Postprint: Research on Development Patterns of Big Data Application Fields Based on Bibliometric Laws

Authors: Zhang Jiaojiao, Liu Yun, Cheng Yijie

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose/Significance] To investigate the research status and development trends in the field of big data applications, and to reveal its developmental patterns. [Method/Process] A data retrieval strategy for big data application technologies was formulated, and literature research data from 1990-2015 was selected. Utilizing bibliometric software VP (Vantage Point), statistical analysis software SPSS, and Microsoft Excel as analytical tools, the study examined and analyzed from three dimensions—publication volume distribution, journal distribution, and author distribution—whether the literature development in this field conforms to Price' s law of scientific literature growth, Bradford' s law, and Lotka' s law. [Results/Conclusion] Papers related to the field of big data applications, beginning in 1990, experienced a period of steady development before demonstrating explosive growth starting in 2012. The literature development conforms to Price' s law of scientific literature growth. The journal distribution of the literature sample basically conforms to Bradford' s law, and a core journal cluster has formed, including BMC BIOINFORMATICS, SENSORS, among others. The author distribution in this field deviates significantly from Lotka' s law, and a core author group with substantial influence has not yet formed.

Full Text

Research on Developing Regulations of Big Data Application Technology Based on Bibliometrics Laws

Zhang Jiaojiao, Liu Yun, Cheng Yijie
School of Management and Economics, Beijing Institute of Technology, Beijing 100081

Abstract

[**Purpose/Significance**] This study investigates the research status and development trends of the big data application field, revealing its underlying developmental patterns.
[**Method/Process**] A data retrieval strategy for big data application technology was developed, selecting relevant literature from 1990–2015. Using bibliometric software Vantage Point (VP), statistical analysis software SPSS, and Microsoft Excel, the study examined whether the field' s literature development conforms to Price' s law of scientific literature growth, Bradford' s law, and Lotka' s law from three perspectives: publication volume distribution, journal distribution, and author distribution.
[**Result/Conclusion**] Big data application-related papers began in 1990 and, after a period of steady development, experienced explosive growth starting in 2012. The literature development conforms to Price' s law of scientific literature growth. The journal distribution of the literature sample basically follows Bradford' s law, and a core journal group has formed, including *BMC Bioinformatics*, *Sensors*, among others. The author distribution in this field differs significantly from Lotka' s law, indicating that a highly influential core author group has not yet formed.

Keywords: Big data application; Price' s law of scientific literature growth; Bradford' s law; Lotka' s law

1 Introduction

In recent years, "big data" has attracted significant attention from governments, industries, and academia worldwide. Globally, leveraging big data to promote economic development, improve social governance, and enhance government services and regulatory capabilities is becoming a major trend. Developed countries have successively formulated and implemented strategic documents to vigorously promote big data development and application [**?**]. Initiatives such as the United Nations' "Global Pulse" program [**?**], the U.S. "Big Data" strategy [**?**], Japan' s "ICT Comprehensive Strategy for 2020" [**?**], and South Korea' s Big Data Center strategy [**?**] have all launched the big data strategic agenda. Numerous multinational corporations have also entered the big data research and development field, including traditional data analytics companies like Teradata, SAP, and SPSS, as well as big data resource enterprises such as Google and Facebook [**?**]. In 2008, *Nature* magazine published a special issue on "Big Data," systematically introducing the potential value and challenges of big data based on the actual research status across multiple disciplines [**?**]. In 2011, *Science* magazine published the special issue "Dealing with Data," marking the formal entry of "big data" onto the scientific research stage and its emergence as a hot topic across various disciplines [**?**]. The era of large-scale data production, sharing, and application is approaching [**?**].

