Patent-Based Import Substitution Analysis with Additively Regularized Topic Models

Maria Milkova[0000-0002-9393-1044]

Central Economics and Mathematics Institute of Russian Academy of Science,
47 Nakhimovsky Prospect, Moscow, 117418, Russia
m.a.milkova@gmail.com

Abstract. The rapid accumulation of textual data forces the use of various methods to present the structure of available information. One of these methods is topic modeling. We apply Additively Regularized Topic Models (ARTM) for analyzing an import substitution program based on patent data. The program includes plans for 22 industries and contains more than 1500 products and technologies for the proposed import substitution. The use of patent search based on ARTM allows to search immediately by the blocks of a priori information - terms of industrial plans for import substitution, and at the output get a selection of relevant documents for each of the industries. This approach allows not only to provide a comprehensive picture of the effectiveness of the program as a whole, but also to obtain more detailed information about which groups of products and technologies have been patented. It is important that topic modeling also solves the problem of synonymy and homonymy of words.

Keywords: Topic search, Topic modeling, Import substitution; Patent search, Patent analysis, Additively Regularized Topic Models, ARTM.

1 Introduction

Currently, in the Digital Age the information accumulation process is rapid and the desire to develop effective ways to perceive the essence and to screen out unnecessary information is natural. The disordered nature of working with information, the lack of necessary skills and tools is a key factor preventing the recognition of future innovations and the prediction of their consequences [1]. Thus, it is necessary to use an approach to the perception of information that would allow us to present a road map, the structure of the direction being studied.

Considering in this work information in text form, we note that its overabundance is presented not only on the Internet, but also in the scientific community [2], the legal field [3], literature [4].

Various clustering methods have been well studied to obtain information about the structure of large amounts of text data: bibliometric analysis [5], clustering social networks users [6], analysis of discourse and sentiment of messages [7, 8], analysis of legal documents [9], etc. However, the changing digital reality requires from us a

Proceedings of the 10th International Scientific and Practical Conference named after A. I. Kitov "Information Technologies and Mathematical Methods in Economics and Management (IT&MM-2020)", October 15-16, 2020, Moscow, Russia

revision of approaches to semantic compression of information. Firstly, with regard to textual data, it is necessary to take into account the problem of synonymy and homonymy of words. Secondly, when searching, it is necessary to take into account information that we already have.

These requirements are satisfied by topic modeling - a modern tool that determines the structure of the collection of text documents by identifying hidden topics in the documents, as well as terms (words or phrases) that characterize each of the topics. In probabilistic topic modeling a document can with certain probabilities relate to several topics at once, just as a term can define a particular topic with different probabilities. Each document is described by a discrete distribution on topics, and each topic is described by a discrete distribution on terms. Presenting the results in this form allows to get a roadmap of the direction you are interested in and significantly increases the accuracy and fullness of the search [10]. Over the past decade, the concept of topic search has been developing [11, 12]. This type of search helps to identify the topics of real interest, observing the most informative terms in the estimated topics.

This article demonstrates an example of applying topic search in patent analysis - an integral part of both Foresight research, individual research on the prospects of innovative development, technological trends in various fields, etc. Our research contributes to this field by providing semi-supervised topic search based on different a priori information.

2 Literature Review

Topic modeling has been intensively developing since the late 90s. An important milestone in the development of probabilistic text modeling is the Probabilistic Latent Semantic Analysis (PLSA) model described in [13]. PLSA was based on the principle of maximum likelihood and was developed as an alternative to classical text clustering methods based on calculating distance functions.

However, PLSA had a number of significant limitations [14], which were eliminated in the Latent Dirichlet Allocation model (LDA) proposed in [15]. LDA is a generative probabilistic model, in which documents are presented as a probabilistic mixture of hidden topics (each word in a document is generated by some latent topic), while the distribution of words in each topic is explicitly modeled, as well as the prior distribution of topics in the document.

