Webometric Analysis of Russian Scientific and Education Web

Denis Kosyakov[12], Andrey Guskov[234], and Egor Bykhovtsev[123]

[1] Institute of Petroleum Geology and Geophysics, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia
[2] State Public Scientific Technological Library, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia
[3] Institute of Computational Technologies of SB RAS, Novosibirsk, Russia
[4] Novosibirsk State University, Novosibirsk, Russia
kosyakovdv@ipgg.sbras.ru
http://www.ipgg.sbras.ru

Abstract. The aim of this work is to create a publicly available database with webometric indicators for research and higher education organizations in Russia updated on monthly schedule accessible through the project's website http://www.webometrix.ru. This paper describes the set-up of the project including initial gathering and actualization of organizations and their web domains list, sources of data, measuring of indicators' values, analytics available on projects website. Starting from 613 institutions of Russian academies of sciences in January 2015 from the middle of 2015 we gather data for 2201 organizations including research and higher education institutions. Continuous data for more than a year allowed us to assess the reliability of indicators used and to draw some conclusions about Russian scientific and educational web space.

Keywords: webometrics, informetrics, websites, rankings.

1 Introduction

In recent years, webometric studies based on the web search engine usage [1][2][3], become a recognized method of measurement of academic institutions websites quality and impact. However, as other informetric studies, this method remains quite controversial due to gaps in research base, quality of measurement instrumentation and weight and meaning of measured indicators. We suppose that there is a lack of nationwide recurring measurements and juxtaposition between webometric and other kinds of scientometric assessments.

This research is based on monthly webometric data collection for websites of over than 2200 Russian research organizations and higher education institutions. Deep analysis of time series of measurements of particular webometric indicators in some cases combined with the parsing of website structure, examining its peculiarities, and correlating with usage statistics [4] allowed us to examine in details the significance of each of indicators, propose some justification methods and to compare different approaches to calculation of webometrics rankings.

In this report, we consider the principles and architecture of the webometric data collecting system. It contains a webometric indicators database with monthly data since January 2015, and web interface (http://www.webometrix.ru), that allows anyone to perform an analysis of trends and evolution of the scientific websites. Institutions can use it to examine the position and dynamics of their website, to compare it with the others and, as a result, to find the ways of its improvement. In the final part of the report, we made an overview of the Russian Academic and Education Web. We also compared webometric rankings with bibliometric data on institutions' academic output and website usage statistics.

2 Area of Study

We began data collection in January 2015 for institutions under the supervision of Russian Federal Agency of Scientific Organizations (so-called academic institutions, since before they were subject to state academies). Official websites and corresponding distinct domains were determined for 613 organizations, most of them being the research institutions. In July 2015 we have added Russian higher and further education institutions, non-academic research organizations and scientific development and production centers to data collection.

The resulting collection contains data for 2201 organization with distinct DNS domains which also include some regional branches with separate websites and corresponding domains. Of these, 1172 organizations are in the research and development segment and 1029 – in the higher education segment.

At the time of the initial information gathering we could not find a consistent and complete list of such organizations and therefore the Scientific Digital Library database combined with Russian Science Citation Index (RSCI) located at http://elibrary.ru was used. Since eLibrary.ru database covers most of the scientific publications of the Russian authors, we assume that it refers almost all organizations with employees that are somehow engaged in scientific activities. Corresponding website addresses were obtained via Google and Yandex search by title with visual verification.

According to Russian State Statistics Agency, there were 2827 research and development organizations in Russia at the end of 2015. However, some of them do not maintain an official website or publish scientific works either due to the restricted field of research or pure technological and construction character of its activities. In higher education segment there are 609 state institutions and universities and 437 private ones with the total of 1046.

Among higher education institutions we allocated well separated classes associated with both the tendencies of development of the Russian higher education in recent years – the federal and national research universities, and with the legacy of Soviet Union – classical, technical, medical, humanitarian, educational, economic, legal and agricultural universities. In R&D segment we allocated the institutions under the control of Federal Agency of Research Organizations (academic institutions) and national research centers, as these classes are funded under government programs in a special way. The rest of the research institu-

tions are under the control of different federal authorities, corporations, and also various forms of private and public organizations.

