UDC 004.8
DOI: 10.15587/1729-4061.2023.273929

A METHOD FOR ASSESSING THE PRODUCTIVITY TRENDS OF COLLECTIVE SCIENTIFIC SUBJECTS BASED ON THE MODIFIED PAGERANK ALGORITHM

Yurii Andrashko
PhD, Associate Professor
Department of System Analysis and Optimization Theory*
Oleksandr Kuchanskyi
Corresponding author
Doctor of Technical Sciences, Head of Department**
E-mail: kuchansky@knu.ua
Andrii Biloshchytskyi
Doctor of Technical Sciences, Professor,
Vice-Rector for Science and Innovation
Astana IT University
Mangylyk El ave., EXPO Business Center, block C.1,
Astana, Republic of Kazakhstan, 010000
Department of Information Technologies***
Oleksandr Pohoriliak
PhD
Department of Probability Theory and Mathematical Analysis*
Myroslava Gladka
PhD, Associate Professor**
Ganna Slyvka-Tylyshchak
Doctor of Physical and Mathematical Sciences, Head of Department
Department of Probability Theory and Mathematical Analysis*
Dmytro Khlaponin
PhD, Associate Professor
Department of Political Sciences and Law***
Ivan Chychkan
PhD, Associate Professor**
*Uzhhorod National University
Narodna sq., 3, Uzhhorod, Ukraine, 88000
**Department of Information Systems and Technology
Taras Shevchenko National University of Kyiv
Volodymyrska str., 60, Kyiv, Ukraine, 01033
***Kyiv National University of Construction and Architecture
Povitroflotsky ave., 31, Kyiv, Ukraine, 03037

The task of calculating the productivity of collective scientific subjects is a relevant issue in scientometrics. This study formalized the problem of assessing productivity trends of collective scientific subjects. The TWPR-CI method for calculating the performance based on the modified PageRank algorithm is described. Formulas for calculating productivity have been derived that make it possible to take into account a change in the productivity of collective scientific subjects over time. The indicators of the basic average absolute change in performance and the chain average relative change in performance were chosen as the basis. To select promising, from the point of view of scientific work, collective subjects, preference is given to those whose basic average absolute change in productivity is positive or the chain average relative change in productivity exceeds unity. Verification of the method for assessing performance trends of collective scientific entities based on the modified PageRank algorithm using the public dataset Citation Network Dataset was carried out. The dataset includes more than 5 million scientific publications and 48 million citations. The citation of scientific publications of 27,500 collective scientific subjects for the period from 2000 to 2022 was analyzed. For this period, for 15 selected collective scientific subjects, performance is calculated using the TWPR-CI method, as well as estimates of productivity trends based on their average relative change. There are three classes of collective scientific subjects according to productivity trends. The results indicate the relevance of the proposed method for quantifying the productivity trends of collective scientific entities (higher education institutions, scientific institutes, laboratories, and other institutions engaged in scientific activities)

Keywords: PageRank algorithm, scientific work, collective scientific subject, scientometrics, scientific productivity

Received date 13.12.2022
Accepted date 15.02.2023
Published date 28.02.2023

How to Cite: *Andrashko, Y., Kuchanskyi, O., Biloshchytskyi, A., Pohoriliak, O., Gladka, M., Slyvka-Tylyshchak, G., Khlaponin, D., Chychkan, I. (2023). A method for assessing the productivity trends of collective scientific subjects based on the modified PageRank algorithm. Eastern-European Journal of Enterprise Technologies, 1 (4 (121)), 41–47. doi: https://doi.org/10.15587/1729-4061.2023.273929*

1. Introduction

The task of evaluating and monitoring the productivity of subjects of scientific activity or scientific subjects is a relevant task of scientometrics. One distinguishes collective scientific subjects and individual scientific subjects. A collective scientific subject (CSS) is an institution or organization that conducts scientific, scientific-technical, innovative activities.

CSSs include higher education institutions, research institutes, and research centers, laboratories, etc. Individual scientific subjects are individual scientists. Each individual scientific entity is affiliated with a specific CSS and is associated with its performance [1]. The productivity of scientific subjects is calculated on the basis of scientometrics indicators, which are mostly determined using open and unbiased data sources.

An important place in assessing the productivity of scientific activity is occupied by citation indices, for example, h-index [2]. However, all citation indices have some limitations and take into account only a fraction of the citations. All citations that fall beyond the core of the index are not taken into account in the calculation. To solve this drawback, methods are being developed that take into account all citations of authors without exception. In particular, such methods are the method of PageRank [3] and its modifications. These modifications relate to both the calculation formula and the purpose of the method. In particular, with the help of modifications of the PageRank method, one can calculate the reputation of scientific journals, the index of the importance of a scientist in the scientific network [4], etc.

