ISSN 0005-1055, Automatic Documentation and Mathematical Linguistics, 2011, Vol. 45, No. 1, pp. 8–14. © Allerton Press, Inc., 2011.
Original Russian Text © V.G. Shamaev, 2011, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 2, 2011, No. 1, pp. 24–30.

Studies of the Document Flow for the Physical, Mathematical and Several Other Sciences, as Reflected in the VINITI Abstract Journal of the Russian Academy of Sciences

V. G. Shamaev

Abstract—We analyzed the frequency distribution of the documentary sources of serial form for arrays in physics, mathematics, and astronomy from the VINITI database. Under certain restrictions, we consider Bradford's law and its application for the selection of sources in the formation of lists of refereed publications or to obtain an adequate view of the state of any field of science.

Keywords: abstract journal, informational databases, Bradford's rule, Zipf's rule.

DOI: 10.3103/S0005105511010122

INTRODUCTION

The empirical Titius–Bode rule, by which the planetary distances in our solar system are expressed in geometric progression, was formulated nearly 250 years ago. This rule has played a prominent role in the history of the discovery of minor planets and asteroids. It also helped the young Gauss, who was working on a method for finding the lost minor planet Ceres and a method for determining an elliptical orbit based on three observations. In the 20th century, Zipf's rule in linguistics and the scattering law of scientific publications of Bradford in computer science were written using simple mathematical expressions, which also stimulated numerous studies.

There is something in mathematics that fascinates people. For example, the amazing number π expresses the ratio of the circumference to the length of the diameter of a circle. The history of the number π goes in parallel with the development of the whole of mathematics. In its decimal places people have tried to find different rules, but the order the numbers follow is random. Fans of the TV show Lost have observed that the famous set of numbers from the series (4, 8, 15, 16, 23, 42) in this sequence is not found among the first 200 million numbers after the decimal point. Their total number, 108, is very important in the mythology of Buddhism and Hinduism, as is, for example, the Pythagorean formula; the expression "Pythagorean Pants" is known by every student.

In the Russian Academy of Sciences, as well as in many well-known academic and scientific institutions, there is a continuous stream of letters from people who have found the answer to the mystery of the world around us and of the universe as a whole, based on an analysis of the sequence of digits with the same number of patterns or that found in a numerical sequence, by which it is possible to explain any natural phenomenon. If one successfully applies these rules and they can make predictions and look into the future then such a rule will be called a fundamental law and thus remain until an epiphany comes when the "prediction" does not come true.

There are several thousands of far-fetched "laws" but for some we must make an exception, because they have made a great contribution to science. Such laws include the astronomical law, or rather the Titius–Bode rule. Johann Titius, a German astronomer and mathematician had no doubt that there is harmony in the solar system, which testifies to its divine origin. In 1766, he revealed a simple pattern in the increases of the mean radii of the orbits of the planets in our solar system, which can be expressed by formula

$$r = 0.4 + 0.3 \times 2^n,$$

where r is the average radius of a planet's orbit in astronomical units (AE) and n takes the values: $-\infty$, 0, 1, 2, 3, 4, 5 for all the then known planets, viz., Mercury, Venus, Earth, Mars, Jupiter, and Saturn. However, in the place where there should have been a planet with $n = 3$, there was a gap. i.e., the relative positions of celestial bodies changes, following this rule. Perhaps Titius' discovery would have remained unnoticed by his contemporaries, but in 1772 the German astronomer Johann Bode according to some accounts became acquainted with Titius' rule and according to others rediscovered it. He promoted this rule, which gradually came to be called the "Titius–Bode rule." In 1781 William Herschel discovered the planet Uranus, whose distance from the Sun differed little from the formula given by Titius (Fig. 1).

