Visualization of Research Trending Topic Prediction: Intelligent Method for Data Analysis

Michael Charnine [1], Alexey Tishchenko [2] and Leon Kochiev [1]

[1] *FRC CSC of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow, 119333, Russia*
[2] *The Presidential Academy, RANEPA, 84 Prospekt Vernadskogo, Moscow, 119571, Russia*

Abstract

This paper presents the results of a method for the visualization of the long-term prediction of research trending topics. Meaningful topics were identified among the words included in the titles of scientific articles. The title is the most important element of a scientific article and the main indication of the article's subject and topic. We treated the titles' words, which occur several times in articles cited in the analyzed collection, as the research trending topics. The longevity of the citation trend growth was the target for the machine learning algorithms. The CatBoost machine learning method, which is one of the best implementations of decision trees, was used. We conducted experiments on a scientific dataset that included 5 million publications from the top conferences in artificial intelligence and data mining areas to demonstrate the effectiveness of the proposed model. The accuracy rate of three-year forecasts for a number of experiments from 1997 to 2014 was about 60%. To visualize the forecast, the t-SNE and Word2Vec methods were used. Clusters of trending keywords on the semantic map helped to accurately identify promising directions. Two examples of forecast visualizations for the topic "Intelligent methods for data and image analysis" are presented. The presented visualizations serve as the analytical method for predicting topic trends and promising directions.

Keywords

Visualization, long-term prediction, research trending topics, decision tree, CatBoost, scientific papers, dynamics of topic trends, Big Data

1. Introduction

The long-term prediction of research trending topics helps to efficiently navigate and evaluate scientific articles, identify promising directions, find breakthrough ideas, and focus efforts on the most fruitful direction. The visualization of predicted topics as clusters of trending keywords on the semantic map can help to more accurately identify promising directions.

The long-term prediction of research trending topics can be done using analyses of bibliographic collections of the millions of scientific articles that are freely available on the Internet. This is an example of human–machine interaction, when a computer processes information entered by people in the form of scientific articles to predict trends and promising directions. Knowing these trends helps authors write new articles and focus their efforts on the most fruitful direction.

Trend prediction has become extremely popular in many industrial sectors and in scientific literature. It is useful for strategic planning, decision making, computer-assisted education, and exploring new research directions. Modern machine learning tools are capable of processing tremendous volumes of scientific data and can help find promising directions for future research and potentially lead to breakthrough discoveries in any field of science.

GraphiCon 2021: 31st International Conference on Computer Graphics and Vision, September 27-30, 2021, Nizhny Novgorod, Russia
EMAIL: mc@keywen.com (M. Charnine); alexeyseti82@yandex.ru (A. Tishchenko); kochiev.lg@phystech.edu (L. Kochiev)
ORCID: 0000-0003-0450-5156 (M. Charnine); 0000-0002-5834-5760 (A.Tishchenko); 0000-0003-3167-8490 (L. Kochiev)

Finding topics from a collection of research publications is helpful for summarizing large-scale text collections and predicting future topic trends. This can be beneficial for many applications, such as forecasting future trends of the IT industry. In order to forecast future topic trends (topic forecasting), it is first necessary to discover meaningful topics from a set of documents (topic mining or discovery). Finding meaningful topics from a set of documents has been studied in existing research [3].

In this work, we discovered meaningful topics among the words included in the titles of scientific articles. The title is the most important element of a scientific article and the main indication of the article's subject and topic [4]. Therefore, we used the title as the topic summarization of each paper. We treated the title words/keywords, which occur over five times in the titles of cited articles of the analyzed collection, as the research trending topics. In this way, we moved from topics to relevant words, and instead of the trends of topics, we explored the trends of the corresponding keywords. Unrelated words, such as prepositions, were filtered out before extracting keywords from the titles.

We studied keyword trends via the dynamics of various indicators in groups of articles containing these keywords. The most important indicator of a trending keyword is the keyword citation count (KCC). To calculate this indicator, we first find all articles with this keyword in a particular year, and then we count all citing links to those articles. The values of this indicator are different in different years. For each word, we calculated the duration of its trend growth, which is equal to the number of years of continuous growth of its average citation count. This longevity/duration of trend growth was the target for the machine learning algorithm.

Thus, for each word/keyword included in the article titles in the collection, the time series were built from various indicators/features of this word, including its citation count. The regression model and a forecast of the duration of trend growth were built using this time series and machine learning. The duration of trend growth is a scalar function that machine learning models learned for each word/term. When creating the time series and calculating prediction errors, the parameter of the maximum forecast duration (1, 3, 5, and 7 years) was used, which reduced the real and forecast duration if their values exceeded the value of this parameter.

