Received 4 May 2015
Revised 26 September 2015
Accepted 3 November 2015

Differences of Pareto principle performance in e-resource download distribution

An empirical study

Qiandong Zhu
Jinan University Library, Guangzhou, China, and
Huimin Xiang
College of Natural Resources and Environment, South China Agricultural University, Guangzhou, China

Abstract

Purpose – The purpose of this paper is to explore whether the databases from a certain library are Pareto-compliant or not? If so, to what extent is the Pareto principle performance evident among these databases? The other purpose is to determine the differences in Pareto principle performance according to time change and database type.

Design/methodology/approach – Data on full-text downloads from six e-resources – Elsevier ScienceDirect (SD), Wiley Blackwell, Springer Journal, EBSCO Business Source Premier (BSP), American Chemical Society and American Institute of Physics (AIP) – for the period 2007-2013 were analysed; 42 samples were collected from these databases. The proportion of frequently downloaded journals from databases was selected as an indicator to determine differences in Pareto principle performance according to time change. The difference between the proportion of frequently downloaded journals and the classic proportion of 20 per cent was used as indicator to determine difference in Pareto principle performance related to database type.

Findings – There are 33 samples (78.57 per cent) which exhibited the Pareto principle. Four databases – Elsevier SD, Wiley Blackwell, EBSCO BSP and AIP – constantly exhibited the Pareto principle. The differences were not significant according to time change. The two multi-discipline databases – Elsevier SD and Wiley Blackwell – fluctuated more moderately than the two single-discipline databases – EBSCO BSP and AIP. Multi-discipline and single-discipline databases showed some differences in Pareto principle performance; however, these differences were not remarkable.

Originality/value – The Pareto principle confirmed that there were frequent and infrequent downloads of e-journals from e-journal databases. It was of great importance to analyse these to improve digital resources acquisition and user service.

Keywords Academic libraries, Acquisitions, Electronic libraries, Electronic resources, e-journals, Electronic resources management

Paper type Research paper

Introduction

In recent years, budget pressure has increased for librarians in Chinese academic libraries because international publishers are substantially increasing prices of electronic resources. The publishers insist on price increases on the grounds of their dominance in digital content publishing. In July 2010, Elsevier required a 14 per cent increase per year for a new three-year contract, regardless of the fact that the average

The Electronic Library
Vol. 34 No. 5, 2016
pp. 846-855

0264-0473
DOI 10.1108/EL-05-2015-0068

increase had already reached 10 per cent during the previous three-year contract. Chinese libraries launched a reaction against the price increase. They signed a petition entitled – *An Open Letter to International Publishers* – hoping that the publishers would consider the country's economic condition because China is still a developing country. The letter also appealed to publishers to set reasonable and flexible subscription prices for new contracts. Many of the top libraries were involved in the activity, including the National Science and Technology Library of China, the National Library of China, the Peking University Library and 33 other libraries in China. In the USA, Elsevier's price increases led to protests from Harvard University, the Massachusetts Institute of Technology and Cornell University, among others. In 2010, Nature Publishing Group required a 400 per cent rise in subscription prices from the University of California system. The University of California said that if Nature Publishing Group did not continue to maintain their current prices, it would organise a joint boycott of their products. In addition to a boycott, libraries devised various ways to respond to price increase trends. Library consortia began to conduct productive activities, such as group discounts for purchasing electronic journal databases, financing open access journal publishing and establishing their own institutional repositories. Libraries are also participating in an interlibrary loan system which helps share expensive resources.