The disadvantage of the traditional PageRank method is that when calculating the performance of CSS, the citation network is considered statically, without taking into account the dynamism of the development of the academic space. Of particular interest is the construction of such a modification of the PageRank method, which would allow calculating the performance of the newly created CSS or those entities that have just started research activities. After all, traditional methods of evaluation, including the traditional PageRank method for new CSS, are not calculated. To calculate productivity estimates in this case, there should be a sufficient history of scientific publications of individual scientific subjects that are affiliated with the data of CSS. It is also of interest to evaluate performance trends. It is important to know the dynamics of the development of CSS. Evaluation of CSS performance trends is important in practice. This information makes it possible to form a rating of CSSs without restrictions and influences the choice of the most promising of them for their involvement in research consortia [5]. That is why the development of a method for assessing CSS performance trends based on the PageRank method is relevant.

2. Literature review and problem statement

The main goal of creating the classic PageRank method is to calculate the user's influence in social networks or determine the reputation of web pages [3]. Numerous modifications of this method make it possible to calculate the ranking of scientific journals by rating, in particular, the assessment of the impact of the EigenFactor article [6], the indices of the influence of journals according to the Scimago Journal Ranking [7]. Another modification of the PageRank method, which for evaluating the journal takes into account the h-indexes of the authors of the papers published in it, is described in [8]. Each of these modifications has some improvements over the classic PageRank method. In particular, work [9] describes the modified method HR-PageRank, which by statistical estimates exceeds the PageRank method. To calculate the performance of CSS, one can choose the HR-PageRank method but in practice it is based on the calculation of h-indices. The disadvantage of using h-indices to evaluate performance is the loss of information about a part of citations that is located outside the core of the calculation. This is especially important for assessing the trends in the performance of CSS, for which it is necessary to take into account as much as possible all the results of scientific activity without exception.

It is important that in order to adequately assess the performance of CSS, it is necessary to establish links between scientists in the form of a network of scientific cooperation and links between citations of scientific publications in the form of a citation network. In [10], the structure of such networks for evaluating the performance of CSS is investigated. Nevertheless, such a study is only the first step to build a full-fledged adequate assessment of the performance of CSS.

Another area that is intensively used to assess the performance of CSS is the use of the Labeled Latent Dirichlet Allocation (Labeled-LDA) thematic model. This model is built in conjunction with PageRank to evaluate CSS within specific subject areas [11, 12]. However, the intensity of citations and production of new scientific publications by the authors has a certain dynamics in each specific scientific direction. This cannot be taken into account in these methods.

To adequately assess the CSS performance using the PageRank method, coefficients are introduced that correct certain parameters of the classic PageRank method. Thus, certain refinements of estimates are obtained, taking into account the obsolescence of scientific publications, the intensity of citations, etc. In particular, work [13] describes the modification of the PageRank method, which takes into account only quotes of scientific publications that were made during the last 10 years at the time of calculation. The results obtained somewhat improve the estimates of the new CSS. However, this does not solve the question of the adequacy of such an assessment. Even if the newly created CSS has powerful scientific results that have been recorded, for example, over the past five years. Such a modification will still give preference to those CSSs, which are associated with the maximum number of scientific publications and quotes on them. That is, promising new CSSs with such a calculation of performance are likely not to be noticed. The way out of this is to introduce an assessment of the productivity trend, taking into account the change in this indicator over time. It is possible to objectively reduce the impact of these shortcomings on the result of calculating the CSS performance by the weighted method PageRank. This method takes into account all publications and citations of the CSS but with a certain coefficient, which is determined by the time interval from the moment of calculation. The longer this time period, the smaller the coefficient of the corresponding publication. This method is described in works [14, 15]. The construction of such a method for evaluating CSS, which would take into account performance trends based on a modification of the PageRank algorithm, remains unresolved. The use of modification of the weighted PageRank method using citation intensity is a further development in this class of methods [16]. This makes it possible to significantly increase the adequacy of the assessment of performance and productivity trends for CSS. Nevertheless, in [16], the assessment of productivity is recorded at the time of calculation annually and does not make it possible to see the trend of changes in productivity gains. This is important for understanding the development of the scientific activities of CSS as a whole. Thus, it suggests that it is expedient to conduct a study on the development and verification of such a method that would make it possible to effectively assess the trends in the productivity of CSSs.

