

Usage patterns of scholarly articles on Web of Science: a study on Web of Science usage count

Xianwen Wang¹  · Zhichao Fang¹ · Xiaoling Sun¹

Received: 4 January 2016 / Published online: 9 August 2016
© Akadémiai Kiadó, Budapest, Hungary 2016

Abstract Usage data of scholarly articles provide a direct way to explore the usage preferences of users. Using the “Usage Count” provided by the Web of Science platform, we collect and analyze the usage data of five journals in the field of Information Science and Library Science, to investigate the usage patterns of scholarly articles on Web of Science. Our analysis finds that the distribution of usage fits a power law. And according to the time distribution of usage, researchers prefer to use more recent papers. As to those old papers, citations play an important role in determining the usage count. Highly cited old papers are more likely to be used even a long time after publication.

Keywords Article usage · Usage count · Altmetrics · Usage metrics · Web of Science

Introduction

The revolutionary transition from print publishing to electronic publishing makes article level usage data available. Although many publishers chose not to disclose the information about the usage data of articles before, some of them begin to provide multiform and detailed usage metrics for researchers nowadays.

The definition of usage may be rather broad. Citation could be regarded as an official, real and essential form of usage of previous studies, while views and downloads are more of unofficial, potential, unessential/peripheral forms of usage.

Compared to citation count-based metrics, which were proposed about 90 years ago (Gross and Gross 1927), article-level usage data is a kind of relatively young metric. The

✉ Xianwen Wang
xianwenwang@dlut.edu.cn; xwang.dlut@gmail.com;
<http://xianwenwang.com>

¹ WISE Lab, Faculty of Humanities and Social Sciences, Dalian University of Technology, Dalian 116085, China

access to the large volume of usage data brings opportunities to researchers to study the usage patterns of scholarly articles that could not be done with bibliographic data or citation data. Nowadays, studies on these “potential usage data” have attracted much attention from researchers. There are at least three main research directions of usage data analysis. Firstly, the obsolescence of articles, which is used to be weighed by circulation and citation information (Gosnell 1944; Egghe and Rousseau 2000), could be measured by usage rate now (Moed 2005; Kurtz and Bollen 2010; Wang et al. 2014a, b). Some researchers further looked into the possibilities of adjusting usage statistics by citation statistics or even replacing citation measures with usage measures (Ladwig and Sommese 2005; Duy and Vaughan 2006). Following this idea, the second research direction emerges. Gradually, usage metrics are regarded as potential complements, perhaps even alternatives, to citation for research evaluation (Bollen and Sompel 2008; Wan et al. 2010; Glänzel and Gorraiz 2015). This assumption is on the basis of the advantages of usage data over citation data, such as faster availability, more convenient data collection and broader usage scope (Brody et al. 2006; Gorraiz et al. 2014). Moreover, the significant correlation between specific usage types, especially downloads, and citations investigated by many researches lays the foundation for this assumption (O’ Leary 2008; Schloegl and Gorraiz 2011; Subotic and Mukherjee 2014). However, the relationship between these two metrics is complex (Line and Sandison 1975), so there also exist different opinions on this question. Even for download, the correlation between downloads and citations may be only moderate or a bit low when being calculated at the article level (Schloegl and Gorraiz 2010, 2011). Finally, in addition to these studies combined with traditional metrics data, usage data is introduced to do some researches that cannot be accomplished by traditional methods. Some researchers turned to work backwards and explore user behaviors through the usage records (Davis and Solla 2003; Davis and Price 2006; Wang et al. 2012, 2013a, b). Which paper was used, what time it was used, where the user is from, these data recorded by systems and disclosed by publishers are deemed to contain a wealth of information that could help us understand various aspects of user behaviors.

All of the above researches are based on the usage data retrieved from one single publisher or repository like Elsevier, Springer, PLoS, Nature, and arXiv.org, etc. The collection scopes of these data sources are too limited to carry out complete and large-scale studies. For this reason, the results of similar questions based on different data sources are sometimes conflicting. Therefore, we need a more comprehensive usage data source.

