CONSTRUCTING SCIENTIFIC PUBLICATION PROFILES BASED ON TEXTS AND COAUTHORSHIP CONNECTIONS (IN THE FIELD OF CONTROL THEORY AND ITS APPLICATIONS)

D. A. Gubanov* and V. S. Melnichuk**

*,***Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia,
**Bauman Moscow State Technical University, Moscow, Russia

*✉ dmitry.a.g@gmail.com, **✉ vs.melnichuk09@gmail.com

Abstract. The calculation of scientific publication profiles is crucial in the systematization of scientific knowledge and support for scientific decision-making. This paper proposes a method for forming publication profiles in the field of control theory, based on the integration of text analysis and coauthorship network analysis. We describe a basic algorithm that analyzes publication texts by a thematic classifier as well as its enhanced version that considers network connections within a heuristic approach. The methods are examined using expert assessments and quantitative metrics; according to the examination results, combining textual and network data significantly improves the accuracy of publication profiles. Hypotheses about a relationship between the thematic similarity and network proximity of publications are tested, and the approach proposed is validated accordingly. In addition, directions for further research are identified.

Keywords: publication network, publication profile, control theory, graph neural networks, text analysis.

INTRODUCTION

Thematic analysis of scientific publications is an important tool for justifying scientific decisions and identifying trends in various fields of knowledge [1–6]. One of the most common approaches to text analysis is thematic modeling [7], which is used to calculate scientific publication profiles. However, when abstracts or texts of publications have limited length and/or contain imprecise terms, using only textual information may lead to low accuracy of profiles.

The inclusion of network data, such as coauthorship or citation connections, has already demonstrated its utility in several disciplines: considering the structural relationships between publications can improve the quality of classification and more adequately reflect hidden thematic dependencies [8–10]. In particular, *graph neural networks* (GNNs) [11, 12] have proven to be an effective network analysis tool, as they simultaneously cover both node features and graph topology.

This study aims to develop and evaluate improved methods for constructing publication profiles that combine text and network information analysis. The main results of the work are summarized below:

- A basic algorithm for calculating publication profiles based on a thematic classifier in the field of control theory and its applications is presented.
- "Advanced" algorithms considering network data are developed. In particular, they are a heuristic method that extends the basic profile with connected publications (through a coauthorship or citation) and a method based on GNNs that deeply integrates structural information on the connections between publications.
- The efficiency of these algorithms is evaluated, showing their higher accuracy compared to the ones using only textual information.
- Relationships between the thematic similarity of publications (assessed by their profiles) and network characteristics (e.g., common neighbors in a graph) are investigated. As part of this analysis, several hypotheses are formulated and tested.

The following sections of the paper describe in detail the implementation of the algorithms, the metrics used, and the experimental results.

1. METHODS

1.1. The Basic Algorithm for Profile Calculation

In this paper, for each scientific publication l, we construct a *basic profile* $p(l)$ using the thematic classifier of ISAND, an information system for scientific activity analysis [13]. This classifier is based on the principles outlined in [14] and represents a hierarchical ontology of themes in control theory and its applications. Its fragment is shown in Fig. 1, where long theme names are abbreviated; the full version of the classifier can be found at: https://www.ipu.ru/sites/default/files/page_file/ClassifierCS.xlsx. The original form[1] was used for marking up publications by a group of experts; the results of applying the basic and network algorithms for marking up publications were then compared with the expert markup.

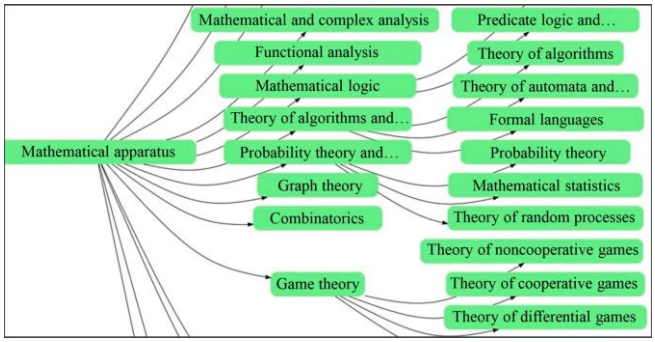

Fig. 1. The thematic classifier (the ISAND ontology).

A publication profile $p(l)$ is a stochastic vector $(p_{l1}, p_{l2},\ldots, p_{ln})$ in which each component p_{li} is the normalized frequency of terms related to theme i in publication l.

