Semantic Approach to Visualization of Evolution Dynamics of Topic Trends in Space of Scientific Publications Using t-SNE and Web-based 3D Graphics

M. Charnine[1], E. Sokolov[1], A. Klokov[2]
mc@keywen.com|evgeny.sokolov@phystech.edu|aaklokov@yandex.ru
[1]FRC CSC of the Russian Academy of Sciences, Moscow, Russia;
[2]Moscow Institute of Physics and Technology, Moscow, Russia

This paper describes a semantic approach to visualization of 3D cyberspace of Artificial Intelligence (AI) publications and their topic trends evolution using web-based 3D graphics. The purpose of research is to group AI publications with same subject into clusters for further visualization of topic trends dynamics. An unsupervised method and algorithm for visualizing the dynamics of topic trends by generating a time series of 2D and 3D semantic visual maps with predictive information is described. The method includes semantic similarity measure and citation prediction for documents, topic modeling and clustering, dimensionality reduction, virtual reality technology, representation of dynamics using time filters. As an example of particular implementation, the method is demonstrated on AI collection data using technologies of neural network prediction, LDA clustering, t-SNE dimensionality reduction, WebVR visualization. Cluster dynamics associated with scientific trends is analyzed. The growth in number of clusters and their consolidation during the period from 1954 to 1993 is demonstrated. It is shown that 3D visual map better preserves articles similarity and high-dimensional clusters structure than 2D visual map. The proposed cyberspace implemented by WebVR and interactive 3D graphics can be considered as a dynamic learning environment that is convenient for discovering new significant articles, ideas and trends.

Keywords: *virtual reality, web-based 3D graphics, WebVR, scientific papers, topic modeling, dynamics of topic trends, semantic similarity, visualization, visual map.*

1. Introduction

Trend prediction has become an extremely popular in many industrial sectors and scientific literature. It is beneficial for strategic planning and decision making, and facilitates exploring new research directions. Modern Artificial Intelligence tools are capable of processing tremendous volumes of data, reaching the human-level performance for various applications. Achieving high performance in unsupervised prediction and visualization of emerging trends in scientific literature can discover promising directions for future research and potentially lead to breakthrough discoveries in any field of science. Visual analytics helps us to make more reliable forecasts and increase the credibility of the model used. Visual analytics not only lets us predict future data values based on historic data over time, but we can also include underlying factors and filters, which may improve the accuracy of our forecasts. In our previous works [3,6], we proposed a model of cyberspace in which the most cited and significant scientific papers are represented by spheres/objects of a large size and the distance between papers is based on their semantic similarity. This work is a step forward in the same line of research. We offer here a new way to visualize the dynamics of topic trends using the proposed cyberspace for the purpose of predictive visual analytics. In scientific literature researchers analyze trends for individual words/terms, for sets of similar terms (topics), as well as for sets/clusters of semantically similar documents and texts. In this paper we present a new way to visualize trends using semantic space of scientific documents (cyberspace) with time-varying clusters/topics. The dynamic changes of these topical clusters and their interconnections in the cyberspace can be viewed as an animation.

Cybernetic worlds and virtual reality (VR) are often used for educational and research purposes. VR is essentially a computer-generated reality, also referred to as cyberspace, virtual environment, simulations, and artificial worlds [10]. The idea of creating cyberspace of scientific papers is not a new idea. One of the first implementations of such cyberspace is the Semantic Constellation created by Chaomei Chen [4,5,2]. In the Constellation the papers are presented by spheres in 3D space and the distance between papers is based on a computed measure of semantic similarity between them.

Based on the achievements of AI in recent years, in our previous works [3, 6], we have proposed 3D space of scientific papers that is similar to Chen's Constellation, but we have improved visualization features and semantic accuracy. In addition to that, in this paper, we have improved the dynamics representation and prediction features of the proposed cyberspace.

In recent years, a great progress has been made both in the development of new visualization tools such as WebVR [9] and in the analysis of semantic textual similarity. Latent Dirichlet Allocation (LDA), Word2Vec [1,8] and Word Mover's Distance (WMD) [11] methods were often used to evaluate the semantic similarity of texts. LDA is unsupervised topic modeling and text-mining tool that is frequently used for discovery of hidden semantic structures in a text body. The Word2Vec method evaluates semantic similarity of different words, while the WMD method is able to evaluate semantic similarity of different phrases without common words. In our approach the visualization and cyberspace construction is implemented using t-distributed Stochastic Neighborhood Embedding (t-SNE) method, that is a clustering and visualization tool.