Web of Science (WoS) is one of the most famous scientific citation index databases in the world. Before September 2015, only metadata (provided directly by publishers) and citation data are available for users. On 26 September 2015, Web of Science released 5.19 version and added item level usage metrics on the Web of Science platform, which is called “Usage Count”. The count reflects the number of times the article has met a user’s information needs as demonstrated by clicking links to the full-length article at the publisher’s website or by saving the metadata for later use. Displayed on the summary page as well as the full record, the usage count shows the number of times the full text of a record has been accessed or a record has been saved in the last 180 days (rolling) as well as since February 1, 2013. (http://wokinfo.com/media/pdf/wos_release_519.pdf).

“Usage” defined by Web of Science has special connotation that differs from general definitions. As mentioned above, the definition of “usage” could be rather broad. From a narrow point of view, citation can be considered as the authentic “usage”, because only citation demonstrates a real sense of usage that leaves behind certain evidences when the user’s identity could be tracked. In a broad point of view, article views and downloads represent potential and unessential/peripheral forms of usage. And in a broader sense, some

even consider altmetrics as usage metrics, but this opinion had been criticized by Glänzel and Gorraiz (2015). In general, “usage” often refers to HTML views and PDF downloads. For example, the usage metrics disclosed by publishers usually denoted by views, downloads or a sum of the two. Nevertheless, compared to citations, views and downloads do not reflect the real “usage”, but the potential “usage”, because there is no evidence that users have really used the articles after viewing and downloading, and it is difficult to confirm if all the references cited in an article have been read by the author. However, the behaviors of viewing and downloading are prepositive procedures of real usage, and they suggest the usage motives of users, namely potential usage. Therefore, most publishers classify views and downloads into “usage metrics”. Similarly, Web of Science defines “usage” as “clicking” and “saving”, because it just offers the metadata of articles (not the full texts) and the access to publishers, so the number of views and downloads of full texts are unavailable. For the users of Web of Science, the behaviors of clicking and saving also reflect their usage motives, as they imply the interest and actions of users, which also have research significance. As the derivations or analogs of views and downloads instead of brand new usage data, “clicking & saving” on Web of Science are more peripheral than views and downloads, which can be considered as “possibly potential usage”.

The usage metrics provided by various publishers have been widely studied. Now, the availability of usage data on the Web of Science platform provides a kind of brand new and all-around data source for us to explore the usage metrics. In this study, our research questions are: what are the usage patterns of scientific literature on Web of Science? Are there any differences between cited times and WoS usage counts of papers published at different time periods? And what usage preference do researchers have when they search articles on Web of Science?

Data source

Five journals in the field of Information Science and Library Science are selected as our research objects, including *Journal of the Association for Information Science and Technology* (JASIST), *Scientometrics*, *Journal of Documentation*, *Information Processing and Management*, and *Research Evaluation*.

These five journals are typical in the field of Information Science and Library Science with high influence, especially in the field of scientometrics. Another reason for choosing these journals is that they all have a long history, which ensures there are enough papers with enough accumulated time for analysis.

Some journals have their journal names changed several times in history, for example, the name of JASIST varied four times in the past years. The journal was established in 1950 entitled *American Documentation*. In 1970, it changed its name to *The Journal of the American Society for Information Science*, and to *Journal of the American Society for Information Science and Technology* in 2001. It obtained its current name as *Journal of the Association for Information Science and Technology* in January 2014. So, in Web of Science, we conduct the search including all the names of the journal. For another journal, *Research Evaluation*, although it started its first issue in 1991, the first record we retrieve from Web of Science was published in 2000. Table 1 shows the data set, only the document type of article is included in the search.

As Table 1 shows, the time span of the data set ranges from 1945 to September 2015. In this study, to better understand the usage of papers published in different periods, we set

Table 1 Data set

Journal	Start year in WoS	Number of papers
JASIST	1956	4466
Scientometrics	1978	3777
Journal of Documentation	1945	1410
Information Processing and Management	1975	2113
Research Evaluation	2000	409

three time periods, which are 1945–1999 (old), 2000–2012 (recent), and 2013– Sep 2015 (current). Old papers were published a long time ago. Recent papers are younger than old papers, and they are in the transitional period from activity to senescence. Current papers are those newly published, which are the youngest that just become active and wait for recognition.

All the metrics data are collected from Web of Science, including cited times and usage counts.

