Bibliometric-enhanced legal information retrieval: combining usage and citations as flavors of impact relevance

Wiggers, G.; Verberne, S.; Loon, W.S. van; Zwenne, G.J.

Received: 7 April 2023 | Accepted: 14 May 2023

DOI: 10.1002/asi.24799

RESEARCH ARTICLE

JASIST WILEY

Bibliometric-enhanced legal information retrieval: Combining usage and citations as flavors of impact relevance

Gineke Wiggers[1,2] | **Suzan Verberne**[3] | **Wouter van Loon**[4] | **Gerrit-Jan Zwenne**[5]

[1]eLaw—Centre for Law and Digital Technology, Leiden University, Leiden, The Netherlands

[2]Wolters Kluwer (Legal Intelligence), Alphen aan den Rijn, The Netherlands

[3]LIACS—Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands

[4]Department of Methodology and Statistics, Leiden University, Leiden, The Netherlands

[5]eLaw—Center for Law and Digital Technologies, Leiden University, Leiden, The Netherlands

Correspondence
Suzan Verberne, LIACS—Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands.
Email: s.verberne@liacs.leidenuniv.nl

Abstract

Bibliometric-enhanced information retrieval uses bibliometrics (e.g., citations) to improve ranking algorithms. Using a data-driven approach, this article describes the development of a bibliometric-enhanced ranking algorithm for legal information retrieval, and the evaluation thereof. We statistically analyze the correlation between usage of documents and citations over time, using data from a commercial legal search engine. We then propose a bibliometric boost function that combines usage of documents with citation counts. The core of this function is an impact variable based on usage and citations that increases in influence as citations and usage counts become more reliable over time. We evaluate our ranking function by comparing search sessions before and after the introduction of the new ranking in the search engine. Using a cost model applied to 129,571 sessions before and 143,864 sessions after the intervention, we show that our bibliometric-enhanced ranking algorithm reduces the time of a search session of legal professionals by 2 to 3% on average for use cases other than known-item retrieval or updating behavior. Given the high hourly tariff of legal professionals and the limited time they can spend on research, this is expected to lead to increased efficiency, especially for users with extremely long search sessions.

1 | INTRODUCTION

Legal information retrieval is a form of professional information retrieval. Legal research (hereafter "research") can consist of several different aspects, for example, finding the current law that governs a specific situation (in the form of legal codes or case law), finding information on how to interpret the law in the situation of their client (interpretative literature), and finding arguments as to how the law should be (normative legal scholarship). Legal information retrieval distinguishes itself from other types of information retrieval (IR) by (1) the requirements of the users, and (2) the documents in the collection.

For legal users, missing an item that turns out to be valuable has a very high negative impact whereas in

Gineke Wiggers is a PhD candidate at Leiden University and Business Analyst at Legal Intelligence.

searching scientific articles it has a low negative impact (Konstan et al., 1997). False positives (reading an irrelevant article) have a medium negative impact, and correct negatives (correctly removing articles from the results list) have a low/medium positive impact (a cost of 5 min and medium/high impact respectively for search for scientific articles).

The collections of legal IR systems have diverse document types. Lengths may vary 57 words[1] to 161 pages.[2,3] There is also a variation from the structured form of legal codes to the free form of blog posts. It is noticeable that the reliance on legal codes and previous cases for argumentation means that there are a lot of references in legal documents. The various references can be mapped to provide an overview of the relations between documents.

It appears though, that this information is not always used to the fullest extent possible (Geist, 2016). The scholarly field of IR focuses a lot on state-of-the-art web-search, where using citations in relevance ranking algorithms has been a standard (Manning et al., 2008) since the introduction of PageRank (Page et al., 1999). But these state-of-the-art techniques are not always implemented IR systems other than web-search (e.g., legal or archeological[4]), who may still rely heavily on older methods.

This article covers the analysis of usage and citation data in a legal IR system. In preliminary work (Wiggers & Verberne, 2020) we addressed two data oriented questions: (1) How soon after publication are citation metrics a reliable predictor of total citations for use in ranking variables? and (2) To what extent are usage and citations correlated? We present the updated results of that data analysis in Section 3. This article builds further on those findings and describes the process of balancing the usage and citations to create a bibliometric ranking variable, as well as balancing this variable with other existing variables in the ranking algorithm, such as a term-frequency based variable. Specifically, we propose a bibliometric ranking variable in Section 4.1, along with a cost based evaluation metric in Section 4.2, the results of which are presented in Section 5.

The term "ranking variable" or "boost function" refers to one factor in the relevance ranking, whereas the term "ranking algorithm" refers to the whole model for relevance ranking. In this article, "ranking" or "relevance ranking" is used to refer to the presenting of search results in decreasing order of likelihood of relevance for the user, where relevance is defined in Section 2.1. Ranking on, for example, date falls outside the scope of this article.

This article addresses the following research question: *Can bibliometrics improve common ranking algorithms in legal information retrieval?* We quantify the improvement of a ranking algorithm as a reduction in the *cost* of user sessions, where the cost is measured as an aggregate of the actions a user does during a session (querying, reformulating, filtering, inspecting documents) multiplied by the average time associated with these actions (see Section 4.2). The contributions of this article compared to prior work are: (1) we show that ranking algorithms in legal IR can be improved using bibliometrics; (2) we show how such a bibliometric ranking variable can be created; and (3) we set an example of cost-based evaluation of live, domain-specific search engines.

In this article, we use data from the Legal Intelligence IR system, the largest legal IR system in the Netherlands. This IR system is based on Apache SOLR. However, we aimed to describe all the steps taken to tune the boost function (e.g., the speed with which citations gather), so that the bibliometric boost factor can also be tuned and evaluated for other domain specific IR systems. Depending on the technology behind the system, the step of hard-coding the boost factor to use only the highest of the two scores can also be replaced by applying a machine learning algorithm (e.g., learning to rank) to the ranking algorithm as a whole and inputting the two scores as two separate variables. The learning to rank function will then determine the optimal way to use these variables as a boost in that situation.

2 | BACKGROUND

2.1 | Impact and relevance

Relevance, in the broadest sense, is a term used to describe "Connection with the subject or point at issue; relation to the matter in hand." (OED, 2019). In everyday language, it is used to describe the effectiveness of information in a given context (Saracevic, 1975, p. 203) It can also be considered as "a measure of the effectiveness of a contact between a source and a destination in a communication process" (Saracevic, 1996, p. 325).

