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

A.G. Kravets[1], S.Yu. Biryukov[2], D.G. Skorikov[2], D.N. Marinkin[3,4]

[1]Volgograd State Technical University, Lenin av. 28, Volgograd, Russia, 400005
[2]Volgograd State University, Universitetskiy 100, Volgograd, Russia, 400062
[3] Perm State National Research University, Bukireva 15, Perm, Russia, 614990
[4]Perm State University for the Humanities and Education, Sibirskaya 24, Perm, Russia, 614990

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.

1. 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 the 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]

Perplexity [3] is a criterion for the numerical estimation of the quality of a probabilistic model, equal to the exponent of minus the averaged log-likelihood:

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

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:

$$p(w) = n_w/n$$

2. 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) [4].

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) [5].

The text of the application is sent to the system via the web interface [6]. The most important information is stored in the "Claim" section. It is the novelty of this information that should be checked by the expert [7].

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. And 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. And the third problem it is expert subjectivity and in this decision-making process as it is the riskiest.

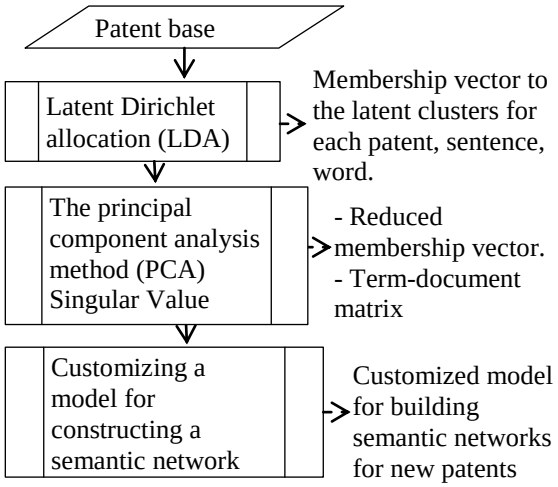

Figure 1. Algorithm for processing the existing patent database.

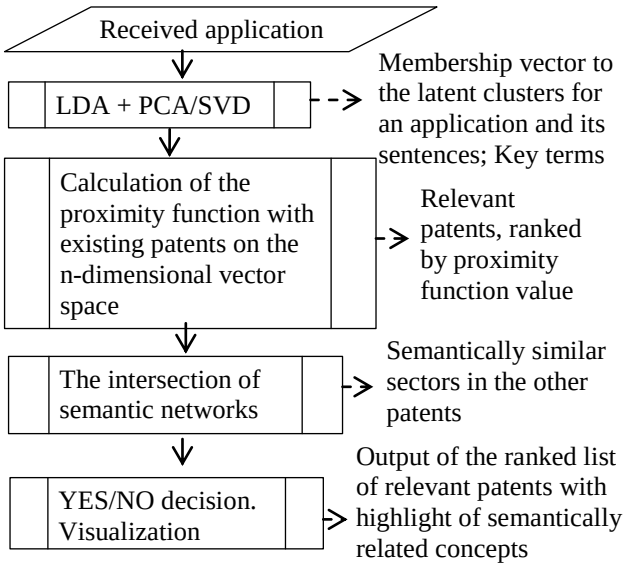

Figure 2. Received application processing algorithm.

3. Models training and experiments

3.1. 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. In order 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).

For each class, a training (4,000 patents) and test (1,000 patents) datasets were created.

To train the model, the Gensim library was used, the resulting models were visualized using the pyLDAvis library.

3.2. 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);
- a number of recoverable topics (K).

3.3. 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 the default.

The results of the experiments series are shown in Figure 3. It can be seen that the increase in the iterations increases the 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.

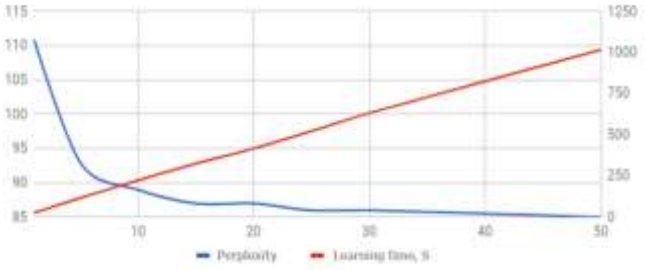

Figure 3. Perplexia and training time diagram.

3.4. Evaluation of the hyperparameters impact to 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 [8, 9] 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 perplexia 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.

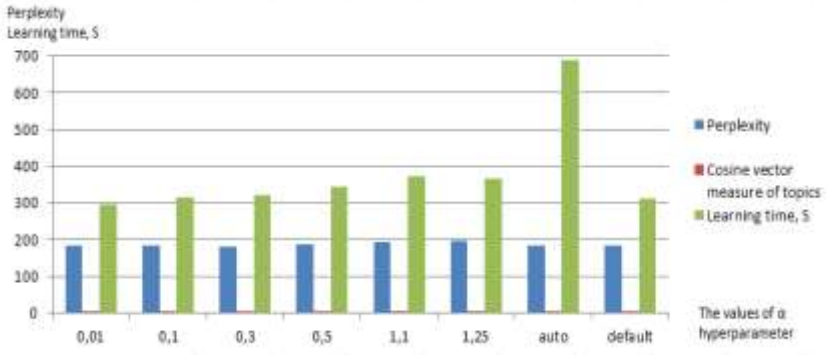

Figure 4. Models behavior with changes of hyperparameters.

As a result, the best value of the parameter α from the presented set is the 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, perplexia 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.

3.5. The number of hidden topics search

The purpose of the second set of experiments is 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 can discover.