Results

Power law distribution of usage

The phenomenon of usage distribution at the journal level showing quite skewed distribution and fitting to power law distribution was observed by Jung et al. (2013). Figure 1 shows the distribution of publication usage data at the article level from 1945 to Sep 2015. In the plot, x represents the usage count of the papers, and $p(x)$ represents the probability of

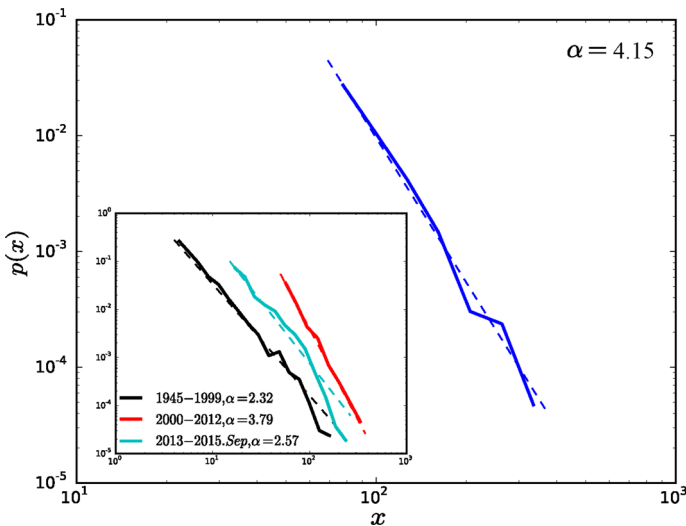


Fig. 1 Power law distribution

x. In the inset, data is classified into three groups according to the published year. All the distributions fit power law (dashed lines), with the fitted parameter α . The blue line indicates the overall usage distribution, from which we could see, a small part of papers accumulated most usage counts, while most other papers were rarely used. There also exists Matthew effect in the behavior of using papers, that is, the papers which have been highly used are more likely to be used in the future. In the inset, the black line indicates the usage distribution of old papers (published from 1945 to 1999), and these papers have relatively low usage times. The turquoise line represents the usage distribution of the current papers (published from 2013 to 2015). The red line represents the usage distribution of recent papers (published from 2000 to 2012), and these papers have the largest number of usage. In contrast, the recent papers have larger usage times than the old papers, indicating that researchers are likely to use more recent publications. Typically, in power law distribution, the exponent falls in the range $2 < \alpha < 3$, though not always. The usage distributions also have similar exponents, except the overall distribution has a larger exponent.

Citation versus usage

Figure 2 shows the scatter plot of cited times and usage counts of papers divided into three groups according to the publication time, which are 1945–1999, 2000–2012, and 2013–2015. The green circles represent old papers published from 1945 to 1999, blue asterisks represent recent papers published from 2000 to 2013 and red circles represent current papers published from 2013 to 2015. An obvious stratification could be seen from Fig. 2. The inset scatter plot is the enlargement of the part near the beginning, which reveals more obvious stratification of cited times and usage counts. Old papers have more cited times than usage times, current papers are highly used but lowly cited, while recent

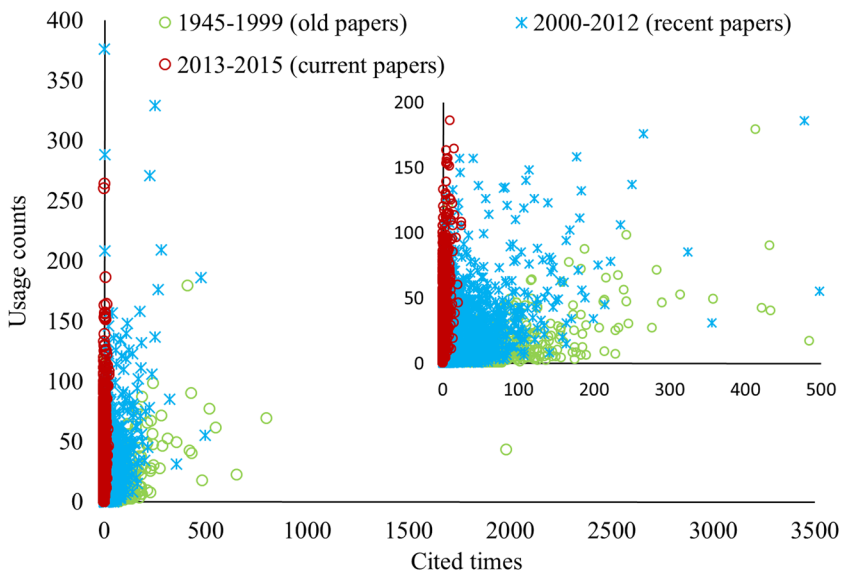


Fig. 2 Scatter plot of cited times and usage

papers are in the middle of them. The relationship between cited times and usage counts of recent papers are relatively balanced comparing with old and current papers.