Big data application (BDA) refers to the process of using big data thinking

and methodologies to leverage big data analytics results, providing auxiliary decision-making for users and uncovering potential value [**?**]. An increasing number of problems can be solved through big data applications. Its utility extends beyond science and technology to public management, basic and applied research, and business domains, where big data can introduce new concepts, thinking patterns, and problem-solving methods or perspectives [**?**]. Gao Xiaoping [**?**], Li Huan [**?**], and Gao Xia [**?**] respectively applied big data thinking and methods to national governance reform and innovation, technology management innovation platform construction, and technology evaluation approaches. From the perspective of basic and applied research, Li Zhenhao [**?**] and Huang Shaofang [**?**] applied big data to quality control of traditional Chinese medicine and informatization of geological data archives, respectively. N. O. E. Olsson and H. Bullberg [**?**] incorporated big data applications in project evaluation processes, S. R. Sukumar and R. Natarajan [**?**] applied big data methods to healthcare research, and M. C. Ebach and M. S. Michael [**?**] examined the relationship between historical science and big data. In the business domain, T. H. Davenport [**?**], J. Frizzo-Barker [**?**], and C. F. Hofacker [**?**] conducted series of studies on international business decision-making, business scholarships, and consumer behavior using big data. Evidently, "big data" has been widely applied across many fields and has essentially transformed into a new way of thinking and problem-solving approach [**?**].

Big data applications yield multifaceted benefits, optimizing government efficiency, management decision-making, market supervision, public services, urban infrastructure, and social security, while also generating substantial benefits for the economy, education, culture, health, and diplomacy. China ranks first globally in internet and mobile internet user scale, possessing abundant data resources and application market advantages. Key technologies for big data have achieved breakthroughs, with numerous internet innovation enterprises and applications emerging, and some local governments have already launched big data initiatives [**?**]. Fully leveraging China's data scale advantages to achieve synchronous improvement in data scale, quality, and application levels, and to explore and release the potential value of data resources, will facilitate better utilization of the strategic role of data resources. Therefore, research on big data applications in science and technology, public management, basic and applied research, business domains, as well as in economic, educational, cultural, health, and diplomatic fields, holds significant research value [**?**].

This study focuses on big data application, exploring the developmental patterns of literature in this field and verifying relevant developmental laws, with the aim of systematically revealing the technological characteristics, patterns, and trends of BDA to provide references for subsequent research.

Fund Project: This paper is a research outcome of the National Natural Science Foundation of China projects "Research on Influencing Factors and Dynamic Mechanisms of Patent Application Growth in China" (Project No.: 71273030) and "Research on Policy Coordination Mechanisms/Process Mod-

els and Effect Evaluation of National Innovation System Internationalization" (Project No.: 71573017).

Authors: Zhang Jiaojiao (ORCID: 0000-0001-8281-710X), Master' s student, E-mail: zhangjjcx@163.com; Liu Yun (ORCID: 0000-0002-2888-8932), Professor, Doctoral supervisor; Cheng Yijie (ORCID: 0000-0002-9365-055X), Doctoral student.

Received Date: 2016-08-01
Published Date: 2016-10-27

2 Research Methods and Data Sources

Opinions vary regarding the origin of the big data concept. Some believe it was first proposed by the globally renowned consulting firm McKinsey, others attribute it to J. R. Mashey, Chief Scientist at SGI, who published an article titled "Big Data and the Next Wave of Infrastres" at a USENIX conference in 1998 [**?**], while still others credit B. Inmon, the "father of data warehousing," from the 1990s. Based on the Web of Science database, this study formulated a relevant data retrieval strategy, treating all papers involving the BDA concept or using big data thinking and methods to solve problems as subjects for analysis. The retrieved data spans from 1990 to 2015, totaling 1,701 items. Web of Science is a comprehensive literature retrieval tool with data sourced from academic journals, monographs, conference proceedings, patent literature, and scientific books from over 40 countries and regions, covering disciplines including biology, agriculture, medicine, chemistry, physics, earth science, and life science. It is the most comprehensive database internationally for collecting scientific paper citations. This study selected the Science Citation Index (SCI) as the basic data source, formulated corresponding retrieval strategies, and conducted further research and analysis.

Since their inception, scientific literature has grown over time. Price identified growth patterns through his examination of the accumulation process of scientific literature, a finding collected in his representative work *Science Since Babylon*, which profoundly influenced subsequent scientific literature research [**?**]. Price calculated the growth rate of abstract journals, plotting scientific literature volume on the vertical axis and historical years on the horizontal axis, connecting data points from different eras with a smooth curve that approximately characterizes the growth pattern of scientific literature over time. This is the famous Price curve, with the mathematical expression:

$$F(t) = ae^{bt}$$

where $F(t)$ represents the volume of scientific literature, a is the literature volume at the initial statistical moment (when $t = 0$), e is the base of natural logarithms, and b is the continuous growth rate of journals, a time constant.