The theory of relevance has several dimensions (Saracevic, 1996, 2007), including algorithmic relevance (the matching of query terms with the terms in the documents returned), topical relevance ("aboutness" [Bruza & Huibers, 1996], which means that the topic of the query has to match the topic of the results returned), cognitive relevance (the relation between what the user already knows and what is in the document, e.g., novelty), situational relevance (the relation between the task that the user is trying to complete, and the information in the document, e.g., how the information is presented), motivational relevance (the relation between the intent of the

user and the documents returned, e.g., satisfaction), and, for legal IR, bibliographic relevance ("isness," "the degree to which the documents retrieved actually are those requested by the user"; Van Opijnen & Santos, 2017). For information to be as effective as possible, it would satisfy all these dimensions.

For scientific documents, citations are commonly used as a proxy for *impact* (Garfield, 1979). The use of citations and statistical methods to analyze the impact of books, articles and other publications is referred to as bibliometrics. Garfield describes the use of references "to support, illustrate, or elaborate on a particular point" (Garfield, 1979, p. 23). This understanding of impact relates strongly with Saracevic's theory of relevance, stating: "Communication of knowledge is effective when and if information that is transmitted from one file results in changes in another. Relevance is the measure of these changes." (Saracevic, 1996, p. 325). A citation can be seen as a token that the cited document has resulted in changes in the citing document, and has therefore been both relevant and impactful.

When placing citation measures in the schema of relevance of Van Opijnen and Santos (Van Opijnen & Santos, 2017), it is important to consider the role of citations. Snel (2016) analyses citations in the legal domain and defines three reasons for citations: to provide context for the research, to legitimize statements made in the research, and to allow others to check the quality of the research. We argue that the reasons to cite embody novelty, utility, authority of the source or author, legal hierarchy and bibliographical "isness" (Van Opijnen & Santos, 2017). When placing it in the schema of Van Opijnen and Santos it touches cognitive, situational, bibliographical and domain relevance,[5] while not covering any of these spheres of relevance completely. Garfield acknowledges that citation metrics as a measure of importance are incomplete by stating "... there are undoubtedly highly useful journals that are not cited frequently" (Garfield, 1972, p. 476).

From an IR perspective, Oard and Kim (2001) have created a framework that describes the different types of user behavior that could be monitored for implicit feedback on the relevance of documents. They have subdivided the behaviors into four groups: examine (read, view, select), retain (print, bookmark, save), reference (copy-paste, reply, cite) and annotate (mark up, rate, publish). Haustein et al. (2016), expanded upon by Erdt et al. (2016) from a bibliometric perspective, created a framework for user interactions with search results (called "acts"), and have three groups with increasing level of engagement: accessing, appraising and applying. Accessing covers views (part of the examine category for Oard and Kim) as well as downloads and prints (part of the retain category for Oard and Kim). Appraisal acts represent comments and links (part of the reference category for Oard and Kim) and rating (part of the annotate category for Oard and Kim). The applying acts represent citations (part of the reference category for Oard and Kim).

Usage of documents (clicks in the search engine) could be an additional source of information for measuring impact on readers (Haustein, 2014; Piwowar, 2012). Clicks indicate perceived relevance by the user based on the document representation in the search results list. Clicks are part of the examine category from Oard and Kim, and part of the accessing category from Haustein et al. Like citations it touches upon multiple spheres of relevance without covering any of those completely. Altmetrics (in this research defined as "short for alternative (to citation) metrics—and as such a misnomer—refers to a new group of metrics based (largely) on social media events relating to scholarly communication"; Haustein et al., 2016; e.g., saving documents or sharing documents with other users) were also considered, but the data available for these metrics was insufficient to provide conclusive results due to the modest user group. Library holdings (Maleki, 2022) were considered but only data from academic and public libraries is available,[6] not the data from the libraries in law firms, which represent another part of the user group, leading to incomplete coverage/representation. For that reason, we aim to introduce a ranking variable for legal IR systems that incorporates both usage and citations as indications of impact for users. These two factors have as additional benefit that they are widely available in IR systems, allowing this work to be easily replicated in other systems.

In referring to the role bibliometrics play in ranking algorithms we use the phrase "impact relevance," to refer to the concept described above which encompasses several spheres of impact as defined by Van Opijnen and Santos (Van Opijnen & Santos, 2017) without covering any completely.

2.2 | Legal IR

Legal IR systems still rely heavily on algorithmic and topical relevance.[7] This does not encompass all aspects of relevance for the user, as described above. As Barry (1994) points out, this may lead to poor user satisfaction.

Work has been done on citation analysis in legal documents in general (Giménez-Toledo et al., 2016; Hicks, 2004; Zuccala & Cornacchia, 2016), as well as specifically tailored to the Dutch legal domain (Opijnen van, 2014; Soetenhorst, 2017; Stolker, 2015; Winkels et al., 2011; Winkels et al., 2013, 2014; Winkels & Ruyter

de, 2011), but this work is either done in the context of bibliometrics rather than IR or is restricted to one document type (e.g., only case law or only journal articles).

It has been suggested (Bock, 2000) that the main focus in legal IR should lie on high recall (that all, or nearly all, relevant documents are returned to the user).[8] Manning et al. (2008, p. 156) suggest that paralegals will tolerate a fairly low precision (meaning that the result list includes many irrelevant documents next to the relevant documents) to obtain this high recall. Precision and recall are often presented as a trade-off, where one can only be increased at the expense of the other (Manning et al., 2008, p. 144). For example, by returning more results, and thus also results that may be irrelevant, the user is more certain that they did not miss relevant results. On the other hand, by showing less results, the user is confronted with less irrelevant results.

Geist (2016) observes that although high recall is in theory preferred, the reality of the time pressure that all legal professionals perform under means that precision is required. He calls it the "completeness ideal" and the "research reality."[9]

The "completeness ideal" suggests that legal professionals do not stop their research until they have achieved full recall (seen all relevant documents). But the "research reality" suggests that there is a point where the legal professional is "sure enough" and will stop. Where this stopping point is depends on the user (e.g., a novice versus a senior lawyer, or a general practice lawyer versus a highly specialized lawyer) and the case at hand. A good relevance ranking can provide users with both high recall and high precision. Such a relevance ranking would put the documents the ranking algorithm considers most likely to be relevant (considering all aspects of relevance) at the top of the results list, followed by documents that are less likely relevant.

2.3 | Citations and usage in bibliometrics

The use of citations as a proxy for impact was introduced by Garfield (1979). 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." (Kurtz & Henneken, 2017) Piwowar (2012) describes citations and usage as different flavors of impact.