Bode took practical steps to find a planet with $n = 3$ and collected data from an informal community of 24 astronomers. However, a planet (or rather a "small planet," as we now call them) was discovered by the Italian astronomer Giuseppe Piazzi, who was not included in the "consortium" in 1801. It was named

Ceres, but soon at the same distance three small planets, Pallas, Juno and West were found and this gave rise to the hypothesis that the fifth planet was ruptured into pieces, which form an asteroid belt at a distance of 2.8 AE between the orbits of Mars and Jupiter. The Titius–Bode rule is not a law similar to, for example, the laws of planetary motion of Kepler's or Newton's three laws of mechanics, namely, a rule, or rather a relationship that was derived from analysis of data that was available at the time about the distances of the planets from the Sun. The Titius–Bode rule has no known theoretical justification. The most likely explanation, other than the assumption of a coincidence, is that at the stage of the formation of the solar system gravitational perturbation caused protoplanets, which formed a regular structure of alternating areas that might or might not exist as a stable orbit.

Such laws (rules) are also found in other sciences, e.g., in physics we have the planetary model of the atom by Rutherford and the inexplicable, without recourse to anthropic principles, values of fundamental constants; in history we have the theory of Academician Fomenko, in linguistics Zipf's rule occurs, and in informatics, Bradford's rule.

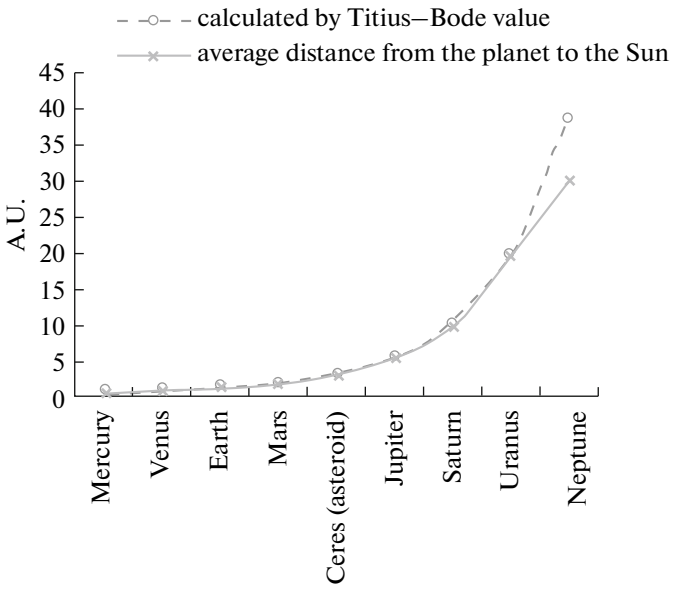

Fig. 1. The positions of planets in the solar system, calculated according to the Titius–Bode law and derived from astronomical observations.

Bradford's Scattering Law of Scientific Publications

We now turn to Bradford's rule. It is formulated as follows: journals (periodicals) for a specific area of science can be divided into three zones, each of which contains about one third of the articles of their total number in all the journals. The first zone is the core (the "nuclear" journals), i.e., the major journals in the area under consideration, the second area is the "profile" journals, containing many articles on the area under consideration, and the third zone is journals in which articles on the treated area may be found. The number of journals in the areas of concern was $1 : x : x^2$. Bradford formulated his rule in 1934 after studying a bibliography of geophysics of 326 journals relevant to the topic. He observed that the first zone A of 9 journals contained 429 articles, the second zone B of 59 journals had 499 articles, and for the third zone C in the remaining 258 journals 404 articles were found, i.e., x_A of the first zone was 6.6, while x_B was 4.4. It is easy to observe that this rule is a little strained even for the most Bradford cases, but as a working hypothesis, perhaps, it may be used.

If we strictly follow Bradford's rule, one can write: $nx^2 + nx + n = N$, where N is the total number of sources and n is the number of sources in the first zone.

This equation is of degree 2, its discriminant under our conditions is obviously <0, and therefore the solution we obtain has two real roots. One of them is negative and does not meet the model, and the other is

$$x_{\text{creat}} = -\frac{1}{2} + \sqrt{\frac{N}{n} - \frac{3}{4}}.$$

This gives us the theoretical value of the Bradford factor and its empirical value we can obtain from consideration of VINITI database.