Bradford' s law and Lotka' s law are both important laws of bibliometrics, together with Zipf' s law forming the three major bibliometric laws. Bradford' s law can be expressed as: if scientific journals are arranged in descending order according to the number of papers published on a specific discipline, journals can be divided into three zones specifically oriented toward that discipline: a core zone, a related zone, and a non-related zone. Each zone contains an equal number of articles, with the number of journals in the core, related, and non-related zones following a relationship. Bradford' s law was created based on the dispersion of scientific papers in journals but can be extended to many different applications, providing guidance for identifying core journals, formulating literature procurement strategies, optimizing collections, examining work performance, understanding reader preferences, and retrieving and utilizing literature [**?**].

Lotka' s law was discovered by American Alfred Lotka in 1926. It is considered the first law to reveal the relationship between author frequency and literature volume, describing the frequency distribution pattern of scientific productivity. The generalized Lotka' s law can be expressed by the following formula:

$$f(x) = \frac{c}{x^n}$$

where $f(x)$ represents the number of authors (or author frequency) who have published x papers in a specific discipline or subject area during a certain period; c and n are two constants greater than zero [**?**]. Lotka's law is generally described as: the number of authors publishing two papers is approximately 1/4 of those publishing one paper; authors publishing three papers are approximately 1/9 of those publishing one paper; authors publishing N papers are approximately $1/n^2$ of those publishing one paper; and authors publishing only one paper account for about 60% of all authors (i.e., taking $c = 1$, $n = 2$). This study adopts this method for analysis.

When employing these three bibliometric methods, this study also utilized bibliometric software VP (Vantage Point), statistical analysis software SPSS, and Microsoft Excel to conduct statistics, analysis, and verification of the obtained data at the levels of publications, journals, and authors.

2.1 Publication Volume Analysis

The number of published papers reflects, to a certain extent, the research level and development trends of a given field during a specific time period. The distribution of literature quantity across variables such as time, region, and type constitutes one of the basic characteristics of literature samples and represents one of the most fundamental and straightforward analytical items [**?**]. This study examined changes in literature statistics for the BDA field from 1990 to 2015, as shown in Figure 1 [Figure 1: see original paper].

The earliest recorded literature dates to 1990, a paper titled *Novel Applications of Halogenation Reactions in Atomic Spectrometry* by Hungarian scientist T. Kantor from the Budapest University of Technology and Economics, who applied big data to research on halogenation reactions in atomic spectrometry, representing an excellent early example of big data application in scientific research. Based on the obtained two-dimensional time-literature volume data table, we can further derive trends in publication volume over time and changes in growth rates (see Figure 1). The analysis reveals three distinct phases: (1) From the first paper in 1990 through 1997, literature volume remained at approximately 10 papers annually, indicating that big data application technology research was in its initial development stage; (2) From 1998 to 2011, literature volume increased compared to the previous stage, maintaining a relatively low growth rate except for occasional years with higher or negative growth, allowing us to define 1990–2011 as the preliminary stable development period for BDA; (3) From 2012 to 2015, literature volume increased dramatically, reaching 509 papers in 2015 alone, nearly equivalent to the total of 603 papers from the initial development stage (1990–2011). During these four years, the annual growth rate increased to approximately 50%, with the line chart showing a basically linear upward trend and strong growth momentum, reaching 92.02% in 2014 and an average annual growth rate of 66.27% over the four-year period, clearly marking this as the rapid development stage for BDA.