Librarians are increasingly considering the return on investment (ROI) analysis of electronic journal databases to be of great importance because it supports informed budget distribution and purchasing decisions. ROI is the benefit to the investor resulting from an investment into a certain resource (Investopedia, 2015). A popular and common indicator used for ROI analysis is the annual total of full-text downloads. However, this indicator has some shortcomings. For example, the annual total of downloads can only reflect the use of the entire database. Total downloads do not show downloads for individual journals, which would be more useful, because it can help librarians in charge of digital resources acquisition to more fully describe usage. Using individual journal data, librarians are able to develop a practical and accurate acquisition plan. Therefore, librarians cannot solely rely on the annual total downloads to evaluate their acquisition of databases. They should explore details about download distribution and attempt to fully understand the use of certain databases. In fact, full-text downloads from databases show a noticeable centralising trend. In other words, the total number of downloads mainly depends on a small number of frequently used journals. It has been demonstrated that the Pareto principle exits in the use of scholarly electronic resources. The Pareto principle originated from work by the Italian economist Vilfredo Pareto. He made the famous observation that 20 per cent of the population owned 80 per cent of the land in Italy (Pareto, 1896). In the late 1940s, Joseph Juran formulated the Pareto principle and named it after Vilfredo Pareto. Juran used the Pareto principle to determine that a few projects provide the bulk of the improvement and that these few received top priority (Juran and Godfrey, 1998). The Pareto principle states that, for many events, roughly 80 per cent of the effects come from 20 per cent of the causes. In library terms, 20 per cent of the journals (referred to as *frequently downloaded journals*) provide 80 per cent of downloads. Therefore, the inclusion of analysis of the Pareto principle in download distribution can reflect the actual usage model of databases, which supports the librarian's decision for acquiring and retaining database subscriptions. If a library can guarantee subscription to frequently downloaded journals, it will meet 80 per cent of users' full-text demands and save the budget for other urgent resources.

Literature review

A number of studies have demonstrated the centralisation of usage in databases, irrespective of libraries or library consortia. A survey of Elsevier's ScienceDirect (SD) usage at ten British universities found that the use of databases achieved a high level of centralisation; moreover, 5 per cent of the journals accounted for a third to one-half of the total page views. The use of e-journals was, thus, in accordance with the Pareto principle, often called the *80/20 rule* (Nicholas *et al.*, 2010). Emrani *et al.* (2010) analysed data on COUNTER-compliant use of Elsevier's SD journals by an Iranian national consortium for the period 2004-2009. The results showed that the use followed the Pareto principle (Emrani *et al.*, 2010). One case study sampled and analysed three database providers at the Russian State Higher School of Economics University and found that 80 per cent of full-text requests in EBSCO, JSTOR and ProQuest were from 18, 16 and 35 per cent of its journals, respectively. The journal download distribution from EBSCO and JSTOR met the Pareto rule or was even higher than the classic 80/20 proportion. However, the ProQuest database with 80 per cent of article downloads from 34 per cent of its journals did not support the Pareto rule (Fan, 2007). Li and Zhang (2009) found that the top 35 frequently downloaded journals only accounted for 1.9 per cent of the 1,861 journals available in Elsevier's SD at the Hebei Normal University. Furthermore, the top 35 downloaded journals produced 21,889 downloads, responsible for 27.6 per cent of the total 79,187 downloads (Li and Zhang, 2009). Related research also demonstrated the centralising distribution of downloads from electronic databases (Guo, 2009; Liu *et al.*, 2010).

While these studies reported the existence of centralising distribution and the Pareto principle, their research aim was somewhat different. One trend is to investigate information seeking behaviour in a certain library or among different institutions and subjects (Li and Zhang, 2009; Nicholas *et al.*, 2010). Nicholas' research is one of the very few studies to investigate subject and institutional differences with regard to the information seeking and usage of UK researchers. Li and Zhang's (2009) study is meaningful for libraries who are conducting user services and information literacy education.

The other trend is to detect the user patterns in databases to help libraries choose better cost-benefit license models when acquiring e-resources (Emrani *et al.*, 2010; Fan, 2007). Emrani *et al.* (2010) proposed that institutions should be grouped into three or four categories based on their subject fields and amount of use. Fan (2007) appealed to libraries to highlight the importance of use efficiency based on the analysis of use data. However, very few studies focus on whether the use of a certain database showed the Pareto principle or diversity of the Pareto principle in terms of time change and database type. The details and differences of the Pareto principle with respect to time change and database type can offer more information for libraries to make acquisition decisions.

Research question, method and data sources

The major research questions of this study are as follows:

RQ1. Are the databases from the studied library Pareto-compliant? If so, to what extent is the Pareto principle performance evident among these databases?

RQ2. What is the difference when Pareto-compliant databases are analysed over time? The same question applies to when Pareto-compliant databases are analysed according to database type.