1.2. Advanced Profile Calculation Methods

Consider a graph $G(V, E)$ in which each vertex $l \in V$ corresponds to one publication and each edge $(l, m) \in E$ denotes a coauthorship connection. In other words, publications l and m have a non-empty intersection of their authors, $|K(l) \cap K(m)| > 0$, where $K(l)$ is the set of authors (coauthors) of publication l. In the original graph, each vertex l is initialized by the basic profile vector $p(l)$.

[1] URL: https://docs.google.com/forms/d/e/1FAIpQLSfR47ZQyjI9wrMgqRPP85j_uZCeUI95dNFnMR-2ruCfq3XtIg/viewform

1.2.1. A Heuristic Method

To improve the accuracy of the basic profile of publication l, consider publications associated with l by a coauthorship connection and issued within a fixed period $\delta \in \mathbb{N}$ (time window). The default value is δ = 4 years, but it can be adjusted based on empirical data. In several disciplines, including control theory, a window of 3–5 years is common to assess scientific activity. In particular, according to the methodological recommendations of the Higher Attestation Commission (VAK RF), it is necessary to consider publications issued in the last five years. Thus, δ = 4 years is a reasonable choice from the viewpoint of assessing the scientific activity of researchers.

An *extended profile* $p_e(l)$ is a weighted combination

$$p_e(l) = \alpha p(l) + (1-\alpha)\frac{\sum\limits_{m \in L_\delta(l)} w_{lm}\, p(m)}{\sum\limits_{m \in L_\delta(l)} w_{lm}},$$

where $L_\delta(l)$ denotes the set of publications connected to l and issued within the time window δ, and $\alpha \in (0, 1]$ is the coefficient regulating the contribution of the original and network profiles. The value of α can be chosen empirically, e.g., based on the results of cross-validation on a delayed sample.

The coefficient $w_{lm} \in [0, 1]$ reflects the contribution of publication m to profile l. In this case, the share of common authors is taken into consideration:

$$w_{lm} = \frac{|K(l) \cap K(m)|}{|K(m)|}.$$

Note that if $\sum\limits_{m \in L_\delta(l)} w_{lm} = 0$ (publication l has no connected publications for the last δ years), the profile $p_e(l)$ coincides with $p(l)$ by definition. Generally speaking, the heuristic method smoothens the "noise" in the basic profiles as well as generates a network profile for publications with missing or uninformative abstracts.

1.2.2. A Method Based on Graph Neural Networks

To further improve the accuracy of profiles, we apply a graph neural network (GNN) trained on the graph $G(V, E)$. Initially, each node $i \in V$ receives a feature vector $\mathbf{h}_i^{(0)} = p(i)$. At the kth layer of the

GNN, the vector $\mathbf{h}_i^{(k)}$ is recalculated considering the neighbors $\mathcal{N}(i)$ by the formula

$$\mathbf{h}_i^{(k)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{c_{ij}} W^{(k)} \mathbf{h}_j^{(k-1)} \right)$$

with the following notation: $\mathbf{h}_i^{(k)}$ is the new representation (profile) of node i in the kth layer of the neural network; $W^{(k)}$ is the trained weight matrix (model parameters); σ is a nonlinear activation function (e.g., ReLU); c_{ij} is a normalization coefficient regulating the contribution of neighbor nodes (e.g., $c_{ij} = \sqrt{\deg(i)\deg(j)}$); finally, $\mathcal{N}(i)$ is the set of neighbors of node i in the publication graph.

Passing through several layers of the GNN gives the resulting vector $\mathbf{h}_i^{(K)}$, which can be treated as the profile of publication i "deeply integrated" in the network topology. With appropriate training (using a quality metric and a target function), this approach can identify and consider complex dependencies between publications, which often improves the accuracy and informativeness of publication profiles.

Well, in this paper we propose the following approaches to extend the basic profile of scientific publications:

- **the heuristic method**, which specifies a linear combination of a publication profile with the profiles of connected papers;
- **the GNN method**, which aggregates features in a more sophisticated way based on a training sample.

Both approaches enrich the basic profiles by considering indirect thematic connections through a coauthorship, which facilitates the analysis of scientific publications (classification and recommendation of relevant papers). The heuristic method is characterized by high interpretability and simple implementation; however, its accuracy may be limited in complex network structures with unobvious connections. In turn, the GNN method identifies complex dependencies between publications, thereby being preferable whenever the in-depth analysis of coauthorship structures is required. However, this method may need large datasets and significant computational resources for effective model training.