The VINITI database is one of Russia's largest databases on natural and engineering sciences [1]. It includes materials from the VINITI Abstract Journal (AJ) on various thematic fragments from the early 1980's. (More accurately, see Table. 1). The database includes 29 thematic fragments, which are separated into more than 230 editions, i.e., there can be some degree of sampling on this database to perform various infometric research on different areas of science and technology.[1]

But first, we consider the filling of VINITI Abstract Journal for some thematic fragments from the time of its foundation, i.e., since 1953 [2]. Figures 2 and 3 show the cumulative number of documents with the increase over the years in various fields of science and technology, as reflected in the AJ during the entire period of its existence. Each year the folding graphs of the filling of the database given in [3] were made, the sharp decline of filling in 1992 and further gradual degradation in many thematic fragments (geophysics, mechanics, mechanical engineering, avionics, and others) are clearly seen, but the cumulative graphs (Figs. 2, 3) more clearly show the perspective of the progressive filling of thematic fragments highlighted here. They clearly show what happened in the way of a decline in documents over the past 20 years and clearly

[1] Infometriya means a discipline whose subject is the quantitative measurement of stored and used information.

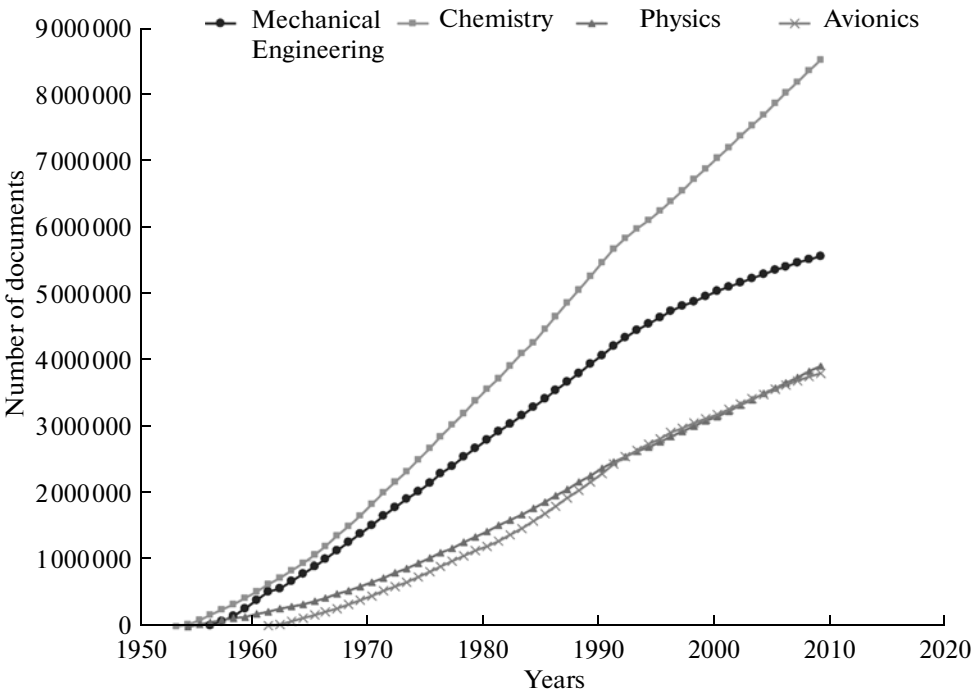

Fig. 2. The cumulative number of documents reflected in AJ issues with a large number of articles.