Old papers have a long time to accumulate citations and make impacts. But compared to cited times they have accrued, their usage counts in WoS are considerably less. This result shows a dramatic inconsistency with many previous studies that based on other usage data sources. Moed (2005), Guerrero-Bote and Moya-Anegón (2014) examined the statistical relationships between downloads (from ScienceDirect) and citations (from SCI and Scopus), and found that the number of downloads is much greater than that of citations. Similar results were also reported by CIBER Research Limited (2011) and other scholarly information metrics services like PlumAnalytics.com. There are at least three reasons for this inconsistency. Firstly, because Web of Science has not provided usage data until February, 2013, for those papers published prior to that, the usage counts are underestimated. The older the paper is, the greater the degree of underestimation would be. For example, for a paper published in 2000, it has approximate 16 years to accumulate the citations and only less than 3 years to accumulate the usage counts; Secondly, it is not easy for old papers to draw researchers' attention. Compared with recent and current papers, old papers have disadvantages in visibility, accessibility and freshness, etc., people have less chance to find old papers. For example, search results of Web of Science are acquiescently sorted from newest to oldest according to publication date. Generally, people won't notice the records published tens of years ago and listed in the last pages. Thirdly, the usage data from the Web of Science platform does not appear to be sufficiently comprehensive. According to our study of the patterns of people's visits to academic articles, Web of Knowledge contributes a lot but not the most in directing people to scholarly articles (Wang et al. 2016). Web of Science is not the most favored platform for researchers to search, view and download primary research articles, meaning that relative limited usage logs would be accumulated by the platform. Current papers received the most usage counts. Researchers are naturally more interested in new findings and inclined to use current papers to track the research frontiers and hot topics. Besides, some journals and publishers urge authors to ensure the novelty of their research, which may be embedded in the references of papers, and leads to more usage counts of current papers. High usage counts could not instantly turn into high cited times immediately. Time delay that mainly resulted from publication delay and citation gap causes citations need a long time to accumulate (Amat 2007; Peng and Zhu 2012). Therefore, newly published current papers could hardly be cited. Recent papers have quite a long time to be cited and meanwhile the advantage of enough novelty. So, recent papers are both highly cited and used.

Average usage of papers published in each year

In order to explore the tendency of the usage counts of publications in different time periods, we calculate the median usage of annual papers published on each journal, as shown in Fig. 3. The median usage counts have been keeping rising trends since the establishment of journals. Because the usage data hasn't been released until February 1, 2013, all the usage behaviors studied in this paper occurred from February 2013 to September 2015, which means that five curves in Fig. 3 reflect the usage preferences of users in a period of 2 years and 8 months. It is also for this reason that the annual median usage falls dramatically since 2013. The papers published from 2013 to 2015 have less time than 2 years and 8 months to accumulate the usage, in this specific period, the later the paper published, the less the chance for being used. Therefore, we only take the variation trend up to 2012 into consideration to investigate researchers' using behaviors to