3. The aim and objectives of the study

The aim of our study is to develop a method for assessing performance trends of collective scientific subjects based on a modified PageRank algorithm. This will make it possible to evaluate performance in dynamics.

To accomplish the aim, the following tasks have been set:

– to formalize the task of assessing productivity trends of collective scientific subjects;

– to verify the method of evaluating performance trends of collective scientific subjects based on the modified PageRank algorithm.

4. The study materials and methods

The object of the study is the method of estimating CSS performance trends based on the modified PageRank algorithm.

The study proposes to combine scientometrics methods and methods for processing dynamic series. Scientometrics methods in the study are used to assess scientific performance at certain points in time. In particular, it is proposed to use the following scientometrics methods: the traditional PageRank method, methods for calculating the intensity of citations of scientific publications, and the method of weighing the influence of citations of scientific publications on the result of calculating scientific productivity over time.

To determine the change in estimates over time, it is proposed to use statistical indicators of the growth rate of scientific productivity.

In particular, it is proposed to calculate the basic average changes in performance, as well as chain changes in the CSS performance.

Combining the advantages of the proposed methods, it is proposed to develop a method for assessing CSS performance trends based on a modified PageRank algorithm. Using this method taking into account the modified PageRank algorithm will make it possible to more effectively calculate the performance of CSS in dynamics.

It is possible to formulate a research hypothesis: using the method of evaluating CSS performance trends based on the modified PageRank algorithm will increase the efficiency of their assessment.

This method takes into account the weighing of the influence of citations of scientific publications on the result of the calculation by time and the intensity of citations. Improving the efficiency of calculating productivity trends is ensured by the increased sensitivity of the modified method to new citations of publications.

The obtained information is important in practice for solving the problem of selecting promising CSS for the implementation of scientific consortia.

To verify the developed method, the public database Citation Network Dataset was analyzed [17].

The dataset includes more than 5 million scientific publications, 48 million citations, and is posted in the public domain. Details of the collection of this dataset are described in [18]. For more than 27,500 CSSs, performance estimates for the period from 2000 to 2022 were calculated. This study is a continuation of the study reported in work [16].

5. Results of the study of constructing a method for assessing productivity trends of collective scientific subjects

5. 1. Formalization of the problem of assessing productivity trends of collective scientific subjects

Denote by $U=\{U_1, U_2, ..., U_s\}$ a set of CSSs. Let $P=\{p_1, p_2, ..., p_n\}$ be the set of all scientific publications. $A^h=\{a_1^h, a_2^h, \ldots, a_{d_h}^h\}$ – a set of individual subjects that are affiliated with the collective subject U_h, $h=\overline{1,s}$, $d_h U_h$, $h=\overline{1,s}$, d_h – the number of individual subjects that are affiliated with the corresponding collective subject U_h.

Let the matrix M of citations between scientific publications of different CSSs be given, which is defined as $M=\{c_{hg}\}_{h,g=1}^{s}$, where $c_{ij}\in[0,1]$ is the probability of citation of the publication p_i in the publication p_j, $i,j=\overline{1,s}$, $M\geq 0$, $\sum_{h=1}^{s} c_{hg}=1$, $g=\overline{1,s}$. Denote through q_h^k the coefficient that determines the performance of CSS U_h at k-th step. According to the Time-Weighted PageRank method with citation intensity (TWPR-CI) [16], for $k=0$ the initial coefficients are calculated by the formula:

$$q_h^0=\beta\cdot\operatorname{arctg}\left(\frac{\sum_{j=1}^{d_h^t} c^t\left(a_j^h\right)}{\lambda\left(t_N-t_\delta^h\right)\sum_{j=1}^{d_h^t} p^t\left(a_j^h\right)}\right)+$$

$$+(1-\beta)\cdot\sum_{k=t_\delta^h}^{N}\frac{\left(k-t_\delta^h+1\right)\cdot c^{t_k}\left(a_j^h\right)}{x_i}, \qquad (1)$$

where the first part of the sum determines the weighted intensity of citation of publications by authors who are affiliated with CSS U_h at the time $t\in T$, d_h^t is the number of individual subjects who are affiliated with CSS U_h at time t, $p^t(a_j^h)$ is the number of scientific publications published by an individual subject a_j^h at time t, $c^t(a_j^h)$ – the number of citations of scientific publications of authors who are affiliated with CSS U_h at time t, $\lambda>1$ – some coefficient, t_δ^h – the moment from which the calculation of the intensity of citations of scientific publications for CSS U_h, $x_i=\sum_{k=t_\delta^h}^{T_N}\left(k-t_\delta^h+1\right)$, $h=\overline{1,s}$, $\beta\in[0, 1]$ begins. All other coefficients are calculated iteratively by the formula:

$$q_h^{k+1}=\alpha M q_h^k+\frac{1-\alpha}{h}\mathrm{E},$$

where E is the unit matrix, α is the damping factor.

