Expert Bibliometrics: An Application Service for Metric Studies of Information

Adilson Luiz Pinto[1(✉)], Rogério de Aquino Silva[1], André Fabiano Dyck[1,2], Gustavo Medeiros de Araújo[1], and Moisés Lima Dutra[1]

[1] PGCIN, Federal University of Santa Catarina, Florianópolis, Brazil
adilson.pinto@ufsc.br
[2] SETIC, Federal University of Santa Catarina, Florianópolis, Brazil

Abstract. Following the current trend of developing software as a service, we propose a software application to handle and support metric studies of information by applying the main laws of Bibliometrics, such as Lotka, Bradford, and Zipf, by means of integrating content and format in a single analysis process. Our application manages metrics according to the relevance of data, dispersion, rule of three and square root, in order to generate new indexes by relying on theories and laws already established among the scientific community. In addition, the developed tool has a pleasant aesthetic, along with a low cognitive effort for the user. To achieve such a scenario, a standardization of the interface combined with the fluidity of navigation within the application were used. The proposed application is suitable for those who work with academic and/or scientific issues, offering quick results compared to manual work, in which a lot of time is spent, either creating a system or analyzing spreadsheets. The result is a beta tool, available online at http://expertsbibliometrics.ufsc.br/, with login and password, respectively: 'ebbc2020@ufsc.br' and 'trocar123'.

Keywords: Bibliometric software · Metric studies application · Lotka application · Bradford application · Zipf application

1 Introduction

The development of tools to support the various types of metric studies presents a successful case among the scientific environment, since every knowledge area makes use of them to organize and understand its specific body of knowledge. These tools help facilitate the understanding of technical and scientific scenarios and trends. Typically, they are shaped in the form of software applications, which handle metric criteria according to specific needs. In this sense, there are some initiatives, such as Vantage Point [1]; CopalRed [2]; IN-Spire [3]; InCites [4]; SciMAT [5]; CiteSpace [6]; BibExcel [7]; SciVal [8]; Sci2 Tool [9]; Publish or Perish [10]; Network WorkBench [11]; VOS Viewer [12]; and Sitkis [13] that stand out in the global scope of metric studies of information. Those tools work with content-clustering systems, frequency of occurrences and average values for processing their calculations.

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2021
Published by Springer Nature Switzerland AG 2021. All Rights Reserved
E. Bisset Álvarez (Ed.): DIONE 2021, LNICST 378, pp. 54–63, 2021.
https://doi.org/10.1007/978-3-030-77417-2_4

This group of applications usually calculate frequency for co-words, authority citation, journal citation, bibliographic coupling, clustering, h-index, co-authorship, performance of institutions, geolocation, among others. However, to the best of our knowledge, there is a lack of tools that apply the laws of Bibliometrics and some bibliometric standards that gave rise to these studies, when considering the complexity of representing data and images, for example. The Bibliometric laws were defined based on studies carried out by **Lotka** [14], **Bradford** [15], and **Zipf** [16].

Lotka [14] proposes the 80/20 rule (inverse distribution law), which establishes there is a core group of highly productive authors in a given theme or area of knowledge and their function is to assess the regularity of scientific productivity in that area. The 80/20 rule works with a constant for each theme (C) divided by the square of the total number of publications by author (n^2). In general, this rule first ranks the 20% of most productive authors identified by the amount of their publications. In fact, nowadays it is quite complex to keep this 80/20 ratio up to date. This is due to the fact that it is extremely difficult to keep an updated count of the authors' publications. There are more and more journal editions being published, new journals being created (especially the open access ones), and new sub-areas of knowledge being defined. In addition, there is the fact that one begins to quantify the efficiency of the authors from other types of documents, such as books, works presented at events, and book chapters. This is because not all areas of knowledge give exclusivity to the publication of articles in scientific journals.

$$Y_{(n)} = \frac{C}{n^2}$$

Bradford [15] determines the core group of the most productive journals in a given theme and whose representation is organized based on the number of titles (both within the same theme and within the same database), by zones or subsets (e.g.: zone of 33%) multiplied by the proportionality factor of the number of journals in each of the defined zones. Often the sum of the number of journals is calculated by a decreasing ranking, by means of using three equal dimensions of 1/3 each or 33% in each zone. The calculation is a division of the total by 3, forming a zone with a large percentage (1st scale), but with few titles; another zone with a median representative percentage (2nd scale), but with a considerable number of journals; and a third zone with a small representative percentage (3rd scale), but with so many journals that, as a rule, only count a single appearance on the topic;