Organizations have a substantially different scale and scientific activity. Unfortunately, detailed data on the number of researchers and faculty members at institutions are not available, so we extracted a number of contributing authors for each organization that has publications in 5 recent years. These data do not fully reflect the organization's research staff, however, these numbers allows us to make an adequate assessment. For large research institutions and universities, the number of contributing authors may exceed the number of actual faculty members and research staff because of temporary employees and students. Thus, the Moscow State University has about 9000 faculty members while eLibrary counts 14211 contributing authors. And vice versa for small organizations and universities the number of authors may be less than faculty members and researchers. The resulting treemap is shown in Fig. 1.

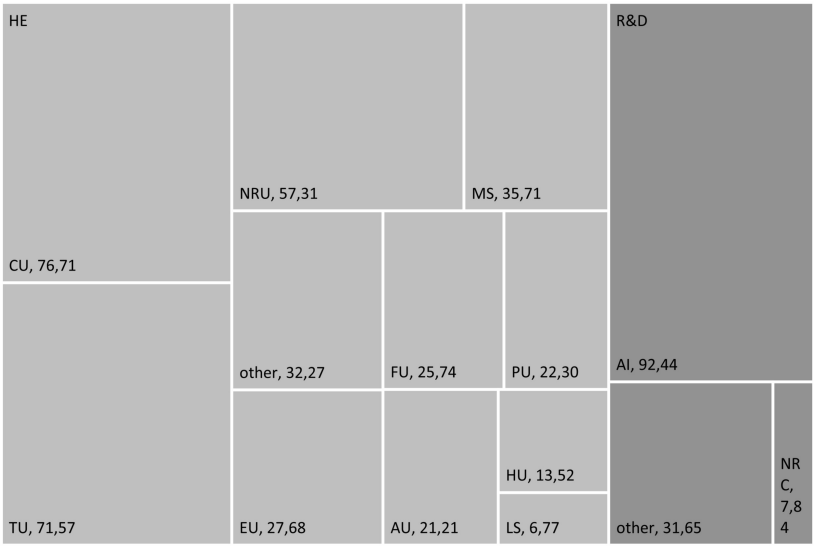

Fig. 1. Russian Higher Education and R&D segments treemap by the number of authors were published in 5 recent years (2011-2015), in thousands. Hereinafter: HE Higher Education segment: FU Federal Universities, NRU National research Universities, CU Classic Universities, TU Technical Universities, EU Economics Universities, MS Medical Schools, PU Pedagogical Universities, HU Humanitarian universities, LS Law Schools, AU Agricultural Universities. R&D Reasearch and Development segment: AI Academic Institutions (under the control of Federal Agency of Scientific Organizations), NRC National Research Centers

Also, from the same source data the number of articles published in the last 5 years and registered in Web of Science and Scopus databases have been extracted (Fig. 2). Detailed data are also shown in Table. 1.

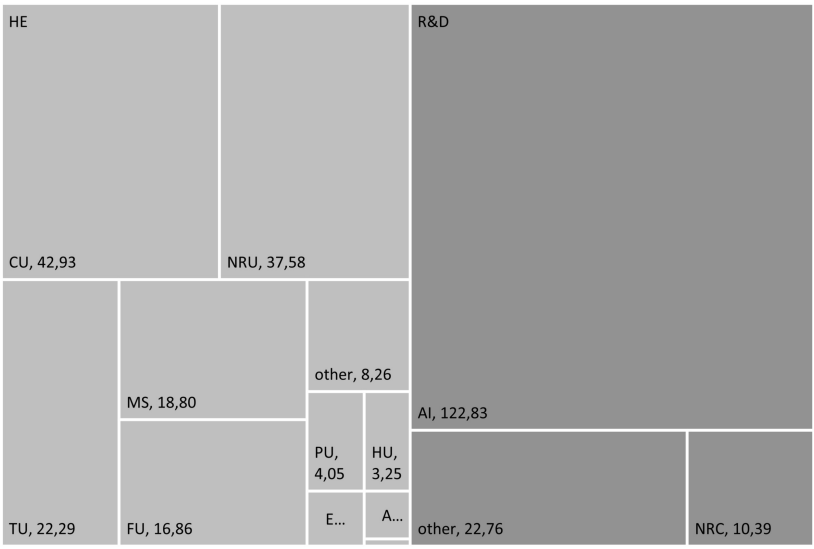

Fig. 2. Russian Higher Education and R&D segments treemap by the number of publications published in 5 recent years (2011-2015) and registered in Web of Science or Scopus databases, in thousands.