In Section 4, the results of the experiments assessing the accuracy of forecasting the long-term growth of trends for 1, 3, 5, and 7 years are presented. It should be noted that the machine learning model makes it possible to evaluate the significance of each indicator/feature using its effect on the forecast accuracy. It was noted that this effect was different for different prognosis durations (1, 3, 5, and 7 years), and for some indicators, their significance increased with an increase in the duration.

Thus, the analysis of the accuracy of the long-term forecast allowed us to find significant indicators/features. In other words, the impact on the accuracy of the long-term forecast is an objective measure for assessing the significance of features.

The contributions of this paper are as follows:

- To the best of our knowledge, we are the first to study the problem of the long-term prediction of thousands of research trending topics from academic big data.
- We conducted experiments on a scientific dataset that included millions publications in artificial intelligence and data mining areas to demonstrate the effectiveness of the proposed model.
- We have developed a trend forecast visualization method that improves forecast accuracy and serves as an analytical method for predicting topic trends and promising directions.

2. Related works

One of the important research topics is the prediction of the trends of scientific development. Below, we will consider the works related to this area and pay special attention to the accuracy of long-term predictions of topic trends, if described in the works.

In their survey [5], Houa and others (2019) analyzed methods and applications in data-driven prediction in the science of science and discussed their significance. They summarized the research issues from three perspectives: the papers' impact prediction, scholar impact prediction, and author collaboration prediction. The authors did not touch on the problem of predicting topic trends, apparently due to the relatively small number of articles on this topic.

The work of Hurtado and others (2015) is devoted to the issues of topic discovery and future trend prediction [6]. In this paper, the authors propose, using association analysis and ensemble forecasting,

to automatically discover topics from a set of text documents and forecast their evolving trends in the near future. The authors also note that there has been significant progress in the area of topic forecasting in time-sensitive domains such as Twitter, but very few works exist in scientific publications.

The work of Prabhakaran and others (2016) is devoted to predicting the rise and fall of scientific topics from trends in their rhetorical framing [7]. The authors note that little is known about the mechanisms underlying topic growth and decline. The authors found that a topic's rhetorical function is highly predictive of its eventual growth or decline. For example, topics that are rhetorically described as results tend to be in decline, while topics that function as methods tend to be in early phases of growth.

The work of Shen and others (2016) is devoted to modeling topic-level academic influence in scientific literatures [8]. Tn this paper, the authors introduce J-Index, a quantitative metric of modeling paper's academic influence. For each paper, J-Index considers its citation number, the strength of each citation and the novelty of all papers where it is cited. The authors propose the Reference Topic Model (RefTM) to measure the novelty of each paper as well as the citation strength among them. RefTM can effectively discover topics of high quality, model paper novelty and predict citation strength.

The work of Chen and others (2018) is devoted to modeling scientific influence for research trending topic prediction [9]. The authors note that providing insights into the topics that have the potential to become trending topics in the future is important for researchers to catch up with the rapid progress of research. The authors use the publications before 2015 to predict the trending topical words in 2016.

The work of Wang and others (2019) is devoted to detecting hot topics from academic big data [10]. In this work a DeepWalk-based keyword extraction algorithm is proposed to detect popular topics in diverse academic fields dynamically. Also the authors propose a keyword extraction algorithm to extract keywords from multiple articles, which enables them to detect new topics in emerging academic areas. The authors adopt the K-Means algorithm to cluster hot keywords to get hot topics. For the future work, the authors plan to predict academic hotspots in the future, which is of great significance for researchers to grasp the future research direction.

The above works discussed the issues of detecting hot topics and predicting topic trends, but did not touch the issues with the accuracy of long-term forecasting of thousands of research trending topics from academic big data, which is our main goal.

Various machine learning methods are used for forecasting. Decision trees and neural network methods demonstrate good prediction and classification quality compared to such machine learning methods as Random Forest, Support Vector Machine, Naive Bayes, and K Nearest Neighbor. The paper [11] compared ensembles of decision trees and neural networks for one-day-ahead streamflow prediction. The results obtained in this study indicate that ensemble learning models yield better prediction accuracy than a conventional artificial neural network model, such as a multilayer perceptron. Moreover, artificial neural network ensembles are superior to tree-based ensembles.

In this work, we used the CatBoost machine learning method, which is one of the best implementations of decision trees [12]. CatBoost is a high-performance open source library for gradient boosting on decision trees. CatBoost is the implementation of ordered boosting, a permutation-driven alternative to the classic algorithm, which was created to fight a prediction shift caused by a special kind of target leakage present in all currently existing implementations of gradient boosting algorithms.