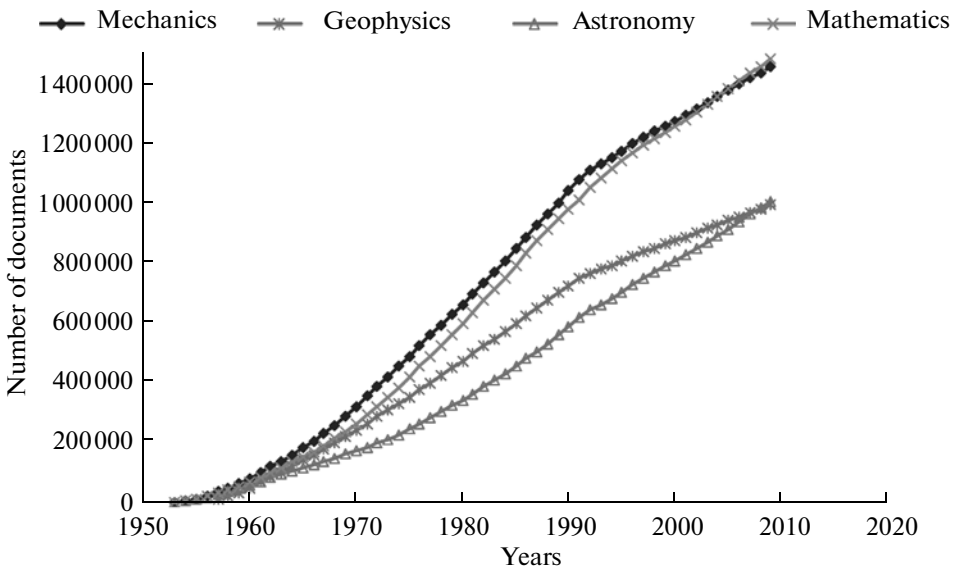

Fig. 3. The cumulative number of documents reflected in AJ issues with an average number of articles.

show the actual number and what the sum total of normal development would be.

On each curve there is some recession in 1992, but for some thematic fragments it has a minimal value, while for others it is catastrophic. For example, engineering reduced its annual size from the late 1980s to 2009 by almost three times, automation and electronics engineers (avionics), geophysics, mechanics by more than two times, while astronomy, by 1.3 times. This is a depressing statement for the first part of the list, especially when extrapolated to the date of the cumulative curves based on the period from 1953, the early registration of the VINITI Abstract Journal, until 1991.

Now we will analyze the array of databases in physics for the years 1986–2010 with division into 5-year

Table 1. Thematic fragments of the VINITI database

Title of the thematic fragment	Beginning year of generation	Title of the thematic fragment	Beginning year of generation
Automation and electronics	1981	Medicine	1998
Astronomy	1989	Metallurgy	1981
Biology	1981	Mechanics	1985
Computer Sciences	1997	Ensuring safety during an emergency	1987
Genetics	1981	Environmental protection	1981
Geography	1991	Welding	1981
Geology	1985	Transportation	1984
Geophysics	1986	Physics	1983
Mining	1981	Physics–Chemical Biology and Biotechnology	1981
Publishing & Printing	1985	Chemistry	1981
Informatics	1982	Industrial Economics	1985
Corrosion and Corrosion Protection	1981	Energy Economy	2002
Medicinal Plants	1991	Electrical engineering	1981
Mathematics	1997	Energy	1981
Engineering	1981		

periods and the comparable arrays of the astronomy and mathematics databases (Table 2).

We apply Bradford's rule to 5-year-old samples from the database of VINITI Physics for the years 1986–2010. Each sample has a significant number of sources (from 2500 to 5000) and documents (from 300000 to 420000), incorporating fairly strictly selected subjects supported by the rubricator. Sources are located in descending of number placed in the document databases (ranked); we divided them into three equal parts by the number of incoming documents and found the numerical data for each of the three Bradford areas (Table 3).

We see that the coincidence of x_{emp} and x_{theor} is absent, suggesting that practical data sets are poorly approximated by a quadratic function; here as a third member we must use x^3.

We have previously noted [3] that at the turn of 1991–1992 there was a dramatic change in the situation in the country and in VINITI. Therefore, unfortunately, as the spectrum of sources and the total number of documents and filling of the individual sections underwent significant changes, which probably is reflected both in the number of sources in the zones and changes in the parameter x. If we calculate the area C as obtained with x_{emp} from the first two zones, denoting it as C_{calc}, the total number of sources needed to cover a significant portion of the thematic fragments becomes first, fairly compact, and second, includes almost all the articles on selected topics (at a level of more than 90% of the original, Table. 3) and, third, is measurable in actual control of completeness[2].