The achievements of AI of the last years described above were used in our works [3,6], but a key question remained unanswered: How to represent the dynamics and forecasts in the cyberspace of scientific papers? This question is very important because it helps to discover promising directions for future research. The answer to this question is given in this article in which we propose a new method for visualizing dynamics of topic trends in the cyberspace of scientific papers.

In Section 3 we describe new unsupervised method and algorithm for visualizing the dynamics of topic trends by generating a time series of 2D and 3D semantic visual maps with predictive information. The method includes semantic similarity measure and citation prediction for documents, topic modeling and clustering, dimensionality reduction, virtual reality technology, representation of dynamics using time filters. As an example of particular implementation, the method is demonstrated using technologies of neural network prediction, LDA clustering, t-SNE dimensionality reduction and WebVR visualization. The method is demonstrated on the AI collection data described in the next Section.

2. AI collection (Data Set)

In our experiments, we analyze DBLP citation network, which is a collection of articles on Artificial Intelligence from 1936 to 2017, compiled by aminer.org and referred to here as 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 V10 version released in October 2017. This data set consists of 3,079,007 articles and 25,166,994 citation relationships. For each article there is a title, authors, year of publication and links. Some articles have a brief description (abstract). We have processed all such articles with abstracts. In total, 118,768 articles with abstracts from 1949 to 1993 were processed. The considered data set does not contain abstracts for articles after 1993. That's why we have analyzed and demonstrated the growth in number of articles clusters and their consolidation only during the period from 1954 to 1993. In this paper, the AI collection was analyzed in different directions: topical clusters and their dynamics were discovered (Section 3), prediction of cited articles (Section 4) were calculated, calculations were visualized in 2D (Section 5) and 3D (Section 6).

3. Unsupervised method for visualizing the dynamics of topic trends

Our goal is to create an unsupervised method and algorithm for visualizing the dynamics of topic trends by generating a time series of 2D and 3D semantic visual maps with predictive information. The method includes semantic similarity measure and citation prediction for documents, topic modeling and clustering, dimensionality reduction, virtual reality technology, representation of dynamics using time filters. This method/algorithm consists of the following steps.

1) Data retrieval and extracting the following fields from the received data: article titles, brief descriptions, and publication years.
2) Remove all the numbers and bring the text to lower case.
3) Extract tokens from the text.
4) Stemming and removing stop-words.
5) Citation prediction for documents.
6) Topic modeling, clustering and calculating semantic similarity measure.
7) Dimensionality reduction and coordinates calculation by compressing n-dimensional vector space into 2D or 3D.
8) Generating a time series of 2D and 3D semantic visual maps using virtual reality technology, time filters and citation prediction information.

Each step of the above method can have multiple different implementations. Below is an example of particular implementation in which the method is demonstrated on AI collection data using technologies of neural network prediction, LDA clustering, t-SNE dimensionality reduction, WebVR visualization. This particular implementation of the method/algorithm consists of the similar steps from 1 to 8 supplemented with detailed information.

1) AI collection data retrieval and extracting the following fields from the received data: article titles, brief descriptions, and publication years.
2) Remove all the numbers and bring the text to lower case.
3) Extract tokens from the text.
4) Stemming and removing stop-words.
5) Citation prediction for documents using neural network technology described in the Section 4.
6) Topic modeling, clustering and calculating semantic similarity measure. Topic modeling is done using Latent Dirichlet Allocation (LDA). As a result, we obtain θ vectors, which show how topics are distributed in each document, and β distributions, which words are more likely to be in certain topics. In our case, there are 8 distinct topical clusters described in Section 5. Thus, we use LDA to get a list of topics and a list of words specific to each topic, i.e. an actual topic description. All training is conducted without a teacher (unsupervised), because input data consists from unmarked texts of articles.
7) Dimensionality reduction and coordinates calculation by compressing n-dimensional vector space into 2D or 3D using t-SNE algorithm. T-distributed Stochastic Neighborhood Embedding (t-SNE), a clustering and visualization method [7] has rapidly become a standard tool in a number of natural sciences. T-SNE is used to facilitate visual inspection and provides two and three dimensional embeddings of high-dimensional data which preserve data similarity. T-SNE is a fruitful technological method with a lot of options for further development. It is efficiently used to visualize complex multidimensional objects including scientific publications. However, t-SNE is usually not used to visualize dynamic processes. In this paper we develop an approach for using t-SNE to visualize the dynamics of topic trends of scientific publications.
8) Generating a time series of 2D and 3D semantic visual maps using virtual reality technology, time filters and citation prediction information. Visual maps generation is done using time filters with the ability to change time intervals in order to track the dynamics of topical clusters development (see Section 5 and Section 6).