As Kousha and Thelwall (2014) 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 (Stolker, 2015). Thelwall (2020) states that "Research targeting commercial, government, or non-governmental organisations may be more likely to be cited by grey literature than by journal articles ..." For this reason, as well as the work of Wiggers, Verberne, and Zwenne (2022), we count citations from all documents in the system, including government publications and gray literature.

2.4 | Correlation between usage and citations

Because some readers are also authors, a correlation between usage and citations counts is expected. Priem and Hemminger (2010) considered that in an online world, readership information is readily available and may provide an early alternative to citation metrics for use in researcher evaluation. Perneger (2004) analyzed the correlation between usage and citations in the medical domain (a domain which, like the legal domain, has an 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. (2006), 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 (2014, 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." Rousseau and Ye (2013) therefore propose the term "influmetrics."

2.5 | Usage in evaluation

Next to using clicks as a sign of impact in bibliometrics, clicks are also used as implicit feedback of relevance for the evaluation of IR systems (Oard & Kim, 2001; the *examine behavior* category on the object level). Cooper and Chen (2001) describe how multiple reasons exist for clicking on an article, but all have an implicit assumption of relevance of the item for that particular user at that point in time.

This implicit feedback model can be used to create a test collection, but as Hersh (1994) describes, user satisfaction is not static, but also influenced by the context (situational relevance), meaning test collections alone are not a complete representation of user satisfaction. Baskaya et al. (2012) analyzed search behavior for 60, 90,

and 120 s time frames and found that the more time a user has, the less important the search strategy becomes. But when under time constraint, which is the case for legal professionals, the behavior of the user plays an important role in the retrieval success. This suggests that measuring user satisfaction requires a combination of user success and user behavior clues.

User effort is a user-centric evaluation measure that considers, for example, the time it takes a user to complete a task in an IR system, or the amount of queries the user enters to complete a task (Baskaya et al., 2012; Tamine-Lechani et al., 2010). This is a measure of effectiveness of both the user and the system, often combined with a questionnaire asking the users as to their satisfaction level and their perception of success (e.g., their level of certainty of the answer or whether the user thinks they found all relevant results).

Järvelin et al. (2008) developed the Discounted Cumulative Gain (DCG) measure of ranking quality further to the sDCG, a session based DCG score, where the user effort like reformulating the query is factored into the discounting of the gain.

Järvelin (2009) state that such a cost/benefit model should contain at least the following elements:

- Search key generation cost: the mental effort required to create the query;
- Query execution cost: the cost of conducting the query and waiting for the results;
- Result scan cost: the cost of scanning the results and deciding on the next step (e.g., clicking on the document or reformulating query);
- Next page access cost: the cost of loading the next page of results;
- Relevant document gain: the gain of finding a relevant document.

Järvelin (2009) suggest to sum all costs, and calculate each cost linearly per unit (second, number of occurrences). This sum of costs is then offset to the gains of the relevant documents found.

McGregor et al. (2021) differentiate between load, effort, and cost. Load is taken to refer to the total amount of resources used to complete the task, internally and externally. Effort represents the internal resources spent (e.g., cognitive effort), while cost represents the external resources spent (e.g., time or money). Cost can be measured in time-orientated cost or interaction orientated/count based costs.

Maxwell (2019) has described a complex searcher model. His work distinguishes between good abandonment (where a user is satisfied) and bad abandonment (where a user stops out of frustration). However, as shown by the work of Geist (2016) we can assume that a legal professional will not stop searching until they reach a point in the "research reality" (Geist, 2016) trade-off where they are satisfied enough to stop, given that their professional reputation is on the line.

The underlying assumption of these cost/benefit models is that if the ratio of cost to gain decreases, that is to say the effort required of the user to reach the gain decreases, this increases effectiveness. When expressed in time, effectiveness can be referred to as efficiency. Effectiveness and efficiency are considered to contribute to overall user satisfaction (Dan & Davison, 2016; Hersh, 1994).

3 | DATA ANALYSIS

In this section, we discuss the data analysis that preceded the creation of the bibliometric ranking variable. We address two questions:

1. How soon after publication are citation metrics a reliable predictor of total citations for use in ranking variables?
2. To what extent are usage and citations correlated?

The KNAW, the Koninklijke Nederlandse Akademie van Wetenschappen,[10] has indicated that it can take up to 2 years for documents in the humanities to gather sufficient citations for research evaluation (KNAW, 2005). For this reason, we decided to use documents from the Legal Intelligence system from the first half of 2017 for our analysis.[11]

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. This resulted in a set of 470,938 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. Using the document identifier, we extract the usage data (clicks) from the search engine logs.

3.1 | How soon after publication are citation metrics a reliable predictor of total citations for use in ranking variables?

3.1.1 | Citation data

After accumulating all citations (excluding self-citations), we see that 235,609 documents have received citations.

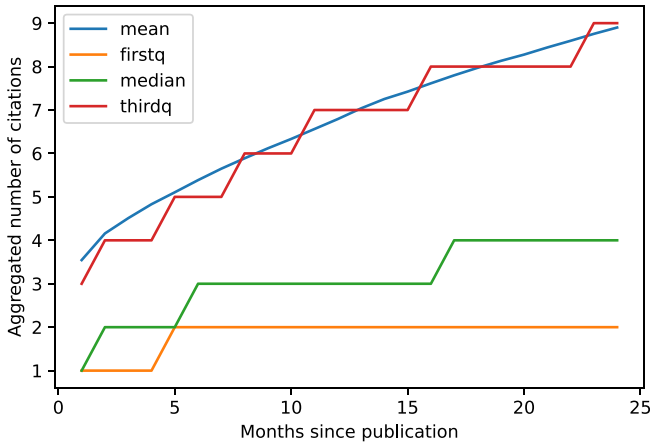

FIGURE 1 Aggregated citations per month after publication.

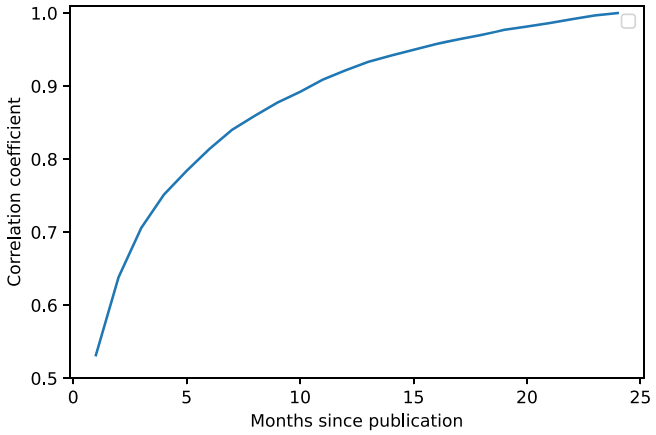

FIGURE 2 Correlation per month of citations up to and including that month with citations after 24 months.