$$p : p1 : p2 : 1 : n : n^2$$

Zipf [16] proposes metrics for supporting thematic issues of publications and revealing what the interdisciplinary contents of documents are. The focus of its applicability relies on the analysis of journal keywords, in order to identify which journals are the most representative ones in the complete universe of existing journals. However, this law's initial proposal aimed to identify the ranking position of frequent words in a given text in relation to the most frequent ones. At that time, all words were counted, including articles and prepositions. In current applications, only the keywords of the articles are worked on, thus facilitating an analysis more directed to the understanding of the syntheses created by the authors to shortly and objectively represent their texts via keywords. Zipf's Law works with a constant extracted from the principle of least effort (c),

which is obtained by multiplying the ranking of the most frequent words in the text (r) by the frequency of occurrences of the words in the text (f). Recently, this same law was applied, with variations, in "stability zones of appearance of terms", with the first zone being reflected as "trivial", the second zone as "interesting data" and the third zone as "noise". It is possible to apply this same approach in the way we understand the information resulting from searches on search engines.

$$(r)(f) = c$$

In this paper, we propose an application service to work with metric studies of information that uses not only the three basic laws of Bibliometry, but also that includes the application of other techniques. This proposal includes the application of two other techniques for analyzing scientific impact, which use other metrics in addition to calculating the frequency of terms. These techniques are the ones proposed by **Price** [17] and **Platz** [18].

Price [17] is an analysis model that identifies the elite of most cited authors, by using the sum total of citations received by the authors. To identify the amount of these elite authors, the square root of the total number of authors is calculated.

$$E = \sqrt{n}$$

Platz [18] works with a visibility index related to the presence of journals in their scientific contexts, which is calculated by dividing the total number of citations received (from other journals) by the total number of articles published by this same journal.

$$V = \frac{In\_cb}{A}$$

The proposed analytical tool has the potential to become one of the most relevant in its niche, whether by calculating scientific productivity or by calculating the impact of publications. Moreover, it comprises the fusion of content for analysis extracted from different databases, such as Web of Science, Scopus, LISA, and Google Scholar, for example. The proposed tool, Expert Bibliometrics, can be accessed at http://expertsbibliometrics.ufsc.br/,with login and password, respectively:**'ebbc2020@ufsc.br'**and'**trocar123**'.

2 Goals

The need for an software application that could integrate content and different formats of corpora in a single analytical dashboard, along with the need for an analytical tool that checks the laws and the main standards of metric studies of information, were the main motivating factors behind this proposal. Furthermore, our expectation is to go beyond the mere representation of indicators by analyzing frequency of terms. We intend to do this by means of applying dispersion techniques, rules of three and square roots, in order to generate new indexes based on theories and laws already established in the scientific environment, but little represented in software applications.

Thus, the main objective of this proposal is to develop a software tool to analyze data through the application of the main laws and bibliometric standards. Specifically, the proposal aims to:

a) Analyze content extracted from several extentions of databases, like CSV format separated by comma and in BIBTEXformat, and aggregate them into common fields;
b) Apply bibliometric laws (Lotka, Bradford and Zipf), depending on the relevance of the data;
c) Apply other bibliometric techniques, such as Price and Platz's theories, for authority and journal citations; and
d) Generate analysis results by more than one representation, such as data clustering based on the graph theory.

3 Development

The main features of the Expert Bibliometrics tool are:

a) It is a software application made available as a service, accessible from any browser, on any operating system;
b) The input data to be processed are the result of the search in several journal databases (Web of Science, Scopus, LISA, IEEE, etc.), in which the user undertook a previous search. To achieve this scenario, the same search terms are used by Expert Bibliometrics. When this occurs, that is, when analyzing data from several journals, the user needs to ensure that his previous searches in each of the databases have used the same search terms. After that, the raw data is downloaded to the user's local computer. The user needs to be aware of the standards defined for the column names in the downloaded data, so that it is possible to select which columns will be analyzed together, when applicable, e.g.: Author, Authors, and/or AU. Through this combination, the system will be able to uniformly reformat the data;
c) It provides an option called "Organize Table", which standardizes the representation of the uploaded files. For example, one file could use "semicolons" to separate fields, another one could use "comma space-space" or "comma space", and still other ones could use "semicolon" to separate different fields and "comma" within the same field. Care must be taken not to separate the surname from the Author's first name, for example. This software feature makes a quantification of commas within each field, to know whether to separate the columns with a comma or semicolon. This arrangement is done before the file is effectively separated into columns;
d) It allows uploading files in CSV format separated by commas and in BIBTEX format. The user needs to know the content of these files, in order to be able to select the correct columns to be analyzed;
e) It can analyze files with only a few dozen records, as well as files with thousands of records. The processing time is proportional to the number of records to be analyzed;
f) It is able to analyze a wide variety of files to apply the laws of Lotka, Bradford and Zipf. The results are presented in the form of: (i) a synthetic summary; (ii) graphical visualizations of the percentages of publications; and (iii) in the form of graphs, in which the clusters of authors are shown.

g) In the next version of the tool, still under development, it will also be possible to analyze visibility applications.

