Research of the LDA algorithm processing results on high-level classes of patents

Alla Kravets
Volgograd State Technical University;
Dubna State University
Volgograd, Dubna,Russia
agk@gde.ru

Svyatoslav Biryukov
Volgograd State University
Volgograd, Russia
bir.slav@yandex.ru

Denis Marinkin
Perm State National Research University;
Perm State University for the Humanities and Education
Perm, Russia
mdn444@yandex.ru

Vladislav Gneushev
Volgograd State Technical University
Volgograd, Russia
guin0@yandex.ru

Dmitriy Skorikov
Volgograd State University
Volgograd, Russia
skdmitri@mail.ru

***Abstract*—The purpose of the article is to study the similarity of extractable topics from different high-level classes of patents and the possibility of classifying these documents according to the generally-trained model. The optimal number of topics can be selected from the interpretation of the resulting topics for the coherence of words in the topic and the reflection of the general discourse. In the presented dataset only general themes are known, is not possible to suggest which sub-themes can discover. In the course of the research, the dynamics of the change in the models' quality with the change of parameters, according to which relatively optimal parameters are chosen, is considered.**

Keywords—patents, machine learning, LDA algorithm, hyperparameters, model quality

I. INTRODUCTION

The latent Dirichlet allocation [1] (LDA) is a generative model used in computer training and information search, which makes it possible to explain the supervision results with the help of implicit groups so that it is possible to identify the reasons for the similarity of some parts of the data. For example, if words collected in documents are observed, it is argued that each document is a mixture of a small number of topics and that the appearance of each word is related to one of the topics of the document. In the LDA, each document can be viewed as a set of different topics. This approach is similar to probabilistic latent semantic analysis (pLSA), with the difference that the LDA assumes that the distribution of topics has a sparse Dirichlet prior. In practice, the result is a correct set of topics.

Thematic model (topic model) is a model of a collection of text documents that determines which topics each document in the collection belongs to. The algorithm for constructing a thematic model receives a collection of text documents as input. At the output for each document, a numeric vector is drawn, composed of membership degree assessments of this document to each of the topics. The dimension of this vector, equal to the number of topics, can either be specified at the input or be determined automatically by the model. [2][3]

Perplexity [4] is a criterion for the numerical estimation of the quality of a probabilistic model, equal to the exponent of minus the averaged log-likelihood:

$$P = exp(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw} ln\, p(w|d)) \qquad (1)$$

where n is the length of the collection in words.

Perplexity depends on the power of the dictionary and the distribution of word frequencies in the collection:

$$p(w) = n_w/n \qquad (2)$$

II. BACKGROUND

Over the past year, according to open data from World Intellectual Property Organization (WIPO)[5], 3,127.9 thousand patent requests and 1,553.3 thousand utility model requests have been received, which is 8.3% and 28.9% more than the previous year, respectively. And this trend is going on for several years. Due to the growth in the count of requests for patents, the load on the patent office's examining the application materials also increases. Sometimes the deadline for the examination of the application reaches several years, this situation is harming, mostly to the high-tech business. After a formal examination, an expert sometimes spends tens of hours examining the merits of one application and analyze thousands of existing patents during the examination [6].

In this regard, there is a need to develop various decision support systems that would allow inventors to evaluate their application at the stage of its preparation, and experts to evaluate the application already taking into account the results of the pre-patent search. Arguably, one of the main tasks that arise at this stage is the task of pre-patent search - the search for existing patents that could potentially refute the novelty of the application.

Many scientists are addressing the issue of automating the pre-patent search and the search for patents, which refute the novelty of the application. Methods were proposed based on machine learning [7], on the analysis of syntactic relations [8], on the analysis of citation graphs and patent classes [9], on the formation of a search query from an application and on the use of the ranking function BM25 [10]. However, all existing methods do not show a significant increase in recall and accuracy compared to the traditional method based on the comparison of TF * IDF vectors [11].

The unique statistical-semantic method developed in our previous research [12] significantly (by 23-25%) increases recall and precision.