4. Prediction of cited articles

The most cited and significant articles in the AI collection were discovered using a deep learning neural network algorithm. The algorithm was created using Python library Keras which allows you to select the number of neural network layers and specify the number of neurons in each layer, the activation function, etc.

The proposed algorithm predicts the probability of citing in the next 3 years based on historical data about titles or abstracts of the articles. The neural network architecture of the proposed algorithm consisted of 4 layers: the first is the layer of embeddings obtained using the Word2Vec architecture, the second layer is biLSTM. The third layer was a linear layer and solved the problem of classification. The last layer consisted of one output neuron. The prediction accuracy for titles of 2017 was ~ 0.6 and for abstracts of 1993 ~ 0.63 by the ROC AUC metric.

For visualization, 20 articles with abstracts and largest citation forecast were selected. These articles are shown as polyhedra octahedrons in Fig. 4. The coordinates of these articles in 3D space were calculated by t-SNE algorithm. The neural network architecture of the proposed deep learning algorithm is presented in the Fig. 1.

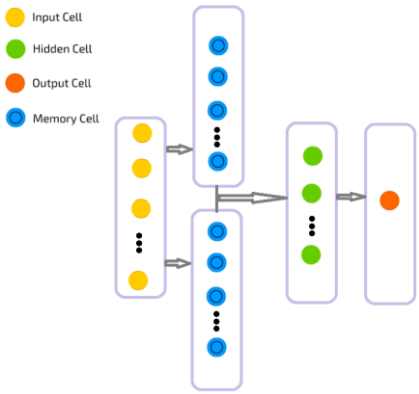

Fig. 1. Neural network architecture.

5. Creating 2D visual map

2D visual maps are created by compressing n-dimensional vector space into 2D using t-SNE algorithm. Visualization is carried out by generating a time series of 2D semantic visual maps using time filters with the ability to change time intervals in order to track the dynamics of topical clusters development. Visualization is done using an interface written in JavaScript.

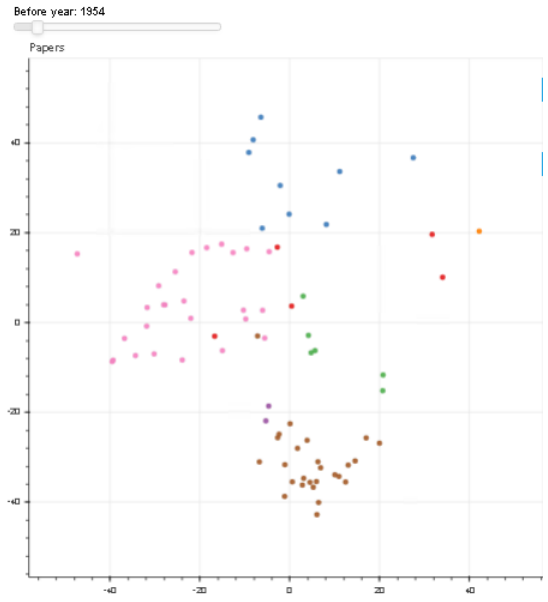

Fig. 2. 2D cluster structure in 1954.

Fig. 2 presents 2D visual map of the AI collection and 2D cluster structures in 1954. In 1954 the first 3 clusters begin to form: pink cluster corresponds to computing systems and algorithms; brown – to digital signal processing; blue – to programming languages and Natural Language Processing.

