **Conditional independence hypothesis.** According to that hypothesis, each topic generates tokens regardless of the document $p(w\,|\,t) = p(w\,|\,t, d)$. In a nutshell, this implies that each token describing a client appeared there due to a certain pattern shared among many clients rather than due to this particular client.

One could see that all of the previously stated hypotheses are fulfilled in that setting.

Thus, topic modelling can provide us with knowledge of the bank client base without aggregating too much unnecessary details on each of the clients.

According to the law of total probability and the assumption of conditional independence,

$$p(w\,|\,d) = \sum_{t \in T} p(t\,|\,d)\, p(w\,|\,t). \tag{1}$$

The probabilistic model (1) describes how the collection D is generated from the known distributions $p(t\,|\,d)$ and $p(w\,|\,t)$.

Learning a topic model is an inverse problem, i.e., the distributions $p(t\,|\,d)$ and $p(w\,|\,t)$ must be found, given the collection.

This problem is equivalent to finding an approximate representation of the matrix of counts $F = \big(\hat{p}(w\,|\,d)\big)_{W \times D}$, $\hat{p}(w\,|\,d) = \frac{n_{dw}}{n_d}$, as a product $F \approx \Phi\Theta$ of two unknown matrices — the matrix Φ of *term probabilities for the topics* and the matrix Θ of *topic probabilities for the documents*:

$$\Phi = (\phi_{wt})_{W \times T}, \qquad \phi_{wt} = p(w\,|\,t);$$
$$\Theta = (\theta_{td})_{T \times D}, \qquad \theta_{td} = p(t\,|\,d).$$

Matrices F, Φ, and Θ are *probability matrices*, i.e., they have non-negative and normalized columns f_d, ϕ_t, and θ_d, respectively, representing discrete distributions. Usually the number of topics $|T|$ is much smaller than the collection size $|D|$ and the vocabulary size $|W|$.

The topic model reveals a hidden thematic structure of the collection and finds a decomposition for each document by a set of its topics.

In the fundamental paper *probabilistic latent semantic analysis*, PLSA [9], the topic model (1) is learned by log-likelihood maximization with linear constraints.

The *likelihood* is the probability of the observed data as a function of model parameters Φ and Θ. Due to the independence assumption, the probability of the observed data is equivalent to the product of the probabilities of words in the documents:

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

Taking the logarithm, the expression above becomes a sum and the terms that don't depend on the model parameter can be dropped because they don't affect optimization.

We have a log-likelihood maximisation problem subject to the linear constraints of non-negativity and normalisation:

$$L(\Phi,\Theta)=\sum_{d\in D}\sum_{w\in d}n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td}\ \rightarrow\ \max_{\Phi,\Theta}; \quad (2)$$

$$\sum_{w\in W}\phi_{wt}=1;\ \phi_{wt}\geq 0; \quad \sum_{t\in T}\theta_{td}=1;\ \theta_{td}\geq 0. \quad (3)$$

We formulate KarushKuhnTucker conditions and use fixed-point iteration method in order to find a local minimum of the problem 2. Additive regularization of topic models (ARTM) is based on maximizing the log-likelihood (2) and a weighted sum of regularizers $R_i(\Phi,\Theta)$:

$$\sum_{d\in D}\sum_{w\in d}n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td}+\sum_{i=1}^{k}\tau_i R_i(\Phi,\Theta)\ \rightarrow\ \max_{\Phi,\Theta}; \quad (4)$$

subject to constraints (3), where the τ_i are non-negative regularisation coefficients.

The ability to impose additional functional restriction (regularisation) allows us to improve the quality of the model in various tasks. The approach provides us with an ability to formulate problem restrictions in mathematical form. Moreover, the most known topic models(PLSA, LDA) can be obtained within the ARTM framework.

In case of the multimodal data ARTM approach can be easily generalised by constructing objective function of weighted log-likelihoods (4).

$$\sum_{m,d}\sum_{w\in W_m}\tau_m n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td}+R(\Phi,\Theta)\rightarrow\max_{\Phi,\Theta}. \quad (5)$$

C. Classification based on client profile

The previously defined set $D=\{d_i\}_{i=1}^n$ a set of bank clients and lets define $Y=\{y_i\}_{i=1}^n$ the set of their target characteristics. We aim to build a model $f(w):D\rightarrow Y$ that uses the distribution $p(t|d)$ of clients profile as the initial feature space and maximise the accuracy:

$$\frac{1}{n}\sum_i\big[f\big(p(t|d_i),w\big)==y_i\big]\rightarrow\max_w. \quad (6)$$

III. Data Preparation

The original data was provided by our industrial partner and not allowed to be public. The client data consists of fields: client id, transaction time, transaction sum, transaction code. Additionally, we had tables containing clients date of birth and gender.