The features of Expert Bibliometrics can be classified into three categories (Figs. 1 and 2):

(a) **Import**. Feature that receives raw data from different periodical databases.

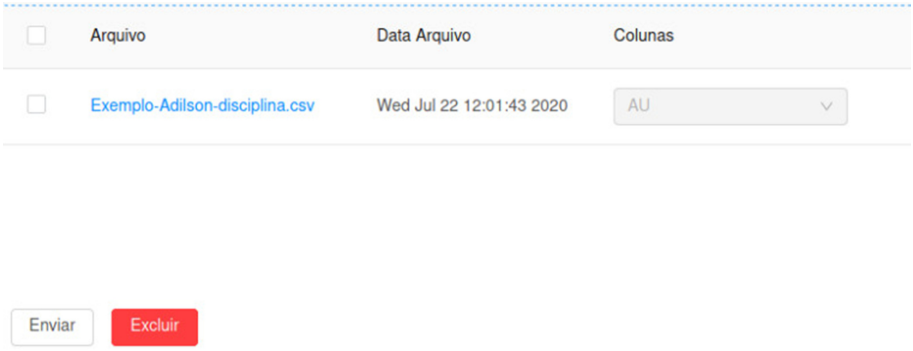

Fig. 1. Files uploaded to be analyzed

(b) **Analysis**. A application of the laws of Lotka, Bradford and Zipf on imported data, and.

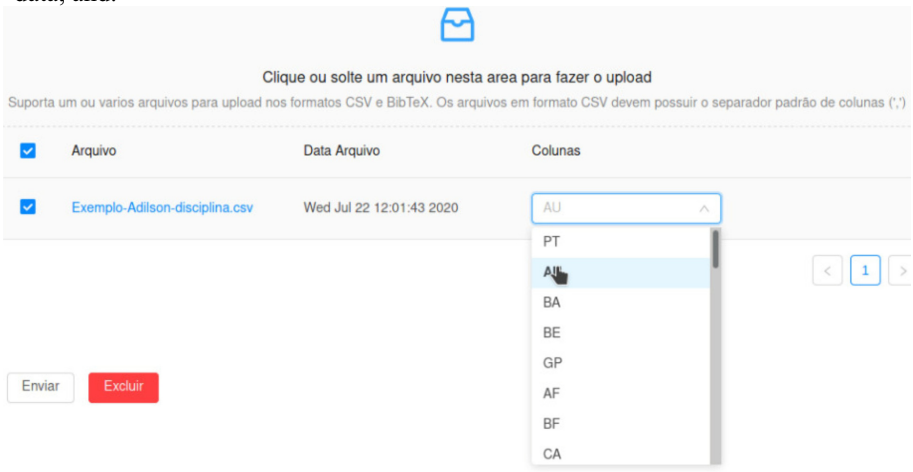

Fig. 2. File and column selection (AU) for analysis of authors

(c) **Report**. Presentation of the results.