The following six electronic journal databases were selected as research objects: Elsevier's SD, Wiley-Blackwell, Springer journal packages, EBSCO Business Source Premier (BSP), American Chemical Society (ACS) and American Institute of Physics (AIP). These databases are representative of electronic resources used in academic libraries in China. Elsevier's SD, Wiley-Blackwell and Springer are multi-discipline databases, popular among academic libraries in China. EBSCO BSP is a discipline database related to economics, management and commercial disciplines. ACS focuses on chemistry and chemical engineering. AIP is a database relevant to physics. These are discipline-specific databases. Currently, EBSCO BSP is an index and abstract platform with full-text. It does not provide an option for title-by-title selection. Except for EBSCO BSP, the other five resources are all publisher-based e-journal platforms.

First, full-text downloads from these six databases were analysed, and the databases whose distribution followed the Pareto principle were identified. Second, these Pareto principle-compliant databases were further analysed to determine the difference in their centralising performance according to time change. Third, the difference between multi-discipline and single-discipline databases was also analysed. The suppliers of these six databases provide libraries with a COUNTER-compliant usage report, namely, Journal Report 1 (Number of Successful Full-Text Article Requests by Month and Journal). The usage reports were obtained from the management websites of these databases for the period 2007-2013. Annual downloads from each database were sampled and a total of 42 samples were involved. This study focuses on the distribution of journals; therefore, the target journals for sampling are journals that produced at least one download in a year. The journals with zero downloads were not included. In the sample, journals were organised in descending order, according to their annual full-text downloads. In descending order from the first journal, all journal downloads were added up until the total of downloads occupied 80 per cent of the database downloads. These journals were defined as frequently downloaded journals and the rest as infrequently downloaded journals. Thus, the proportion of frequently downloaded journals equals the number of frequently downloaded journals divided by the total number of journals. Likewise, the proportion of infrequently downloaded journals equals the number of infrequently downloaded journals divided by the total number of journals.

The Pareto principle states that roughly 80 per cent of the effects come from 20 per cent of the causes. This study proposed that if a sample's proportion of frequently downloaded journals is in the 10-30 per cent range, the sample is considered to meet the Pareto principle.

To measure differences of the Pareto principle, according to database type, we proposed a K value where:

K = Proportion of frequently downloaded journals − the classic proportion of 20 per cent.

The average value of the seven-year K value was regarded as a performance indicator for databases. The average value was named M in this study:

$$M = \text{Total of } K/\text{Total of years}$$

Results

Download distribution of the six electronic databases

The 42 samples from Elsevier's SD, Wiley-Blackwell, Springer, EBSCO BSP, ACS and AIP from 2007 to 2013 were analysed, and their detailed download distributions are presented in Table I. In 2007, except for ACS, the databases followed the Pareto distribution. In 2008, four of the journal databases followed the Pareto distribution. Springer and ACS databases did not. In 2009 and 2010, excluding the ACS database, the journal databases followed the Pareto distribution. From 2011 to 2013, ACS did not comply with the Pareto principle and neither did Springer in 2013. From 2007 to 2013, 33 samples followed the Pareto distribution, which accounted for 78.6 per cent of the total 42 samples. The samples that did not meet the Pareto principle accounted for 21.4 per cent. The ACS database was distinguished from the other five databases as none of its seven samples complied with the Pareto principle during the time frame under study. Two samples from Springer in 2008 and 2013 did not have the Pareto distribution. In general, most of the samples on database download distribution were consistent with the Pareto principle.

Differences in Pareto principle performance according to time change

Among the six journal databases, the proportion of frequently downloaded journals in Elsevier's SD was 27.7, 21.0, 23.7, 22.7, 20.6, 21.9 and 21.8 per cent from 2007 to 2013. The value varied from 2007 to 2013, but it fluctuated around 20 per cent showing an obvious centralising trend. The proportion of frequently downloaded journals in Wiley-Blackwell also varied during the period 2007-2013. However, the value continually approached 20.0 per cent, i.e. towards the classic proportion of the Pareto principle. This indicated that download distribution was relatively stable in Wiley-Blackwell. The EBSCO BSP database fluctuated between 15.6 and 22.6 per cent, which implied comparatively small changes annually. The AIP download distribution was also stable during the past seven years. Figure 1 demonstrates these results.