We were able to enrich the data by a hierarchy of MCC codes clustering similar MCC codes together and giving them readable labels such as "taxi", "petrol stations" and so on. During preprocessing we applied two methods to encode the sum spent in each transaction. First, we used the sum as term-frequency for each transaction. More spent on MCC code - more frequent that transaction in user profile. This might lead to a distortion when a rare but expensive purchase could count the same way as many regular cheap transactions. To compensate for that and to enrich the "dictionary" of our data we separated each MCC code into quantile: we introduced new tokens. These tokens are encoding the MCC code of transaction and quantile of the sum spent on that token: below average spent, average, above average.

Finally, we introduce the hierarchy of the MCC codes as new modalities for our model. The original hierarchy contained In this article, we use embeddings containing only MCC codes and Small group modality. Other modalities are used to check the sanity of an obtained topic model - MCC codes from really distant and uncorrelated groups should not be in the same topic. For example, a model with topics containing codes from "Travel" and "Home renovation" groups would be discarded in the training process.

IV. Experiment

We test the ability of our approach to create bank client vector representations on the same level as the actual data. Along with the main goal we test the impact of data preprocessing on our task. To train the standard topic model we need to determine crucial hyperparameters such as: number of EM algorithm steps, number of topics, regularizer coefficients by measuring model coherence score as it is known to correlate with interpretability [1]. To make a comparison between different topic models fare we fix the main hyperparameters: number of EM algorithm steps and number of topics, while we allow regularizer coefficients tweaking according to a metrics of our choice. Upon investigating the dataset we found that we obtain best models around 30 topics and we fixed this hyperparameter for all of the topic models in this paper. Here we provide topics related to Vacation and replace MCC tokens by interpretative groups of such to demonstrate what we call an interpretative topic in table I An inquisitive reader could mention that the probability of expenses does not add up to one. This is due to a long tail in the distribution that is not representative of the main theme of the topic and not shown here. The resulting clusterisation of the clients performed by first topic model can bee seen on Fig. 1. The experiment is built in the following way: first, we create a baseline vector representation of the clients based on their transaction data

TABLE I
VACATION TOPIC

Expenses group	**probability**
Plane tickets	0.575
Duty-free	0.177
Theatres	0.0094
Hotels	0.049
Attractions	0.0038
Drug stores	0.0022
Car sharing	0.009
Gender	**probability**
Male	0.393
Female	0.607
Age group	**probability**
17-23	0.138
24-35	0.442
36-54	0.376
55+	0.043

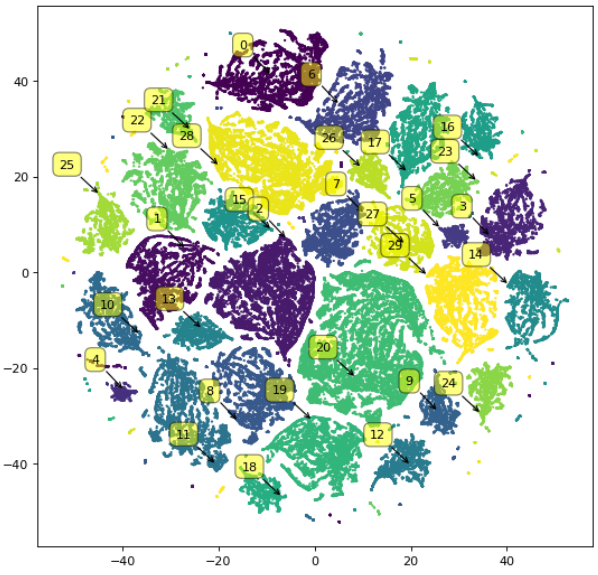

Fig. 1. Visual representation of the consumption profiles.

and its counterpart embedding produced by a topic model. Next, we predict the accuracy of gender and age prediction using CatBoost [17] models adjusting topic model embeddings on the training dataset. Performance results for various embeddings presented in Table II. In the table II models

TABLE II
MODEL COMPARISON WITH BASELINES FOR VARIOUS TYPES OF PREPROCESSING

Model type	*gender accuracy*	*age accuracy*	micro f1 score
categorised MCC	**0.730**	**0.465**	0.439
ARTM Coherence	0.665	0.439	0.414
ARTM DemoOpt	0.685	0.45	0.427
ARTM DemoOpt	**0.728**	**0.457**	**0.435**
ARTM Coherence	0.658	0.405	0.377
MCC	0.719	0.452	0.424

with "DemoOpt" tag are optimised for better performance in gender and age prediction, while a models with "Coherence" tag were optimised coherence score making them to be more interpretable [1]. As we can see we were able to obtain an embedding that scored better than the initial preprocessed data. We were not able to repeat that for the data separated by spending categories, however we were also able to improve the standard model baseline.