3 Metrics and Data Sources

As in [2] we consider those main metrics for our study:

1. Domain size that is measured as a number of hits in corresponding search engine for a request limited by domain URL.
2. A number of documents in popular formats ps, pdf, doc(x), ppt(x), xls(x) measured by narrowing previous request by the appropriate filter.
3. A number of scholarly papers indexed by Google Scholar in which can be full-text documents or web pages with the correct publication metadata.
4. A number of external references hyperlinks from other domains to the target domain pages (and a number of such domains).

The first two metrics are collected via Google, Bing and Yandex search engines and the last one with the help of Ahrefs and Majestic SEO search optimization and backlinks tracking services. Additionally, we obtain basic usage statistics

Table 1. Decomposition of Russian Higher Education and R&D segments

segment / Class	Number of organizations	Percentage of authors	Percentage of papers in RSCI	Percentage of papers in WoS / Scopus
HE	1029	74.8%	80.0%	50.3%
FU	10	4.9%	5.2%	5.4%
NRU	29	11.0%	10.2%	12.0%
CU	72	14.7%	15.8%	13.7%
TU	131	13.7%	14.5%	7.1%
EU	85	5.3%	6.7%	0.7%
HU	61	2.6%	3.3%	1.0%
MS	50	6.8%	5.5%	6.0%
PU	48	4.3%	5.0%	1.3%
LS	22	1.3%	1.6%	0.1%
AU	49	4.1%	5.0%	0.5%
others	472	6.2%	7.2%	2.6%
R&D	1172	25.2%	20.0%	49.7%
AI	613	17.7%	15.0%	39.1%
NRC	35	1.5%	0.8%	3.3%
others	524	6.1%	4.2%	7.2%
Total	2201	100%	100%	100%

from SimilarWeb service, that uses data extracted from four main sources: 1) a panel of web surfers made of millions of anonymous users equipped with a portfolio of apps, browser plugins, desktop extensions, and software; 2) global and local ISPs; 3) web traffic directly measured from a learning set of selected websites and intended for specialized estimation algorithms; 4) A colony of web crawlers that scan the entire Web. Comparison of this data to the data, gathered from corresponding Google Analytics and Yandex Metrika site counters for several academic websites, participated in our research [4] shows its sufficient accuracy.

4 Data Collection and Processing Pipeline

Data are collected by PowerShell scripts that request Yandex and Bing web services and process web pages obtained from Google, Google Scholar and SimilarWeb using Internet Explorer automation with operators visual control. Ahrefs and Majestic data are collected by their bulk export features. Results are stored in the MongoDB database.

During data collection we met with 2 types of errors and failures mainly in interpreting of web pages received as a response to a query: a) errors caused by wrong response parsing due to unexpected changes in response details, b) changes in search engine's database caused by reindexing of web sites and global index rebuilding.

The first type of distortions may be detected by low and high pass filter, which compares the measured value with the average of several previous values. Such combination of band filter with moving average allows us to detect single isolated outliers. If errors were detected after the data collection cycle and cannot be corrected by recollection, data may be recovered by linear interpolation of neighbor values. We retain originally collected values also.

The situation is much worse with effects caused by reindexing of some websites and global rebuilding of search engines' indices that occur relatively often [5] and affect measurements dramatically. For example, in July 2015 Bing counted 2.8 millions of pages in the domain of Institute of Astronomy of the Russian Academy of Sciences (inasan.ru), but in subsequent months this indicator falls back to several thousand of pages. We analyzed these effects in details in [6]. These effects may last more than one month and so we need more sophisticated logic to determine and justify them and it is one of directions of future investigations.

For each domain 26 metrics are gathered with 9 main indicators and 17 supplementary. We retain exact timestamp for each value collected. The monthly cycle lasts for more than a week resulting in 57 226 values. The projects website provides the following basic functionality:

- Ranking of organizations web domains by one of the indicators and its change in time in the tabular form.
- Ranking of organizations web domains by every indicator values for the single month.
- Dynamics of totals, means and medians of selected indicators for all or selected part of domains during selected time period.
- Comparison of different series indicator month as a scatter chart.
- Detailed info for a single domain.

5 Analysis of Research and Education Web Space

Data collected for the majority of Russian research and higher education organizations for a period of 10 months allow us to make some brief review of basic characteristics of Russian research and higher education web space in terms of size and quality. By size we mean a number of pages and documents and by quality a number of papers in Google Scholar index, Yandex thematic citation index, and a number of visits per month. As we show in [6] backlinks data is quite controversial and cannot be used in thorough analysis without complex cleaning.