Based on the 28 samples from four databases, it was concluded that Elsevier's SD, Wiley-Blackwell, EBSCO BSP and AIP always followed the Pareto principle. ACS was excluded because it was not compliant with the Pareto principle. Springer was excluded because its sample did not meet the Pareto principle criteria in 2008 and 2013. The former four databases were further analysed to determine differences in their Pareto principle performance, according to database type: multi-discipline or single-discipline.

Differences in Pareto principle performance according to database type

The differences in Pareto principle performance related to database type were measured by the difference between the proportion of frequently downloaded journals and the classic proportion of 20 per cent. The difference was named the K value, and the average value of the seven-year K value was named the M value in this study. From 2007 to 2013, the M values of Elsevier's SD, Wiley-Blackwell, EBSCO BSP and AIP were 2.8, 2.1, 3.1 and 4.6 per cent, respectively. The M values of two multi-discipline databases – Elsevier's SD and Wiley-Blackwell – exhibited a more stable fluctuation than the two single-discipline databases: EBSCO BSP and AIP. Thus, multi-discipline databases and single-discipline databases demonstrated a slight difference in Pareto principle performance; however, this difference was not very significant. Of course, there were only four databases selected for analysis. If more databases and samples are involved,

e-journal databases	2007		2008		2009		2010		2011		2012		2013	
	FDJ	IDJ	FDJ	IDJ	FDJ	IDJ	FDJ	IDJ	FDJ	IDJ	FDJ	IDJ	FDJ	IDJ
Elsevier's SD	27.67	72.33	21.03	78.97	23.72	76.28	22.69	77.31	20.59	79.41	21.93	78.07	21.84	78.16
Wiley-Blackwell	20.72	79.28	22.95	77.05	17.79	82.21	21.72	78.28	19	81	21.36	78.64	15.42	84.58
Springer journals	19.9	80.1	8.2	91.8	22.2	77.8	21.49	78.51	19.7	80.3	21.85	78.15	31.12	68.88
EBSCO BSP	17.7	82.3	21.72	78.28	22.57	77.43	15.64	84.36	15.64	84.36	16.55	83.45	17.4	82.6
ACS	37.84	62.16	40.54	59.46	35.56	64.44	36.17	63.83	32.65	67.35	34.69	65.31	31.37	68.63
AIP	18.2	81.8	18.2	81.8	18.2	81.8	16.67	83.33	11.11	88.89	10	90	15.38	84.62

Notes: FDJ = frequently downloaded journals; IDG = infrequently downloaded journals

Table I.
Downloads distribution of six electronic databases from 2007 to 2013

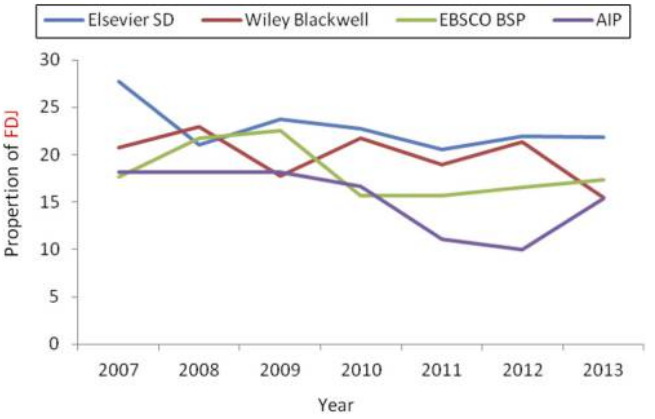

Figure 1.
Proportion of FDJ variance between 2007 and 2013

the differences in Pareto principle performance between the two types of database needs further analysis to determine (Table II).