Under this condition on the considered material, with the previously mentioned restriction, Bradford's rule works. As in the case of the Titius–Bode law we

[2] The last point requires explanation. This lies in the fact that today the completeness of the reflection of a source often is not checked in an AJ, and if it is not verified, then missed articles in the future may be placed with great difficulty. At the same time they are placed, naturally, not with all the basic array of articles, but, at best, in the next issue of the AJ. If this occurs and the database is not corrected then the print or electronic edition is not correct.

Table 2. Filling of the Physics (1986–2010), Astronomy (1996–2010), and Mathematics databases (1997–2010)

Years	Number of documents	Number of sources
Physics		
1986–1990	422585	4277
1991–1995	364679	4721
1996–2000	376555	4641
2001–2005	310609	2973
2006–2010	357159	2631
Astronomy		
1996–2010	254828	4435
Mathematics		
1997–2010	257960	3799

Table 3. Bradford zone parameters for the Physics database (1986–2010)

Years	Zone A	Zone B	Zone C	x_{emp}	x_{theor}	Zone C_{calc}	Number of sources in zones A, B and C_{calc}	Number of documents in zones A, B and C_{calc}	% from total count
1986–1990	37	115	4125	3.11	10.22	358	510	381535	90.3
1991–1995	31	141	4549	4.55	11.81	642	814	328818	90.2
1996–2000	20	112	4509	5.60	14.71	627	759	361734	96.1
2001–2005	15	77	2881	5.13	13.55	395	487	290651	93.6
2006–2010	17	70	2544	4.12	11.91	288	375	329233	92.2

Table 4. Bradford zones parameters for the Astronomy (1996–2010) and Mathematics (1997–2010) databases

Years	Zone A	Zone B	Zone C	x_{emp}	x_{theor}	Zone C_{calc}	Number of sources in zones A, B and C_{calc}	Number of documents in zones A, B and C_{calc}	% from total count
Astronomy									
1996–2010	7	49	4379	7.00	24.66	343	399	228378	89.6
Mathematics									
1997–2010	62	205	3532	3.31	7.28	679	946	238076	92.3

now check this rule on other thematic fragments included in the VINITI database, and thus confirm, or find the unacceptability of this model for other subjects. Consider the data presented in Table 4.

The data presented in Table 4 also indicate the accuracy of Bradford's rule for the above conditions. Here it is important to understand that this rule gives an empirical basis for limiting the number of sources that should be handled in order to obtain sufficiently complete retrospective reviews, as well as the current state of any field of science or technology without fear of missing anything substantial. However, if there is reliable statistical data then the level at which we should put a restriction on the number of sources used is clear.

Now we consider the cumulative thematic fragments of astronomy and mathematics, generated not by year, but by rank, i.e., the summation is performed over the frequency of occurrence of sources in the ordered array of sources for the period. The ordinate shows their cumulative growth and on the horizontal axis the source of the rank on a logarithmic scale is shown in order to obtain a visual representation.

The graphs show details that are difficult to observe in the tables. First, the first two curves have a concave form up to a certain value $\ln R$, and, consequently, to a certain rank of the source. Following this is a straight section and then a convex one. Second, the curves begin with different values on the ordinate, indicating different zero-points. Third, they increase differently. The first part of the curve describes the different values of $\ln R$ for zone A in astronomy and mathematics, which follows from Table 4, viz., 7 and 62 sources, respectively. The steeper middle part of the curve for mathematics is characterized by a rapid increase in the volume of documents, i.e., it shows a slower decline in the number of documents depending on rank at this site. The final part of the curve, becomes nearly parallel to the abscissa and characterizes the convergence of the frequency range; the large bulge in the curve for mathematics at the right end of the curve indicates a greater loss of documents compared with their possible (extrapolated) number. Thus, the graphical representation gives us a quick qualitative assessment of the appropriate resources.