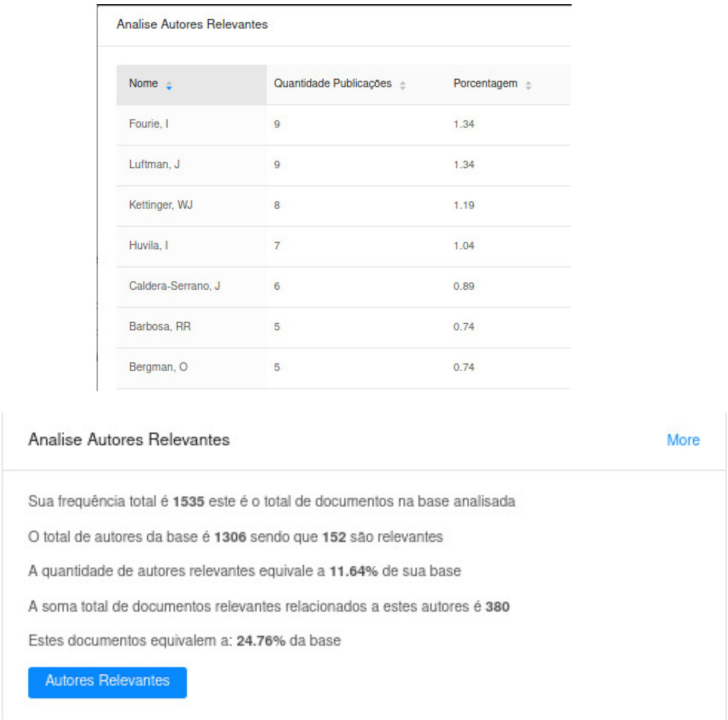

Fig. 3. Summary report of the analysis of relevant authors.

4 Results

The tool's dashboard was developed with the focus on offering the user a standardized interface with pleasant aesthetics, low cognitive effort, with the least possible number of clicks per functionality, and with smooth navigation in the software.

The proposed application service, Expert Bibliometrics 1.0, has the following modules:

a) User authentication module;
b) Module for loading multiple files in CSV format separated by comma and in BIBTEX format, extracted from different databases;
c) Module for analyzing author relevance;
d) Module for analyzing the more representative journals;
e) Module for analyzing keyword relevance;
f) Analysis report module, in table format (Fig. 3);
g) Analysis report module, in graph format (Fig. 6); and
h) Authors' analysis report module, in graph format (Fig. 6).

5 Discussion

We undertook a search in the Web of Science database, with the following configuration:

- search expression: “information management”;
- field: Topic; range: 2008 to 2017;
- collections: Science Citation Index Expanded (SCI-EXPANDED), Social Sciences Citation Index (SSCI), Arts & Humanities Citation Index (A & HCI);
- category: Information Science Library Science

When performing the search, we obtained 672 results.

In a non-automated scenario, after obtaining these results and exporting them, there is a manual work of handling on author and journal data. After that, the application of the formulas of each law (e.g. Lotka and Bradford) begins, as well as a co-authoring analysis. In this specific scenario, both tasks could take a considerable amount of time to generate results. When using the Expert Bibliometrics tool, this time will be considerably reduced. Expert Bibliometrics provides: (i) the necessary formulas for bibliometric laws integrated in its source code; (ii) uploading the search result files from external databases; (iii) the “Author Analysis” or “Journal Analysis” options, to automatically obtain the results of the application of Lotka’s and Bradford’s laws, respectively.

The “Author Analysis” feature, which uses Lotka’s law applied to the results of the aforementioned search, for a given set of criteria, produces the data presented in Fig. 4 as a result.

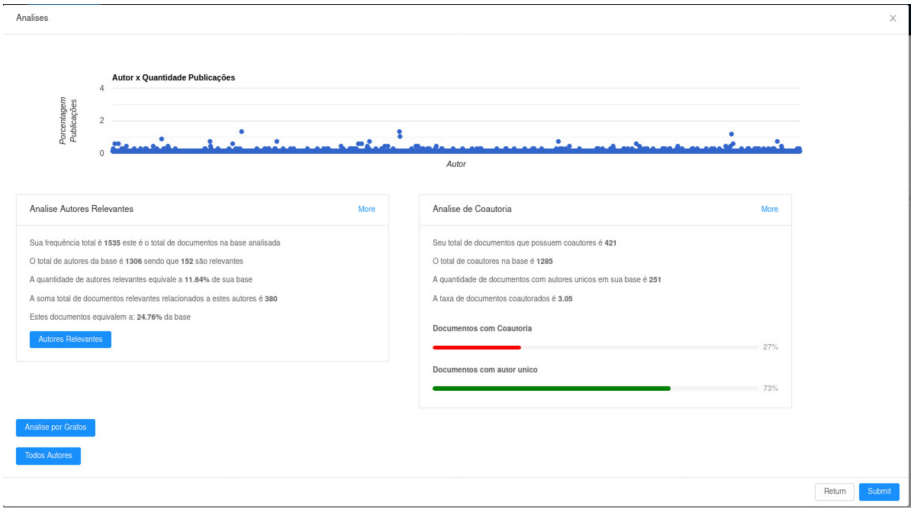

Fig. 4. Analysis of relevant authors and analysis of co-authorship

The analysis of relevant authors shows that the total number of documents analyzed in the database was 1535. The total number of authors in the database is 1306, of which

152 are relevant. The number of relevant authors is equivalent to 11.64% of the total number of documents within the database. The total sum of relevant documents related to this total number of authors is 380, which is equivalent to 24.76% of the base documents. In addition, it is also possible to click on the "Relevant Authors" button to obtain a list of the names of the relevant authors.