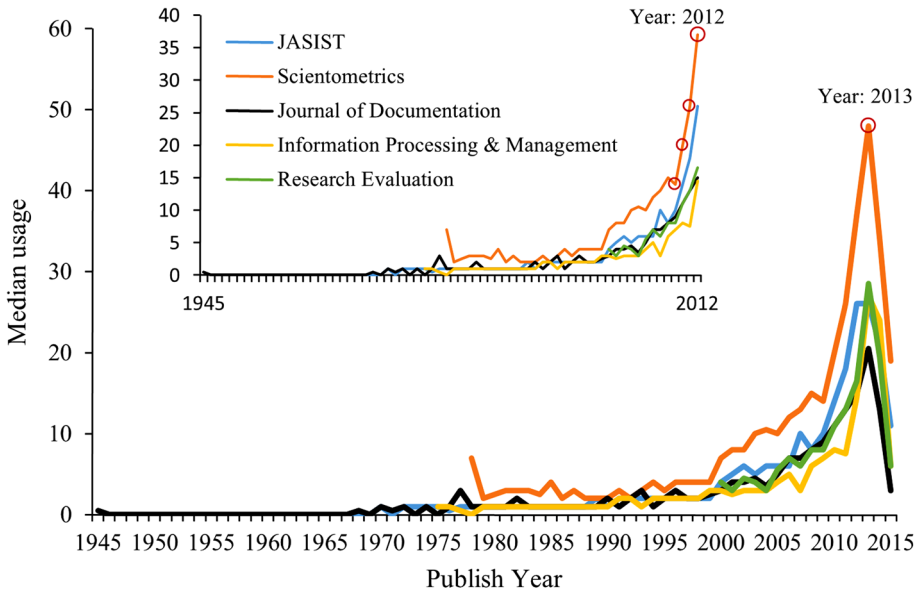


Fig. 3 Annual median usage of papers published on each journal

the papers published more than 2 years ago, as the inset of Fig. 3 shows. All the curves show upward trend, which means that more recent papers are used more often than older ones.

One way ANOVA

In previous studies, the positive effects generated by downloads on citations have been discussed based on the correlation between them (Moed 2005; Schloegl et al. 2014; Jung et al. 2015). In turn, does the citations of an article that it has accumulated have any relationship with its subsequent usage counts? Concerning this question, we find that as to old papers published from 1945 to 1999, their usage counts are generally low, but there still exists differences among them. Some of them are used from time to time, while others fall silent. Could this difference be differentiated by the citations that old papers have accumulated? In order to test the hypotheses, all old papers are divided into two groups, the highly cited and lowly cited papers. Papers with cited times greater than or equal to 17 (75th percentile, Q3) are classified into the highly cited group, while the papers which are cited less than 7 (50th percentile, Q2) are classified into the lowly cited group.

As Fig. 4 shows, the median value of the highly cited papers (left box plot) is much greater than that of lowly cited papers (right box plot). The right box plot is much shorter and lower than the left box plot, which suggests that the usage counts of overall lowly cited papers are much fewer than that of highly cited papers and have a high level of agreement with each other, and the comparatively tall left box plot means that the usage counts of highly cited papers vary over a wide range.

Then, a one-way ANOVA is conducted to compare the effect of high or low citation on the usage of papers. The alpha level is set to 0.01 instead of 0.05 to avoid possible Type I error because of the repeated tests (over year and adjusted). ANOVA is calculated on the

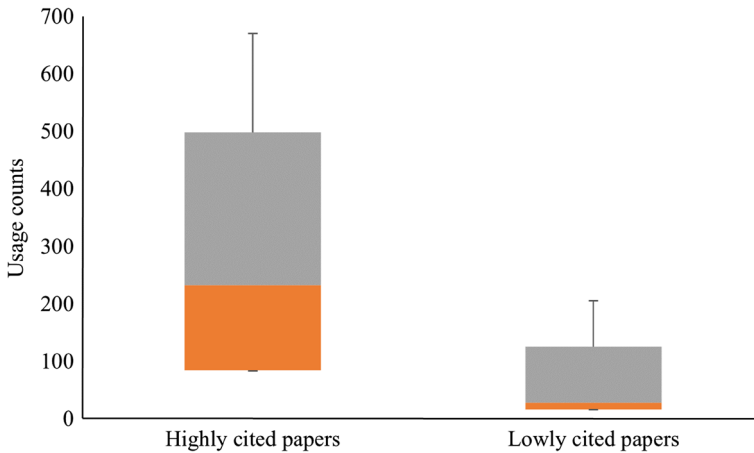


Fig. 4 Box plot of usage counts for highly and lowly cited papers

difference of usage between highly cited and lowly cited papers. The result is significant, $F(1, 3832) = 468.91$, $p = .000$, which means that there is significant difference between highly cited and lowly cited papers. As time goes by, compared to lowly cited old papers, highly cited papers accumulate more usage counts. Citation is a critical factor that has an impact on the chance of being used for old papers. Highly cited papers are more likely to be used in the future than the lowly cited ones due to the impacts they have made, even though they could not reflect the frontiers at that time in their fields anymore (Table 2).