If $\beta=0$, then the value of the citation intensity in calculating the performance of CSS using the TWPR-CI method is not taken into account. To take into account all the components in the calculation of productivity, the coefficient β must be selected from the interval (0.1).

The first way to establish the productivity of collective scientific subject of CSS U_h is to fix for them at the k-th step the values q_h^k, which are calculated by formula (1), $h=\overline{1,s}$. The CSS, which has the maximum value q_h^k at the k-th step is the most productive, $q_h^k\in[0,1]$. And the value $q_h^k=1$ corresponds to the maximum performance of the CSS U_h. The value $q_h^k=0$ corresponds to the zero performance of the CSS U_h. The latter means that either the CSS does not publish articles in journals with a high impact factor, or these articles are not cited by other scientists.

To highlight promising CSSs, it is not enough to calculate their performance at the current time. It is important to understand how performance changes over time.

Let $q_{h,t}^{k}$ be the performance of CSS U_h for the period t, calculated from (1). Let productivity be calculated annually. Then, based on the results of the calculations, a time series $\left\{q_{h,1}^{k}, q_{h,2}^{k}, \ldots, q_{h,T}^{k}\right\}$ will be constructed and the t index will correspond to the year index.

CSS performance trends can be assessed using the following formulas:

$$\bar{q} = \frac{q_{h,T}^{k} - q_{h,1}^{k}}{T-1}, \qquad (2)$$

$$\bar{\bar{q}} = \sqrt[T]{\frac{q_{h,T}^{k}}{q_{h,1}^{k}}}, \qquad (3)$$

$$\bar{q}_L = \frac{\sum_{w=2}^{T}\left(q_{h,w}^{k} - q_{h,w-1}^{k}\right)}{T-1}, \qquad (4)$$

$$\bar{\bar{q}}_L = \sqrt[T]{\prod_{w=2}^{T} \frac{q_{h,w}^{k}}{q_{h,w-1}^{k}}}, \qquad (5)$$

where $\bar{q}$ is the basic average absolute change in the performance of CSS subjects U_h at time T relative to the initial value of performance $q_{h,1}^{k}$, $h=\overline{1,s}$, $\bar{\bar{q}}$ – the basic average relative change in the performance of CSS U_h at time T relative to the initial value of performance $q_{h,1}^{k}$, $\bar{q}_L$ – the chain average absolute change in the performance of CSS U_h at time T, $\bar{\bar{q}}_L$ – chain average relative change in the performance of CSS U_h at time T.

5. 2. Results of verification of the method for assessing productivity trends of collective scientific subjects

To verify the method of evaluating CSS performance trends based on the modified PageRank method, the Citation Network Dataset was analyzed [17]. This public dataset currently includes data on 5,354,309 scientific publications and 48,227,950 citations of these publications for the period from 1815 to 2022. The data included in the dataset are collected from the databases Microsoft Academic Graph [19], DBLP [20], ACM [21], etc. The areas of scientific publications correspond to computer science, artificial intelligence, neural networks, software engineering, etc.

To assess the trends in the performance of CSS, the period from 2000 to 2022 was chosen. Binary reflections between scientific publications and CSS were calculated, citation graphs of scientific publications were constructed, the weights of the arcs of which correspond to the number of citations. For each CSS, performance estimates are calculated using the TWPR-CI, (1), $\beta=1/2$, $\lambda=2$ method. CSS performance estimates are calculated on the basis of an iterative method with an accuracy of $\varepsilon=10^{-5}$. After that, performance estimates were normalized at the maximum value. Performance estimates using (2) to (5) were also calculated. Collective entities that had zero performance values under the TWPR-CI method during the period from 2000 to 2022 were excluded from consideration. A total of 27,500 CSSs from around the world were analyzed. To visualize the results, 15 CSSs were selected for productivity trends that demonstrate different behaviors. Table 1 gives the relative performance estimates of the selected CSSs, normalized by maximum, which are calculated using the TWPR-CI method for $\beta=1/2$, $\lambda=2$ (1).