This means that (470,938 − 235,609 =) 235,329 documents (50%) 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.[12] case law reprints that have their own journal article number are counted as a separate document, even though they relate to the same case, because the journal might have added an interpretative note (annotation) to the case. This means that authors citing that particular version might do so because of the annotation.

However, based on citations in other fields, it is also to be expected that a large number of documents does not generate citations.[13] Of the documents with citations, 195,381 documents have only one citation. For the analysis of how citations aggregate over time, we will use the remaining 40,228 documents that have gathered more than 1 citation since publication. We look at the period up until 24 months after publication.

3.1.2 | Analysis

To analyze how soon after publication citation data becomes reliable for use as a predictor of total citations in ranking variables, we computed the time between the month the cited document became available and the month the citing documents became available. Because we are interested in the pattern of aggregation of citations, Figure 1 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.

Figure 1 shows that documents gather citations much more quickly than after 2 years as the KNAW suggested. Even the documents with a low number of citations receive their first citations in the first months after publication. We hypothesize that this might be because case law has a high recency value, or because case law is reprinted or summarized in legal journals. We found no evidence that this is the cause for these early citations. Even when we exclude case law, or exclude news and reprints, we still see these early citations.

In all situations 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 (Bornmann et al., 2014; Brody et al., 2006), which states that citation counts often show long-tail distributions.

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 of the monotonic relationship between citations and usage and because the data, like all citation data, does not follow a normal distribution but a long-tail distribution with extreme outliers.

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 2 months is valuable. It is also possible to update the data regularly,[14] so increases in citation counts can be incorporated as they occur.

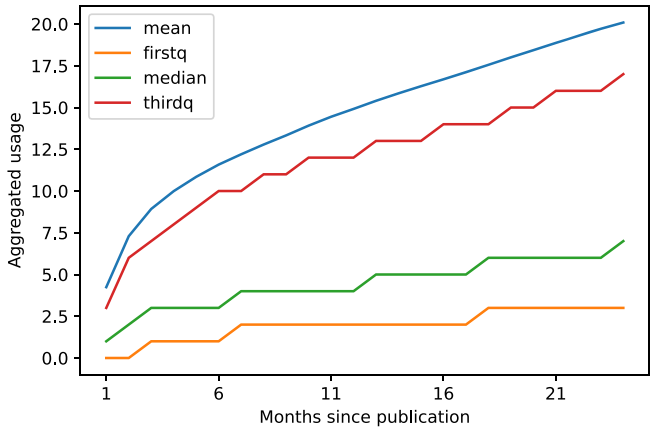

FIGURE 3 Aggregated usage per month after publication.

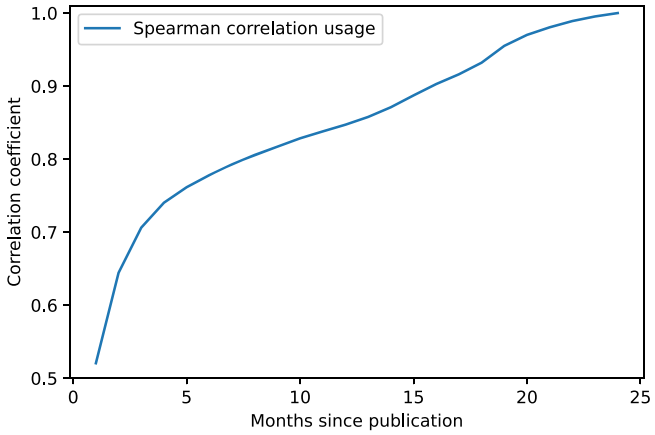

FIGURE 4 Correlation per month of usage up to and including that month with usage after 24 months.

3.2 | To what extent are usage and citations correlated?

3.2.1 | Usage data

After accumulating all usage data for up to 24 months after publication, we see that only 116,637 documents have received usage actions. This means that (470,938 − 116,637 =) 354,301 documents (75%) did not receive any clicks. Like 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 86,717 documents.

Similar to the citation data, we see a difference between the mean (4.24 after 1 month) and the median (1.00 after 1 month) in Figure 3. This is again caused by a long-tail distribution, and is seen throughout the 24 months.

Figure 4 shows a Spearman correlation between usage after 1 month and usage after 24 months of $\rho = 0.52$. The Spearman correlation between usage after 2 months and usage after 24 months is $\rho = 0.64$.

3.2.2 | Analysis

To calculate the correlation between usage and citations, for all documents that have usage, 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 (86,717 documents, see Section 3.1). The Spearman correlation between 1 month of usage and 24 months of citations is $\rho = 0.36$. The highest correlation found between usage and 24 months of citations is $\rho = 0.47$ after 11 months.

If we consider all 470,938 documents, the correlation at 1 month is $\rho = 0.18$ and at 11 months is $\rho = 0.12$. The correlation of usage at 24 months with citations at 24 months is $\rho = 0.07$. However, this also includes documents that have no reference number based on which citations could be retrieved. When we remove those documents, we have a set of 274,663 documents for which citations could be retrieved. With this data set, we have a correlation at 1 month of $\rho = 0.22$, and at 11 months $\rho = 0.24$. The correlation between 24 months of usage and citations at 24 months is $\rho = 0.23$. It is expected that the correlation on the full data set is lower than that of our initial analysis with only documents that have usage actions, given the highly skewed nature of usage and citations. The subset that has usage actions is more likely to also have citations, given that it is not likely a document is cited without being read.

The development of the correlation between usage and citations is as expected. Brody et al. (2006) found that the increase of the correlation between usage and citations is not linear with time, but reaches its highest point after about 6–7 months. In their paper Brody et al. (2006) indicate that after these 6 months the correlation increases by a small amount. The decline in the correlation in Figure 5 can be explained as the usage no longer grows much while the citations do, leading to a lower correlation between the two.

As indicated by Haustein (2014), medium positive correlations (in this article between $\rho = 0.52$ and $\rho = 0.64$), show that citations and usage measure different flavors of impact.

4 | METHODS

In this article, we propose a bibliometric ranking variable. We evaluate this ranking variable with a cost-based model by comparing usage data from before the introduction of this variable, and after the introduction of this variable.