Literature review shows that LDA is the leader among probabilistic topic models due to numerous generalizations, extensions and applications to the analysis of collections of text documents [16-20].

However, in the works [21, 22], in which the view of PLSA and LDA is critically revised, it is noted that the widespread use of LDA is explained rather by its purely mathematical convenience for Bayesian learning. It was emphasized that the prior Dirichlet distributions and their generalizations have no convincing linguistic justification. Moreover, the transition from a generating model to an algorithm for adjusting its parameters requires rather cumbersome calculations, which become much more

complicated when more complex prior distributions are introduced or when several linguistic phenomena are jointly simulated.

For these reasons, the development of the so-called Additive Regularization of Topic Models (ARTM) approach developed in [21] received a powerful impulse. ARTM is a multicriteria approach based on the presentation of the topic modeling problem as an ill-posed optimization problem requiring the introduction of a regularizer - an additional criterion that takes into account the specific features of the applied problem or knowledge of the subject area [21].

Currently, two directions of development of topic models are outlined - based on Bayesian learning (LDA model) and on the basis of Additive regularization. [21] revise topic models previously developed in the Bayesian approach, for each of which a corresponding regularizer is found, which leads to the same or very similar model learning algorithm. Compared to the Bayesian approach, ARTM radically simplifies the inference of the algorithm and allows to combine regularizers in arbitrary combinations. Also, recent studies have shown the superiority of ARTM over LDA in terms of the quality of highlighted topics (see, for example, [23], where ARTM and LDA are compared using the example of monitoring ethnically determined discourse in social networks).

The use of topic modeling for the analysis of patent data has been gaining popularity in recent years. Research publications in the field of patent analysis show the effectiveness of both the use of text mining methods in general [24] and the validity of applying topic modeling [25, 26]. The construction of topic models is used to get an idea of patenting in the industry [27], to identify technological trends [28], to develop individual specialized software products for conducting topic patent analysis [26].

3 Materials and methods

From 2015 to 2018 Minpromtorg of Russia has been approving the import substitution programs [29] in the range of economic industries. The programs govern action plans for import substitution in 22 economic sectors (hereinafter - the Plans). In each of the industries, a list of goods and technologies has been compiled for which its own indicator of the share of imports by 2020 has been established. Simple statistics on the characteristics of the Plans are shown in Table 1.

Table 1. Import substitution plans characteristics.

Characteristic	Value
Number of plan items:	
Min.-Max.	4-602
Mean	70.3
Median	41
Sum	1553
Import share:	

Median (fact)	90%
Median (plan)	15%

Today it is important to present some results of import substitution based on the analysis of patent data. During the implementation of the Program, a number of publications appeared in the scientific community evaluating the possibilities of import substitution for certain goods [30-32]. Despite the crucial importance of conducting a detailed analysis in each of the development areas, it is useful to have a general structure of the results. The approach, covering all sectors at once, will allow both to demonstrate the results of the program as a whole and give a general idea of the state of various sectors of the economy (based on patent data).

To analyze the implementation of the import substitution plan, it is necessary to obtain information on all 1553 points of the Plan, which requires a fundamentally different approach to the patent search.

Currently, there are various approaches for building topic models [33, 34], [14]. In our work we focus on Additive Regularization of Topic Models as it provides a convenient way for semi-supervised learning and greater flexibility in constructing topic models with given properties [8]. In this article we give only the basics of multimodal ARTM, as it was done in [35].

Let us denote a finite collection of documents by D, a finite set of topics by T, and a finite set of modalities by M. In our work we use two modalities: words (unigrams) and most common bigrams. Each modality $m \in M$ has a dictionary of tokens W^m. Each document $d \in D$ is a sequence of n_d tokens from $W = \cup_w W^m$. According to the bag-of-words hypothesis we take into account how many times the token w appears in the document d (n_{dw}).

