Usage and Citation Metrics for Ranking Algorithms in Legal Information Retrieval Systems[*]

Gineke Wiggers[1][0000−0002−1513−2212] and Suzan Verberne[2][0000−0002−9609−9505]

[1] eLAW - Centre for Law and Digital Technology, Leiden University
Steenschuur 25, 2311 ES Leiden, The Netherlands
`g.wiggers@law.leidenuniv.nl`
[2] LIACS - Leiden Institute of Advanced Computer Science, Leiden University
Leiden, The Netherlands
`s.verberne@liacs.leidenuniv.nl`

Abstract. Usage and citation metrics are indicators of interest in documents by users in information retrieval (IR) systems. Our aim is to create an impact relevance variable for ranking functions in legal IR systems. In this paper, we study the development of user clicks and citation counts over time for documents in a Dutch legal search engine, and the relation between citation counts and user clicks. Based on a set of 95,074 documents we find a Spearman correlation with 24 months of citation data of $\rho = 0.39$ after 1 month of usage data, and $\rho = 0.47$ after 12 months.

Keywords: Legal Information Retrieval · Ranking · Bibliometric-enhanced Information Retrieval

1 Introduction

Legal Information Retrieval (IR) systems still rely heavily on algorithmic and topical relevance. This does not encompass all aspects of relevance for the user, as described by Saracevic [15], Van Opijnen and Santos [17], and Wiggers et al. [19]. The impact of a document can also be seen as a form of relevance. For scientific documents, citations are commonly used as a proxy for impact. Citations in legal publications, however, may a different meaning than academic citations [6], for example because legal publications do not only impact scholars but legal practitioners as well. Therefore, usage of documents (clicks in the search engine) could be an additional source of information for measuring impact on readers [7], and thereby another flavour of relevance [13]. For that reason we aim to introduce of a ranking variable for legal IR systems that incorporates both usage and citations as indications of interest for users.

2   G. Wiggers et al.

This paper presents the analysis of usage and citation data in a legal search engine, for the future purpose of transforming raw counts into a ranking variable that reflects the impact relevance of documents for the users. We address the following research questions:

1. How soon after publication are citation metrics informative to be included in ranking algorithms?
2. To what extent are usage and citations correlated?

In this research we will analyse citation and usage (click) data from the Legal Intelligence IR system, the largest legal IR system in the Netherlands. The contributions of this research are (a) an analysis of legal citation data to determine how soon after publication citation data is informative to be included in a ranking function, and (b) an analysis of the correlation between usage and citations of legal publications.

2   Background

2.1   Relevance

In IR, the theory of relevance has several dimensions, including algorithmic relevance, topical relevance, cognitive relevance, situational relevance, and, in particular for legal IR, bibliographic relevance [15, 17, 19]. The practice however, is that legal information retrieval systems rely heavily on algorithmic and topical relevance.[3] As Barry [1] points out, this may lead to poor user satisfaction.

2.2   Citations and Usage

Another form of relevance can be found in the impact of the document. The use of citations as a proxy for impact was introduced by Eugene Garfield [5]. Kurtz and Henneken describe it as: "The measurement of an individual's scholarly ability is often made by observing the accumulated actions of individual peer scholars. A peer scholar may vote to honor an individual, may choose to cite one of an individual's articles, and may choose to read one of an individual's articles." [9]

As described by Wiggers and Verberne [18], citations in legal publications do not measure impact in the same way as in the hard sciences. In the hard sciences, citations are thought to measure impact on the academic community. But because legal scholars and legal professionals read and cite each others publications, usage and citation metrics indicate impact on not only legal scholars, but on the legal field as a whole.

---

[3] As discussed by Mart [11] the algorithms of commercial legal information retrieval systems are trade secrets, but her work and information obtained from Lexis [10] and the system used in our previous research [19], Legal Intelligence, indicate that algorithmic and topical relevance are still the main focus.

To measure this broader impact, citations alone do not provide enough information, since not all legal professionals are also author, and the impact of the publications on these legal professionals will not be visible in citations. Garfield himself acknowledged that: "there are undoubtedly highly useful journals that are not cited frequently" [4, p. 476] "that does not mean that they are therefore less important or less widely used..." [4, p. 476] An example he uses is Scientific American, a journal readers read to keep up to date, but tend not to cite. The impact of these sources can not be captured by citation measurement. Haustein [7] describes that though not all readers cite, these non-citing readers might still use the documents in their daily work. Piwowar [13] describes this as different flavors of impact. This motivates the combination of citation and usage counts in our impact relevance variable for legal IR.