The co-author analysis shows that the total number of documents with co-authors is 421. The total number of co-authors in the database is 1285. The number of documents with single authors is 251. The rate of co-authors is 3.05%. The "Journal Analysis" feature, which uses Bradford's law applied to the search results, with the above criteria, produces the data presented in Fig. 5 as a result.

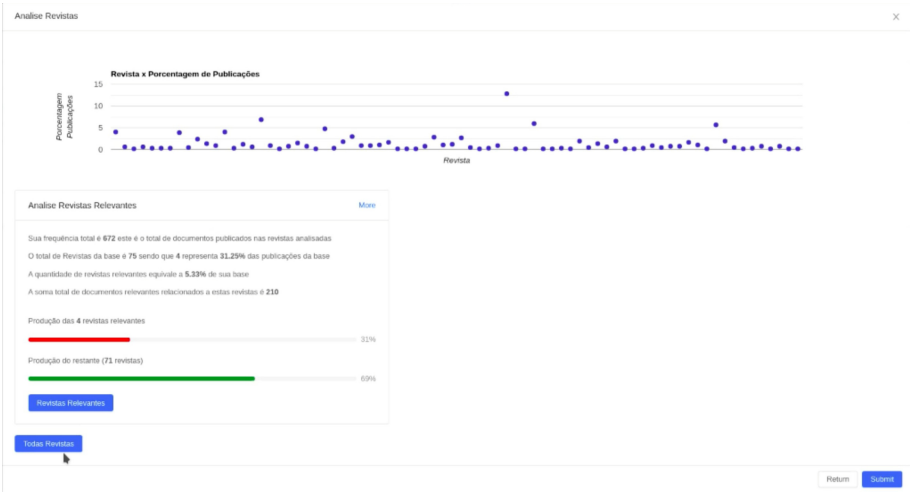

Fig. 5. Analysis of relevant journals

The analysis of relevant journals shows that the total number of documents published in the analyzed journals is 672. The total number of journals in the database is 75, with only 4 of them representing 31.25% of the publications in the database. The number of relevant journals is equivalent to 5.33% of the total base. Finally, the total sum of relevant documents related to these journals is 210.

In addition to the analysis of bibliometric laws, the system is also programmed to present the results in a graph form. This presentation, in addition to being more visually pleasing, also provides a range of indicators of co-authorship, co-relationship, proximity, intermediation.

The use of the proposed tool on the research undertaken in the Web of Science database allowed us to verify the speed and ease in obtaining the desired results. Consequently, we believe that its public availability on an open access portal will bring enormous benefits to researchers in the area of metric studies of information.

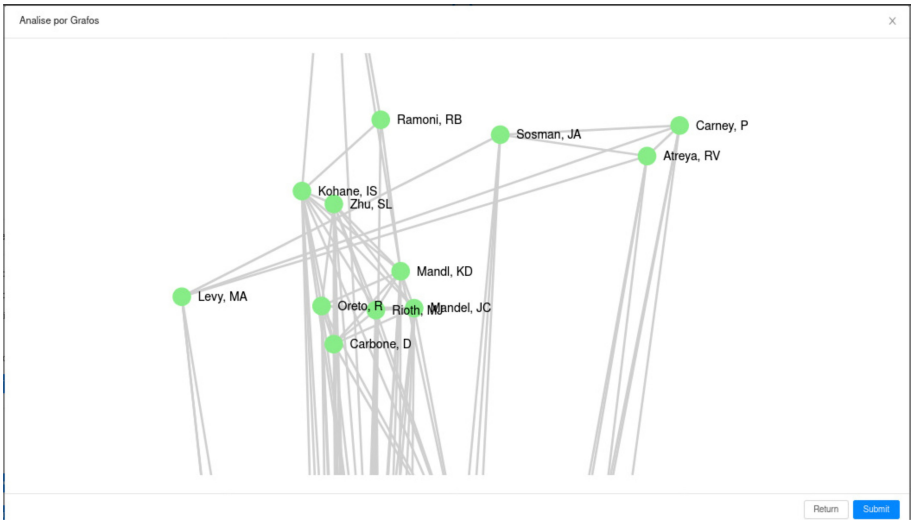

Fig. 6. Analysis of authors in graph format

6 Final Remarks

The proposed software application is currently still in beta. Our expectation is that a tested and stable version will be released in the first months of 2021. New features are being added, such as the analysis of Price's elite authority (in analysis of citation), Platz's theory (for the impact of journals), in addition to processing simpler indexes, such as co-authorship, works by authorship, and citations by authorship.