Another imperfection in the process of analyzing new technical solutions presented in the form of a patent application is a significant time gap (frame) between the grant of a patent and its open publication. Moreover, the priority date is the filing date of the application, which, taking into account the time of the examination, leads to situations of duplication of certain technical solutions by different applicants. The gap also leads to the problem of examination by the second significant criterion (after novelty) - the industrial applicability of the invention. This

complex criterion includes, inter alia, an assessment of the disclosure of the invention and the possibility of technical implementation of the solution proposed in the application. Due to the lag in the provision of information in patent databases, the expert, when making a decision, can classify the invention as a “the perpetual motion machine” and reject it only because the content of the application does not meet the criterion of industrial applicability according to the subjective opinion of the expert. Moreover, rejected applications are not published in accordance with the current regulations.

Among the most common commercial products in this area, there are such services as Thomson Reuters (Thomson Innovation), Questel (Orbit), GridLogics (PatSeer), VantagePoint, STN Analyze Plus, STN Anavist, Invention Machine (Knowlegist, Goldfire), etc., as well as many additional tools: Metheo Patent, TEMIS, TotalPatent, Wisdomain, PatBase, ArchPatent, PatentLens, PatentBuddy, PatentTools, FreePatentsOnline, Intellogist, PriorSmart, MaxVal, BizInt SmartCharts, Espacenet, AmberScope Inn, IPlaim, PatentInspiration.

However, all of the above products search for documents relevant to the application according to the formulated request and do not implement the functionality for determining the patentability of the application. Therefore, it is impossible to consider them as direct analogs of the developed technology.

At the same time, the attention of the expert community is increasingly focused on the implementation of artificial intelligence methods for solving the problems of analyzing technical solutions, managing intellectual property, and other challenges of the current stage of the digital economy and Industry 4.0 [13].

III. Automation of patent information analysis

An automatic positioning system for the application materials to obtain a patent for an invention in the global patent space based on statistical and semantic approaches Cyber Examiner is a system for expert decision-making in the examination of a patent application. A pilot project of The Cyber Examiner system was implemented by the order of the World Intellectual Property Organization (Switzerland) [14].

One of the most important stages in the implementation of the system is the definition of a patents’ list relevant to the submitted application (Fig. 1, 2) [15].

At the first stage of the pre-processing, the existing bases are transformed into the developed uniform format. In the process of transformation to a uniform format, the US and IPC classifications are compared. The next stage is the selection of parts of speech. Then, based on the classes of patents, a plurality of patents is subdivided into subsets for training independent LDA models. During the models' learning process for each patent, the patent's vector(s) of its membership in topics is built. In conclusion, the patent claims points are divided into simple sentences and semantic networks are built on their basis, followed by simplification.

For the received request, at the first stage, it is pre-processed by analogy with the pre-processing of existing patent bases.

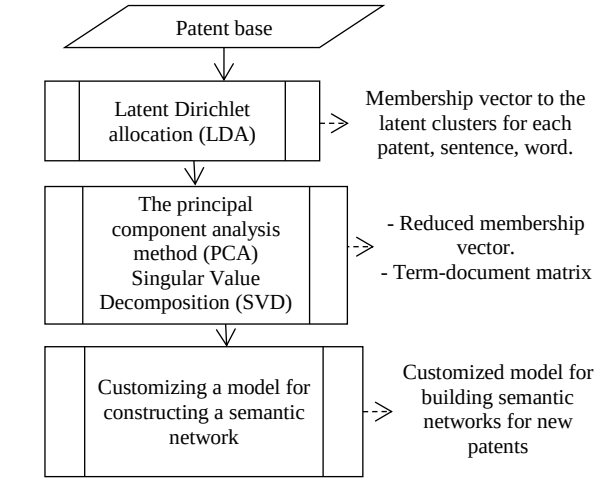

Fig. 1. Algorithm for processing the existing patent database.