In ARTM topic modeling is considered as a special case of approximate stochastic matrix factorization. To learn a factorized representation of a text collection is an ill-posed problem, which has an infinite set of solutions. A typical regularization approach in this case is to impose problem-specific constraints in a form of additive terms in the optimization criterion [36].

Given the $(n_{dw})_{D\times W_m}$ matrix and find its approximate matrix factorization by $\Phi^m = (\phi^m_{wt})_{W_m\times T}$ matrix of token probabilities for the topics and $\Theta = (\theta_{td})_{T\times D}$ matrix of topic probabilities for the documents:

$$\frac{n_{dw}}{n_d} \approx p(w|d) = \sum_{t\in T} p(w|t)p(t|d) = \sum_{t\in T} \varphi_{wt}\theta_{td} \quad (1)$$

where $|T|$ is a number of topics in the model (in our case $|T|$ =22).

Additive regularization narrows the set of solutions of (1) by maximizing the weighted sum of modality log-likelihoods and regularizers $R_i(\Phi, \Theta)$:

$$L = \sum_{m\in M} \tau_m \sum_{d\in D} \sum_{w\in W^m} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta) \rightarrow max_{\Phi,\Theta} \quad (2)$$

under non-negativity and normalization constraints for all columns of Φ^m and Θ matrixes. Regularization coefficients τ_m are used to balance the importance of different modalities.

This optimization problem can be solved using the EM-algorithm. At first, the initial approximation for φ_{wt}, θ_{td} is selected. At the E-step, auxiliary variables $p_{tdw} = p(t|d,w)$ are calculated:

$$p_{tdw} = norm_{t\epsilon T}(\varphi_{wt}\theta_{td}) \tag{3}$$

where operator $norm$ transforms a real vector to a vector representing a discrete distribution (by zeroing out the negative elements and normalizing). At the M-step, φ_{wt}, θ_{td} are specified:

$$\varphi_{wt} = norm_{w\epsilon W^m}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right), \quad n_{wt} = \sum_{d\epsilon D}\tau_{m(w)}p_{tdw}n_{dw} \tag{4}$$

$$\theta_{td} = norm_{t\epsilon T}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right), \quad n_{td} = \sum_{w\epsilon d}\tau_{m(w)}n_{dw}\,p_{tdw} \tag{5}$$

where $m(w)$ is the modality of the term w, $w\epsilon W^m$.

Calculating (3)-(5) continue in a loop until convergence.

Regularizers are aimed at taking into account the linguistic features of the text and increasing the interpretability of topics. The most common regularizers are sparsing, smoothing, and decorrelation regularizers of topics [21].

For our task regularizers need to be constructed for grouping the terms of each of the Plans in its own topic. Thus, we use 22 smoothing regularizers for Φ matrix (both for words and bigrams) that encourages terms from each Plan $w\epsilon Q_i$ to appear in related topic S_i, $i = 1, \ldots 22$.

$$R_i(\Phi) = \tau_1 \sum_{t\in S_i}\sum_{w\in Q_i} ln\,\varphi_{wt} \tag{6}$$

The convenience of ARTM is that regularization term R_i yields a simple additive modification of the M-step. For our task, this modification led to the fact that the τ_1 parameter was added to the frequencies of terms related to the terms from the "white list" (Plan terms) at each iteration of the EM algorithm. The value of τ_1 is selected experimentally.

The main task in constructing the model is to select the regularization strategy - function of the regularization coefficient on the iteration number and model quality criteria. Following [36] we use such quality criteria as perplexity (the degree of convergence of the model with a given dictionary W, $P(D) = \exp\left(-{}^{L}/_{\sum_{d,w} n_{dw}}\right)$, the degree of sparseness of the matrices Φ, Θ (the proportion of zero elements in the matrix), the size of the kernel $|W_t|$ (many words with a high conditional probability, $W_t = \{w \in W | p(t|w) > 0.25\}$), the purity of the topic (how much the terms inside the topic are determining - the total probability of the terms of the kernel of the topic $purity_t = \sum_{w\in W_t} p(w|t)$), the contrast of the topic (how well the topic kernel distin-

The reported study was funded by RFBR according to the research project No. 20-07-22059

guishes it from the rest in that, i.e., the probability of meeting the terms of the kernel in this particular topic $contr_t = \frac{1}{|W_t|}\sum_{w \in W_t} p(t|w)$).