The ability to perform the above calculations and the construction of various graphs occurred only with the beginning of the formation of the databases; the results of processing large volumes of data and serious analysis were representative of not only the advent of computers, but also the good filling of databases for many years. Here again, we cannot say much about the condition of the VINITI database earlier than the last 15 years. The filling of the entire database, including the physics thematic fragment for the years 1983–1995, can be criticized. The name of the same journal is found in dozens of variations and unfilled fields,

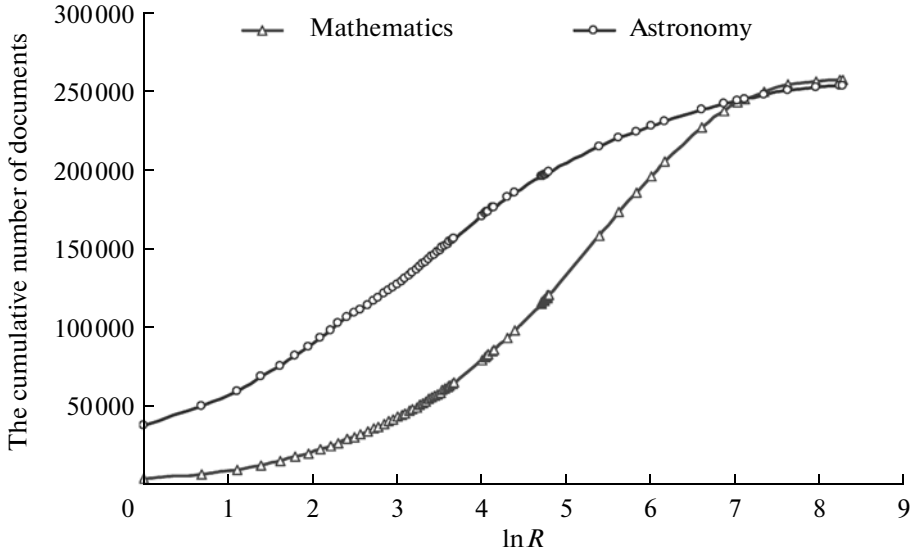

Fig. 4. The cumulative number of documents of type 1 (article in the serial edition) reflected in AJ issues *Astronomy* (1996–2010) and *Mathematics* (1997–2010) depending on rank *R*.

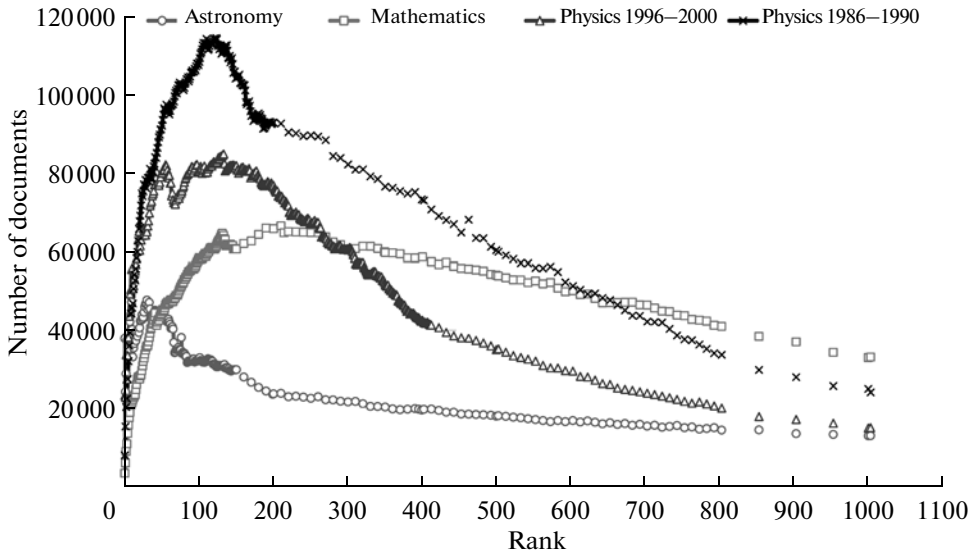

Fig. 5. Product of source rank on the frequency of its occurrence by thematic fragments Astronomy (1996–2010), Mathematics (1997–2010) and two 5-years intervals of Physics (1986–1990) and (1996–2000).

entanglement of fields, etc occur. The treatment of the 5-year sets from 1986 to 1990 and 1991 to 1995 in physics took us about 20 days, as required for our purposes, reliable information on sources and the number of documents in them. Isn't it time to once and for all clean the VINITI database of the 1980s–1990s, are branch departments ready to render all possible help? By the way, we have already discussed this issue [4], but it is time to do something.