Discussion

The Pareto principle shows the presence of frequently and infrequently downloaded journals. Frequently downloaded journals only accounted for 20 per cent of the total journals while producing 80 per cent of the total downloads. Studies have shown that, in the print environment, 15 per cent of the total amount of journals has 80 per cent usage, whereas 15 per cent of journals account for 38 per cent of the total budget (Sennyey and Ellern, 2002). Likewise, 85 per cent of the total number of journals occupied 62 per cent of the total budget but led to only 20 per cent usage (Sennyey and Ellern, 2002). In a digital environment, as long as the electronic journal database download distribution follows the Pareto principle, the same problem will continue to exist. Currently, suppliers and publishers tend to provide libraries with a whole package of e-journal databases, and the price also depends on the whole package. In this situation, each journal price is an average value of the total titles and total price. Thus, the frequently downloaded journals take up 20 per cent of the total titles and budget, producing 80 per cent of total downloads. However, the infrequently downloaded journals are responsible for 80 per cent of the total titles and budget, with an output of 20 per cent of total downloads. Thus, the frequently downloaded journals show remarkable ROI performance, whereas the infrequently downloaded journals show poor ROI performance.

In recent years, e-journal database prices have been rapidly increasing every year. Therefore, it becomes increasingly urgent for librarians to analyse the usage statistics of

Table II.
Values of K and M

e-journal databases	2007	2008	2009	2010	2011	2012	2013	M value
Elsevier's SD	7.7	1.0	3.7	2.7	0.6	1.9	1.8	2.8
Wiley-Blackwell	0.7	3.0	2.2	1.7	1.0	1.4	4.6	2.1
EBSCO BSP	2.3	1.7	2.6	4.4	4.4	3.5	2.6	3.1
AIP	1.8	1.8	1.8	3.3	8.9	10.0	4.6	4.6

each database and figure out a reasonable price model to bargain with suppliers. Ready-made deals sometimes appear as unbounded and excessive supply, but they are not truly suited to and sufficient for users' needs (Schöpfel and Leduc, 2012). This study showed that the Pareto principle performance in multi-discipline databases was more stable than in single-discipline databases, though the difference was not very significant. It means libraries can acquire multi-discipline e-journal databases, according to frequently used titles. Some multi-discipline e-resources, such as Wiley-Blackwell, contain a variety of subjects, and the content is divided into several subject packages. Libraries are able to choose one or several subject packages from a database. If the frequently downloaded journals congregate around certain subjects in a multi-discipline e-resource and its supplier provides an option for subject packages, it is reasonable for libraries under increasing budget pressure to choose subject packages with most of the frequently downloaded journals. This can guarantee that most of users' full-text demands are met and produce maximum returns. The subject packages with less heavily used titles from a multi-discipline e-resource are also useful for certain users. So libraries with adequate budget should purchase these subject packages. There is another situation where frequently downloaded journals scatter among subjects in multi-discipline e-resource. It will be appropriate for libraries to apply title-by-title selection in this case.

It is of great importance for libraries to balance users' full-text demand and a tight budget. These approaches, such as subject package subscription, title-by-title selection and other purchasing modes, are all used by libraries to save expenses under budget pressure and produce maximum returns on digital resources. At present, some e-resources, such as EBSCO BSP, do not provide an option for title-by-title selection. This is unreasonable for library practices. Hence, it is urgent for libraries or library consortia to argue for a reasonable price model, especially when usage concentrates on a few frequently downloaded journals. Full-text demand for articles from infrequently downloaded journals can be achieved via interlibrary loans.

This research also detected the differences of Pareto principle performance linked to time change. Not all samples from Springer were Pareto-compliant during the investigation period. Even a Pareto-compliant database, such as the AIP, showed a changing proportion of frequently downloaded journals. Therefore, it is necessary for libraries to analyse usage over a long time to fully and truly understand usage models for databases before making decisions about subscriptions.

Conclusion

These data demonstrated that download distribution in Elsevier's SD, Wiley-Blackwell, EBSCO BSP and AIP were compliant with the Pareto principle, whereas ACS was not Pareto-compliant. The Springer samples of 2008 and 2013 were not Pareto-compliant, but the other samples were all Pareto-compliant. The Pareto principle performance of Elsevier's SD, Wiley-Blackwell, EBSCO BSP and AIP showed small differences over time. They also indicated a difference, according to database type, but that difference was not remarkable. The conclusion is based on usage statistics for six online databases from Jinan University, which is a comprehensive university. It is unclear to what extent the Pareto principle performance would apply to non-comprehensive universities. Each online database has its own full-text journals and the number of full-text journals varies. For

example, the number of full-text journals in Elsevier is close to 2,000, whereas that in AIP is 14. The Pareto principle performance tends to be correlated with the quantity of full-text journals in a database besides their type. The relationship between the Pareto principle and the total number of titles needs to be studied further.