2.3 Legal Information Retrieval

Legal IR systems are a hybrid of academic search and professional search, as they are used by both legal scholars and legal practitioners [18]. This is further reason why not only the impact on the academic community should be considered when using impact metrics in ranking for legal IR, but impact on the legal field as a whole, as both groups will be using the system.

As Kousha and Thelwall [8] indicate, when assessing impact in book-based disciplines, citations in and of books should be included in the citation analysis. The legal domain is one where books still play an important role in the transferring of knowledge [16]. For this reason, books are included in legal IR systems and will be included in this research.

2.4 Correlation Usage and Citations

For the above reasons, we aim to combine metrics for document usage and citations. However, because some readers are also authors, a correlation between usage and citations counts is expected. Perneger [12] researched the correlation between usage and citations in the medical domain (a domain which, like the legal domain, has a largely interwoven group of scholars and practitioners), and found a Pearson correlation coefficient of $r = 0.50$ ($p < 0.001$) between the two variables. Brody et al. [3], using arXiv data, found Pearson correlation coefficients of $r = 0.270$ between 1 month of usage data and 2 years of citation data and $r = 0.440$ between 2 years of usage data and 2 years of citation data. Haustein [7, p. 333] concludes: "medium correlations confirm that downloads measure a different impact than citations. Nonetheless, these should be seen as complementary indicators of influence because a fuller picture of impact is provided if both are used."

4   G. Wiggers et al.

3   Methods

3.1   Data Collection

The KNAW, the Royal Netherlands Academy of Arts and Sciences, has indicated that it can take up to two years for documents in the humanities gather sufficient citations for research evaluation [14]. For this reason, we decided to use documents from the first half of 2017 for our analysis.[4]

From the document index of the legal search engine, we select all documents that were added to the system between January 1st and June 30th 2017. Documents have both a publication date and a date on which they were added in the system. In most cases, these dates will be the same, but in some cases there are small differences (for example when the document is published in folio before being made available online or vice versa). To be able to accurately asses the usage of the documents, we decided to use the date added rather than the publication date. This resulted in a set of 536,635 documents.

For each of these documents, we retrieve a unique document identifier and a reference number. Using the reference number, we conduct a search in the document index, counting how many documents refer to this document in their main text. Using the document identifier, we extract the usage data (clicks) from the search engine logs.

3.2   Data Processing

Citation data After accumulating all citations (excluding self-citations), we see that only 104,048 documents have received citations. This means that $(536,635 - 104,048 =)$ 432,587 documents (81%) did not receive any citations. This might be because some document types (such as books) do not have a reference number that can easily be used for citation extraction[5]. However, based on citations in other fields, it is also to be expected that a large number of documents does not generate citations.[6] Of the documents with citations, 68,781 documents have only one citation. For the analysis how citations aggregate over time, we will use the remaining 35,267 documents that have gathered more than 1 citation since publication (since documents with 0 or 1 citation(s) will generate a flat line). We look at the period up until 24 months after publication.

Usage data After accumulating all usage data for up to 24 months after publication, we see that only 131,494 documents have received usage actions. This means that $(536,635 - 131,494 =)$ 405,141 documents (75%) did not receive any clicks. Similar to the citations above, this highly skewed distribution is as expected. For the analysis of how usage changes over time, we look at documents that have gathered more than 1 usage interaction (click) since publication. This gives us a set of 95,074 documents.

---

[4] Usage data is available from 2017 on. For that reason, it was not useful to use older documents (before 2017).

[5] But the citations mentioned in the books are available.

[6] See, for example Brody et al. [3]

Correlation data To calculate the correlation between usage and citations, we used the document identifiers from the usage data. For these documents, we retrieved the total number of citations after 24 months. We compute the Spearman correlation between the usage at each month and the citations after 24 months. We chose this approach since it is possible that documents that are read are not cited, whilst it is less likely that documents are cited that are not read. Note that the correlation coefficients are possibly higher if the documents that have zero or one click(s) are included.