3. AI collection (data set)

In our experiments, we analyzed the DBLP citation network, which is a collection of articles on artificial intelligence from 1936 to 2020, compiled by aminer.org and referred to here as the AI collection.

The citation data is extracted from DBLP (Digital Bibliography & Library Project dblp.org), ACM (Association for Computing Machinery acm.org), MAG (Microsoft Academic Graph), and other sources.

We used the V12 version released in April 2020. This data set consists of 4,894,081 articles and 45,564,149 citation relationships. For each article there is a title, authors, year of publication and links. We have processed all titles, venues, authors, and citation relationships.

In this study, we used the DBLP information network, which exactly follows the network schema presented in Figure 1. The schema has links between "Term" and "Paper (Title)," "Author" and "Paper," as well as publication "Venue" and "Paper," all of which are observable during the trend prediction process. The title is the topic summarization of each paper. Thus, we treated the title words/keywords, which occur over 5 times in the titles of cited articles of the AI collection, as the trending research topics. There are millions of such keywords, and we have analyzed all of them. In this paper, the AI collection was analyzed in different directions described in the next section.

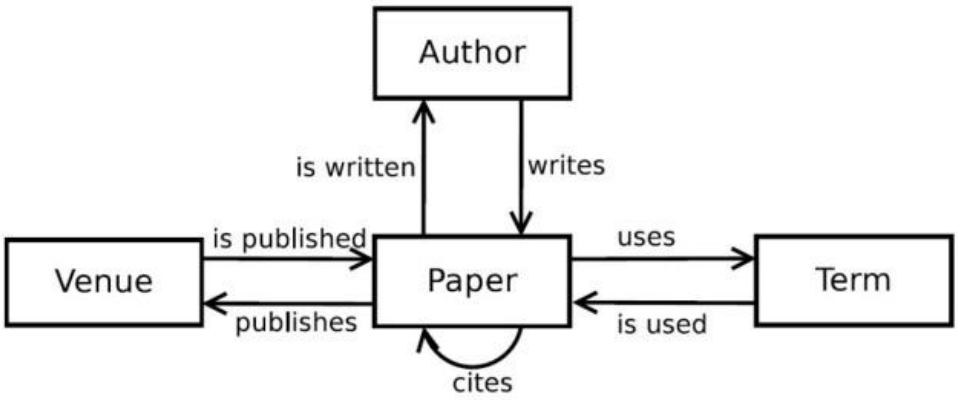

Figure 1: The DBLP network schema

4. Experiments

For this paper, we studied keyword trends via the dynamics of various indicators/features of groups of articles containing these keywords. The most important feature of the trending keyword is the keyword citation count (KCC) being equal to the total citation of the corresponding group of articles. For each keyword, we calculated the long-term growth of its trend that is equal to the number of years of continuous growth of the moving average of KCC trend. This duration of trend growth was the target for the CatBoost machine learning algorithm.

The CatBoost regression model was trained for 20 trend features of the word/keyword, including:

- the number of years of growth of the future trend (this must be predicted);
- current KCC (a number of features for the last 6 years);
- total number of articles with the word (in the current and previous years— a number of features);
- the number of years of growth of the previous trend;
- total citation growth for the previous trend;
- the number of years from the beginning of the trend of the word;
- the number of citing links between articles with the word since the beginning of the trend (in the current and previous years—a number of features);
- the number of years from the trend situation to the current/ base year T.

Thus, for each keyword included in the article titles in the AI collection, time series were built from 20 different indicators/features of this keyword, including its KCC. The regression model and a forecast of the duration of KCC trend growth were built using this time series and machine learning method. The duration of KCC trend growth was measured as the number of years of growth of the moving average of KCC trend. The model looked for the first instance of a downward trend of a moving average and reported the number of years of growth up to that point. The moving average of KCC trend was always calculated for 3 years, and it is different from the parameter D of the maximum forecast duration.

When compiling the time series and calculating prediction errors, the parameter D of the maximum forecast duration (1, 3, 5, and 7 years) was used, which limits the real duration and predicted duration to the value of D. The parameter T was also used to indicate the current/base year. When training the model, all information related to the time after T was deleted. The accuracy and forecast error of the trained model was checked at the time T + D on completely different data than in training.

Figure 2 shows the results of experiments on estimating errors in forecasting the long-term growth of trends for 1, 3, 5, and 7 years. In Figure 2, the vertical axis represents the forecast error, while the horizontal axis represents current/base year of the experiment. Data Series1 represents a 7-year forecast and Data Series4 represents a 1-year forecast.