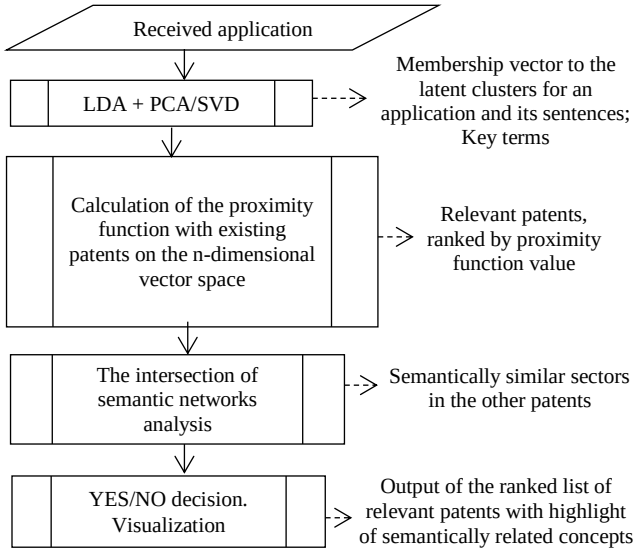

Fig. 2. Received application processing algorithm.

In the second stage, the LDA models are selected and the affiliation of the received application to the topics of each model is calculated. Next, there is a calculation process of the proximity between the application and existing patents obtained in the first stage which based on the similarity between their distributions by topic.

It the third stage, many of the closest patents come.

In the third stage, there is a process of building the semantic network of application formulas is constructed and compared with the semantic networks of existing patents from the resulting set. As a result of this comparison, there is the selection of existing patents, which could refute the application.

The text of the application is sent to the system via the web interface [16]. The most important information is stored in the “Claim” section. It is the novelty of this information that should be checked by the expert [12].

There are three major problems of expert decision-making in the examination of a patent application. First - it is very large volumes of unstructured information, that is, the information stored in the form of texts, images from different sources often have a completely different structure. The second problem is also informational - is information

incompleteness, that is, lack of access to certain patent databases, open-source, citation indexes, which require additional connection costs, for example. The third problem is expert subjectivity and in this decision-making process as it is the riskiest.

IV. Models Training and Experiments

A. Initial data and pre-processing

As initial data, the texts of the five high-level classes of patents were used:

- A (HUMAN NECESSITIES),
- B (PERFORMING OPERATIONS; TRANSPORTING),
- G (MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING),
- H (PHYSICS),
- F (ELECTRICITY).

The source files are in XML format, from which the "Claim" section was extracted for model training. The crypt for XML files parsing was developed.

The extracted claims were collected in a single string. To increase the statistical significance, the formulas cross-referred clauses were refined by the referred text (as "according to clause 1").

Thus, the patent document was a string consisted of a set of claim's clauses, disclosed if necessary until the first cross-reference to other clauses.

The order of text processing included the following steps:

1) tokenization (built-in Python tools);
2) lowercase (built-in Python tools);
3) discarding tokens that are less than two characters long (because the expressed content of the formula elements was found) (built-in Python tools);
4) removal of punctuation and stop words (Nltk package);
5) lemmatization of words (Pymorphy2 package).
6) For each class, training (4,000 patents) and test (1,000 patents) datasets were created.
7) To train the model, the Gensim library was used, the resulting models were visualized using the pyLDAvis library.

B. Experiments' conditions

The purpose of the first set of experiments is to study the dependence of the model achieved quality and the training time on the parameters' values.

A series of experiments are carried out with the implementation of LDA in the library Gensim (a function version with parallel learning). The following parameters can be set:

- number of training iterations (passes) through the collection (P);
- hyperparameters of the model (the value of the parameter α, the parameter β was duplicated);
- number of recoverable topics (K).

C. Experiments on the definition of the optimal number of iterations

Of the five training samples A-Train.Sample, B-Train.Sample, F-Train.Sample, G-Train.Sample and H-Train.Sample, the general dataset was combined, on which the model with the following parameters was trained:

- the number of latent topics: 2;
- the number of iterations for the documents collection: 1, 5, 10, 15, 20, 25, 30, 50;
- other parameters by default.

The results of the experiments' series are shown in Figure 3. It can be seen that the increase in iterations increases training time. With the number of iterations of more than 8, the time costs are incomparably increased in comparison with the accuracy. In the subsequent experiments, we will use the parameter value equal to 10 iterations in the collection.

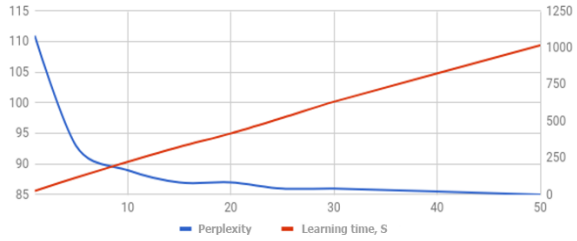

Fig. 3. Perplexity and training time diagram.