4 Model construction and Results

It was collected patents for inventions and utility models issued over a 3.5-year period (January 2016-June 2019) - a total of 152718 documents: 120768 inventions and 31950 utility models. For building the ARTM the Python and the open source library BigARTM were used [37].

The model was built on the basis of Titles and Abstracts of patents presented in the form of unigrams (i.e. single words) and most frequency bigrams (two-word phrases with a frequency of occurrence in the Title and Abstract of more than or equal to 2). The experimentally chosen modality weight τ_m was: 1.0 for words and 5.0 for bigrams. The coefficient of 22 smoothing regularizers was $\tau_1 = 1e + 7$.

The final model had the following quality metrics: Perplexity is 630.7, the proportion of sparse elements in unigram matrix $\Phi^1 = 0{,}994$, in bigrams $\Phi^2 = 0{,}998$, $\Theta = 0{,}818$. The kernel size is $|T| = 628$, the average purity is 0.992, and the average contrast is 0.976. Total number of iterations: 40.

Based on the ranged probabilities of Θ columns, we selected patent documents for each of 22 industries (threshold=0.6) in accordance with a topic characterized by a set of words and phrases from the corresponding Plan. In addition to standard automatically calculated metrics, the quality of the model was also evaluated using assessors, which determine how relevant the selected document is. The value q = 1 was set in accordance with the patent document if the patent exactly corresponded to one of the import substitution items declared in the Plan; q = 0.5 was assigned if the patent is associated with one of the points of the Plan; q = 0 - if it did not correspond to any of the items in the Plan. This technique has been successfully used in [8, 10].

For documents with values q = 1, q = 0.5, a key phrase was selected that characterizes document belonging to the item of the Plan (Table 2). Thus, each industry was characterized by import substitution categories (key phrases), total number of categories (k), average and total mark ($\bar{q}, \sum q$), and total score $score = {\sum q \cdot k}/{N}$, where N - total points of the Plan. Industry ranking results are presented in Fig. 1.

Table 2. Patentable import substitution categories.

Industry	Import substitution category (RU)
Car industry	internal combustion engine
Civil aircraft industry	-
Baby goods	furniture for children; games and toys; sport complexes; baby clothes; children's creativity
Light industry	non-woven materials; protective clothing; wool processing

Timber industry	cellulose treatment; paper, cardboard
Machinery for food and processing industry	grain processing
Medical industry	sterilization and disinfection; endoscopic devices; injection needles; implantable pumps
Oil and gas engineering	hydrotreating catalysts; drilling of the wells; hydrocracking catalysts; hydrocarbon processing; hydraulic fracturing; catalytic cracking catalysts
Conventional Arms Industry	cartridges; sports weapon
Electronic industry	-
Agricultural and forestry engineering	bearings; combine harvester; baler
Machine tool industry	milling machine; lathe; boring machine; spindles; finish grinding; waterjet cutting; cnc machines
Builds. materials and builds. construction	ceramic mass for tiles; thermal insulation materials; crushed stone and mastic asphalt concrete
Road construction technique	road surface; hydraulic equipment; front loaders; bulldozers; front loader; excavator; trailer and semi-trailer; crane chassis; municipal engineering
Shipbuilding industry	mover; flange screw
Transport machine building	cistern wagon; brake system; wagon trolleys; covered wagon
Heavy engineering	support mountain; refrigeration units
Pharmaceutical industry	inosine + nicotinamide + riboflavin + succinic acid; bismuth potassium ammonium citrate; drotaverine; yohexol; lopinavir + ritonavir; ethyl methyl hydroxypyridine succinate; rocuronium bromide; digoxin; 1 carbamoylmethyl 4 phenyl 2 pyrrolidone; fenspiride; isoniazid; lappaconitine hydrobromide; standard immunoglobulin; bromodihydrochlorophenylbenzodiazepine; desmopressin; fingolimod; anastrozole
Chemical industry	paints and varnishes; sealing materials; epoxy composite; adhesive materials; polyethylene terephthalate; ultra high molecular weight polyethylene; polymer composites
Non-ferrous metallurgy	aluminum alloy; aluminum, electrolysis; aluminum ligature; aluminum hydroxide; aluminum powder; aluminium foil; aluminum rods; anode mass
Ferrous metallurgy	refractories; tubing; threaded connections; drill pipes; pipes based on chromium-nickel alloys; casing
Power engineering	current transformers