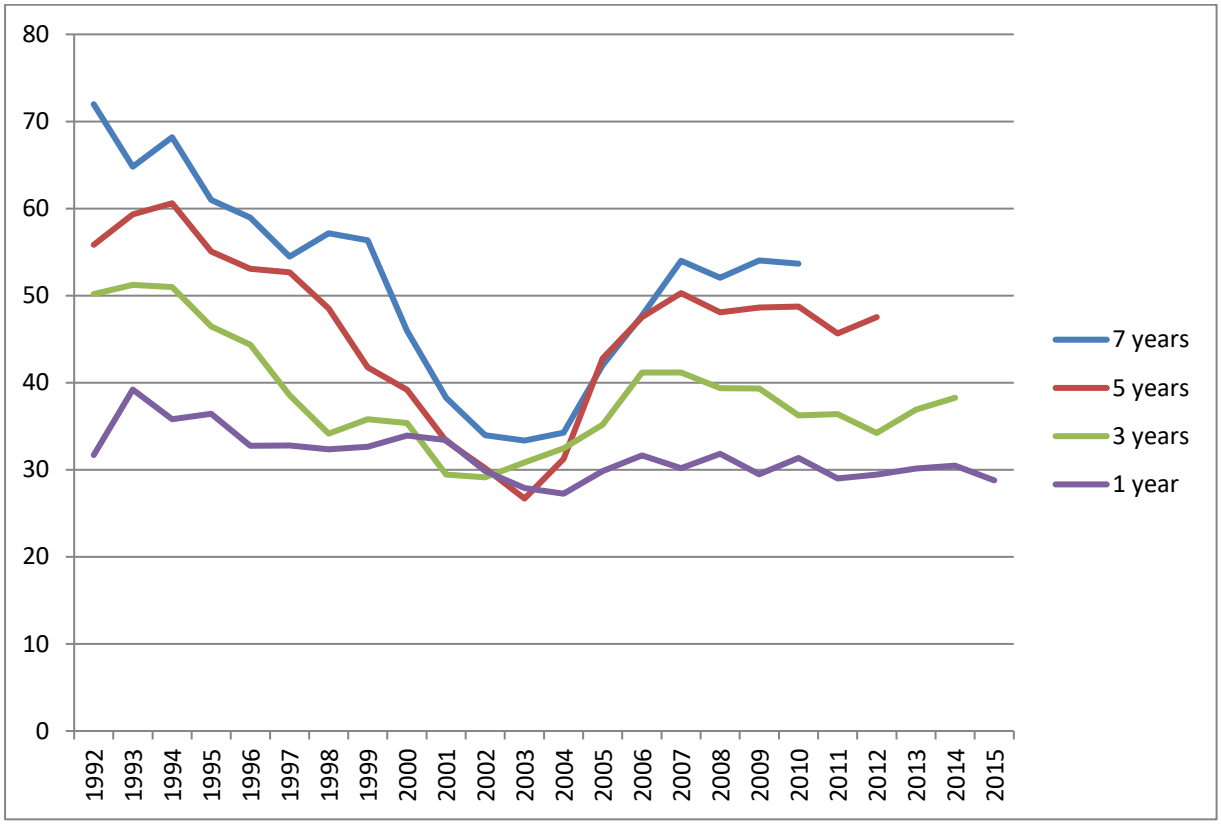

Figure 2: Errors in forecasting the long-term growth of trends in different years

Figure 2 shows that the longer the forecast, the higher its error rate. The error rate of three-year forecasts for a number of experiments from 1997 to 2014 was about 40%, whereas the accuracy rate was 60%. The maximum error rate of seven-year forecasts for a number of experiments from 1997 to 2010 was about 58%. The error rate of seven-year forecasts from 2000 to 2004 was less than 40%.

The number of keywords with long-term trends for 1, 3, 5 and 7 years in different base years is shown in Figure 3. In Figure 3, the vertical axis represents the number of keywords with long-term trends, while the horizontal axis represents the current/base year of the experiment. Data Series1 represents a 7-year forecast and Data Series4 represents a 1-year forecast. A trend is considered long-term if its duration exceeds parameter D.

Figure 3 shows that the more long-term trends considered, the higher the accuracy. In the early years of the AI collection, there were fewer articles, fewer words, and correspondingly fewer growing trends. The number of trends has been decreasing in recent years, as no data is yet available.