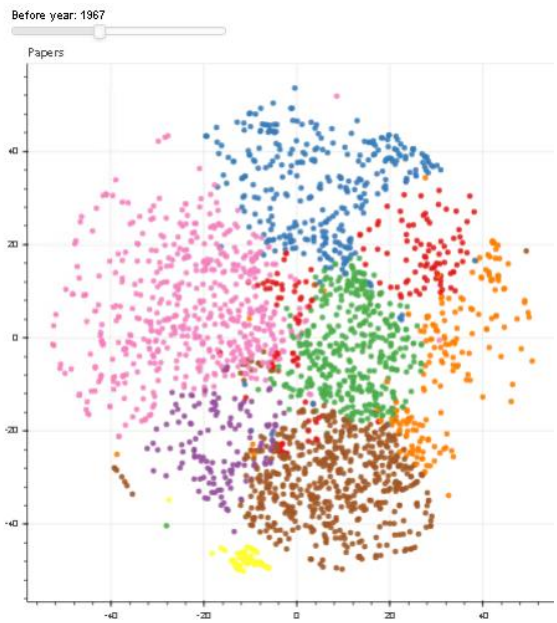

Fig. 3. 2D cluster structure in 1967.

Fig. 3 presents 2D visual maps of the AI collection and 2D cluster structures in 1967. In 1967 already all main clusters had been formed. To those clusters that began to stand out in 1954, the following were added: purple, corresponding to robotics; green – optimization methods and algorithms; orange – theoretical problems in computer science and computational complexity; red, still very rarefied – neural networks and computer networks; yellow – an erroneous cluster of German articles.

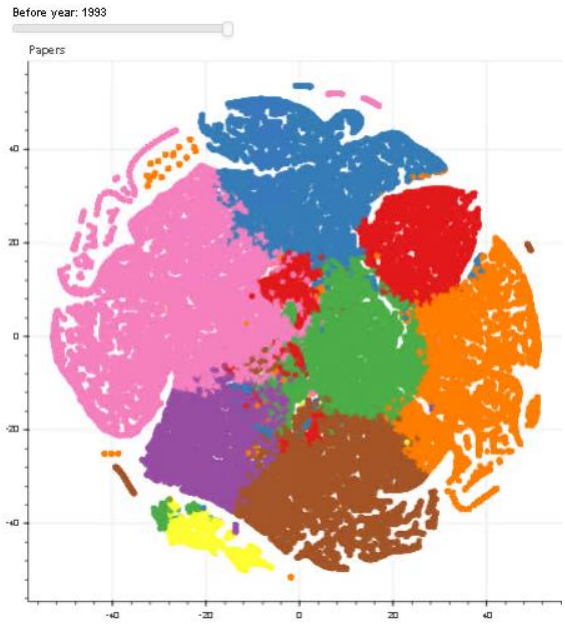

Fig. 4. 2D cluster structure in 1993.

Fig. 4 presents 2D visual maps of the AI collection and 2D cluster structures in 1993. In 1993 the red cluster (networks) consists of two large parts, which is the result of the distortion of high-dimensional space where red cluster is connected. The comparison of visualization in 2D and 3D shows that 3D visual map better preserves high-dimensional clusters structure than 2D visual map.

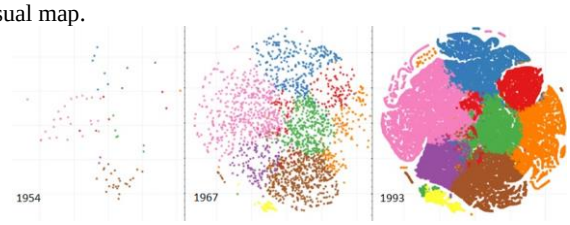

Fig. 5. Time series of 2D semantic visual maps.

Fig. 5 demonstrates that a time series of 2D semantic visual maps can be a good representation of the dynamics of topic trends. Such sequence of visualizations can be presented in the form of animation or a movie. Similarly, the time series of 3D visual maps allows us to present the dynamics in more detail.

6. WebVR visualization of 3D map

We use WebVR technology for creating 3D visual map of the AI collection (see Section 2) based on the results of t-SNE algorithm.

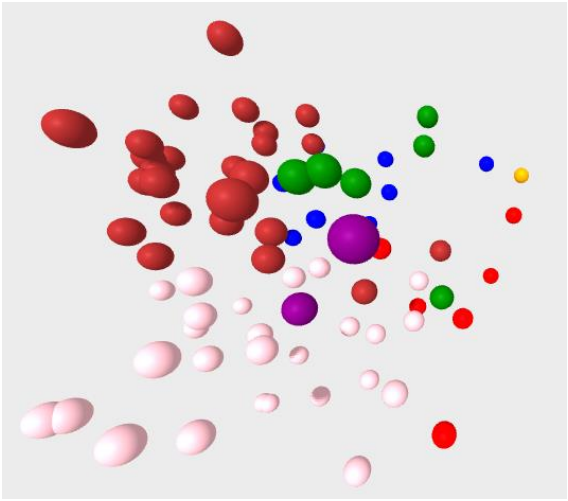

Fig. 6. 3D cluster structure in 1954.