Table 2 gives performance estimates, rounded to the second decimal place, calculated using formulas (2), (5): $\bar{q}$ – basic average absolute change in the productivity of collective scientific subjects, $\bar{\bar{q}}_L$ – chain average relative change in the productivity of collective scientific subjects.

Table 1

Relative performance estimates, normalized to the maximum, of some collective scientific subjects, calculated using the TWPR-CI method for $\beta=1/2$, $\lambda=2$. The rows indicate the years, the columns are the indices of collective scientific subjects (indicated after the table)

No. / year	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2000	0.787	0.893	0.930	0.926	0.892	0.410	0.701	0.922	0.783	0.951	0.944	0.961	0.631	0.295	0.895
2001	0.636	0.816	0.861	0.867	0.788	0.251	0.528	0.852	0.839	0.917	0.894	0.923	0.484	0.053	0.795
2002	0.567	0.745	0.814	0.838	0.695	0.283	0.445	0.794	0.911	0.892	0.842	0.890	0.452	0.094	0.705
2003	0.557	0.681	0.778	0.826	0.615	0.313	0.378	0.750	0.927	0.863	0.793	0.859	0.427	0.173	0.295
2004	0.528	0.635	0.752	0.835	0.552	0.242	0.350	0.727	0.941	0.809	0.752	0.834	0.456	0.202	0.171
2005	0.534	0.595	0.728	0.840	0.501	0.205	0.376	0.712	0.948	0.792	0.716	0.820	0.472	0.226	0.246
2006	0.553	0.566	0.713	0.843	0.468	0.205	0.422	0.698	0.956	0.642	0.720	0.808	0.507	0.187	0.299
2007	0.560	0.541	0.703	0.843	0.443	0.239	0.461	0.692	0.960	0.399	0.677	0.794	0.488	0.180	0.320
2008	0.576	0.530	0.709	0.833	0.428	0.234	0.487	0.681	0.964	0.253	0.707	0.769	0.466	0.172	0.253
2009	0.582	0.515	0.709	0.824	0.418	0.335	0.506	0.676	0.969	0.238	0.725	0.755	0.486	0.168	0.213
2010	0.574	0.514	0.711	0.818	0.416	0.423	0.524	0.670	0.970	0.239	0.754	0.752	0.578	0.171	0.209
2011	0.563	0.511	0.716	0.806	0.412	0.496	0.535	0.661	0.971	0.231	0.779	0.749	0.603	0.167	0.253
2012	0.562	0.515	0.722	0.793	0.405	0.492	0.538	0.651	0.971	0.221	0.798	0.738	0.646	0.173	0.305
2013	0.554	0.517	0.724	0.783	0.393	0.503	0.534	0.644	0.971	0.221	0.798	0.734	0.663	0.174	0.369
2014	0.549	0.517	0.722	0.772	0.384	0.535	0.532	0.631	0.971	0.200	0.807	0.725	0.691	0.172	0.407
2015	0.552	0.529	0.731	0.760	0.377	0.523	0.535	0.621	0.971	0.198	0.819	0.723	0.704	0.164	0.462
2016	0.551	0.519	0.731	0.746	0.369	0.530	0.537	0.611	0.970	0.200	0.826	0.719	0.716	0.165	0.462
2017	0.564	0.521	0.739	0.732	0.363	0.583	0.546	0.606	0.970	0.233	0.843	0.722	0.722	0.158	0.480
2018	0.577	0.525	0.748	0.718	0.358	0.611	0.546	0.596	0.970	0.266	0.844	0.712	0.734	0.150	0.492
2019	0.595	0.523	0.751	0.703	0.352	0.627	0.546	0.584	0.970	0.296	0.844	0.709	0.745	0.147	0.486
2020	0.610	0.527	0.751	0.692	0.350	0.630	0.544	0.568	0.969	0.327	0.844	0.700	0.754	0.143	0.496
2021	0.597	0.512	0.740	0.680	0.337	0.619	0.529	0.552	0.967	0.316	0.836	0.688	0.744	0.135	0.481
2022	0.583	0.498	0.730	0.668	0.324	0.606	0.515	0.538	0.966	0.304	0.829	0.676	0.734	0.130	0.467

Note: 1 – University of Bern, Switzerland; 2 – Institute of Computer Science, Jagiellonian University, Poland; 3 – Max Planck Institute for Informatics, Germany; 4 – Argonne National Laboratory, USA; 5 – Faculty of Physics, University of Kyoto, Japan; 6 – Faculty of Zoology, University of Oxford, UK; 7 – University of Fribourg, Switzerland; 8 – Danish School of Education, Aarhus University, Denmark; 9 –AT&T Wireless Services, USA; 10 – University of Liverpool, UK; 11 – Catholic University of Louvain, Belgium; 12 – ECEC Faculty, University of Illinois, USA; 13 – Applied Research Associates Inc.; 14 – Adobe Systems Incorporated; 15 – Fraunhofer Development Center X-ray Technologies, Germany.