First of all, let us look at the size of Russian scientific and education web and its dynamics (Fig. 3). At first sight, we can deduce that total size of the segment under consideration increases with the exclusion of Bing data, that can be justified by some Bing engine peculiarities. But if we take into account the total size of Russian web space which can be measured as the .ru zone size we can see a more complicated situation (Fig. 4). While the size of overall Russian web in Google index increases in more than 150% and in Bing nearly doubles,

Yandex oscillates near 400 millions of pages. At the same time, academic share in Bing index decreased from 18% to 7%, in Google remains almost the same (7% to 6%) but in Yandex it tripled from 6% to 18%.

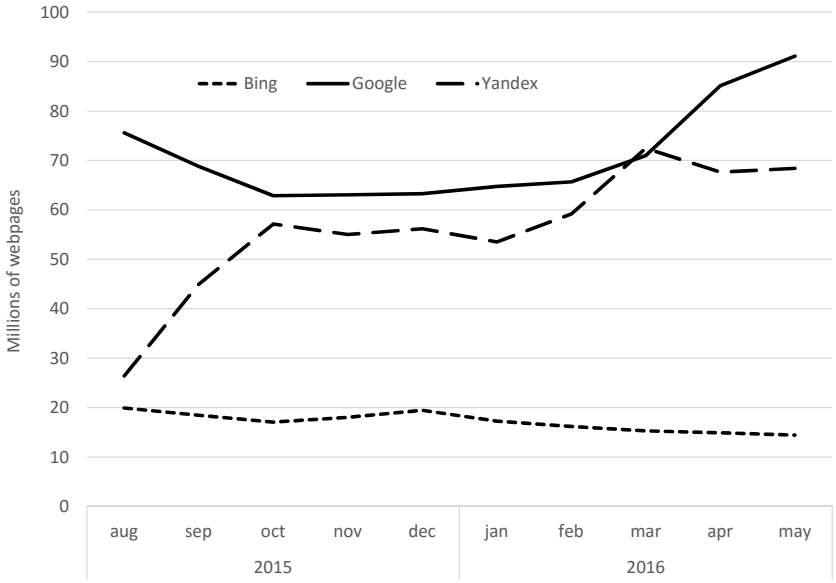

Fig. 3. Total number of pages in Russian research and education web space by month

While considering indicators based on search engines we will use aggregated values calculated as the maximum of values, obtained from Google, Yandex and Bing onwards. The total size of the Education and Scientific web measured by this indicator has grown from 78 million pages in August 2015 to more than 100 million in May 2016. The share of HE decreased by almost 3% from 74.3% to 71.6%. The fastest growing classes were TU, CU and NRU. The share of R&D is divided almost equally between AI and other R&D organizations (14%) with a minor proportion of NRC (0,3%). During the studied period the proportion of AI decreased from 17% with a simultaneous growth of others from 8%. Web space partition in classes is shown on Fig. 5.

HE average domain size increased from 111 to 146 thousand of pages, R&D - from 14 to 19.5 thousand, the most active growth of average domain size from 303 to 525 thousand of pages was in FU and NRU domains grew from 272 to 371 thousand of pages. The average size of domains LS, EU, MS and PU has not changed, and the HU and AU even decreased. In R&D the average size of AI and NRC domains have not changed, the increase was only observed among others.