Conclusion and discussion

In this study, by using the newly released article usage counts from the Web of Science platform, firstly, we analyze the usage patterns of scientific papers on WoS platform in the field of Information Science and Library Science. It is proved that the distribution of usage counts obeys the power law. Most usage counts are contributed by a small number of papers. A few papers are used frequently, while most others are little-visited.

Secondly, we investigate the relationship between the WoS cited times and WoS usage counts of papers published in three different periods (old, recent and current). We find that researchers prefer to use the newly published papers on Web of Science. Current papers published in recent 3 years have the most usage with few citations, because papers reach their citation peak 2–4 years after publication. Old papers accumulated more citations than recent and current papers, but are used less than the latter two groups.

Table 2 One way ANOVA

	Mean (SD)		<i>F</i>	Significance
	Highly cited group	Lowly cited group		
All_views	6.84 (11.93)	1.46 (2.63)	468.91	$p = .000$

Thirdly, for the old papers published tens of years ago, highly cited papers (WoS citation) are used more frequently than lowly cited papers. Citation is an important factor to reflect the usage of papers even published tens of years. Highly cited old papers have distinct advantages of being used by the following researchers.

There are some limitations in this study. Firstly, only 5 journals in the field of Information Science and Library Science are investigated, therefore, the conclusions about the researchers' preference might be different in other fields, which needs further experiments using data from different fields. Secondly, Web of Science only provides the usage count of each article recorded since February 2013. Therefore, in this study, the usage data we collect and analyze has a time period of only 2 years and 8 months. A long-term dataset should be tracked and harvested to support our conclusions.

Last but not least, as previously stated, the usage count defined in this paper refers to the number of times the article has met a user's information needs, as demonstrated by clicking links to a full-length article at the publisher's website or by saving the metadata for later use. The usage count in this sense represents a potential usage count, but does not reflect real article "usage". We don't know the accurate proportion of researchers who use Web of Science as the preferred platform to search for scholarly content. As one of the referral sources of views and downloads, WoS usage count data of "clicking and saving" is less representative than those "usage data" of "views and downloads" provided by publishers' platforms, which leads to the limitation to infer the overall usage patterns of scientific literature based on the WoS usage count data.

Although with these limitations, we hope that this first try to explore the usage patterns of WoS records would provide a new perspective to promote the research on usage metrics.

Acknowledgments The work was supported by the project of "National Natural Science Foundation of China" (61301227), the project of "Growth Plan of Distinguished Young Scholar in Liaoning Province" (WJQ2014009), the project of "the Fundamental Research Funds for the Central Universities" (DUT15YQ111) and Liaoning Province Higher Education Innovation Team Fund "Research on Responsible Innovation" (WT2015002). We appreciate the reviewers' suggestions and comments, some of their comments are used in this paper.

References

- Amat, C. (2007). Editorial and publication delay of papers submitted to 14 selected Food Research journals. *Influence of online posting. Scientometrics*, 74(3), 379–389.
- Bollen, J., & Sompel, H. V. D. (2008). Usage impact factor: the effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology*, 59(1), 136–149.
- Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8), 1060–1072.
- CIBER Research Limited. (2011). *The journal usage factor: exploratory data analysis*. Accessed April 11, 2016 from: http://www.projectcounter.org/documents/CIBER_final_report_July.pdf.
- Davis, P. M., & Price, J. S. (2006). eJournal interface can influence usage statistics: implications for libraries, publishers, and Project COUNTER. *Journal of the American Society for Information Science and Technology*, 57(9), 1243–1248.
- Davis, P. M., & Solla, L. R. (2003). An IP-level analysis of usage statistics for electronic journals in chemistry: Making inferences about user behavior. *Journal of the American Society for Information Science and Technology*, 54(11), 1062–1068.
- Duy, J., & Vaughan, L. (2006). Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination. *The Journal of Academic Librarianship*, 32(5), 512–517.
- Eghe, L., & Rousseau, R. (2000). Aging, obsolescence, impact, growth, and utilization: Definitions and relations. *Journal of the American Society for Information Science*, 51(11), 1004–1017.