Zipf's Law

In the late 1940's J. Zipf, using a vast array of texts, showed that the distribution of words is subject to a simple law, which stated as follows. If a text is large enough to calculate the frequency of occurring words *F* and we arrange them in order of their frequency of occurrence *R*, i.e., that encountered for the words of the text, for every word the product of its rank (ordinal number) and the frequency of its occurrence in the text has approximately the same value for every word of this text. This can be expressed using the formula $F \times R = C$, where *F* is the frequency of occurrence of words in the text; *R* is the rank (number) of words in the list, and *C* is an empirical and supposedly constant value. The resulting dependence is graphically expressed as a hyperbola. When we replace one text with others, including those in other languages or

other subjects, the general nature of the distribution does not change.

J. Zipf and other researchers found that the distribution is subject to not only natural languages, but also other biological phenomena and social cultural measures, such as the distribution of scientists by the number of their published articles[3].

Some authors have stated that Zipf's law can be used in the preparation of thesauri on various topics. We do not understand why you need this law, because in the preparation of these distributions the frequency of occurrence alone is important; for infrequent, but necessary cases, the thesaurus terms rather the opinions of experts. However, we could not resist checking Zipf's law using our files of documents (Fig. 5).

The distribution we received is such that it is better to talk about weak subordination to Zipf's law, only for some thematic areas, rather than following it.

CONCLUSIONS

Johann Titius and Johann Bode certainly believed in a mathematical pattern in the structure of the solar system and many other scientists continue to engage in this kind of search for patterns in different areas of science. The problem is that no one went further to find the real attraction and they did not try to find a physical cause for why the orbits of nearby planets are subject to their laws. Without physical justification "laws" and "rules" of this kind are pure numerology; in this they differ from the laws of classical mechanics of Isaac Newton, the laws of celestial mechanics of J. Kepler, or the well-known laws of physics for gasses by Charles, Gay-Lussac, and Boyle. It is true that a satisfactory theory of the formation of the solar system does not exist, despite the already large number of open extra-solar planetary systems and near-stellar disks, viz., flattened gas–dust clouds around young stars. Similarly, the laws of Bradford, Zipf, and Lotka are sometimes accurate and sometimes violated; these rules may be violated, or rather act only within certain limits with no generalizations.

[3] A. Lotka in 1926, while exploring the frequency of publications of authors of scientific articles in different areas, showed that this dependence can be represented by $X^n Y = C$, where X is the number of publications of an author whose rank is equal to Y, where C and n are constants that depend on the field of knowledge.

ACKNOWLEDGMENTS

We thank A.N. Sedyakina for help in the editing of this article.

REFERENCES

1. Arskii, Yu.M. et al., *Bank dannyh VINITI. Sostoyanie i perspectivy razvitiya* (Databank of VINITI. States and Perspectives of Development), Moscow: VINITI, 2006.
2. Cherny, A.I., *Vserossiyskiy institut nauchnoy i tehnicheskoy informacii* (Russian Institute for Scientific and Technical Information), Moscow: VINITI, 2005.
3. Shamaev, V.G. and Shamaev, N.V., Processing of the Documentary Flow in Branch Departments of Scientific Information VINITI of RAS and its Interpretation in the Physical Representations, *Nauchnaya i tecknicheskaya informaciya*, 2010, vol. 2, no. 8, pp. 19–27.
4. Shamaev, V.G., Summary Tome "Physics" of Abstract Journal of VINITI of RAS: The Problems of Existence and Development, *Nauchnaya i tecknicheskaya informaciya*, 2010, vol. 1, no. 10, pp. 21–27.