4   Results and Analysis

4.1   Development of Citation Counts Over Time

To analyse how soon after publication citation data becomes relevant for use in ranking algorithms, we computed the time between the month the cited document became available and the month the citing documents became available[7]. Because we are interested in the pattern of aggregation of citations, this plot only shows documents that have more than 1 citation. We plotted the aggregated number of citations over time for the mean, median, first and third quartile.
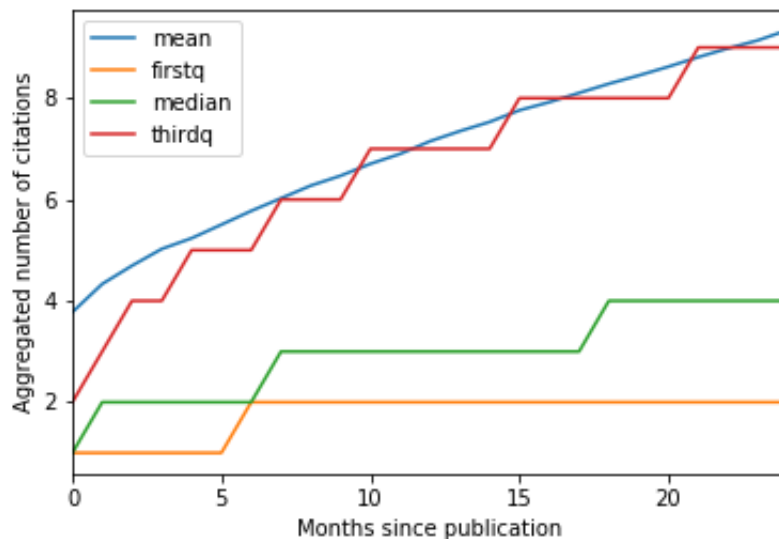

Fig. 1. Aggregated citations per month after publication

---

[7] For explanation why the date added is used rather than the publication date see Methods section

6   G. Wiggers et al.

Figure 1 shows that documents gather citations quite quickly, and are informative for use in IR much more quickly than after 2 years, as the KNAW suggested. Even the documents with low number of citations receive their first citations in the first months after publication.

The data shows a large difference between the mean and the median. This is likely caused by a large number of documents with limited citations, and a small number with a very large number of citations. This is as expected based on bibliometric theory [2, 3], which states that citation counts often show long-tail distributions.
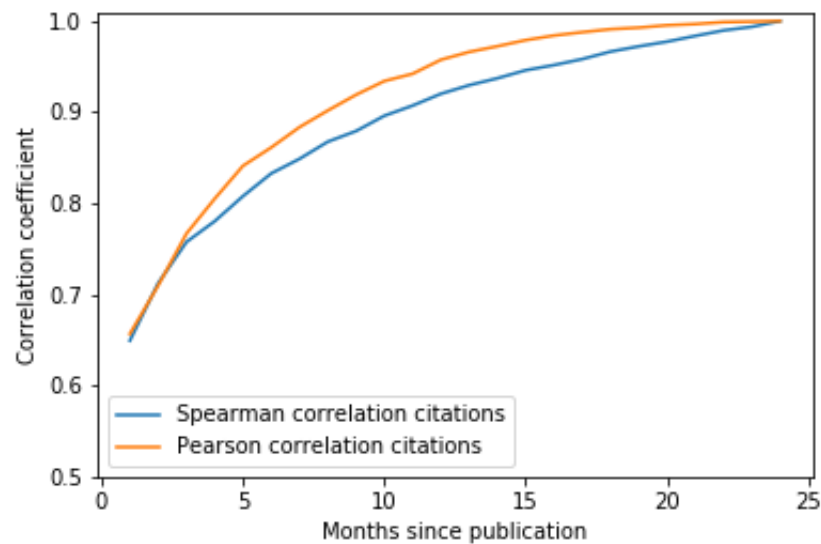
Fig. 2. Correlation per month of citations up to and including that month with citations after 24 months

Figure 2 shows the correlation between citation counts at each month after the documents are made available and citation counts at 24 months. A month after publication (for documents published in January 2017 this means citation data up until the end of February 2017, since some documents were published at the very end of January) we find a Spearman correlation of $\rho = 0.65$. We chose Spearman correlation because the data, like all citation data, does not follow a normal distribution but a long-tail distribution with extreme outliers. However, as figure 2 shows, initially the Pearson correlation gives similar results.