V. CONCLUSION

We described an approach of building topic models that can provide their user with an insight into the transaction data. This approach is addressing a problem of usefulness and interpretability of vector embeddings in Banking. The obtained model hyperparameters were selected according to the external metrics providing an opportunity to tune the resulting embeddings to contain the information not presented in the actual data. We demonstrated that the vector representations provided by the model kept all the crucial features to tell a story about the client to an inexperienced user, while keeping high interpretability unavailable for most of the better performing ML models.

As we can see from the table II not only topic modelling but just a preprocessing described in the article can give a boost to data applications. The properly tuned topic model could perform better than uncategorised data but could not beat a baseline for the categorized MCC code. We tend to think that this could be attributed to one of the two factors. First, our self-posed restriction on model topic number could be crucial when the token dictionary increased by our preprocessing. Second, we suspect that our model architecture does not allow to capture some of the data specifics that became apparent to CatBoost classifier on categorized data.

In the future, we will focus on improving our model to account for transaction co-occurrences and would also be able to deal with heavy data imbalance that we faced in our dataset.

REFERENCES

[1] Bulatov V.G. Vorontsov K.V. Alekseev, V.A. Intra-text coherence as a measure of topic models' interpretability. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue*, pages 1–13, 2018.

[2] Darya Polyudova Eugenia Veselova Artem Popov, Victor Bulatov. Unsupervised dialogue intent detection via hierarchical topic model. In *RANLP*, pages 932–938, 2019.

[3] Leonardo Baldassini and Jose Antonio Rodríguez Serrano. client2vec: towards systematic baselines for banking applications. *arXiv preprint arXiv:1802.04198*, 2018.

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

[5] Mahil Carr, Vadlamani Ravi, G Sridharan Reddy, and D Veranna. Machine learning techniques applied to profile mobile banking users in india. *International Journal of Information Systems in the Service Sector (IJISSS)*, 5(1):82–92, 2013.

[6] Chiranjit Chakraborty and Andreas Joseph. Machine learning at central banks. 2017.

[7] Joe J Gladstone, Sandra C Matz, and Alain Lemaire. Can psychological traits be inferred from spending? evidence from transaction data. *Psychological science*, page 0956797619849435, 2019.

[8] Aboobyda Jafar Hamid and Tarig Mohammed Ahmed. Developing prediction model of loan risk in banks using data mining. *Machine Learning and Applications: An International Journal (MLAIJ) Vol*, 3, 2016.

[9] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[10] Timothy Hospedales, Shaogang Gong, and Tao Xiang. Video behaviour mining using a dynamic topic model. *International journal of computer vision*, 98(3):303–323, 2012.

[11] Anastasia Ianina, Lev Golitsyn, and Konstantin Vorontsov. Multi-objective topic modeling for exploratory search in tech news. In *Conference on Artificial Intelligence and Natural Language*, pages 181–193. Springer, 2017.

[12] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[13] Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

[14] Piotr Ładyżyński, Kamil Żbikowski, and Piotr Gawrysiak. Direct marketing campaigns in retail banking with the use of deep learning and random forests. *Expert Systems with Applications*, 134:28–35, 2019.

[15] Clément Lesaege, François Schnitzler, Anne Lambert, and Jean-Ronan Vigouroux. Time-aware user identification with topic models. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 997–1002. IEEE, 2016.

[16] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM, 2007.

[17] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, pages 6638–6648, 2018.

[18] Martin Reisenbichler and Thomas Reutterer. Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics*, 89(3):327–356, 2019.

[19] Klaus-Dieter Schewe, Roland Kaschek, Claire Matthews, and Catherine Wallace. Modelling web-based banking systems: Story boarding and user profiling. In *International Conference on Conceptual Modeling*, pages 427–439. Springer, 2002.

[20] Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. A combination approach to web user profiling. *ACM Trans. Knowl. Discov. Data*, 5(1):2:1–2:44, December 2010.

[21] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. Bigartm: Open source library for regularized multimodal topic modeling of large collections. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 370–381. Springer, 2015.

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

[23] Jongwook Yoon, Seok Hwang, Dan Kim, and Jongsoo Yoon. A balanced view for customer segmentation in crm. *AMCIS 2003 Proceedings*, page 67, 2003.