The following is a non-exhaustive list of new features to be added to the initial proposal:

a) The "Graph Analysis" report will show the power of influence of each author (analysis of force), based on changes in the diameters of the nodes.
b) The "Analysis by Graphs" report will allow the selection of authors to be analyzed, which will allow a faster result to be obtained.
c) The option to select authors will also be offered in the main menu of the tool.
d) The "Graph Analysis" report will allow a click on the author's name to display his/her affiliation.
e) From the identification of affiliation, the analysis of authors will show which universities cooperate and collaborate with each other. In addition, it will be possible to know which authors interact with which universities.
f) A feature will be implemented that will allow the results of the analyzes to be sent by e-mail. Some possible export formats that will be used: PDF, CSV, or LaTeX.
g) The user will be offered the option to define filters by year or other time intervals, so that it is possible to analyze the number of publications per year, for each author.
h) Analysis of citations, citations of the most cited authors, and the most cited journals will be made available.
i) The user will be able to save his analyzes, in order to be able to access them later.

References

1. Porter, A.L.; Cunningham, S.W.: Tech Mining: Exploiting New Technologies for Competitive Advantage. Wiley-Interscience (2004)
2. Bailón-Moreno, R.: Ingeniería del conocimiento y vigilancia tecnológica aplicada a la investigación en el campo de los tensio activos: desarrollo de un modelo cienciométrico unificado. Doctoral Tesis. Universidad de Granada, Granada (2003)
3. PNNL. IN-SPIRETM: Visual document analysis. U.S. Department Energy (2019). https://in-spire.pnnl.gov/index.stm. Accessed 07 Aug 2020
4. Clarivate Analytics. InCities Help (2020). http://help.incites.clarivate.com/inCites2Live/indicatorsGroup/aboutHandbook.html. Accessed 07 Aug 2020
5. Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F.: SciMAT: a new science mapping analysis software tool. J. Am. Soc. Inf. Sci. Technol. **63**(8), 1609–1630 (2012). https://doi.org/10.1002/asi.22688. Accessed 07 Aug 2020
6. Chen, C.: CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J. Am. Soc. Inf. Sci. Technol. **57**(3), 359–377 (2006)
7. Perssonl. O., Danell, R., Schneider, J.W.: How to use Bibexcel for various types of bibliometric analysis. ISSI Newslett. **5**(1), 5–24 (2009). https://homepage.univie.ac.at/juan.gorraiz/bibexcel/ollepersson60.pdf. Accessed 07 Aug 2020
8. ELSEVIER. SciVal (2020). https://www.scival.com/landing. Accessed 07 Aug 2020
9. Linnemeier, M.: Science of Science (Sci2) tool manual (2014). http://sci2.wiki.cns.iu.edu. Accessed 07 Aug 2020
10. Harzing, A.W.: The Publish or Perish Book. Tarma Software Research Pty Ltd, Melbourne (2010)
11. Börner, K., Chen, C., Boyack, K.: Visualizing knowledge domains. Ann. Rev. Inf. Sci. Technol. **37**(1), 179–255 (2003)
12. van Eck, N.J., Waltman, L.: VOS: a new method for visualizing similarities between objects. In: Decker, R., Lenz, Hans -J. (eds.) Advances in Data Analysis. SCDAKO, pp. 299–306. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-70981-7_34
13. Singh, N.: Complementing Bibliometrics with Network Visualization to Support Scientific Spheres. IFLA WLIC, Atenas (2019)
14. Lotka, A.J.: The frequency distribution of scientific productivity. J. Washington Acad. Sci. **16**(12), 317–323 (1926)
15. Bradford, S.C.: Sources of information on specific subjects. Eng. Illustrated Weekly J. **137**, 85–86 (1934)
16. Zipf, G.K.: Human Behaviour and The Principle of Least Effort. Addison-Wesley Press, Boston (1949)
17. Price, D.J.S.: Networks of scientific papers. Science **149**(july), 510–515 (1965)
18. Platz, A.: Psychology of the scientist: XI Lotka's law and research visibility. Psychol. Rep. **16**(2), 566–568 (1965). https://doi.org/10.2466/pr0.1965.16.2.566. Accessed 07 Aug 2020