2. EXPERIMENTS

To evaluate the effectiveness of these methods for calculating publication profiles in the field of control theory and applications, we used a sample of 20 thousand items (the publication database of the Trapeznikov Institute of Control Sciences, the Russian Academy of Sciences). The dataset included texts (for the construction of textual features) and coauthorship information (for the construction of network features).

Two types of profiles were calculated:

– the basic profiles (by abstract texts only),

– the extended profiles (considering the network structure, i.e., coauthorship connections).

The following criteria were used to evaluate the quality of the resulting profiles in quantitative terms:

- expert assessments: subject matter experts assessed the relevance of the themes assigned to each publication;
- quality metrics: the values of *Precision@k*, completeness, and F_1-measure were calculated. To determine *Precision@k*, the k most probable themes from the profile were selected and compared to the reference themes identified by the experts.

The next section presents the results of experiments, i.e., the analysis of the distances between different profiles and the test of hypotheses about a relationship between the thematic similarity and network proximity of publications.

3. RESULTS

With the extended algorithm, which considers network information, the results were significantly improved in accuracy relative to the basic approach. For example, the heuristic method (Fig. 2) achieved *Precision@k* = 37%, whereas the basic method provided only 25% (on the sample marked by experts). This indicates the high utility of network features for the formation of thematic profiles.

In this study, a *graph neural network* (GNN) architecture was developed. It includes three sequential layers of a *graph convolutional network* (GCN) and Dropout regularization layers. To evaluate the quality of the model, we partitioned the sample (several hundred publications) in a proportion of 70%/15%/15% into training, validation, and testing subsamples. The training was carried out for 100 epochs, with the final model parameters selected from the best results achieved on the validation subsample. The average value of *Precision@k* (for k = 3) across all runs was 39%, constituting a 2% increase over the heuristic method used previously. Therefore, the accuracy of prediction was improved.

3.1. Analysis of the Distances between Profiles

For a more detailed comparison, we analyzed the distribution of distances between the extended and