- Glänzel, W., & Gorraiz, J. (2015). Usage metrics versus altmetrics: Confusing terminology? *Scientometrics*, 3(102), 2161–2164.
- Gorraiz, J., Gumpenberger, C., & Schlögl, C. (2014). Usage versus citation behaviours in four subject areas. *Scientometrics*, 101(2), 1077–1095.
- Gosnell, C. F. (1944). Obsolescence of books in college libraries. *College and Research Libraries*, 5, 115–125.
- Gross, P. L. K., & Gross, E. M. (1927). College libraries and chemical education. *Science*, 66, 385–389.
- Guerrero-Bote, V. P., & Moya-Anegón, F. (2014). Relationship between downloads and citations at journal and paper levels, and the influence of language. *Scientometrics*, 101(2), 1043–1065.
- Jung, Y., Kim, J., & Kim, H. (2013). Stm e-journal use analysis by utilizing kesli usage statistics consolidation platform. *Collnet Journal of Scientometrics & Information Management*, 7(2), 205–215.
- Jung, Y., Kim, J., So, M., & Kim, H. (2015). Statistical relationships between journal use and research output at academic institutions in South Korea. *Scientometrics*, 103(3), 751–777.
- Kurtz, M. J., & Bollen, J. (2010). Usage bibliometrics. *Annual Review of Information Science and Technology*, 44, 3–64.
- Ladwig, J. P., & Sommese, A. J. (2005). Using cited half-life to adjust download statistics. *College and Research Libraries*, 66(6), 527–542.
- Line, M. B., & Sandison, A. (1975). Practical interpretation of citation and library use studies. *College and Research Libraries*, 36(5), 393–396.
- Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology*, 56(10), 1088–1097.
- O' Leary, D. E. (2008). The relationship between citations and number of downloads in Decision Support Systems. *Decision Support Systems*, 45(4), 972–980.
- Peng, T. Q., & Zhu, J. J. (2012). Where you publish matters most: A multilevel analysis of factors affecting citations of internet studies. *Journal of the American Society for Information Science and Technology*, 63(9), 1789–1803.
- Schloegl, C., & Gorraiz, J. (2010). Comparison of citation and usage indicators: The case of oncology journals. *Scientometrics*, 82(3), 567–580.
- Schloegl, C., & Gorraiz, J. (2011). Global usage versus global citation metrics: The case of pharmacology journals. *Journal of the American Society for Information Science and Technology*, 62(1), 161–170.
- Schloegl, C., Gorraiz, J., Gumpenberger, C., Jack, K., & Kraker, P. (2014). Comparison of downloads, citations and readership data for two information systems journals. *Scientometrics*, 101(2), 1113–1128.
- Subotic, S., & Mukherjee, B. (2014). Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles. *Journal of Information Science*, 40(1), 115–124.
- Wan, J. K., Hua, P. H., Rousseau, R., & Sun, X. K. (2010). The journal download immediacy index (DII): experiences using a Chinese full-text database. *Scientometrics*, 82(3), 555–566.
- Wang, X., Mao, W., Xu, S., & Zhang, C. (2014a). Usage history of scientific literature: Nature metrics and metrics of Nature publications. *Scientometrics*, 98(3), 1923–1933.
- Wang, X., Peng, L., Zhang, C., Xu, S., Wang, Z., Wang, C., et al. (2013a). Exploring scientists' working timetable: A global survey. *Journal of Informetrics*, 7(3), 665–675.
- Wang, X., Wang, Z., Mao, W., & Liu, C. (2014b). How far does scientific community look back? *Journal of Informetrics*, 8(3), 562–568.
- Wang, X., Wang, Z., & Xu, S. (2013b). Tracing scientist's research trends realtimely. *Scientometrics*, 95(2), 717–729.
- Wang, X., Xu, S., & Fang, Z. (2016). Tracing digital footprints to academic articles: An investigation of PeerJ publication referral data. *arXiv preprint arXiv:1601.05271*.
- Wang, X., Xu, S., Peng, L., Wang, Z., Wang, C., Zhang, C., et al. (2012). Exploring scientists' working timetable: Do scientists often work overtime? *Journal of Informetrics*, 6(4), 655–660.