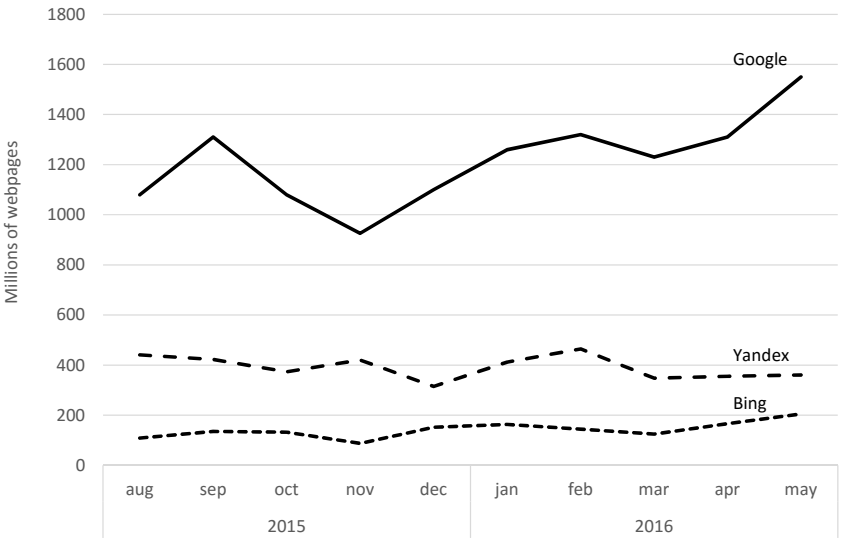

Fig. 4. Total number of pages in .ru zone by month

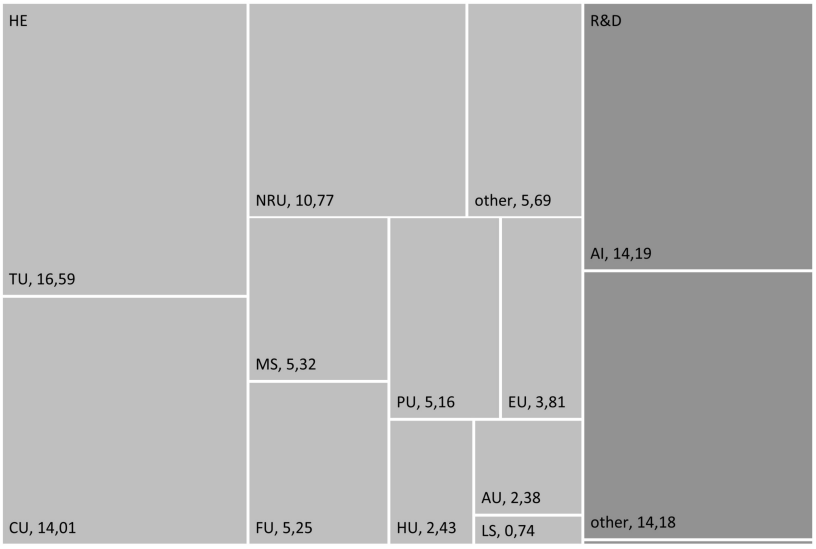

Fig. 5. Russian Higher Education and R&D segments treemap by the number of web pages indexed by main search engines, in millions.

The total number of documents increased from 6 to 7.7 million (Fig. 6). The main growth in both absolute terms and in share occurred in the NRU and the CU classes, final distribution is shown in Fig. 7. The average number of documents in FU class increased from 40 to 50 thousand, NRU – from 30 to 42, CU – from 13 to 18. In the R&D average number of documents grows slowly and is slightly more than 1000 documents in a domain. It should be noted that the bulk of documents are on the top domains for each group as mean values significantly higher than the median, and even the upper quartile.

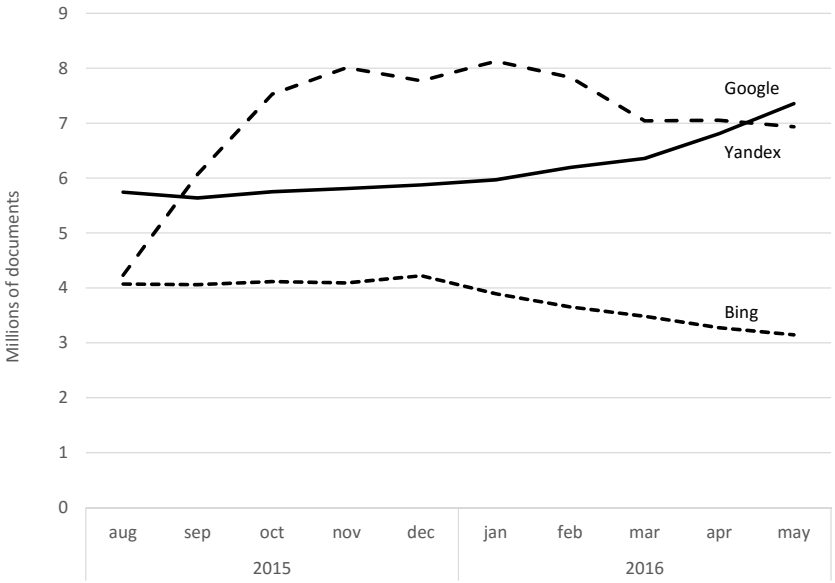

Fig. 6. Total number of documents in Russian research and education web space by month

Number of publications that are indexed by Google Scholar has changed slightly from 664 up to 693 thousand. The bulk of publications indexed are in the NRU, CU, and AI domains (Fig. 8) and located on a small number of leading websites. Main growth was observed in FU domains.