Two months after the cited document has become available, the Spearman correlation is $\rho = 0.71$. For research evaluation purposes, this correlation may not be sufficient. But for information retrieval, where we would like to be able to

reasonably estimate the impact of a document as early as possible, a correlation of $\rho = 0.71$ at two months is valuable. It is also possible to update the data regularly[8], so increases in citation counts can be incorporated as they occur.

4.2   Development of Usage Over Time

Similar to the citation data, we see a difference between the mean (7.11 after 1 month) and the median (2.00 after 1 month) in figure 3.[9] This is again caused by a long-tail distribution, and is seen throughout the 24 months.
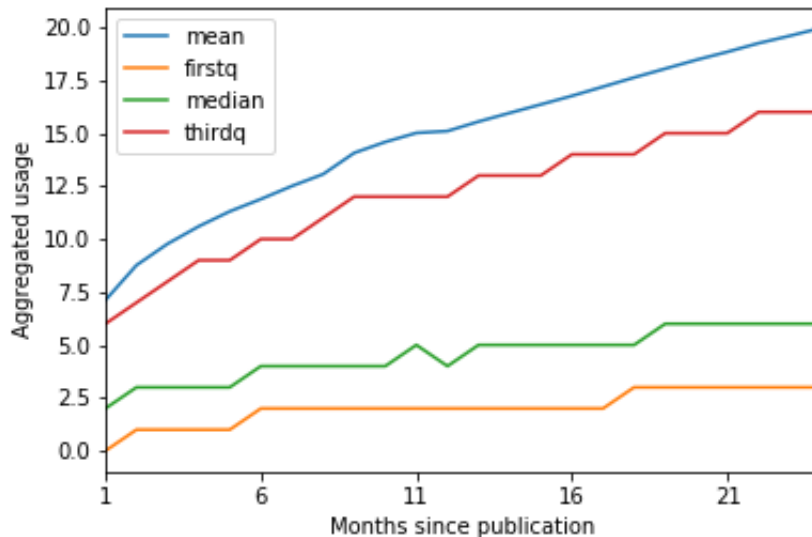
Fig. 3. Aggregated usage per month after publication

Figure 4 shows a Spearman correlation between usage after 1 month and usage after 24 months of $\rho = 0.63$. The Spearman correlation between usage after two months and usage after 24 months is $\rho = 0.69$.[10]

---

[8] e.g. monthly

[9] The bump visible in the line of the mean between 9 and 11 months, and the decrease visible in the line of the median at 12 months, are the result of errors in the underlying data. In future work, we will research the cause of these errors and correct for them.

[10] The bump visible in the line of the Spearman correlation between 9 and 11 months is the result of errors in the underlying data, as was also visible in figure 3. In future work, we will research why these data errors are visible only in the Spearman correlation, and not in the Pearson correlation.