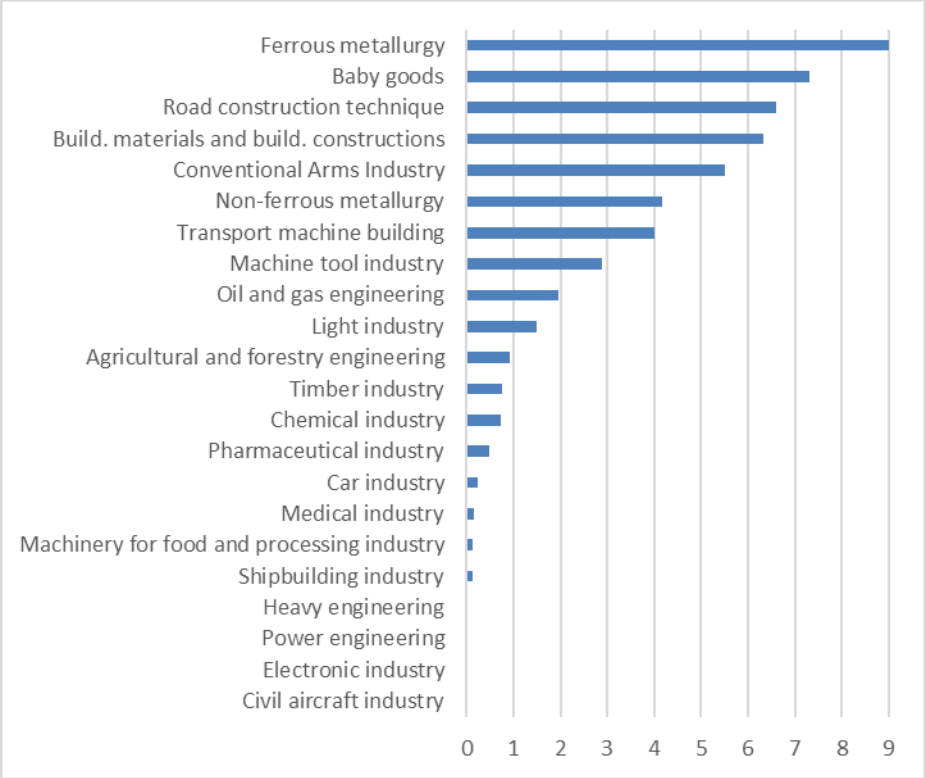

Fig. 1. Industry ranking results

Thus, in the context of granted patents for inventions and utility models, industries that are demonstrated the best indicators of import substitution: Ferrous metallurgy, Baby goods, Road construction technique, Building materials and building constructions, Conventional Arms Industry, etc.

Industries that are currently unable to comply with the import substitution plan (based on patent documents): Civil aircraft industry, Electronic industry, Power engineering, Shipbuilding industry, etc.