Finally, let us take a look at site traffic of research organizations and higher education institutions. The total site traffic has increased from 70 to 97 million sessions per month, most of the growth occurred in the sites of AI and other R&D organizations (in general in the R&D segment we can observe almost two-time growth) as well as in the NRU, FU, PU and CU sites – growth was from 40% to 58%. Final distribution is shown on Fig. 9. The highest average number of 0.5 million of sessions per month was in FU and NRU classes. In R&D NRC shows the highest values of about 153 thousand sessions, slightly less than the

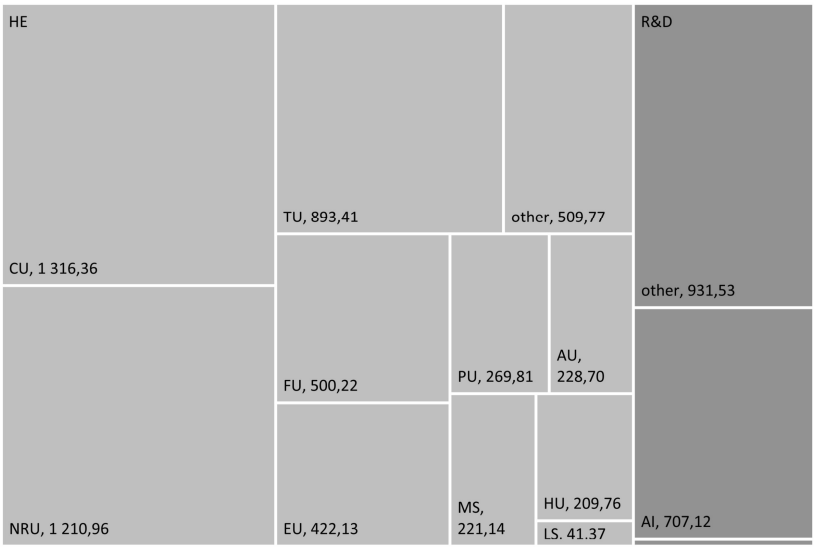

Fig. 7. Russian Higher Education and R&D segments treemap by the number of documents indexed by main search engines, in thousands.

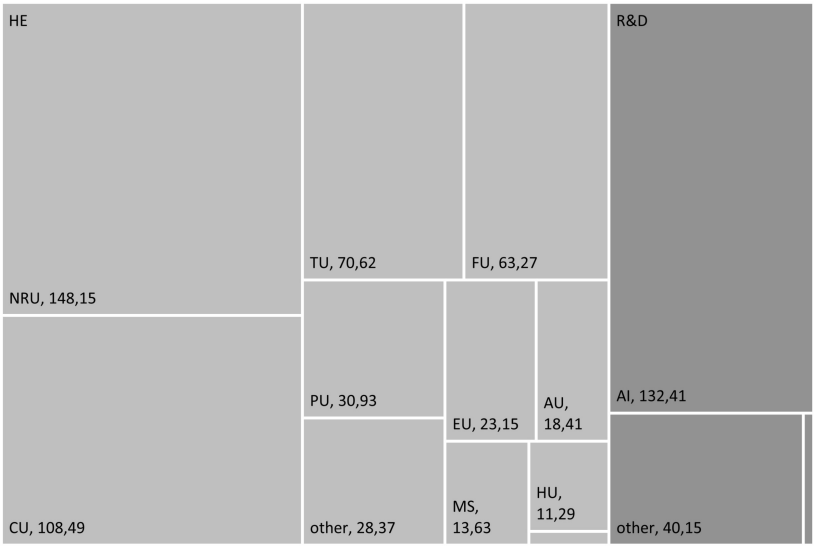

Fig. 8. Russian Higher Education and R&D segments treemap by the number of publications indexed by Google Scholar, in thousands.

CU (186 thousand sessions). The most effective were NRC sites for which 100 pages indexed by search engines resulted in more than 1800 sessions per month. AI showed the lowest efficiency of all of the classes with 47 sessions per 100 pages indexed. In HE segment NRU were the best (150 sessions) and the other average values were about 100 sessions except MS (70) and the PU (82).