Let's make the assumption that the more topical diverse (for a certain K), the more successful is the topics' definition. 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 perplexia 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.

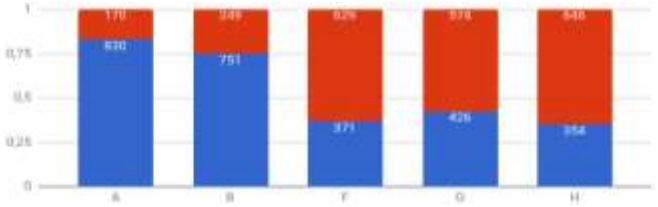

Figure 5. The distribution of patents classes on 2 topics.

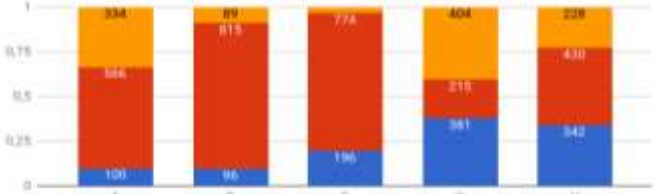

Figure 6. The distribution of patents classes on 3 topics.

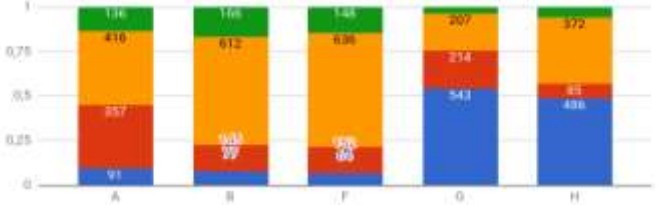

Figure 7. The distribution of patents classes on 4 topics.

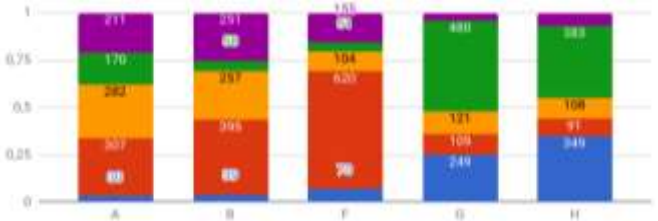

Figure 8. The distribution of patents classes on 5 topics.

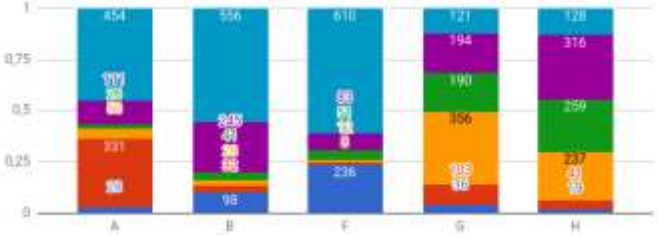

Figure 9. The distribution of patents classes on 6 topics.

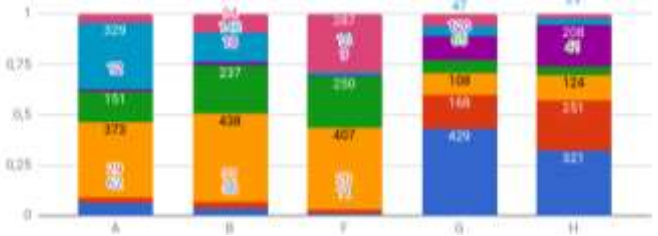

Figure 10. The distribution of patents classes on 7 topics.

4. Results and discussion

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

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 the 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.

5. Conclusions

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 [10].

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.

6. Acknowledgments

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

7. References

[1] Blei, D.M. Latent Dirichlet Allocation / D.M. Blei, A.Y. Ng, M.I. Jordan // Journal of Machine Learning Research. - 2003. – Vol. 3(4-5). – P. 993-1022.

[2] Machine Learning [Electronic resource]. – Access mode: http://www.machinelearning.ru/wiki/ (01.12.2019).

[3] Brown, P.F. An Estimate of an Upper Bound for the Entropy of English // Computational Linguistics. – 1992. – Vol. 18. – P. 2007.

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

[5] Korobkin, D. Methods of statistical and semantic patent analysis / D. Korobkin, S. Fomenkov, A. Kravets, S. Kolesnikov // Communications in Computer and Information Science. – 2017. – Vol. 754. – P. 48-61.

[6] Kravets, A. “Smart Queue” Approach for new technical solutions discovery in patent applications / A. Kravets, N. Shumeiko, B. Lempert, N. Salnikova, N. Shcherbakova // Communications in Computer and Information Science. – 2017. – Vol. 754. – P. 37-47.

[7] Kravets, A.G. On approach for the development of patents analysis formal metrics // Communications in Computer and Information Science. – 2019. – Vol. 1083. – P. 34-45.

[8] Kravets, A.G. Patent application text pre-processing for patent examination procedure / A.G. Kravets, A.G. Mironenko, S.S. Nazarov, A.D. Kravets // Communications in Computer and Information Science. – 2015. – Vol. 535. – P. 105-114.
[9] Kravets, A.G. Cross-thematic modeling of the world prior-art state: rejected patent applications analysis / A.G. Kravets, A.D. Kravets, V.A. Rogachev, I.P. Medintseva // Journal of Fundamental and Applied Sciences. – 2016. – Vol. 8(SI 3). – P. 2542-2552.
[10] Fomenkova, M. Extraction of Knowledge and Processing of the Patent Array / M. Fomenkova, D. Korobkin, A.G. Kravets, S. Fomenkov // Communications in Computer and Information Science. – 2019. – Vol. 1084. – P. 3-14.