5 Discussion

Intellectual property in the form of patents plays a vital role in today's economy. However, the constantly growing volume of information, including patent information, significantly complicates its effective monitoring and analysis. Currently, many search and analytical systems (for example, Yandex.Patents, Google Patents, Patseer) use the advanced achievements of computational linguistics, including the methods of text semantic analysis. Modern search engines are able to find similar

patents, related patents (which mention one or another document of interest to the user, or other documents to which he refers). The search for similar patents is carried out not only by keywords, but also by meaning. It should be noted that modern patent search and analytics systems are designed to obtain information about objects one at a time. If there are many objects of interest, the search requires significant time investment. To analyze the implementation of the import substitution plan, it is necessary to obtain information on all 1553 items of the Plan, which obviously requires a fundamentally different approach to the implementation of patent search.

The purpose of topic modeling of patent documents is to simplify access to documents of interest from the perspective of import substitution. The constructed model allows you to get a general picture of the implementation of import substitution as a whole, within the considered time window.

It is important that the resulting structure allows, if necessary, to detail the results. For example, to identify the share of individual patent holders who will not be able to become the main agents for capturing market niches and will not be able to compete with large foreign companies; share of non-valid patents, etc. This approach is a kind of "close-up" of patent search, which can serve both the final goal or a starting point for a more detailed analysis.

6 Conclusions

The results demonstrate the effectiveness of the new patent search method based on topic modeling. The approach allows to search by blocks of a priori information (in our case, points from all twenty-two industrial import substitution plans at once) and, at the output, receive a selection of relevant documents for each of the industries. Applying the topic modeling also solves the problem of synonymy and homonymy of words.

In today's constantly changing digital reality, the rate of information accumulation is so rapid that it requires to revise our approaches to semantic compression of information. In order to comprehensively cover and analyze the entire spectrum of ongoing changes, it is necessary to make increased demands on the methods of information retrieval. An innovative search approach must flexibly take into account the large amount of already accumulated knowledge and a priori requirements for results. The results, in turn, should immediately represent a roadmap of the studied direction with the possibility of as much detail as necessary. The topic modeling approach allows us to take into account all these requirements and thereby streamline the nature of working with information, increase the efficiency of knowledge extraction, and avoid cognitive biases in the perception of information, which is important both at the micro and macro levels.

Acknowledgments

The reported study was funded by RFBR according to the research project No. 19-010-00293.