Table 2

Performance trend estimates

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\overline{q}$	–0.01	–0.02	–0.01	–0.01	–0.03	0.01	–0.01	–0.02	0.01	–0.03	–0.01	–0.01	0.00	0.00	–0.02
$\overline{\overline{q}}_L$	0.99	0.97	0.99	0.99	0.96	1.02	0.99	0.98	1.01	0.95	0.99	0.98	1.00	1.01	0.97

Visually assessing the change in productivity gains among all CSSs, we can distinguish subjects whose productivity remains unchanged during the observation period (Fig. 1), increases (Fig. 2), or decreases (Fig. 3). Fig. 1–3 show the relative estimates of CSS performance, normalized by maximum, by the TWPR-CI method for $\beta=1/2$, $\lambda=2$, which demonstrate the indicated dynamics for 23 periods corresponding to the observation years: 1 corresponds to 2000, 2 – 2001, and so on.

If the basic average absolute change in the productivity of CSS $\overline{q}<0$, then the productivity of the subject decreases over time, if $\overline{q}>0$, accordingly, increases, if $\overline{q}\approx 0$, there is no change in productivity. Similarly, if the chain average relative change in the productivity of CSS $\overline{\overline{q}}_L<1$, then the productivity of the subject decreases over time, if $\overline{\overline{q}}_L>1$, accordingly, increases, if $\overline{\overline{q}}_L\approx 1$, then the change in productivity does not occur.

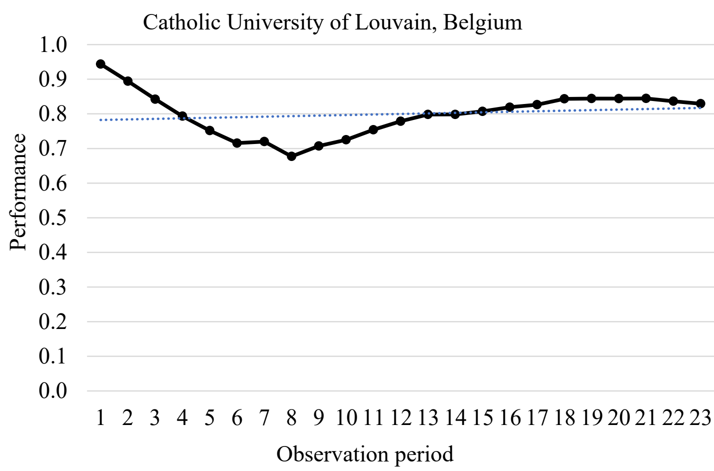

Fig. 1. Relative performance estimates, normalized by maximum, for the Catholic University of Louvain, Belgium

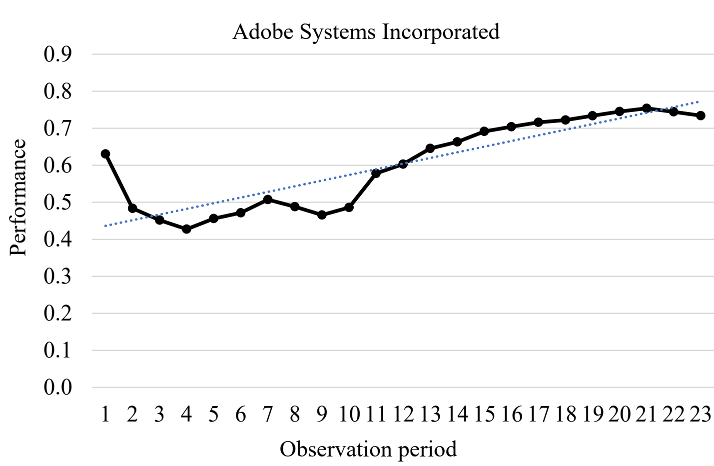

Fig. 2. Relative performance estimates, normalized by maximum, for Adobe Systems Incorporated