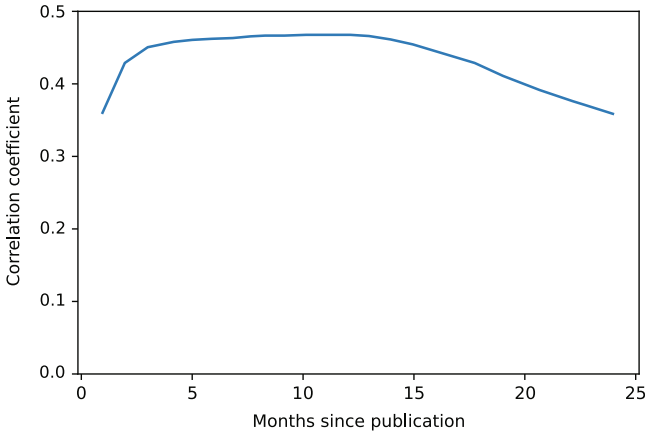

FIGURE 5 Correlation per month of usage up to and including that month with citations after 24 months.

4.1 | Our proposed bibliometric ranking variable

Given the two different flavors of impact that usage and citations represent, both variables are valuable to include as impact relevance factors in a ranking algorithm. 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") of recent remarkable case law. Using the lowest of the two values would also disregard these publications. For that reason the ranking variable determines the highest of the two scores for each individual document, and calculates the document's score with that, 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.

4.1.1 | Normalization

The normalization of the raw citation and usage counts of the publications is based on the NCS (normalized citation score) of the CWTS (the Centre for Science and Technology Studies) (Waltman et al., 2011) and the work of Rehn et al. (2014) on the normalization of citations. Normalization is needed because not every document (type) is likely to gather the same amount of citations. For example, because one law area is larger than another (in our data ranging from 11 to 196,002 documents, see Table 1), or because one document type is more likely to be cited (Snel, 2018). The method normalizes for time (based on year/month of publication), law area (as reported by publisher of the document, including government documents) and document type (Table 2). We decided to apply the same normalization to the usage counts since it regards the same documents, and the effects that influence likelihood of citation are also applicable to the likelihood of usage count.

TABLE 1 Law area sizes in number of documents.

Law area	Number of documents
General	196,002
Tax law	47,120
Intellectual property law	26,830
Contract law	25,764
Private law	20,440
Constitutional law	17,984
Labor law	17,558
Company law	15,435
EU law	13,011
Criminal law and procedure	11,107
Banking law	9794
Environmental and zoning law	9774
Civil procedure	9238
Family law	7563
International public law	6162
Health law	5999
Unassigned	3709
Education law	3700
Insolvency law	2871
Human rights	2698
Telecom/ICT/media law	2476
Competition law	2018
Transportation law	2007
Nationality and migration law	1620
Construction law	1514
Insurance law	1404
Tenancy law	1107
Private international law	549
None	374
Foreign law/religious law	22
Sports law	11

We calculate the normalized citation score by dividing the number of clicks/citations of the document ($citations_d$) by the average number of clicks/citations for documents that have gathered at least one click/citation and that were published in the same month of the same year, in the same law area, with the same document type ($citations_a$):

$$W_d = citations_d / \ citations_a, \quad (1)$$

Our normalization is a slight variation from the NCS in that only documents that have gathered at least one click/citation are counted for the average, as a large number of documents will gather no clicks/citations. Leaving the large number of unused/uncited documents in the denominator would potentially lead to all averages nearing zero.

This method will result in a normalized score that is a positive number or zero. Documents that have no usage or citations themselves are given a score of zero. Documents that have a score of 1 have the same number of citations as the average used/cited document of the group. Documents with a score of 2 have twice the number of citations than the average in the group. To limit outliers caused by the Matthew effect (Merton, 1988) we cap the normalized score at 2. This means that all documents that have a score of 2 or higher, are given a score of 2. It is capped at 2 since the average is 1 and the score cannot be negative. 2 gives the same distance from neutral (1) to positive (2), as there is from neutral (1) to negative (0).

The choice to cap at 2 rather than use a log of the score was made for multiple reasons: (1) the normalized score 1 indicates that the document performed as average. This score of 1 should remain the median, in order to be able to push down lower scoring documents and boost higher scoring documents. (2) A document that is cited more than twice the average number for the group should not necessarily be boosted more than a document that was cited twice the average number for the group. The distinction whether a document was cited more than average or less than average is more important than the number of citations it got. In this sense citation metrics for IR differ from citation metrics for research evaluation. (3) the boost based on citations or usage should never exceed other similarity functions, such as TF-IDF (term frequency-inverse document frequency, see e.g., Manning et al., 2008). A log based normalization risks that outliers exceed the maximum, in our data even a log10 scale exceeded the chosen maximum of 2 for certain extreme outliers.

TABLE 2 Document type sizes in percentage of documents assigned to this type.

Document type	Percentage of documents
Government other[a]	54.0%
Laws	1.5%
Journal other	6.5%
Journal notices	0.2%
Case law	8.9%
Reference guide	8.7%
Books	5.2%
Blogs	1.2%
Newsletters	1.2%
Commentaries	1.2%
Websites	0.7%
Other	10.4%

[a]The category "Government other" refers to government documents that are not case law nor laws, for example, government reports and parliamentary proceedings.

4.1.2 | The bibliometric boost function

To incorporate this usage and citation data in the ranking algorithm, we define an impact variable I that has limited influence in the first period after publication of a document, when the data cannot yet provide a reliable prediction of the impact the document will have, and increases in influence as the data about the document increases and predictions become more reliable. One way to achieve this is to use an initial constant c, and allow the normalized usage and citation scores to impact this over time

$$I_d = c + ((\beta - (s/(t_d + \alpha))) * (W_d - 1)). \quad (2)$$

Thus, to incorporate the increasing influence of citations over time, we take the normalized score of the document (W_d), ranging from 0 to 2 (see Section 4.1.1), and subtract 1, to get a score ranging from -1 to 1.[15] The multiplication by -1 allows the normalized score of the document to add or subtract points from the initial constant c over time.

To model the influence over time, we use a time factor (t_d), the number of days since publication of the document. t_d has to be a positive number. To change the speed with which the variable increases power, we can increase α. The higher α, the steeper the increase in the early days.