References

1. Milovidov, V. Hearing the sound of the wave: what makes it difficult to anticipate innovation? [In Russian]. Forsajt 12(1), 88–97 (2019) doi:10.17323/2500-2597.2018.1.88.97
2. Nedumov, YA.R., Kuznecov, S.D. Issledovatel'skij poisk nauchnyh statej [In Russian]. Trudy ISP RAN, 30 (6), 171-198 (2018) doi: 10.15514/ISPRAS-2018-30(6)-10.
3. Pagallo, U., Palmirani, M., Casanovas, P., Sartor, G., Villata, S. Introduction: Legal and Ethical Dimensions of AI, NorMAS, and the Web of Data. In: Pagallo, U., Palmirani, M., Casanovas, P., Sartor, G., Villata (Eds). Lecture Notes in Artificial Intelligence Springer (2018) doi: 10.1007/978-3-030-00178-0_1 .
4. Moretti, F. Distant reading. London: Verso (2013).
5. Gibson, Je., Dajm, T., Garses, Je., Dabich, M. Bibliometric analysis as a tool for identifying common and emerging methods of technological Foresight. Forsajt 12(1), 6-24 (2018) doi: 10.17323/2500-2597.2018.1.6.24
6. Halibas, A.S., Shaffi, A.S., Mohamed, M.A. Application of text classification and clustering of Twitter data for business analytics. Majan International Conference (MIC), Muscat, 1-7 (2018) doi: 10.1109/MINTC.2018.8363162
7. Krishna, A., Aich, A., Akhilesh, V., Hegde, C. Analysis of Customer Opinion Using Machine Learning and NLP Techniques. International Journal of Advanced Studies of Scientific Research 3(9), 128-132 (2018) https://ssrn.com/abstract=3315430
8. Apishev, M., Koltcov, S., Koltsova, O., Nikolenko, S., Vorontsov, K. Mining Ethnic Content Online with Additively Regularized Topic Models. Computación y Sistemas 20(3), 387–403 (2016) doi: 10.13053/CyS-20-3-2473.
9. Sulea, O-M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L.P., Genabith, J. Exploring the Use of Text Classification in the Legal Domain. In: Proceedings of the 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts (ASAIL), London, United Kingdom (2017) arXiv:1710.09306v1.
10. Janina, A.O., Vorontsov, K.V. Multimodal topic models for exploratory search in a collective blog [In Russian]. Mashinnoe obuchenie i analiz dannyh 2(2), 173-186 (2016) doi: 10.21469/22233792.2.2.04.
11. Grant, C.E., Clint P. G., Virupaksha, K., Nirkhiwale S., Wilson, J.N., Wang, D.Z.: A Topic-Based Search, Visualization, and Exploration System. In: FLAIRS Conference, pp. 43-48. AAAI Press, Massachusetts (2015).
12. Eisenstein, J., Chau, D.H., Kittur, A., Xing, E.P.: TopicViz: interactive topic exploration in document collections. In: Proceeding of CHI EA'12, pp. 2177-2182. Association for Computing Machinery, New York, NY, USA (2012) doi: 10.1145/2212776.2223772.
13. Hofmann,T. Probabilistic Latent Semantic Analysis. Uncertainty in Artificial Intelligence, UAI'99, Stockholm (1999) doi: 10.1145/312624.312649.
14. Daud, A., Li, J., Zhu, L., Muhammad, F.: A generalized topic modeling approach for maven search. In: Li, Q., Feng, L., Pei, J., Wang, S.X., Zhou, X., Zhu, QM. (eds) Advances in Data and Web Management. APWeb/WAIM 2009. LNCS, vol 5446, pp. 138-149. Springer, Heidelberg (2009) doi: 10.1007/978-3-642-00672-2_14.
15. Blei, D., Ng, A., and Jordan, M. Latent Dirichlet allocation. Journal of Machine Learning Research, 3 (2003) doi: 10.1162/jmlr.2003.3.4-5.993.
16. Chemudugunta C., Smyth P., and Steyvers M. Modeling general and specific aspects of documents with a probabilistic topic model. In: Advances in Neural Information Processing Systems. - MIT Press, Vol. 19, 241–248 (2006).