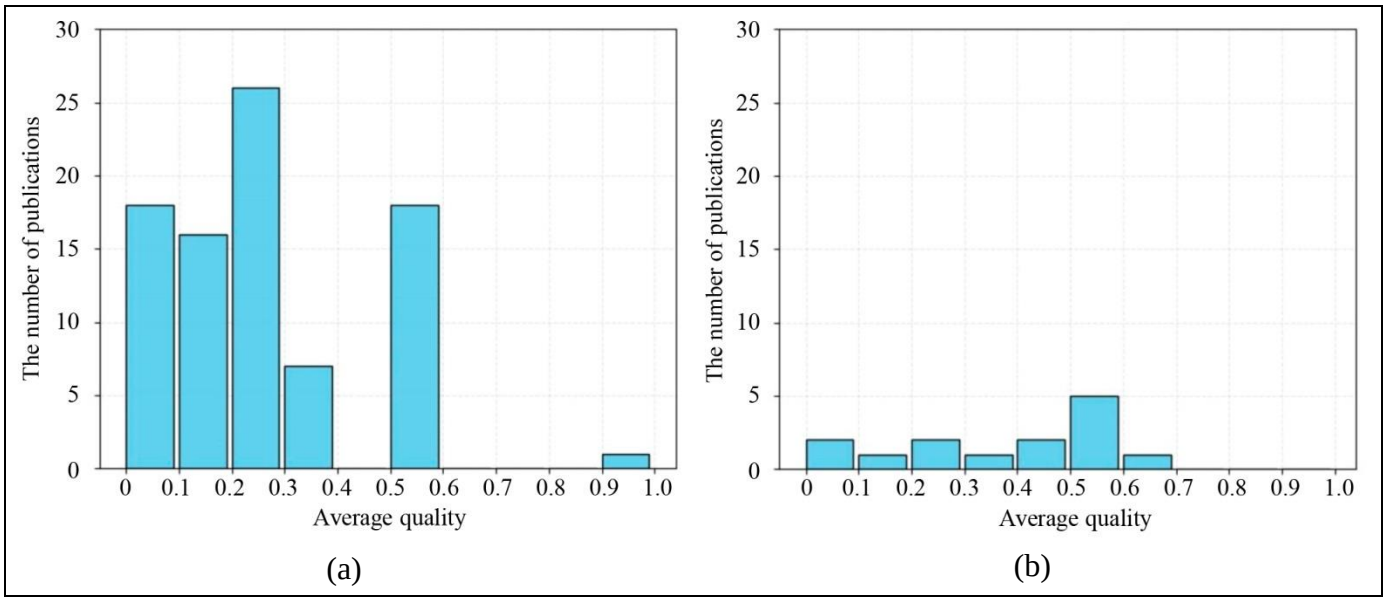

Fig. 2. Quality evaluations of the methods for constructing publication profiles: (a) basic (BaseAlgo, 25% accuracy over abstracts) and (b) network (HAdvAlgo, 37% accuracy over abstracts). The horizontal axis corresponds to the average quality of profiles whereas the vertical axis to the number of publications with such quality.

basic profiles for the entire dataset of publications (Fig. 3). Since the metric $d \in [0, 1]$ (*L1*) was used, two characteristic peaks can be observed on the graph:

- The first peak (for $d = 0$) corresponds to cases when an author has only one publication. In such a situation, the network profile almost coincides with the basic one.
- The second peak (for high values of d) is observed for publications where network information (coauthorship) strongly affects the final profile and/or textual data are too scarce (complicating the adequate calculation of the basic profile).

In particular, if an abstract is very short or contains few relevant terms, the basic profile may be weakly informative. In these cases, the impact of network data turns out to be the most significant, and the distance between the profiles (basic and extended) increases noticeably.

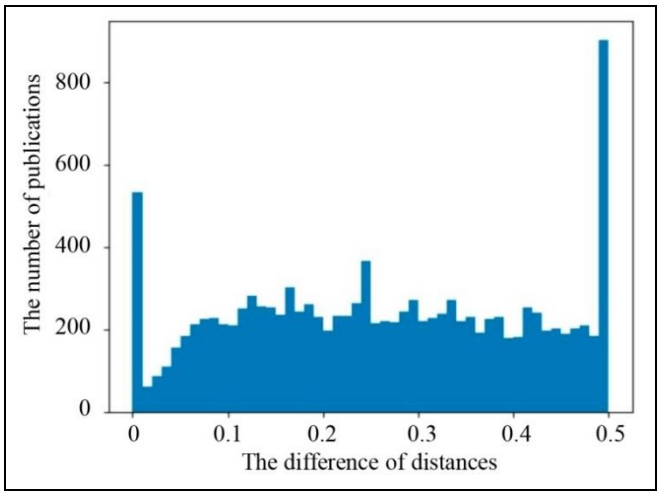

Fig. 3. The distribution of the distances between basic and extended profiles.

3.2. Test of Hypotheses about a Relationship between the Thematic Similarity and Network Proximity of Publications

Hypotheses concerning a relationship between the thematic similarity (by profile) and network proximity (by the coauthorship graph) of publications were also investigated.

Hypothesis 1: *The profiles of randomly selected publications differ from each other.*

The calculations confirmed this hypothesis: the average distance between the profiles of random publications was about 0.9 (with values ranging from 0 to 1), indicating a significant diversity of themes in the field of control theory.

Hypothesis 2. *The closer the content of the abstracts of two randomly selected publications, the closer their profiles will be.*

Tests using vector representations (*embeddings*) showed no significant correlation. Note that different language models were applied to construct them: RuSciBert, SciBert, and Sentence Embeddings. Probably, textual abstracts were too short or heterogeneous to ensure consistency with the thematic profiles extended with network information. A more detailed analysis of the nature of the discrepancies (the number of terms, language variation, etc.) is needed in the future.

Hypothesis 3. *The more terms an abstract has, the higher the correlation between the profiles and vector representations* (*embeddings*) *of these abstracts will be*.

This hypothesis was confirmed: the correlation coefficient increased from 0.25 (with five terms) to 0.88 (with eight terms). The result emphasizes the importance of the completeness and accuracy of abstracts in the sense of terms used.

Hypothesis 4. *The similarity of the profiles of two publications depends on:*

- the existence of a coauthorship connection,

- the coincidence of the authors' composition.

This hypothesis was confirmed: for pairs of publications with common authors, the average distance between the profiles (0.63) turned out to be noticeably smaller than for all other pairs (0.88).

Hypothesis 5. *The smaller the period between publications is, the closer their profiles will be*.