8   G. Wiggers et al.

In figure 4 the difference between the Spearman correlation and Pearson correlation is more pronounced. In this research we work with the Spearman correlation because the data has a longtail distribution with extreme outliers.
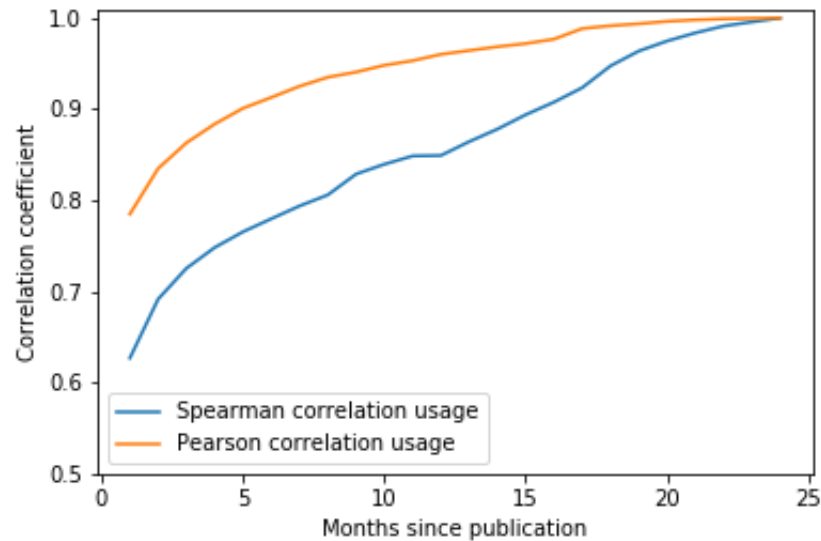
Fig. 4. Correlation per month of usage up to and including that month with usage after 24 months
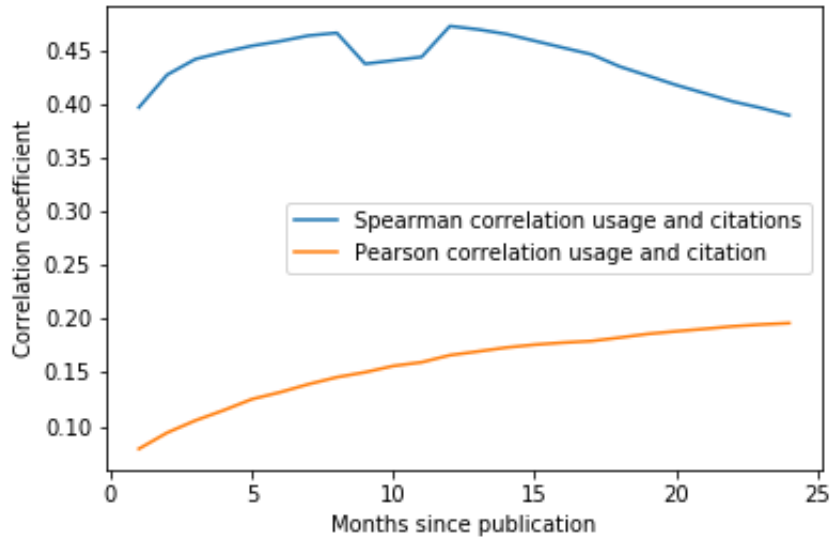
4.3   Correlation Between Usage and Citation Counts

We compute the Spearman correlation between the usage at each month and the citations after 24 months (95,074 documents, see Section 3.2). Figure 5[11] shows both the Spearman and Pearson correlation coefficients, though we focus on the Spearman correlation because of the longtail distribution of the data with extreme outliers, and because the order of magnitude for usage is different than for citations.

The Spearman correlation between 1 month of usage and 24 months of citations is $\rho = 0.39$. The highest correlation found between usage and 24 months of citations is $\rho = 0.47$ after 12 months.

---

[11] The dip visible in the line of the Spearman correlation between 9 and 11 months is the result of errors in the underlying data, as was also visible in figure 3 and figure 4. In future work, we will research why these data errors are visible only in the Spearman correlation, and not in the Pearson correlation.

Fig. 5. Correlation per month of usage up to and including that month with citations after 24 months

The development of the correlation between usage and citations is as expected. Brody et al. [3] estimated that the increase of the correlation between what usage and citations is not linear with time, but reaches it's highest point after about 6-7 months.

As indicated by Haustein [7], medium positive correlations (in this research between $\rho = 0.39$ and $\rho = 0.47$), show that citations and usage measure different flavors of impact.

4.4   Using Citations and Usage in Ranking Algorithms

Given the two different flavors of impact that usage and citations represent, both variables have to be considered in order to include relevance impact in a ranking algorithms. However, since usage and citations are correlated (albeit moderately), it would be unwise to add the two factors as separate boost factors in the ranking algorithm of the search engine, since that would overestimate the impact of the publication. Possible solutions are (a) taking the average of the two impact values, (b) taking the lowest of the two values, or (c) taking the highest of the two values. In a large number of situations the average would give an adequate representation of the impact of a document. However, with the example of the Scientific American in mind, which is highly read but not often cited, there is a risk of disregarding sources which readers use to keep up to date with the field. In Dutch legal publications this might be overviews ('Kronieken')

10   G. Wiggers et al.

of recent remarkable case law. Using the lowest of the two values would also disregard these publications. For that reason we choose the highest of the two scores to incorporate as variable in the ranking function, thereby allowing both documents that are used for research and documents that are used to keep up-to-date to appear high in the ranking.