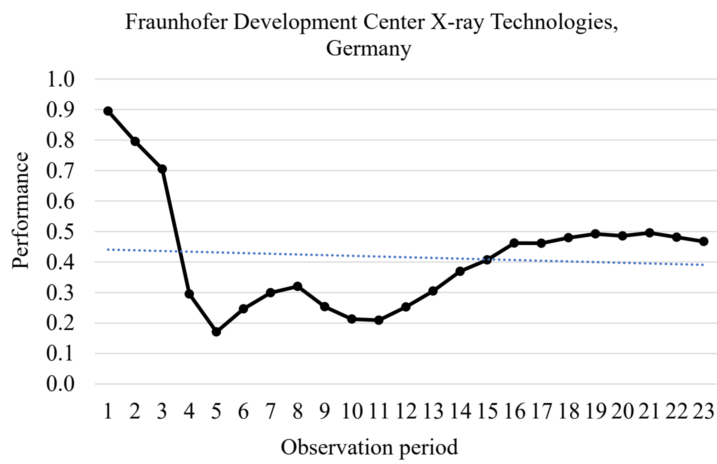

Fig. 3. Relative performance estimates, normalized by maximum, for Fraunhofer Development Center X-ray Technologies, Germany

CSSs, for which $\overline{q}>0$ or $\overline{\overline{q}}_L>1$ are promising in the future as their productivity trends grow. This trend is observed for three CSSs: the Department of Zoology (University of Oxford, UK), AT&T Wireless Services (USA), and Adobe Systems Incorporated.

6. Discussion of the development of a method for assessing productivity trends of collective scientific actors

Our results of the study of trends in assessing the performance of CSS show the presence of both growing and downward trends (Fig. 1–3). The presence of various trends in the performance of CSS can be explained by the fact that the described method makes it possible to comprehensively review the development of CSS productivity and decide on an assessment in terms of changes in productivity over time. The results of the method operation were calculated on 27, 500 CSSs. To visualize the results, 15 CSSs were selected from them, the productivity trends of which demonstrate different behaviors. That is, the goal of the study was achieved: a method for assessing CSS performance trends based on a modified PageRank method that takes into account the intensity of citations and the coefficient of obsolescence of publications has been developed. The method is verified using a large volume dataset.

The study, which is described in this paper, is a continuation of the study reported in [16] in terms of improving the method of calculating productivity. As shown in [16], the proposed TWPR-CI method has advantages over the PageRank and Time-Weighted PageRank method in terms of evaluating the performance of new CSSs during the first 10–12 years. The current study shows the possibility of using this method to assess the performance of CSSs that have different development trends.

The limitation of the study is the structure of the dataset. In the Citation Network Dataset, most scientific publications relate to the field of computer science, artificial intelligence, and software engineering. It can be assumed that in the case of analyzing a dataset for publications in the field of humanities, the results could be different. However, this is a separate research task.

The main disadvantage of the study is that for the identified trends in assessing the performance of new CSSs, no causal relationships of their occurrence have been determined. Also, the results of the study do not make it possible to predict the change of trends in the future.

In this study, $\beta=1/2$ was determined, which was chosen empirically by the researchers. Also a separate task of the study is the calculation of the optimal value of the β parameter for the TWPR-CI method. An important goal in the development of this method is the formation of criteria for the selection of promising CSSs and individual entities that are affiliated with them, for the creation of scientific consortia [22]. It is in this direction that the development of this study can proceed.

7. Conclusions

1. Formalization of the problem of assessing trends in the productivity of CSS was carried out. According to the results of formalization, it was established that in order to assess the trends in scientific productivity, it is necessary to first calculate estimates of the values of CSS productivity. To calculate the scientific performance of CSS, the TWPR-CI method based on the modified PageRank algorithm was used. The use of the proposed TWPR-CI method has qualitative advantages over other methods, in particular PageRank and Time-Weighted PageRank in terms of the sensitivity of evaluating the performance of new CSSs during the first 10–12 calculation periods. Formulas have been derived for assessing performance trends that make it possible to take into account changes in the performance of CSS over time. In the case of ranking a set of CSSs containing new collective subjects, the proposed method makes it possible to better take into account the change in scientific productivity.

2. Verification of the method for assessing CSS performance trends based on the modified PageRank algorithm using a large public dataset was carried out. The dataset includes more than 5 million scientific publications, 48 million citations. The verification results indicate the relevance of the proposed method for quantifying the trends in the productivity of the CSSs (higher education institutions, scientific institutes, laboratories, and other institutions engaged in scientific activities). The calculated indicators should be taken into account when choosing CSSs. In particular, one might prefer those in which the base average absolute performance change is positive or the chain average relative change in performance exceeds unity.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study and the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

All data are available in the main text of the manuscript.