D. Evaluation of the hyperparameters impact on the model quality

The selection of the model's hyperparameters assumes the search for values by scanning certain values in the interval (for example [0,2]) with a small step, which is quite laborious. Authors [17, 18] refer to the empirical selection of these parameters. In the course of the experiments, the empirical values of the hyperparameters were used and the tendency to change the model's perplexity was studied.

Static parameters.

- Training sample: a collection of patents (16 thousand documents);
- Number of topics K: 2;
- Number of iterations P: 10

Variable parameters.

Hyperparameters of the model are:

- α {0.01; 0.1; 0.3; 0.5; 1.1; 1.25};
- auto (the library chooses the best value itself);
- default (default mode is symmetric)

The comparison of the changed parameters is visualized in Figure 4.

As a result, the best value of the parameter α from the presented set is coefficient 1.1. The value of the parameters auto-selection is not allocated by the library, but the learning time has significantly increased. Because on average, perplexity values do not change much for different values of hyperparameters (and possibly will depend on the dataset

and other parameters) in the following experiments, we set up the default value.

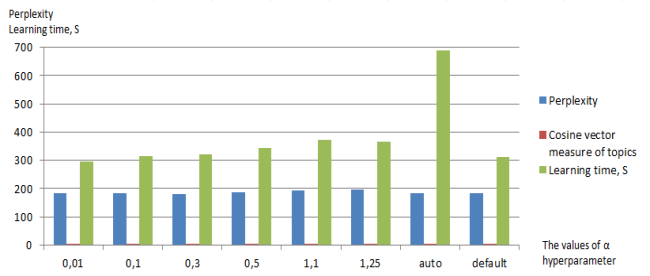

Fig. 4. Models behavior with changes of hyperparameters.

E. The number of hidden topics search

The purpose of the second set of experiments is the research of the similarity of extracted topics of patent classes and opportunities for the generally-trained classification model.

The optimal number of topics can be selected from the interpretation of the resulting topics (for example, expert judgment) for the words coherence in the topic and the reflection of the general discourse. In the presented set of documents only general themes are known, it is impossible to guess which sub-themes could be discovered.

We assume that the more topical diverse (for a certain K), the more successful is the topics' definition. A comparison of the topics vectors similarity is carried out with the cosine measure.

For each model, regardless of the parameters being changed, the following set of characteristics is saved:

- Training data file;
- Number of discoverable topics ;
- Length of the document/dictionary;
- Time of model training;
- The value of perplexity for the model;
- Topics with sets of 30 most popular words for each of them;
- Cosine measure between all topics of the model;
- Visualization of the representation of the topic of the model (pyLDAvis library).

Parameters of the model (Fig. 5 - 10):

- Number of latent topics: 2, 3, 4, 5, 6, 7;
- Number of iterations per document collection: 10
- Other parameters by default.

V. Results and Discussion

Based on the results obtained, the following provisions can be discussed.

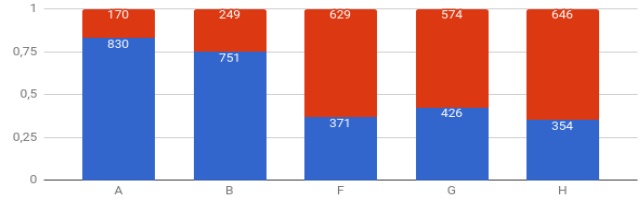

Fig. 5. The distribution of patents classes on 2 topics.

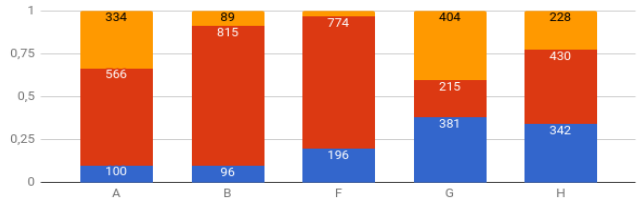

Fig. 6. The distribution of patents classes on 3 topics.