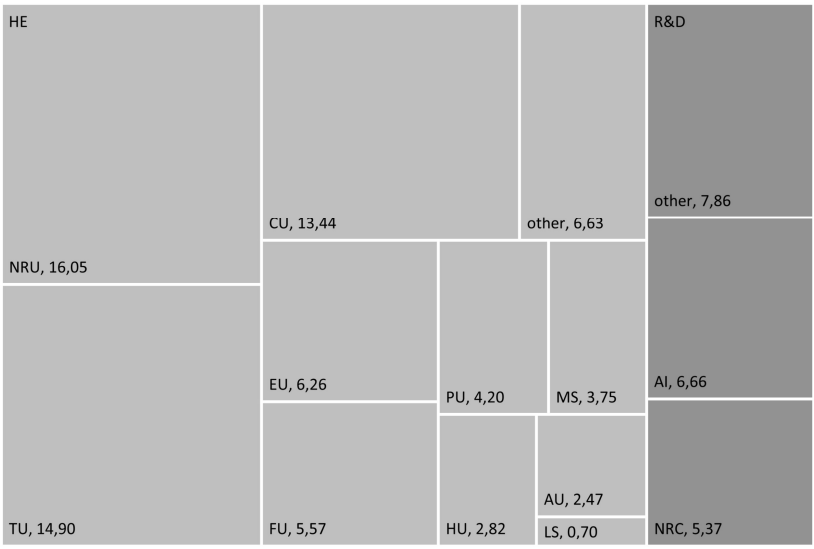

Fig. 9. Russian Higher Education and R&D segments treemap by the number user sessions per month, in millions.

6 Conclusions

A comparison of the segments and classes shares for different indicators allows identifying of possible points of growth. The greatest growth potential is concentrated in the area of open science – access to full-text and metadata of scientific publications. Most of the publications of more than 7 million indexed by Google Scholar in the Russian segment of the Internet resides on the sites of scientific digital libraries eLibrary.ru (about 4 million) and CyberLeninka.ru (slightly more than 1 million). It should be noted that documents, indexed by Google Scholar are highly rated in general Google search results, leading to an increase in the site traffic and contribute to the promotion of scientific results. Organization's web site can provide quite a different context with information on current research projects and different kinds of scientific output to a visitor than a digital library and in most cases, it leads to better results. The total number of publications of the organizations in question only for the last 5 years was more than

3 million of which less than 25% are available online. Noting the great progress and the rapid development of Internet resources of federal and national research universities and the weak, and often even a negative trend in other classes, we can conclude about the high and unrealized development potential in some of the classical universities and other types of higher education institutions.

Finally, one can see a clear backlog of R&D segment combined with high scientific potential, which may be partly explained by a more narrow, niche nature of Web resources. However, the leaders of this segment show good results and demonstrate the broad development opportunities for others.

An analysis of the dynamics of webometric indicators allows a better understanding of trends in the development of the studied web space, neutralize weaknesses inherent in measuring instruments and provide a better picture. Source data and tools located on the project site at http://www.webometrix.ru enable researchers and owners of Internet resources to explore trends in the development of scientific and educational web space, determine the position of specific organizations.

We understand that webometric rankings are quite rough because of nature of measurement instrumentation, but we suppose that conclusions drawn from such assessment may give a rise to efforts to improve web representation of educational and scientific activities.

References

1. Bar-Ilan, J: The use of web search engines in information science research. Annual Review of Information Science and Technology, 38, 231 (2004).
2. Aguillo, I.F., Ortega, J.L., Fernandez, M.: Webometric Ranking of world universities. Higher Education in Europe, 33, 233 (2008).
3. Shokin, Yu.I., Klimenko, O.A., Rychkova, E.V., Shabalnikov, I.V.: Website ranking of scientific institutions of SB RAS. Computational Technologies, 13, 128 (2008).
4. Guskov, A. E., Bykhovtsev, E. S., Kosyakov, D. V.: Alternative Webometrics: Study of the traffic of the websites of scientific organizations. Scientific and Technical Information Processing, series 1, 12, 12 (2015).
5. Van den Bosch, A, Bogers, T, de Kunder, M: Estimating search engine index size variability: a 9-year longitudinal study. Scientometrics, 147, 839-856 (2016)
6. Kosyakov, D. V., Guskov, A. E., Bykhovtsev, E. S.: Russia's academic institutions as mirrored by webometrics. Herald of RAS, 86, 490-499 (2016).