During certain periods, there may be bursts of interest in particular technologies or phenomena in a given research area (e.g., "big data," "machine learning"), which should be reflected in the content of publications. Such bursts may be related to different phases of the popularity cycle of technologies. However, in the field of control theory, this hypothesis was not confirmed. According to the analysis results, for random pairs of publications, there were no significant changes in the level of similarity of their profiles depending on the period. However, for non-random pairs of publications, such dependence takes place. (Pairs of publications connected to each other in the network are called non-random.)

Hypothesis 6. *The similarity of the profiles of two publications depends on the topological strength of their connection*.

The hypothesis was not confirmed: the resulting data (Fig. 4) did not reveal a significant correlation between the number of common neighbors in the network and the distance between profiles. (The distance is zero if the profiles are equal.) The topological strength of a connection means the number of coauthorship connections between publications. (The topological strength of a connection is zero if there exist no coauthorship connections.)

4. DISCUSSION OF THE RESULTS

According to the experimental results, considering network features (coauthorship) improves the accuracy of constructing publication profiles. The heuristic method is valuable for its simplicity and interpretability, which is especially convenient at the stage of initial assessment of profiles and analysis of the subject matter. At the same time, the direct application of this method has demonstrated its disadvantages: if the authors have few papers or the abstracts are too short, the quality of the basic profile will remain low, and even network information does not always compensate for the lack of textual data.

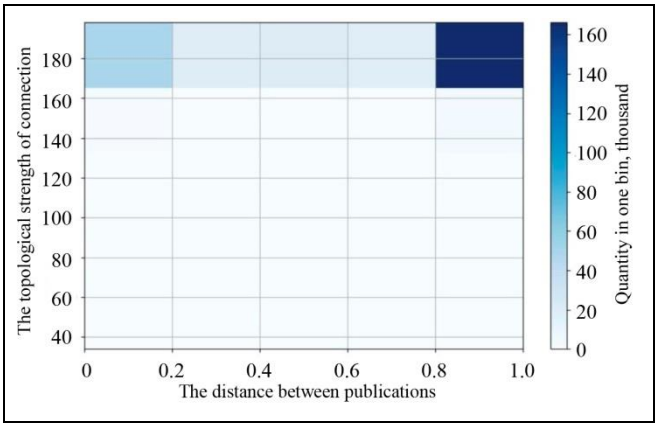

Fig. 4. The dependence of the proximity of publications on the topological strength of the connection between them. A bin is a cell that partitions *N* into *M* equal rectangles.

When testing the hypotheses, it has been established that there exists a significant diversity of themes in the field of control theory (Hypothesis 1) and that "textual similarity" (Hypothesis 2) does not necessarily lead to profile similarity, especially for incomplete abstracts. In addition, the more terms a publication contains, the more the similarity of embeddings will determine the similarity of profiles (Hypothesis 3). The similarity of profiles depends on the existence of a connection between publications (Hypothesis 4). However, the test results of the other hypotheses (5 and 6) have shown that proximity in time or a large number of common neighbors do not guarantee the similarity of publication: additional factors and additional research are required here.

Among the limitations of this study, we mention data sparsity (a small number of publications per one author) and the heterogeneous quality of the abstracts used to construct basic publication profiles. It seems promising to investigate further how the use of citation networks, keywords, longer texts (full-text papers), and advanced GNN models (e.g., *graph attention networks*) can improve the accuracy of publication profiles.

CONCLUSIONS

Thus, the hybrid methods proposed in this paper, combining textual and network features, are significantly superior to the basic (textual) approach in constructing scientific publication profiles. Several hypotheses about the thematic similarity and network proximity of publications have been tested. According

to the test results, in some cases, network connections turn out to be of a much higher utility for theme identification than the content of short abstracts. The results of this study will be used to develop methods for analyzing scientific publications and systematizing knowledge in the field of control theory and applications.