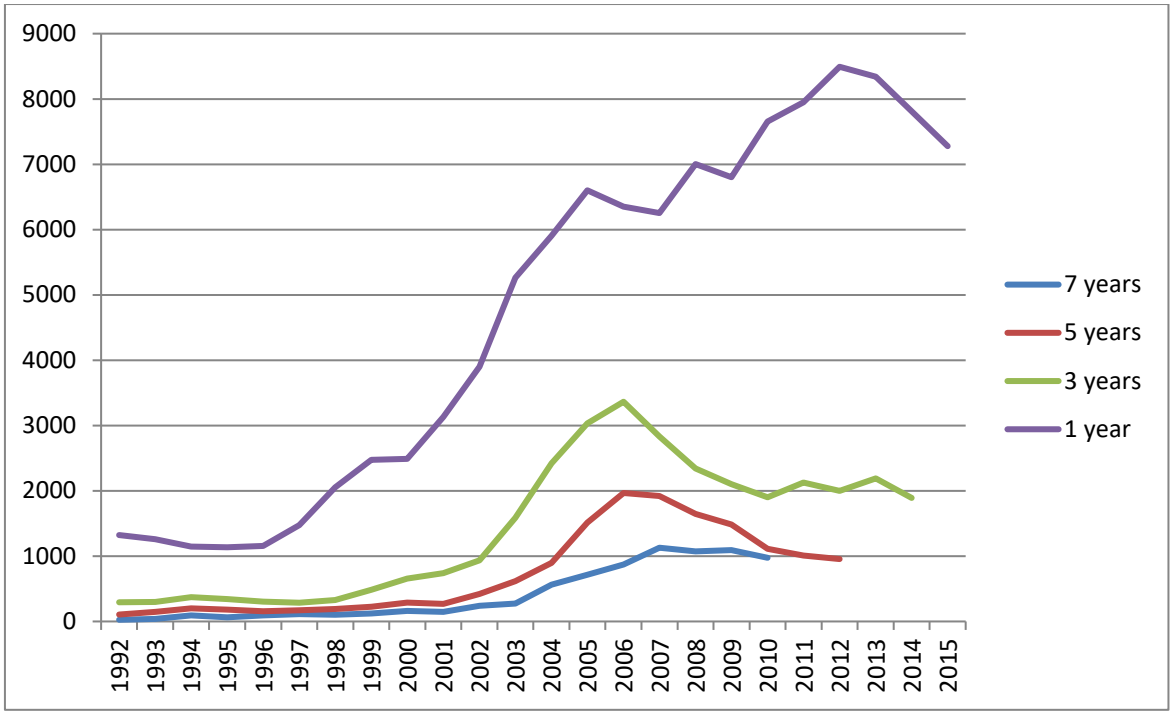

Figure 3: The number of keywords with long-term trends

As a result of a predictive experiment, 5587 keywords with a positive duration of trend growth were identified using only data before 2018. The most long-lasting forecast of the trend growth duration (over 10 years) had the following topics and their representative keywords: artificial intelligence, ai, convolutional networks, cnn, cnn-based, deep network, using deep, via deep, explainable, unsupervised learning, learning-based, comprehension, reading comprehension, adversarial, adversarial learning, lstm , and local differential.

Thus, topics related to artificial intelligence and neural networks (CNN, LSTM) had the most long-lasting predicted trends. This automatic forecast is consistent with the expectations of experts. Also, an interesting trend in scientific literature is the development of "explainable" methods. This trend is very important for the topic of cognitive human–machine interaction because it is often not clear why and how neural networks get their results.

To search for trending keywords in the topic INTELLIGENT METHODS FOR DATA AND IMAGE ANALYSIS, a query was formed containing the following specific words related to this topic: artificial intelligence, graphical, visualization, visual, vision, image, recognition, virtual, reality, augmented, video analytics, data analytic, and robotics.

This query was used for selecting the following trending keywords with the biggest forecasts: artificial intelligence (14.1), image recognition (10.3), computer vision (8.0), biomedical image (7.4), image captioning (7.4), virtual reality (6.4), satellite imagery (5.1), and image dehazing (4.6). The predicted duration of the trend growth after 2018 is indicated in parentheses. All selected trending keywords contain the requested specific words.

Thus, according to the results of the forecast on the topic INTELLIGENT METHODS FOR DATA AND IMAGE ANALYSIS, the longest-growing trends have the sub-topics: image recognition, computer vision, and biomedical image. This automatic forecast is also consistent with the expectations of experts.