To more scientifically and reasonably grasp the developmental patterns of BDA technology literature, this study conducted a time-series statistical analysis of the field' s literature growth patterns. Using SPSS statistical analysis software, we performed curve fitting on the field' s literature growth to verify whether its development pattern conforms to Price' s law of scientific literature growth. Using time (year) as the independent variable and cumulative literature volume as the dependent variable, we conducted curve fitting on relevant statistical data in SPSS, including linear, quadratic, cubic, logarithmic, compound, growth, and exponential models. Based on the fitting results, linear and logarithmic models showed smaller correlation coefficients R^2 (0.714 and 0.440, respectively) and were excluded. Among the remaining models with larger correlation coefficients —quadratic, cubic, compound, growth, and exponential—this study selected the exponential fitting model according to the Price curve expression, yielding the following equation:

$$F(t) = 8.909 \times 0.205t$$

In this model, the initial value of t (year) is 1, with 1990 set as 1 and incrementing sequentially. The fitting model' s analysis results (model summary, ANOVA table, and coefficient table) are shown in Table 1 , and the curve fitting plot is shown in Figure 2 [Figure 2: see original paper].

Analysis of the exponential function model fitting results proceeds as follows:

(1) Model fit reflects the model' s explanatory power for the data. The larger

the adjusted R^2, the stronger the model' s explanatory capability. Table 1 shows that this model' s R^2 is 0.882, indicating good explanatory power.

(2) Variance analysis reflects the overall significance of the model. Generally, the model' s test p-value (Sig.) is compared to 0.05; if less than 0.05, the model is significant. Table 2 shows the model' s significance level is 0.000, less than 0.05, indicating the model is significant.

(3) Regression coefficients represent the coefficient values of each variable in the regression equation, with Sig. values indicating coefficient significance. Table 3 shows the constant term is 8.909, the year coefficient is 0.205, the t-value is 13.715, and significance is 0.000, demonstrating both results are highly significant.

(4) Figure 2 shows fitting situations for linear, quadratic, cubic, and exponential functions, with circles representing actual values. Clearly, the exponential fitting model demonstrates the best fit. Analysis of Tables 1-3 indicates this model has the highest R^2 value and smallest overall p-value, representing the best goodness-of-fit and significance.

Based on these results, we conclude that BDA literature volume is in a period of exponential growth—a rapid development stage. We can thus predict that related research in this field will continue to grow rapidly for some time, with literature volume showing exponential growth (within a certain period). Comprehensive analysis suggests that BDA literature development conforms to Price' s law of scientific literature growth.

2.2 Journal Analysis

British bibliometrician S. C. Bradford first proposed the concept of core journals. Although Bradford' s definition is determined entirely by the number of papers published in journals and thus has certain limitations—while modern core journal definitions consider additional factors such as usage rates (including citation rates, abstract rates, circulation rates) and academic impact—Bradford' s law of scatter remains the theoretical foundation for core journal evaluation [**?**]. This study employs Bradford's law to analyze and investigate the journal distribution patterns of BDA field articles.

Literature in the selected data source comes from 886 different journals, with *Future Generation Computer Systems—The International Journal of Grid Computing and eScience* publishing the most articles (30). Fourteen journals published 10 or more articles (including 10). Specific analysis is as follows:

According to Bradford's law regarding the proportional relationships among journal numbers in the core, related, and non-related zones, this study divided the 886 journals' publication volumes into three roughly equal zones. The first zone comprises approximately 6.4% of total journals (57 journals) publishing 30% of total literature; the second zone comprises approximately 21.3% of journals (189

journals) publishing about 30% of literature; and the third zone comprises approximately 72.2% of journals (640 journals) publishing about 40% of literature. This indicates that BDA literature exhibits a clear core-dense distribution. The distribution ratio of journals across the three zones is 57:189:640, which aligns with Bradford' s law pattern of $1 : n : n^2$, where n is approximately 3.3. Based on this data and analysis, we can determine that the journal distribution of our literature sample basically conforms to Bradford' s law. From only one journal publishing one article in 1990 to a cumulative total of 886 journals publishing 1,701 articles by 2015, the BDA field has formed extensive research coverage and its own core journal group, as shown in Tables 4 , 5 , and Figure 3 [Figure 3: see original paper].