References

Emrani, E., Moradi-Salari, A. and Jamali, H.R. (2010), "Usage data, e-journal selection, and negotiations: an Iranian consortium experience", *Serials Review*, Vol. 36 No. 2, pp. 86-92.

Fan, Z. (2007), "Analyzing e-journal use efficiency with Pareto rule: a case study of the library of State University-Higher School of Economics", *New Century Library*, Vol. 2007 No. 3, pp. 101-104.

Guo, Z.Y. (2009), "Foreign electronic journals usage statistics and analysis", *Capital Medical University (Social Science Edition Supplement)*, Vol. 2009 (Special Issue), pp. 146-148.

Investopedia (2015), available at: www.investopedia.com/terms/r/returnoninvestment.asp (accessed 17 April 2015).

Juran, J.M. and Godfrey, A.B. (1998), *Juran's Quality Handbook*, 5th ed., McGraw-Hill, New York, NY.

Li, L.P. and Zhang, L.J. (2009), "Subscription and usage statistical analysis of the ScienceDirect database", *Library Tribune*, Vol. 29 No. 4, pp. 60-62.

Liu, Y.R., Han, P.M. and Yan, P. (2010), "Statistical analysis and research on utilization of full-text electronic journal databases – take the Elsevier SDOL of Tianjin University as example", *Library Work and Study*, Vol. 2010 No. 1, pp. 73-76.

Nicholas, D., Rowlands, I., Huntington, P., Jamali, H.R. and Hernández Salazar, P. (2010), "Diversity in the e-journal use and information-seeking behaviour of UK researchers", *Journal of Documentation*, Vol. 66 No. 3, pp. 409-433.

Pareto, V. (1896), *Cours d'Economie Politique*, Droz, Geneva.

Schöpfel, J. and Leduc, C. (2012), "Big deal and long tail: e-journal usage and subscriptions", *Library Review*, Vol. 61 No. 7, pp. 497-510.

Sennyey, P. and Ellern, G.D. (2002), "Newsome N. collection development and a long-term periodical use study: methodology and implications", *Serials Review*, Vol. 28 No. 1, pp. 38-44.

Further reading

An Open Letter from Chinese Libraries: Foreign Periodicals and Databases Prices Rise Too High (2015), available at: news.sohu.com/20100906/n274730454.shtml (accessed 17 April 2015).

An Open Letter to International Publishers (2015), available at: www.las.ac.cn/others/gongkai2.jsp (accessed 17 April 2015).

Liu, X.M. and Zhang, J.Y. (2009), "Understanding the influence of digital resources on scientific research", *Journal of Academic Libraries*, Vol. 27 No. 1, pp. 60-63.

The COUNTER Code of Practice (2015), available at: www.projectcounter.org/r3/Release3D9.pdf (accessed 17 April 2015).

UC libraries, faculty protest planned price hike by Nature publisher (2010), available at: www.today.ucla.edu/portal/ut/uc-libraries-faculty-protest-proposed-160117.aspx (accessed 17 April 2015).

Ying, J. and Taniguchi, K. (2005), "How Japanese academic libraries deal with journal inflation", *Journal of Academic Libraries*, Vol. 23 No. 5, pp. 86-89.

Zhang, J.Y. (2009), "Usage analysis on the foreign language electronic journals in Guangzhou Medical College", *Journal of Medical Informatics*, Vol. 30 No. 12, pp. 82-85.

Zhu, Q.D. (2010a), "Pareto principle in the using of full-text electronic journals: a case study of Elsevier SD in Jinan University", *Journal of Modern Information*, Vol. 30 No. 12, pp. 125-129.

Zhu, W.J. (2010b), "SPIE and the other four publishers confirm that they will cooperate with the Massachusetts Institute of Technology on open access policy", *Library and Information Dynamics*, Vol. 2010 No. 3, pp. 8-9.

Corresponding author
Qiandong Zhu can be contacted at: zqd51@126.com