17. Mimno, D., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A. Polylingual Topic Models. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, 880–889 (2009).
18. Ramage, D. Hall D., Nallapati R., and Manning, C.D. (2009) Labeled LDA. A supervised topic model for credit attribution in multi-labeled corpora. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 248–256 (2009).
19. Blei, D. M. and Lafferty, J. Dynamic topic models. In: Proceedings of 23rd International Conference on Machine Learning (ICML), Pittsburgh, Pennsylvania, USA (2006) doi: 10.1145/1143844.1143859.
20. Wang, C., Blei, M. D. and Heckerman, D. Continuous time dynamic topic models. In: Proceedings of Uncertainty in Artificial Intelligence (UAI), Helsinki, Finland (2008) arXiv:1206.3298.
21. Vorontsov, K.V., Potapenko A. A. Additive Regularization of Topic Models. Machine Learning Journal, Special Issue "Data Analysis and Intelligent Optimization", 1-21 (2014) doi: 10.1007/s10994-014-5476-6 .
22. Potapenko, A. A., Vorontsov, K. V. Robust PLSA Performs Better Than LDA. 35th European Conference on Information Retrieval, ECIR-2013, Moscow, Russia, 24-27 March 2013. —Lecture Notes in Computer Science (LNCS), Springer Verlag-Germany, 784–787 (2013) doi: 10.1007/978-3-642-36973-5_84.
23. Apishev,M., Koltsov S., Koltsova, O., Nikolenko, S., and Vorontsov, K. Additive Regularization for Topic Modeling in Sociological Studies of User-Generated Texts. Conference Paper in Lecture Notes in Computer Science (2017) doi: 10.1007/978-3-319-62434-1_14 .
24. Tseng, Y-H., Lin, C-J. Text mining techniques for patent analysis. Information Processing & Management 43, 1216-1247 (2007) doi: 10.1016/j.ipm.2006.11.011.
25. Chen., L., Shang, W., Yang, G., Zhang, J., Lei, X. A topic model integrating patent classification information for patent analysis. Geomatics and Information Science of Wuhan University 41, 123-126 (2016).
26. Tang, J., Wang, B., Yang, Y., Hu, P., Zhao, Y., Yan, X., Gao, B., Huang, M., Xu, P., Li, W., Usadi, A.k.: PatentMiner: Topic-driven Patent Analysis and Mining. In: Proceedings of KDD'12, pp. 1366-1374. Beijing, China (2012) doi: 10.1145/2339530.2339741 .
27. Suominen, A., Toivanen, H., Seppänen, M. Firms' knowledge profiles: Mapping patent data with unsupervised learning. Technological Forecasting and Social Change 115, 131-142 (2017) doi: 10.1016/j.techfore.2016.09.028 .
28. Choi, D., Song, B. Exploring Technological Trends in Logistics: Topic Modeling-Based Patent Analysis. Sustainability 10(8), 1-26 (2018) doi: 10.3390/su10082810.
29. Ministry of Industry and Trade of Russia, Sectoral plans for import substitution in twenty-two industries, https://gisp.gov.ru/plan-import-change/, last accessed 2020/10/10.
30. Jerivanceva T. N. The use of patent analysis to assess the prospects of import substitution on the example of domestic retractors and crosslinking products [In Russian]. Jekonomika nauki 4, 261-275 (2016) doi: 10.22394/2410-132X-2016-2-4-261-275.
31. Jerivanceva, T.N. Assessment of the competitiveness of Russian scientific and technological backlogs in the field of creating medical instruments [In Russian]. Jekonomika nauki 1, 53-69 (2017) doi: 10.22394/2410-132X-2017-3-1-52-68.
32. Andrejchikov, A.V., Teveleva, O.V., Nevolin, I.V., Milkova M.A., Kravchuk, I.S. Methodology for conducting search research to identify opportunities for import substitution of high-tech products based on world patent and financial information resources [In Russian]. Jekonomika i predprinimatel'stvo 4, 157-167 (2019).
33. Milkova, M.A. Topic models as a tool for distance reading [In Russian]. Cifrovaja jekonomika 1(5), 57-69 (2019) doi: 10.34706/DE-2019-01-06 .

34. Boyd-Graber, J., Hu, Y., Mimmo, D. Applications of Topic Models. Foundations and Trends in Information Retrieval 11(2-3), 143-296 (2017) doi: 10.1561/1500000030.
35. Ianina, A., Golytsyn, L., Vorontsov, K.: Multi-objective topic modeling for exploratory search in tech news. In: Filchenkov, A., Pivovarova, L., Žižka, J. (eds) Artificial Intelligence and Natural Language. AINL 2017. Communications in Computer and Information Science, vol 789, pp.181-193. Springer, Cham (2017) doi: 10.1007/978-3-319-71746-3_16.
36. Vorontsov, K., Frei, O. Apishev, M., Romov, P., Suvorova, M., Yanina. A.: Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections. In: Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications (TM '15), pp. 29–37. Association for Computing Machinery, New York, USA (2015) doi: 10.1145/2809936.2809943.
37. Frei, O., Apishev, M.: Parallel non-blocking deterministic algorithm for online topic modeling. In: Ignatov, D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016, Communications in Computer and Information Science, vol 661, pp. 132-144. Springer, Cham (2016) doi: 10.1007/978-3-319-52920-2_13.