2.3 Author Distribution Pattern

Authors constitute one of the important external characteristics of papers and are key to determining paper quality [**?**]. Core author evaluation requires comprehensive consideration of multiple factors and indicators. While Lotka' s law evaluates core authors in a specific field solely from the perspective of publication volume—ignoring the "quality" of authors' publications and contributions of different authors to the same paper—its application to depicting literature and author distribution in a field remains fair to a certain extent.

This study statistically analyzed actual author contributions in the BDA field from 1990–2015 and calculated the estimated number of authors according to Lotka' s law, comparing the relative error between the two. Details are shown in Figure 4 [Figure 4: see original paper].

Figure 4 reveals that among high-productivity author groups, the number of authors is too small. For instance, only one author published 12, 11, and 9 papers, respectively. According to Lotka' s law estimates, the number of authors should be far greater: 26, 31, and 47, respectively. This shows the field' s actual author distribution differs considerably from Lotka' s law estimates. However, 5,916 authors published only one paper, while Lotka' s law estimates 3,809—actual conditions far exceed estimates. In summary, BDA literature author distribution differs significantly from Lotka' s law estimates, indicating that big data application development has not yet reached a mature and stable stage.

3 Research Conclusions and Implications

Through verification of Price' s law of scientific literature growth, Bradford' s law, and Lotka' s law, this study analyzed and investigated BDA literature growth from three perspectives: publications, journals, and authors. Based on the results, we draw the following conclusions and implications:

(1) Publication analysis shows that since 1990, BDA literature experienced steady development for a period before showing rapid advancement from 2012 onward, exhibiting overall exponential growth that conforms to Price' s law of scientific literature growth. This verification method can not only

simply and accurately depict BDA' s past development trajectory but also predict that the field will continue developing along this trend in the coming years, better characterizing its developmental path and overall direction.

(2) From the journal perspective, the BDA field has formed a core journal group, including *BMC Bioinformatics*, *Sensors*, etc., which to some extent represents the field' s development frontier. Additionally, analysis results show that the literature sample' s journal distribution basically conforms to Bradford' s law. Based on this conclusion, we can conduct in-depth research on core journals and their articles to more timely and accurately grasp the field' s cutting-edge developments.

(3) Regarding author distribution, this field' s distribution differs greatly from Lotka' s law, and a highly influential core author group has not yet formed, reflecting to some extent that related research in this field remains immature. Furthermore, we can infer that for an emerging field that is not yet mature and remains in an exploratory stage, forming a more authoritative and influential core research group requires time and accumulated experience.

(4) This study applied classical bibliometric laws to investigate and verify BDA field development patterns. By extension, this verification method of classical laws can also be applied to research on other emerging fields to better grasp their development trends and status.

From the perspective of big data technology' s rapid development in recent years, governments, international organizations, social enterprises, universities, and various disciplinary fields have all paid great attention and hold positive expectations. The most critical aspect concerns big data technology applications across industries. Although challenges exist in BDA technology, we believe this represents an excellent opportunity to promote overall progress in human society. Big data application represents an inevitable trend and a new direction for social progress. This paper reveals the current status and development trends of the BDA field from the perspectives of publications, journals, and authors, hoping to systematically reveal research patterns and predict future trends to provide valuable references for subsequent research.

Due to limited conditions, many factors were not considered during the research process, which may have led to some inaccuracies and incompleteness. Future research will strengthen comprehensive considerations to more accurately and completely grasp BDA field development.

References

[1] State Council Notice on Issuing the Action Outline for Promoting Big Data Development [EB/OL]. [2016-04-21]. http://www.gov.cn/zhengce/content/2015-09/05/content_{10137}.htm.