WebVR provides JavaScript API that allows to build a three-dimensional picture. The easiest way to start working with WebVR is to use existing libraries, such as Three.js, D3.js or other frameworks capable of developing VR environment. WebVR provides the ability to rotate a 3D model, view it from different angles, and thus create a more complete picture of the text collection structure.

The Mozilla VR team developed A-Frame in mid-2015 (https://aframe.io). A-Frame is a WebVR framework that makes implementing virtual reality experiences quicker and easier by enabling to code in HTML. A-Frame is open source and it is primarily maintained by Mozilla and the WebVR community. A-Frame is an entity component system framework for Three.js where developers can create 3D and WebVR scenes using HTML.

Fig. 6 presents 3D visual map of the AI collection and 3D cluster structure in 1954. The coordinates of spheres in Fig. 6 are calculated by t-SNE algorithm which uses semantic similarity measures between documents in the collection. The color of the spheres points to topical cluster. This map is one of the screenshots of the virtual reality scene that was built using WebVR and A-Frame technology. Below is a fragment of HTML-code from which this virtual reality scene was built.

```
<!DOCTYPE html > <html > <head >
<script src= " https: // aframe.io / releases /0.7.0/
aframe.min.js "> </script > </head > <body > <a-scene >
<a-octahedron position="-1.0869318 -1.3309450 0.6411748" radius="0.30" color="pink" shadow></a-octahedron>
<a-sphere position="-1.3742855 -0.3519478 0.8263161" radius="0.10" color="pink" shadow></a-sphere>
. . .
<a-sky color= "# ECECEC "></ a-sky >
</ a-scene > </body > </html >
```

This HTML-code was automatically created based on 3D coordinates calculated by t-SNE algorithm which uses semantic similarity measures between documents in the collection calculated using LDA algorithm (see Section 3). Parameters “color” (that represent topical cluster) in HTML-code are also calculated by LDA algorithm.

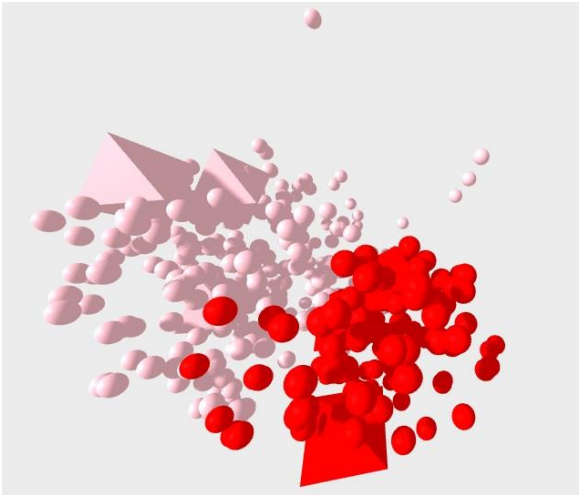

Fig. 7. 3D cluster structure in 1993 filtered by time and topic.

Fig. 7 presents 3D visual map of the AI collection and 3D cluster structure in 1993. This map uses article filters by time (until 1993) and filters by topic (all clusters except pink and red are prohibited).

Red cluster on 2D map of 1993 in Fig. 4 consists of two large parts, which is the result of distortion of high-dimensional space, where red cluster is connected. In Fig. 7, this red cluster consists of one large part, which is more correct. Thus, the comparison of visualization in 2D and 3D shows that 3D visual map better preserves articles similarity and high-dimensional clusters structure than 2D visual map.

In Fig. 7, the articles with largest citation forecasts are presented as polyhedra octahedrons. These forecasts were calculated using a deep learning neural network algorithm described in Section 4. Visualization of this predictive information allows you to better understand the dynamics of future trends and find the most promising topics. Such predictive visual analytics helps us to make more reliable forecasts and increase the credibility of the model used.