5. Visualization

To visualize the forecast, the Word2Vec [1] and t-SNE [2] methods were used. The visualization algorithm consists of the following steps:

1. The collection of scientific articles is analyzed, trending keywords are found, and a forecast of the growth time of their trends is calculated using the method described in Section 4.
2. Using the Word2Vec method, the semantic similarity between trending keywords is calculated, and a similarity matrix for trending keywords is constructed.
3. The coordinates of points on the plane are calculated using the t-SNE method.
4. A visual map is built and keywords are marked on it with dots of different colors, depending on their forecast. The keywords with the longest forecast of the trend growth time are shown in red, and the keywords with the shortest forecast are indicated in black.

An example of the forecast visualization for the topic "Intelligent methods for data and image analysis" is presented below. The t-SNE algorithm works in an ambiguous manner and builds several different semantic maps for a single similarity matrix. Figures 4 and 5 show two such semantic maps.

Figures 4 and 5 show that semantically similar keywords are grouped into clusters on the semantic map. As it turned out, these clusters mainly consist of keywords with similar trend durations. In this way, trending keywords on the semantic map can confirm the duration of each other's trends.

The keywords with the longest trends are marked in red and form an image processing cluster (computer vision, image recognition, biomedical image, etc.). The next cluster is formed by the keywords: virtual reality, augmented virtuality, graphical modeling, and educational robotics. Most of the keywords in this cluster have medium trends and are indicated in blue. The third cluster is formed by keywords with short trends, which are marked in black. This cluster is related to visualization and analysis. The keyword colors and their predictions are calculated independently, so combining keywords into homogeneous clusters helps improve the prediction accuracy for clusters and topics.

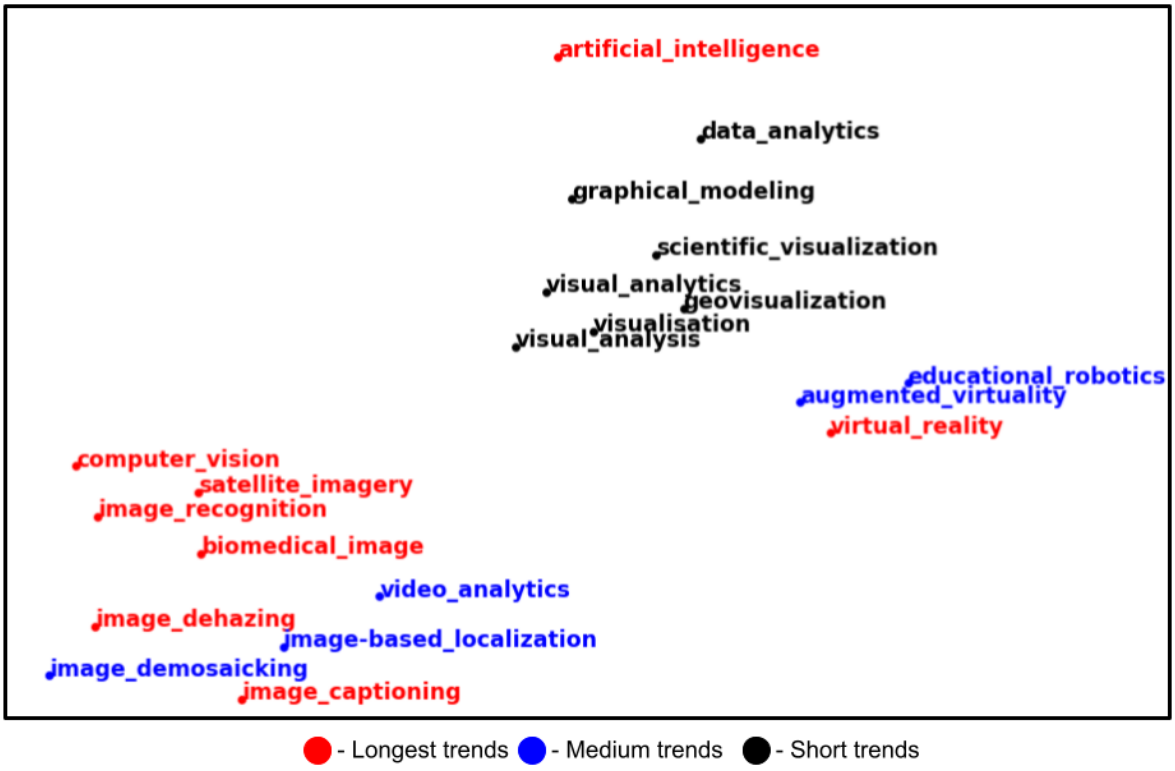

Figure 4: Visualization number one of keywords with long-term trends based on t-SNE projections of the Word2Vec similarity matrix between keywords and a forecast of the growth time of their trends (red: keywords with the longest trends, blue: keywords with medium trends, black: keywords with the shortest trends).