[2] United Nations "Global Pulse" Program Releases "Big Data for Development: Opportunities and Challenges" [EB/OL]. [2016-04-23]. http://www.docin.com/p-750680124.html.
[3] MASHEY J R. Big data and the next wave of infra stress[C]//Computer Science Division Seminar. Berkeley: University of California, Berkeley, 1997.
[4] SU H. The Impact of Big Data on H Company' s Technology Innovation Strategy [D]. Qingdao: Ocean University of China, 2014.
[5] LUO Z C, LÜ Z J. Analysis of Big Data Development and Application in South Korea [J]. Global Science, Technology and Economy Outlook, 2014(3): 22-26.
[6] Nature Specials Archive. Big data [EB/OL]. [2016-04-29]. http://www.nature.com/news/specials/bigdata/i
[7] Science, Special Online Collection. Dealing with data [EB/OL]. [2016-02-11]. http://www.sciencemag.org/site/special/data/.
[8] The Era of Big Data Arrives [EB/OL]. [2016-05-02]. http://www.banyuetan.org/jrt/120922/70953_1.shtml.
[9] ZHANG Y, CHEN M, LIAO X F. Current Status and Prospects of Big Data Applications [J]. Journal of Computer Research and Development, 2013, 50(S1): 216-233.
[10] GAO X P. Seeking Innovation in National Governance with Big Data Technology [J]. Chinese Public Administration, 2015(10): 10-14.
[11] LI H. Research on Technology Management Innovation Platform Construction Under Big Data Background [J]. Scientific Management Research, 2014, 32(3): 44-48.
[12] GAO X. Technology Evaluation Research Based on Big Data [J]. Technology Forecasting and Assessment, 2015(11): 27-30.
[13] CUI Y. Analysis of Patent Business Data Quality Evaluation in the Big Data Era [J]. China Invention & Patent, 2013(9): 68-71.
[14] LI Z H, QIAN Z Z, CHENG Y Y. Innovation Strategy for Quality Control Technology of Traditional Chinese Medicine Based on Big Data [J]. China Journal of Chinese Materia Medica, 2015, 40(17).
[15] HUANG S F, LIU X H. Informatization and Services of Geological Data Archives Based on Big Data [J]. Resources & Industries, 2015, 17(6): 56-61.
[16] OLSSON N O E, BULLBERG H. Use of big data in project evaluations [J]. International journal of managing projects in business, 2015, 8(3): 491-512.
[17] SUKUMAR S R, NATARAJAN R, FERRELL R K. Quality of big data in health care [J]. International journal of health care quality assurance, 2015, 28(6): 621-634.
[18] EBACH M C, MICHAEL M S, SHAW W S, et al. Big data and the historical sciences: a critique [J]. Geoforum, 2016, 71: 1-4.
[19] DAVENPORT T H. How strategists use "big data" to support internal business decisions, discovery and production [J]. Strategy & leadership, 2014, 42(4): 45-50.
[20] FRIZZO-BARKER J, CHOW-WHITE P A, MOZAFARI M, et al. An empirical study of the rise of big data in business scholarship [J]. International journal of information management, 2016, 36(3): 403-413.
[21] HOFACKER C F, MALTHOUSE E C, SULTAN F. Big data and consumer behavior: imminent opportunities [J]. Journal of consumer marketing, 2016,

33(2): 89-97.
[22] JIN L. Big Data and the Transformation of Information Technology Teaching [J]. China Educational Technology, 2013(10): 8-13.
[23] ZHAO S. Price' s Contributions to Scientometrics and Contemporary Significance [C]. The 3rd National Graduate Forum on Philosophy of Science and Interdisciplinary Studies. Beijing: Chinese Society for Dialectics of Nature, 2010.
[24] QIU J P. Informetrics (Part 4): Lecture 4—Literature Information Discrete Distribution Law—Bradford' s Law [J]. Information Studies: Theory & Application, 2000(4).
[25] WANG J, WANG H X. Research on Lotka' s Law—Commemorating the 80th Anniversary of Lotka' s Law [J]. Journal of Intelligence, 2007, 26(4): 94-96.
[26] FU Y Y Z, HUA W N. Bibliometric Analysis of Cloud Computing Research Literature in Web of Science Database [J]. New Century Library, 2013(7): 57-
[27] LIU X L. Evaluation Index System of Chinese Core Journals: Evolution, Problems, and Suggestions [J]. Acta Editologica, 2014, 26(1): 92-95.

Author Contribution Statement

Zhang Jiaojiao: Designed the research framework, conducted the research, and wrote the paper;
Liu Yun: Collaborated on conceptualizing and designing the research framework and revised the paper;
Cheng Yijie: Collected and analyzed data and assisted in writing the paper.

Note: Figure translations are in progress. See original paper for figures.