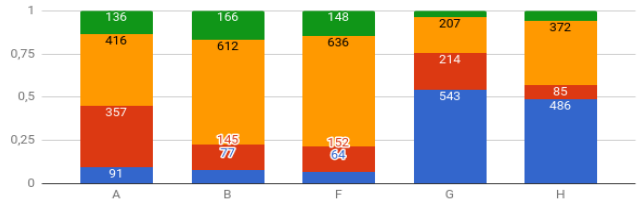

Fig. 7. The distribution of patents classes on 4 topics.

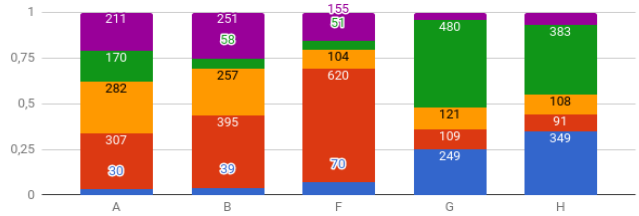

Fig. 8. The distribution of patents classes on 5 topics.

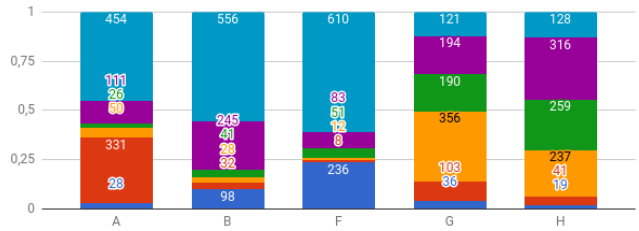

Fig. 9. The distribution of patents classes on 6 topics.

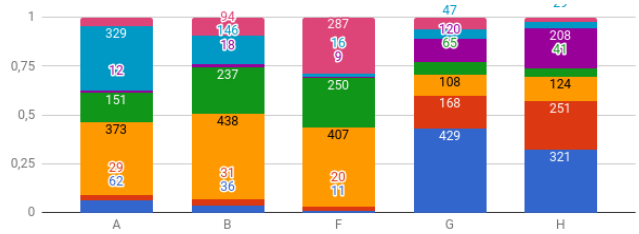

Fig. 10. The distribution of patents classes on 7 topics.

When distributing the presented collection of documents on two topics, it is possible to highlight the evidential similarity between the two classes of patents: A (HUMAN NECESSITIES) and B (PERFORMING OPERATIONS; TRANSPORTING), and the less evidential similarity of classes G (MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING), H (PHYSICS) and F (ELECTRICITY).

When distributing patents' classes on 3 topics, it is obvious that the following classes have common parts: A (HUMAN NECESSITIES), B (PERFORMING OPERATIONS; TRANSPORTING), G (MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING). When distributing patents' classes on 4 topics, it is observed a similar distribution as in the distribution of three topics.

Very close were the results for the classification into 5, 6 and 7 topics, with the only difference that in the distribution of classes of 5 and 7 topics, one can single out the similarity

in one of the topics for classes H (PHYSICS) and F (ELECTRICITY), and in the distribution on 5 topics only for classes A (HUMAN NECESSITIES), B (PERFORMING OPERATIONS; TRANSPORTING), G (MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING), actually as and at experiments 2, 3 and 4.

Thus, we can conclude that with the use of formed from five training samples general model obtained, by the search for a different number of common topics, the next closest classes of considered in this study: A (HUMAN NECESSITIES), B (PERFORMING OPERATIONS; TRANSPORTING), G (MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING). Also, some experiments have shown that classes (PHYSICS) and F (ELECTRICITY) have latent similarities. Also, we can conclude that the distribution of fewer topics gives a more evidential result. So, in the first experiment, classes A and B had an obvious similarity, with a further increase in the number of common topics, this similarity was not lost, but became less noticeable.

VI. Conclusion

As a result of the research done, the quality of LDA algorithm processing results on five high-level classes of Russian-language patents was investigated.

The dynamics of the change in the models' quality is considered when changing the parameters by which relatively optimal parameters are chosen. However, the question of model optimization requires further more detailed research [19].

The comparisons of the selected topics are based on the cosine measure, the results of which can roughly assess the quality of clustering. Because of a large number of topics (Fig. 8 - 10) increases the number of similar vectors. In general, the problem of choosing the number of clusters refers to the content interpretation and involves a deeper study.