To set the maximum value of the variable, we change β or the start value s. This maximum value will have to

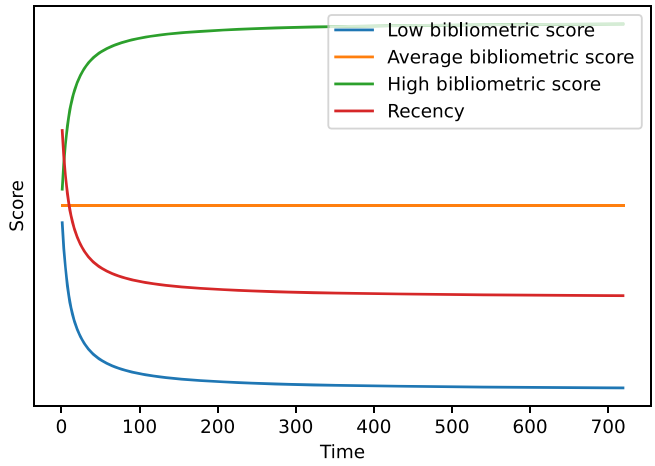

FIGURE 6 A visualization of the ranking variable for a low, average, and high citation/usage score, and the corresponding recency variable.

be capped off at a maximum below the maximum score for text matching (calculated through, e.g., TF-IDF; Manning et al., 2008), to prevent this variable (representing the impact form of relevance) from overruling other variables (representing other forms of relevance).

4.1.3 | The recency variable

To compensate for the limited influence of the citation and usage scores in the beginning, we want the publication date (or enactment date) to weigh heavily in the first month. This gives documents that do not have a reliable prediction of total impact based on citation or usage scores yet the same score as documents with an average citation or usage score. We therefore replace the existing simple recency variable by a new recency variable R_d

$$R_d = c_2 + (s/(t_d + \alpha)). \quad (3)$$

We want this recency variable to decrease in power at the same rate as the citation or usage score increases, to allow for the citation and usage scores to take over, so the

$$s/(t_d + \alpha)$$

is the same as in the bibliometric boost, as shown in Figure 6. The time factor (t_d) again represents the number of days since publication of the document. The c_2 variable is an initial constant that helps tune the recency variable compared with the other variables in the ranking algorithm, such the TF-IDF. This is likely not the same as the c variable in the I_d variable.

The citation and usage information is normalized aggregated information, so it also reflects which documents were important in the past, not just what is important now. The remainder of the recency boost will remain as a tie-breaker.

4.1.4 | The combined ranking function

In the before situation ($A_d + B_d$), the ranking algorithm consisted of the initial ranking function A (a group of additive variables including a term-frequency based variable) to which a simple recency variable B_d was added. Given that the ranking algorithm as a whole is a trade-secret, we are not able to present it here in full. In the after situation ($A_d + R_d + I_d$), A remains the same. Recency variable B is replaced by R, which is defined above. This replacement is needed to ensure that new documents are given the benefit of the doubt. Bibliometric variable I_d is added. The change evaluated is therefore the addition of the bibliometric variable, in tandem with the changes that makes to the recency variable.

4.2 | Evaluation

The aim of the boost function is to improve the ranking so that the most relevant results are presented first. If the users see the relevant results at the top of the results list, it reduces the amount of effort required from the user to find this information and complete their task. It thereby contributes to increasing effectiveness.

To determine whether the boost function achieves this aim we use a cost model inspired by Järvelin (2009) and compare the cost before the introduction of this variable (the intervention) with the cost after the intervention. Azzopardi (2017) and Azzopardi and Zuccon (2016) have used such cost based models to determine the effectiveness of changes to the user interface.

This model will be limited to cost without gain, as there are no relevance judgments to base gain on. However, as shown in Section 2 we can assume that a legal professional will not stop searching until they are satisfied enough (e.g., they have the fact they need, they are certain enough of an interpretation, or they are reassured that no new information has appeared on a topic) to stop.

To calculate the cost related to this assumed gain, we use a time-oriented metric (McGregor et al., 2021). Because of the time pressure legal professionals work under this appears to be the most suitable representation of cost or effort. Because we measure the results in time, we use the term efficiency instead of effectiveness.

The intervention (the moment when the new ranking algorithm went live on production) took place on September 14th 2020 at the close of business day. We took data from the 3 weeks before the intervention (24th of August until 14th of September) and 3 weeks after the intervention (15th of September to 5th of October). The dataset contained data from 27,786 unique users. The average distribution of users in the system based on their affiliation are 42% affiliated to a university, 25% affiliated to the government (including courts and lower government) and 33% affiliated to other organizations (including law firms and legal service providers).

The number of sessions per user ranged from 1 to 201, with an average of 9.8 sessions and a median of 4 sessions.

In the before situation, we have 129,571 sessions, of which 106,852 consist of more than 1 cost-based action (query, click, etc.). Session times (based on max 30 min between two actions) vary between 1 and 82,555 s (or almost 23 h), with a mean of 714.61 and a median of 197.00. In the after situation, we have 143,864 sessions, of which 118,991 consist of more than 1 cost-based action. The session times vary between 1 and 86,125 s (or almost 24 h), with a mean of 774.09 and a median of 205.00. Because of these skewed distributions we work with the median, rather than the mean.

Calculation of cost. We compute the session interaction cost as follows:

- From the system logs we take the date and timestamp, user id, and, where applicable, the position of the document, for events of querying, reformulation of a query, filtering and opening of documents (clicks).
- Using the user id and timestamp, we group different events into sessions, where a group of actions is considered to be one session if there is no more than 30 min (Jansen et al., 2007) between two actions. The difference between a new query and a reformulation of a query is based on the interface and not a determining factor for defining the session.
- Baskaya et al. (2012) use 3 s per action, which they have based on literature. But when we calculated the average time per action based on our data, we found different results, so we are using the average time (per second) found in our data.
- To establish a time cost based on these counts, we multiply the number of occurrences and/or the position of the document by that average time (in seconds) that an action takes. This is done because the cost of some actions are larger than others (e.g., a reformulation takes more time than inspection an additional document). By assigning time cost values to actions, rather than using pure action counts, we can make this distinction visible, especially in situations where the number of occurrences of one action decreases but the other increases.

In the following paragraphs, we specify how we computed the time cost for each action.

Query formulation: Time between login and query. To compute the average time required for query formulation we selected sessions that started with a query (other starting points could be navigation or from an e-mail alert). For those queries, we retrieved the closest login event from the logs, with a maximum of 30 min (our chosen boundary for 1 session). This resulted in 144,479 sessions with a median of 14 s and a mean of 52.68 s.

Inspection: Time between query and first click. To calculate the average time required to inspect a search result, we take from the data query events and click events. From this data we take queries that are followed by a click (as opposed to, for example, a reformulation). We take the time difference between the two events. We then divide the time by the position of the clicked result. We assume that the time spent on inspecting results is spread evenly over the number of items inspected. This gives us an indication of the time spent inspecting each search result, under the assumption of the "cascade model" (Joachims et al., 2017). This model assumes that search engine users scan lists from top to bottom in an exhaustive fashion. This gave us a total of 101,711 query-click pairs, with a median inspection time per result of 5 s and a mean of 17.22 s.