7. Future work

We plan to improve visualization features of proposed cyberspace by visualizing formal links between articles and the most reliable of implicit links. We also plan to visualize multilingual collections for integrating information from different languages. Also, our plans include:

- to improve semantic similarity measure by using Word2Vec, Doc2Vec, WMD and other new methods;
- to improve semantic similarity measure by taking into account the time of the articles and direction of implicit links;
- to check our results on different collections in other subject areas and in other languages.

8. Conclusion

An unsupervised method and algorithm for visualizing the dynamics of topic trends by generating a time series of 2D and 3D semantic visual maps with predictive information is described. The method includes semantic similarity measure and citation prediction for documents, topic modeling and clustering, dimensionality reduction, virtual reality technology, representation of dynamics using time filters. As an example of particular implementation, the method is demonstrated on AI collection data using technologies of neural network prediction, LDA clustering, t-SNE dimensionality reduction, WebVR visualization.

The comparison of visualization in 2D and 3D shows that 3D visual map better preserves articles similarity and high-dimensional clusters structure than 2D visual map. The colors of the articles in visual maps correspond to different topical clusters. The articles with largest citation forecast in 3D visual map are presented as polyhedra octahedrons. This forecast was calculated using a deep learning neural network algorithm. For a better visual representation of trend dynamics, the time series of semantic visual maps can be presented in the form of a 3D film or animation.

The presented algorithm can be a good tool for analyzing the emergence of scientific topics and their trends in any research field. The results obtained will make it possible to notice in time new promising areas for attracting timely financing in order to create new competitive technologies that will be in demand in the market.

9. Acknowledgment

This work is supported by Russian Foundation for Basic Research, grants 19-07-00857, 16-29-09527, 18-07-00909 and 18-07-01111. We are grateful to the Russian Foundation for Basic Research for financial support of our projects.

10. References

[1] O. Abdelwahab and A. Elmaghraby. UofL at SemEval-2016 Task 4: Multi Domain word2vec for Twitter Sentiment Classification. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, 2016, pp. 164–170. https://doi.org/10.18653/v1/S16-1024

[2] M. Dodge and R. Kitchin. Exposing the Second Text of Maps of the Net, Journal of Computer-Mediated Communication 5, 2000, http://www.ascusc.org/jcmc/vol5/issue4/ dodge_kitchin.htm

[3] M. Charnine, S. Klimenko, “Semantic cyberspace of scientific papers,” Proc. 2017 International Conference on Cyberworlds, 20-22 September 2017, Chester, United Kingdom, pp.146-149.

[4] C.Chen, “Structuring and visualizing the WWW with Generalized Similarity Analysis”, Proceedings of the Eighth ACM Conference on Hypertext (Hypertext’97), 1997, Southampton, UK, pp. 177-186.

[5] C.Chen, “Visualization of knowledge structures”, Handbook of Software Engineering and Knowledge Engineering, 2002.

[6] S.Klimenko, M.Charnine, O.Zolotarev, N.Merkureva, A.Khakimova, “Semantic approach to visualization of research front of scientific papers using web-based 3D graphic,” Proc. 2018 International Conference Web3D’18, June 20–22, 2018, Poznan, Poland.

[7] L.J.P. Maaten van der and G.E. Hinton. Visualizating data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2008, pp. 2579–2605.

[8] T. Mikolov, G. Corrado, K. Chen, and J. Dean. Efficient Estimation of Word Representations in Vector Space, 2013, Vol. 1, pp. 1–12.

[9] S. Neelakantam and T. Pant. Learning Web-based Virtual Reality: Build and Deploy Web-based Virtual Reality Technology. 2017.

[10] H.Patel, R.Caedinali “Virtual Reality Technology in Business”, Management Decision, 1994, 32, 7, 5-12.

[11] J. Tian and M. Lan. ECNU at SemEval-2016 Task 1: Leveraging Word Embedding From Macro and Micro Views to Boost Performance for Semantic Textual Similarity. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, 2016, pp. 621–627. https://doi.org/10.18653/v1/S16-1094.

About the authors

Dr. Michael Charnine is a senior researcher of the Institute of Informatics Problems of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences. E-mail: mc@keywen.com.

Sokolov Evgeniy is a postgraduate student of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences. E-mail: evgeny.sokolov@phystech.edu.

Klokov Alexey is a student of Moscow Institute of Physics and Technology with math modeling specialization. E-mail: aaklokov@yandex.ru.