Acknowledgment

This research was supported by the Russian Fund of Basic Research (grant No. 19-07-01200).

References

[1] D.M. Blei, A.Y. Ng, M.I. Jordan and J. Lafferty, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, no. 4-5, pp. 993-1022, 2003.

[2] MachineLearning [Online]. URL: http://www.machinelearning.ru/wiki/.

[3] V.M. Chernov, "Fibonacci, tribonacci, …, hexanacci and parallel "error-free" machine arithmetic," Computer Optics, vol. 43, no. 6, pp. 1072-1078, 2019. DOI: 10.18287/2412-6179-2019-43-6-1072-1078.

[4] P.F. Brown, "An Estimate of an Upper Bound for the Entropy of English," Computational Linguistics, vol. 18, 2007.

[5] World Intellectual Property Organization Fact and Figures [Online]. URL: http://www.wipo.int/edocs/infogdocs/en/ipfactsandfigures 2017

[6] World Intellectual Property Organization Statistic [Online]. URL: https://www.wipo.int/export/sites/www/freepublications/en/intproperty/941/wipo_pub_941_2013.pdf.

[7] Yu.V. Vizilter, V.S. Gorbatsevich and S. Zheltov, "Structure-functional analysis and synthesis of deep convolutional neural networks," Computer Optics, vol 43, no. 5, pp. 886-900, 2019. DOI: 10.18287/2412-6179-2019-43-5-886-900.

[8] E. D'hondt, S. Verberne, W. Alink and R. Cornacchia, "Combining Document Representations for Prior-art Retrieval," CLEF Labs and Workshop, Notebook Papers, Amsterdam, 2011.

[9] M. Verma and V. Varma, "Exploring Keyphrase Extraction and IPC Classification Vectors for Prior Art Search," CLEF Labs and Workshop, Notebook Papers, Amsterdam, 2011.

[10] S.E. Robertson, S. Walker, M.M. Beaulieu, M. Gatford and A. Payne, "Okapi at TREC-4," Proceedings of the 4th Text Retrieval Conference, Gaithersburg, MD, 1996.

[11] F. Piroi, M. Lupu, A. Hanbury and V.A. Zenz, "CLEF-IP: Retrieval in the Intellectual Property Domain," CLEF Labs And Workshops Notebook Papers, Amsterdam, 2011.

[12] A.G. Kravets, "On approach for the development of patents analysis formal metrics," Communications in Computer and Information Science, vol. 1083, pp. 34-45, 2019.

[13] World Intellectual Property Organization Artificial Intelligence and Intellectual Property [Online]. URL: https://www.wipo.int/about-ip/en/artificial_intelligence/.

[14] D. Korobkin, S. Fomenkov, A. Kravets, S. Kolesnikov and M. Dykov, "Three-steps methodology for patents prior-art retrieval and structured physical knowledge extracting," Communications in Computer and Information Science, vol. 535, pp. 124-136, 2015.

[15] D. Korobkin, S. Fomenkov, A. Kravets and S. Kolesnikov, "Methods of statistical and semantic patent analysis," Communications in Computer and Information Science, vol. 754, pp. 48-61, 2017.

[16] A. Kravets, N. Shumeiko, B. Lempert, N. Salnikova and N. Shcherbakova, "Smart Queue" Approach for new technical solutions discovery in patent applications," Communications in Computer and Information Science, vol. 754, pp. 37-47, 2017.

[17] A.G. Kravets, A.G. Mironenko, S.S. Nazarov and A.D. Kravets, "Patent application text pre-processing for patent examination procedure," Communications in Computer and Information Science, vol. 535, pp. 105-114, 2015.

[18] A.G. Kravets, A.D. Kravets, V.A. Rogachev and I.P. Medintseva, "Cross-thematic modeling of the world prior-art state: rejected patent applications analysis," Journal of Fundamental and Applied Sciences, vol. 8, no. SI 3, pp. 2542-2552, 2016.

[19] M. Fomenkova, D. Korobkin, A.G. Kravets and S. Fomenkov, "Extraction of Knowledge and Processing of the Patent Array," Communications in Computer and Information Science, vol. 1084, pp. 3-14, 2019.