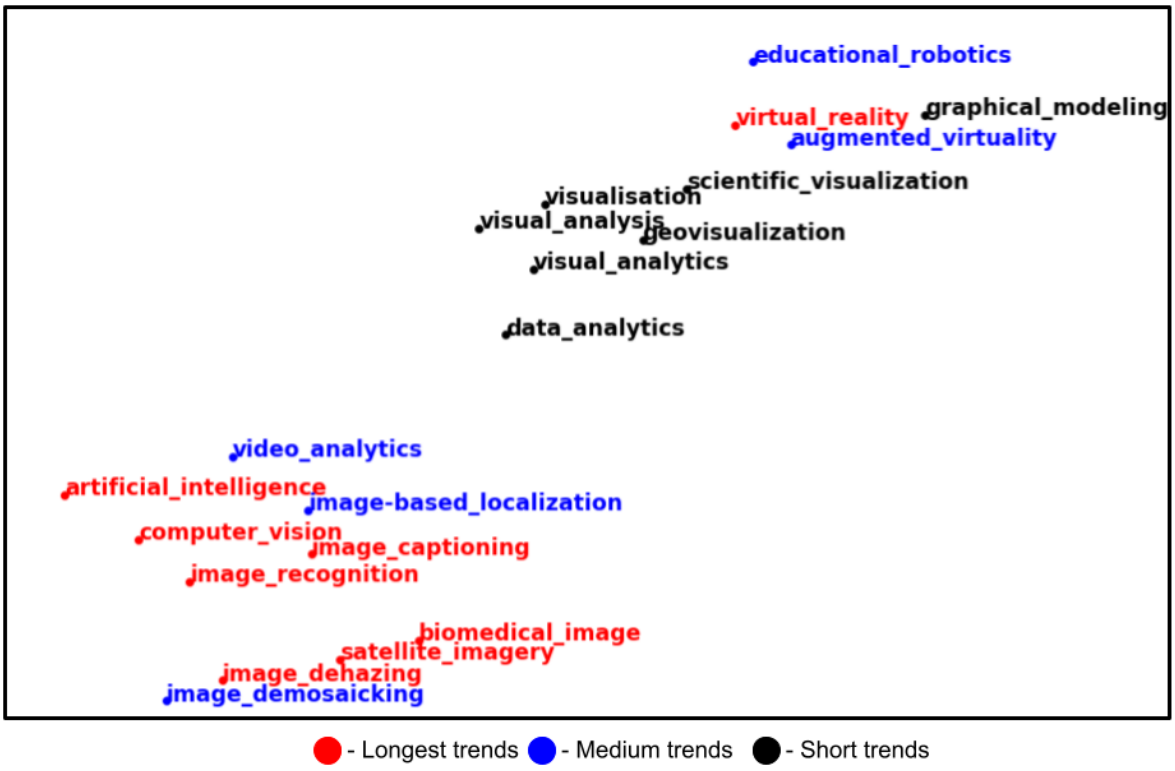

Figure 5: Visualization number two of keywords with long-term trends based on t-SNE projections of the Word2Vec similarity matrix between keywords and a forecast of their trends. The figure illustrates the ability of the model to automatically organize keywords into clusters/topics of the same color with similar trend durations.

As you can see from Figures 4 and 5, the close points/keywords on the semantic map often have the same colors, despite the fact that the color of each keyword and its predicted trend were calculated independently. Spatial proximity of semantically similar keywords is provided using neural network technology, Word2Vec, and t-SNE methods. Semantically similar keywords with the same colors form clusters on the map. The more such keywords are included in a cluster, the larger the cluster size and the more accurate the forecast for the topic corresponding to this cluster. The forecast for a topic is determined as the average of the forecasts for keywords included in the cluster. With the help of such averaging, the forecast accuracy is increased.

The average is used to obtain a generalizing characteristic of some datasets. If the data is more or less homogeneous and there are no anomalous observations (outliers), then the average generalizes the data well, which minimizes the influence of random factors (they cancel each other out during addition). According to the law of large numbers, the average of the results obtained from a large number of trials should be close to the expected value and will tend to become closer to the expected value as more trials are performed [13]. Thus, trending keywords on the semantic map confirm each other and help to more accurately identify promising directions.

Using such semantic maps, we can visually estimate the number of keywords in a cluster of the same color, the semantic similarity of these keywords, and what topic these keywords form, with ease. Such visual analytics makes it possible to more accurately predict the duration of topic trends. It should also be noted that semantic maps detail the strength of semantic relationships between keywords and clusters. Also, the structure of the keywords included in the clusters provides an idea of promising methods and approaches that can be used in this direction. Therefore, the visualizations presented in this paper serve as a proposed analytical method for predicting topic trends and promising directions.