References

1. Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Omirbayev, S., Mukhatayev, A., Faizullin, A., Toxanov, S. (2021). Development of the set models and a method to form information spaces of scientific activity subjects for the steady development of higher education establishments. Eastern-European Journal of Enterprise Technologies, 3 (2 (111)), 6–14. doi: https://doi.org/10.15587/1729-4061.2021.233655
2. Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, 102 (46), 16569–16572. doi: https://doi.org/10.1073/pnas.0507655102
3. Brin, S., Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30 (1-7), 107–117. doi: https://doi.org/10.1016/s0169-7552(98)00110-x
4. Leskovec, J., Rajaraman, A., Ullman, J. D. (2020). Mining of massive datasets. Cambridge University Press, 565. doi: https://doi.org/10.1017/9781108684163
5. Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S. (2019). Improvement of the method for scientific publications clustering based on n-gram analysis and fuzzy method for selecting research partners. Eastern-European Journal of Enterprise Technologies, 4 (4 (100)), 6–14. doi: https://doi.org/10.15587/1729-4061.2019.175139
6. Bianchini, M., Gori, M., Scarselli, F. (2005). Inside PageRank. ACM Transactions on Internet Technology, 5 (1), 92–128. doi: https://doi.org/10.1145/1052934.1052938
7. Assessing universities and other research-focused institutions. Scimago Institutions Rankings. Available at: https://www.scimagoir.com/
8. Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. College & Research Libraries News, 68 (5), 314–316. doi: https://doi.org/10.5860/crln.68.5.7804

9. Zhang, F. (2017). Evaluating journal impact based on weighted citations. Scientometrics, 113 (2), 1155–1169. doi: https://doi.org/10.1007/s11192-017-2510-z
10. Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Mukhatayev, A., Toxanov, S., Faizullin, A. (2020). Methods of Assessing the Scientific Activity of Scientists and Higher Education Institutions. 2020 IEEE 2nd International Conference on Advanced Trends in Information Theory (ATIT). doi: https://doi.org/10.1109/atit50783.2020.9349348
11. Zhang, J., Liu, X. (2022). Citation Oriented AuthorRank for Scientific Publication Ranking. Applied Sciences, 12 (9), 4345. doi: https://doi.org/10.3390/app12094345
12. Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S. (2020). The use of probabilistic latent semantic analysis to identify scientific subject spaces and to evaluate the completeness of covering the results of dissertation studies. Eastern-European Journal of Enterprise Technologies, 4 (4 (106)), 21–28. doi: https://doi.org/10.15587/1729-4061.2020.209886
13. Wang, Y., Zeng, A., Fan, Y., Di, Z. (2019). Ranking scientific publications considering the aging characteristics of citations. Scientometrics, 120 (1), 155–166. doi: https://doi.org/10.1007/s11192-019-03117-9 9
14. Xing, W., Ghorbani, A. (2004). Weighted PageRank algorithm. Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004. doi: https://doi.org/10.1109/dnsr.2004.1344743
15. Manaskasemsak, B., Rungsawang, A., Yamana, H. (2010). Time-weighted web authoritative ranking. Information Retrieval, 14 (2), 133–157. doi: https://doi.org/10.1007/s10791-010-9138-4
16. Kuchansky, A., Biloshchytskyi, A., Andrashko, Y., Biloshchytska, S., Faizullin, A. (2022). The Scientific Productivity of Collective Subjects Based on the Time-Weighted PageRank Method with Citation Intensity. Publications, 10 (4), 40. doi: https://doi.org/10.3390/publications10040040
17. Citation Network Dataset: DBLP+Citation, ACM Citation network. Aminer. Available at: https://www.aminer.org/citation
18. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z. (2008). ArnetMiner. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. doi: https://doi.org/10.1145/1401890.1402008
19. Microsoft Academic Graph. Microsoft. Available at: https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
20. DBLP Computer science bibliography. Available at: https://dblp.org/
21. Association for Computing Machinery. Available at: https://www.acm.org/
22. Xu, H., Kuchansky, A., Gladka, M. (2021). Devising an individually oriented method for selection of scientific activity subjects for implementing scientific projects based on scientometric analysis. Eastern-European Journal of Enterprise Technologies, 6 (3 (114)), 93–100. doi: https://doi.org/10.15587/1729-4061.2021.248040