5   Conclusions

This research demonstrates that legal publications gather citations from the moment they are published. We find a correlation of $\rho = 0.71$ between the citation counts at 2 months and the citation counts at 24 months after publication of the document. We find a correlation of $\rho = 0.69$ between the usage after 2 months and the usage after 24 months after publication of the document. This suggests the early citation and usage data can be used as a predictor for later citations/usage for ranking in legal IR.

Usage and citations show different forms of impact but are correlated (Spearman's correlation between $\rho = 0.39$ and $\rho = 0.47$). This means that usage and citations measure different flavors of impact. This also means that a usage boost should not be added on top of a citation boost, since that would overestimate the impact of certain publications. As solution we suggest to take the highest of the two values.

In future work we will incorporate these metrics in a ranking algorithm. This will include an impact relevance variable that has limited influence at the beginning, when the correlation with later usage/citations may not yet be reliable enough, and increases in influence as the data becomes more reliable.

References

1. Barry, C.: User-defined relevance criteria: An exploratory study. Journal of the American Society for Information Science **45**(3), 149–159 (1994)
2. Bornmann, L., Bowman, B.F., Bauer, J., Marx, W., Schier, H., Palzenberger, M.: Bibliometric standards for evaluating research institutes in the natural sciences. Beyond bibliometrics: harnessing multidimensional indicators of scholarly impact p. 201 (2014)
3. Brody, T., Harnad, S., Carr, L.: Earlier web usage statistics as predictors of later citation impact. Journal of the American Society for Information Science and Technology **57**(8), 1060–1072 (2006)
4. Garfield, G.: Citation analysis as a tool in journal evaluation. Science **178**(4060), 471–479 (1972)
5. Garfield, G.: Citation Indexing: its theory and application in science, technology, and humanities. John Wiley & Sons, Inc., New York, NY (1979)
6. Gingras, Y.: Criteria for evaluating indicators. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact pp. 109–125 (2014)
7. Haustein, S.: Readership metrics. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact p. 327 (2014)

8. Kousha, K., Thelwall, M.: Web impact metrics for research assessment. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact p. 289 (2014)
9. Kurtz, M., Henneken, E.: Measuring metrics - a 40-year longitudinal cross-validation of citations, downloads, and peer review in astrophysics. Journal of the Association for Information Science and Technology **68**, 695–708 (2017)
10. LexisNexis: LexisNexisLawSchools. Understanding the technology and search algorithm behind Lexis Advance (2013), https://www.youtube.com/watch?v=bxJzfYLwXYQ&feature=youtu.be
11. Mart, S.: The algorithm as a human artifact: Implications for legal [re]search. Law Library Journal **109**, 387 (2017)
12. Perneger, T.V.: Relation between online "hit counts" and subsequent citations: prospective study of research papers in the bmj. Bmj **329**(7465), 546–547 (2004)
13. Piwowar, H.: Flavors of research impact through# altmetrics. Research Remix **31** (31)
14. Royal Netherlands Academy of Arts and Sciences: Judging research on its merits – an advisory report by the council for the humanities and the social sciences council (2005)
15. Saracevic, T.: Relevance reconsidered, information science: Integration in perspectives. In: Proceedings of the Second Conference on Conceptions of Library and Information Science. pp. 201–218 (1996)
16. Stolker, C.: Rethinking the Law School: Education, research, outreach and governance. Cambridge University Press (2015)
17. Van Opijnen, M., Santos, C.: On the concept of relevance in legal information retrieval. Artificial Intelligence and Law **25**, 65–87 (2017)
18. Wiggers, G., Verberne, S.: Citation metrics for legal information retrieval systems. In: Proceedings of the 8th International Workshop on Bibliometric-enhanced Information Retrieval (BIR), co-located with the 41st European Conference on Information Retrieval (ECIR 2019), Cologne, Germany, April 14th, 2019. pp. 39–50. CEUR Workshop Proceedings (2019)
19. Wiggers, G., Verberne, S., Zwenne, G.J.: Exploration of intrinsic relevance judgments by legal professionals in information retrieval systems. In: Proceedings of the 17th Dutch-Belgian Information Retrieval workshop. pp. 5–8 (2018)