6. Conclusion

This work has several key contributions. To the best of our knowledge, we are the first to study the problem of the long-term prediction of thousands of research trending topics from academic big data. The visualization of such long-term predictions is a form of cognitive enhancement because it helps to efficiently navigate and evaluate research topics, identify promising directions, and focus efforts on these directions.

We conducted experiments on a scientific dataset that included millions publications in artificial intelligence and data mining areas to demonstrate the effectiveness of the proposed model. The error rate of three-year forecasts of research trending topics in a number of experiments from 1997 to 2014 was about 40%, whereas the accuracy rate was 60%. As a result of the predictive experiment, it was found that the topics related to artificial intelligence and neural networks (CNN, LSTM) had the most long-lasting predicted trends. This automatic forecast is consistent with the expectations of experts. A further increase in the accuracy of the proposed method for the long-term forecasting of trending scientific topics is possible through the use of ensembles of deep learning neural networks, including convolutional neural network (CNN) and long short-term memory (LSTM).

The results of the long-term forecasting of trends for individual keywords were used to build visualizations on the topic "Intelligent methods for data and image analysis". The visualizations are built in the form of semantic maps, wherein keywords with different predictions are indicated in different colors. Visualizations made it possible to combine keywords into clusters and topics, as well as create more accurate predictions for topics.

A very interesting result of this work is that semantically similar keywords are grouped into clusters which mainly consist of keywords of the same color with similar trend durations. Thus, trending keywords on the semantic maps confirm each other and help to more accurately identify promising directions. The visualizations presented in this paper serve as a proposed analytical method for predicting topic trends and promising directions.

7. Acknowledgements

This work is supported by Russian Foundation for Basic Research, grant 19-07-00857. We are grateful to the Russian Foundation for Basic Research for financial support of our projects.

8. References

[1] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, in: Proceedings of Workshop at ICLR, 2013.

[2] L.J.P. van der Maaten, G.E. Hinton, Visualizing Data Using t-SNE, Journal of Machine Learning Research 9 (2008) 2579−2605.

[3] Q. Mei, C. Zhai, Discovering evolutionary theme patterns from text: an exploration of temporal text mining, in: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining, 2005.

[4] H. R. Jamali, M. Nikzad, Article title type and its relation with the number of downloads and citations, Scientometrics 88(2) (2011) 653–661.

[5] Jie Hou, Hanxiao Pan, Teng Guo, Ivan Lee, Xiangjie Kong, Feng Xia, Prediction Methods and Applications in the Science of Science: A Survey, Computer Science Review 34 (2019) 100197. doi: 10.1016/j.cosrev.2019.100197.

[6] J. Hurtado, S. Huang, X. Zhu, Topic Discovery and Future Trend Prediction Using Association Analysis and Ensemble Forecasting, in: 2015 IEEE International Conference on Information Reuse and Integration, 2015, pp. 203-206. doi: 10.1109/IRI.2015.40.

[7] V. Prabhakaran, W. L. Hamilton, D. McFarland, D. Jurafsky, Predicting the rise and fall of scientific topics from trends in their rhetorical framing, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1170–1180.

[8] J. Shen, Z. Song, S. Li, Z. Tan, Y. Mao, L. Fu, L. Song, X. Wang, Modeling topic-level academic influence in scientific literatures, in: AAAI Workshop: Scholarly Big Data, 2016.

[9] Chengyao Chen, Zhitao Wang, Wenjie Li, Xu Sun, “Modeling Scientific Influence for Research Trending Topic Prediction”, in: Proceedings of the AAAI Conference on Artificial Intelligence 32 (1), 2018.
[10] B. Wang, B. Yang, S. Shan and H. Chen, "Detecting Hot Topics From Academic Big Data," in: IEEE Access, vol. 7, pp. 185916-185927, 2019, doi: 10.1109/ACCESS.2019.2960285.
[11] O. Karakurt, H.I. Erdal, E. Namli, H. Yumurtaci Aydogmus, Y.S. Turkan, "Comparing ensembles of decision trees and neural networks for one-day-ahead streamflow prediction", Sci. Res. J., 2013.
[12] L.Prokhorenkova, G.Gusev, A.Vorobev, A.V.Dorogush, A.Gulin, CatBoost: unbiased boosting with categorical features. arXiv preprint arXiv:1706.09516.
[13] Michel Dekking, A Modern Introduction to Probability and Statistics, Springer (2005), pp. 181–190. ISBN 9781852338961.