Dwell time: Time between two clicks after a query. The logs do not contain dwell time, as the system redirects a user to the publisher web-page after the click. We have therefore approximated dwell time by using query–click–click triples, without other events in between. This estimation is noisy, as the user may have navigated further in the publisher web-page, or gone to get a coffee. However, there is no reason to assume that the frequency with which this happens changes at the time of evaluation.

For each of these triples, we calculate the individual's inspection time based on the query–click pair. We then take the time difference between the two clicks, and subtract the individual's inspection time multiplied by the number of documents between the first and second click. This gives us an approximation of the time that an individual spends evaluating the first opened document.

We remove any triples in which the difference between the two clicks is less than 1 s, as that is likely a scenario where the user clicked open all results that appeared relevant in new tabs without actually looking at the content of the results before continuing. This led to a total of 16,611 triples, with a median of 24 s and a mean of 73.92 s.

This method does contain a bias, as the click we are examining is the first click in the pair; never the final, perhaps most satisfying, document. The time spent on a document that is not relevant upon further inspection is likely less than the time spent on a relevant document.

Reformulation: The time between the initial query and a reformulation. To determine the average time spent reformulating a query, we searched the data for query–reformulation pairs, with no other actions in between. In these situations, the user enters a query, scans the results list, and reformulates the query to get more suitable results. We found a total of 33,997 pairs with a median of 18 s and a mean of 73.83 s. It is likely that users inspect some of the results before reformulating the query, at a cost of 5 s per item as determined above. However, the data does not tell us how many results a user has inspected before deciding to reformulate the query. The interface shows 20 results per page, but given the time difference of 18 s between the query and the reformulation it is unlikely that the user inspected all 20 results.

Filtering: The time between a query and selecting a filter. To determine the cost of selecting a filter, and narrowing down the search results in that way, we looked at pairs of query–filtering, with no other actions in between. In these situations, the user conducts a query, sees the results list, and refines the results by selecting one or more filters (e.g., document type, year of publication). This led to a total of 26,438 pairs, with a median of 12 s and a mean of 34.60 s.

4.2.1 | Application

Given that the interface did not change, we expect the time per action to be stable. We averaged these time periods over the entire user population to calculate the average time the action costs. Since we do not have relevance judgments, we cannot determine whether a click is a cost or a gain. We have therefore made two formulas, one including clicks as a cost, and one excluding clicks as a cost.

Using the method described above we come to the following formula for Cost without clicks:

$$Cost = (Q * Tq) + (R * Tr) + (F * Tf) + (I * Ti), \quad (4)$$

where Q represents the number of queries done in the session, R the number of reformulations done in the session, F the number of filters applied, I the number of documents inspected, and the T values the average time for that action. Extended cost uses the same formula, but also includes the number of clicks (C) multiplied by the average time it took the user to conduct a next action after a click (Tc). This gives us the following formula:

$$ExtendedCost = (Q * Tq) + (R * Tr) + (F * Tf) + (I * Ti) + (C * Tc). \quad (5)$$

When we apply the average time per action from the data, we end up with the following formulas to calculate the cost per session:

$$Cost = (Q * 14) + (R * 18) + (F * 11) + (I * 5), \quad (6)$$

and

$$ExtendedCost = (Q * 14) + (R * 18) + (F * 11) + (I * 5) + (C * 24). \quad (7)$$

In the Legal Intelligence system, a functionality for known-item retrieval (navigational search) uses hard boosts to push the document searched for to the top. When a user searches for "civil code article 6:162," that document will be hard pushed to the top, ignoring the position assigned by the ranking algorithm. It is possible that a query results in more than one preferred result. Because of this hard boost, known-item retrieval situations will not be impacted by changes in the ranking algorithm. Therefore known-items sessions will be excluded from the evaluation. We identify known-item sessions as query consisting of either just one action (e.g., updating behavior; Makri et al., 2008, where the user verifies that the legal status of a document is still the same), or one action followed by max one click (e.g., a query and one click), on position 1 or 2.

The use of such a cost model will be limited to within-system comparisons, as usage patterns may differ between systems. With these assumptions, it is possible to create an evaluation metric based only on cost, and compare the average cost of users under two rankings of the same system.

5 | RESULTS AND ANALYSIS

5.1 | Results

Table 3 shows the results of applying the Cost and ExtendedCost formula to the user sessions. Even though, as explained in Section 4.2.1, we have removed known-item retrieval from the evaluation, this table shows a long-tail distribution. This reflects the completeness ideal and research reality as described by Geist (2016): according to the completeness ideal, professional users would inspect all results; but in reality, many users do not.

TABLE 3 Cost per session Before/After.

	Cost		Extended cost	
	Before	After	Before	After
Count	59,081.00	66,519.00	59,081.00	66,519.00
Mean	135.11	131.61	334.71	327.30
Std	164.45	169.49	671.85	773.58
Min	5.00	5.00	51.00	51.00
25%	49.00	49.00	113.00	112.00
50%	87.00	87.00	193.00	189.00
75%	161.00	157.00	356.00	345.00
Max	4977.00	10,788.00	44,610.00	84,097.00

5.2 | Statistical analysis (without clicks)

We model the difference in the logarithm of the cost (*log-cost*) before and after the change to the ranking algorithm. It is important to note that different sessions may correspond to the same user. To take this dependency between the observations into account, we apply a linear mixed model (LMM) with a random effect for user ID. We denote by x_{ij} an indicator variable which takes value 0 if session j of user i took place before the intervention, and 1 if it took place after the intervention. This means the model for the log-cost of session j corresponding to user i is given by:

$$\log - \text{cost}_{ij} = \alpha + \beta x_{ij} + u_i + e_{ij}, \quad (8)$$

where α is the intercept, β is the (fixed) effect of the intervention, $u_i \sim N(0,\sigma_u)$ is the random effect of user ID, and $e_{ij} \sim N(0,\sigma_e)$ the residual. The analysis was performed in R (version 4.0.3) (R Core Team, 2020). Model fitting was performed using `lme4` (version 1.1–27.1) (Bates et al., 2015). Statistical significance was assessed using an approximate t-test with Satterthwaite's degrees of freedom, implemented as the default in `lmerTest` (version 3.1–3) (Kuznetsova et al., 2017). Table 4 shows that the mean log-cost is reduced by 0.022 after the intervention. In terms of the untransformed cost variable, this is equivalent to a reduction of the estimated geometric mean of the cost from 87.3 to 85.4.