REFERENCES

1. Kryzhanovskaya, S.Yu., Vlasov, A.V., Eremeev, M.A., and Vorontsov, K.V., Semiautomatic Summarization of Thematic Samples of Scientific Publications: Problems and Approaches, *Tezisy dokladov 20-oi Vserossiiskoi konferentsii s mezhdunarodnym uchastiem "Matematicheskie metody raspoznavaniya obrazov"* (Abstracts of the 20th All-Russian Conference with International Participation "Mathematical Methods of Image Recognition"), Moscow, 2021, pp. 333–338. (In Russian.)
2. Shibayama, S., Yin, D., and Matsumoto, K., Measuring Novelty in Science with Word Embedding, *PLoS ONE*, 2021, no. 7, pp. 1–16.
3. Yuan, W., Liu, P., and Neubig, G., Can We Automate Scientific Reviewing?, *Journal of Artificial Intelligence Research*, 2022, no. 75, pp. 171–212.
4. Cachola, I., Lo, K., Cohan, A., and Weld, D., TLDR: Extreme Summarization of Scientific Documents, *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4766–4777.
5. Bao, P., Hong, W., and Li, X., Predicting Paper Acceptance via Interpretable Decision Sets, in *Companion Proceedings of the Web Conference 2021 (WWW 21)*, New York: Association for Computing Machinery, 2021, pp. 461–467.
6. Kasanishi, T., Isonuma, M., Mori, J., and Sakata, I., SciReviewGen: A Large-scale Dataset for Automatic Literature Review Generation, *arXiv:2305.15186*, 2023, pp. 1–19. DOI: https://doi.org/10.48550/arXiv.2305.15186
7. Blei, D.M., Ng, A.Y., and Jordan, M.I., Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 2003, no. 3, pp. 993–1022.
8. Hasegawa, T., Arvidsson, H., Tudzarovski, N., et al., Edge-Based Graph Neural Networks for Cell-Graph Modeling and Prediction, *Information Processing in Medical Imaging*, 2023, vol. 13939, pp. 265–277.
9. Xiong, C., Li, W., Liu, Y., and Wang, M., Multi-Dimensional Edge Features Graph Neural Network on Few-Shot Image Classification, *IEEE Signal Processing Letters*, 2021, vol. 28, pp. 573–577.
10. Faber, L., Lu, Y., and Wattenhofer, R., Should Graph Neural Networks Use Features, Edges, or Both?, *arXiv: 2103.06857*, 2021, pp. 1–12. DOI: https://doi.org/10.48550/arXiv.2103.06857
11. Zhou, J., Cui, G., Hu, S., et al., Graph Neural Networks: A Review of Methods and Applications, *AI Open*, 2020, vol. 1, pp. 57–81.
12. Kipf, T.N. and Welling, M., Semi-Supervised Classification with Graph Convolutional Networks, *arXiv:1609.02907*, 2017, pp. 1–14. DOI: https://doi.org/10.48550/arXiv.1609.02907
13. Gubanov, D.A., Kuznetsov, O.P., Sukhoverov, V.S., and Chkhartishvili, A.G., On the Construction of Profiles in the Thematic Space of Control Theory, *Materialy 9-oi Mezhdunarodnoi konferentsii "Znaniya-Ontologii-Teorii"* (Proceedings of the 9th International Conference "Knowledge–Ontology–Theories"), Novosibirsk, 2023, pp. 89–94. (In Russian.)
14. Kuznetsov, O.P. and Sukhoverov, V.S., An Ontological Approach to Determining the Subject Matter of Scientific Text, *Ontology of Designing*, 2016, vol. 6, no. 1, pp. 55–66. (In Russian.)

This paper was recommended for publication by O.P. Kuznetsov, a member of the Editorial Board.

Received November 1, 2024, and revised February 28, 2025. Accepted March 6, 2025.

Author information

Gubanov, Dmitry Alekseevich. Dr. Sci. (Eng.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ dmitry.a.g@gmail.com
ORCID iD: https://orcid.org/0000-0002-0099-3386

Melnichuk, Vladislav Sergeevich. Technician, Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia; bachelor's student, Bauman Moscow State Technical University, Moscow, Russia
✉ vs.melnichuk09@gmail.com
ORCID iD: https://orcid.org/0009-0005-8252-0804

Cite this paper

Gubanov, D.A. and Melnichuk, V.S., Constructing Scientific Publication Profiles Based on Texts and Coauthorship Connections (in the Field of Control Theory and Its Applications). *Control Sciences* **1**, 39–44 (2025).

Original Russian Text © Gubanov, D.A., Melnichuk, V.S., 2025, published in *Problemy Upravleniya*, 2025, no. 1, pp. 46–52.

This paper is available under the Creative Commons Attribution 4.0 Worldwide License.

Translated into English by *Alexander Yu. Mazurov*, Cand. Sci. (Phys.–Math.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ alexander.mazurov08@gmail.com