5.3 | Statistical analysis (including clicks)

We apply the same model to the data with clicks included. Table 5 shows that in this case the mean log-cost is reduced by 0.027 after the intervention. In terms of the untransformed cost variable, this is equivalent to a reduction of the estimated geometric mean of the cost from 205.84 to 200.3.

5.4 | Practical significance

To demonstrate the effect of the change on the user, we have reported the estimated geometric mean.[16] This is the exponent of the arithmetic mean of the log-cost. The geometric mean, as opposed to the arithmetic mean, is used because the statistical analysis is done using a log-cost. Because of this log-cost, we also no longer have the problem of the large difference between the median and the mean, since the distribution of the log-cost is approximately normal. Note that if the distribution of the log-cost was exactly normal, the geometric mean of the untransformed cost would be the same as the median untransformed cost.

We see a difference in the geometric mean of 2 s for the Cost of a search session (a reduction of 2.2%), and 5 s for the ExtendedCost of a search session (a reduction of 2.7%). Though this may appear small, this is of practical significance for legal professionals, who may spend up to a third of their time doing research (Lastres, 2013). At a regular hourly tariff of 300 euros for attorneys, a 2%–3% reduction in search time can have substantial financial impact.

5.5 | Analysis of long sessions

At the extreme end of the long-tail we see user sessions with an Extended Cost of 84,097 s (1401 min, equals 23.36 h). It appears unlikely that a user would be conducting research for 23 h, without pausing for more than 30 min. To investigate this particular behavior, we analyzed the top 1% longest sessions by ExtendedCost. We had two questions: (1) are these sessions conducted by persons, or are they technical processes that are submitting queries for example to monitor response time, and (2) if the sessions are conducted by persons, are these long sessions also exceptions for these persons or are there people who regularly conduct these long sessions.

We found that users associated with these long sessions are customers of the Legal Intelligence system, and are not technical processes. We also found that there are users that have a pattern of extremely long sessions, having multiple such sessions in the span of the 6 weeks in our sample. We therefore have no reason to excluded these long-tail sessions from the data; these are the users for which more effective rankings are potentially the most valuable.

TABLE 4 ANOVA table for the structural part of the model (without clicks).

	Estimate	SE	df	*t*	*p*-value
Intercept	4.469	0.005			
Effect of intervention	−0.022	0.005	125,594.84	−4.644	< 0.001

TABLE 5 ANOVA table for the structural part of the model (including clicks).

	Estimate	SE	df	*t*	*p*-value
Intercept	5.327	0.004			
Effect of intervention	−0.027	0.005	125,593.48	−5.836	< 0.001

6 | CONCLUSIONS

This article shows the steps required to create an impact relevance variable for use in a bibliometric-enhanced ranking algorithm. The variable 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 at about 2 months after publication. We suggest to take the highest of the normalized usage/citation counts as input for the ranking variable. This variable has to be coupled with a recency variable that decreases at the same speed, to give new documents the benefit of the doubt before the usage and citation data becomes available.

Using a cost model, we show that such a bibliometric ranking variable can reduce the time of a search session of legal professionals by 2%–3% for use cases other than known-item retrieval or updating behavior. Though this may seem modest, given the high hourly tariff of legal professionals and the time they may spend on research, this is expected to lead to increased efficiency.

ACKNOWLEDGMENTS

The authors wish to thank Legal Intelligence for providing the data for this article.

ORCID

Gineke Wiggers https://orcid.org/0000-0002-1513-2212
Suzan Verberne https://orcid.org/0000-0002-9609-9505

ENDNOTES

[1] A newspaper article.

[2] A government report.

[3] Note that books are indexed by chapter or paragraph, legal codes are indexed at article level.

[4] Examples from research conducted by the authors.

[5] Domain relevance because of the level of agreement on certain factors, as described in Wiggers et al. (2022b).

[6] Through Worldcat at https://www.worldcat.org/

[7] As discussed by Mart (2017) the algorithms of commercial legal IR systems are trade secrets, but her work and information obtained from LexisNexis (2013) and the system used in our previous article (Wiggers et al., 2022b), Legal Intelligence, indicate that algorithmic and topical relevance are still the main focus.

[8] See also Mart (2017).

[9] "Vollständigkeit(sideal) und Recherche-Realität" (Geist, 2016, p. 158), translation by authors.

[10] The Royal Netherlands Academy of Arts and Sciences.

[11] Usage data is available from 2017 and later. For that reason, it was not useful to use older documents.

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

[13] See, for example Brody et al. (2006).

[14] For example, monthly.

[15] Documents published before 2017, before usage data became available, are given the benefit of the doubt with a usage score of 1. This means that they are treated as if they received the average number of clicks. This is done since documents are likely to gather the most clicks in the period after first publication.

[16] See also Fuhr (2018).

REFERENCES

Azzopardi, L. (2017). Building cost-benefit models of information interactions. In *Proceedings of the 2017 conference on human information interaction and retrieval* (pp. 425–428). ACM.

Azzopardi, L., & Zuccon, G. (2016). Two scrolls or one click: A cost model for browsing search results. In *Advances in information retrieval: 38th European Conference on IR Research* (Vol. 2016, pp. 696–702). ECIR.

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, *45*(3), 149–159.

Baskaya, F., Keskustalo, H., & Järvelin, K. (2012). Time drives interaction: simulating sessions in diverse searching environments. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 105–114). ACM.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bock, A. (2000). Gütezeichen als Qualitätsaussage im digitalen Informationsmarkt: dargestellt am Beispiel elektronischer Rechtsdatenbanken. *S. Toeche-Mittler*. (Vol. 24, p. 376).

Bornmann, L., Bowman, B. F., Bauer, J., Marx, W., Schier, H., & Palzenberger, M. (2014). Bibliometric standards for evaluating research institutes in the natural sciences. In *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact* (Vol. 2014, p. 201). MIT Press.

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.

Bruza, P. D., & Huibers, T. W. C. (1996). A study of aboutness in information retrieval. *Artificial Intelligence Review*, *10*(5), 381–407.

Cooper, M. D., & Chen, H.-M. (2001). Predicting the relevance of a library catalog search. *Journal of the American Society for Information Science and Technology*, *51*(10), 813–827.

Dan, O., & Davison, B. D. (2016). Measuring and predicting search engine users satisfaction. *ACM Computing Surveys (CSUR)*, *49*(1), 1–35.

Erdt, M., Nagarajan, A., Sin, S. C. J., & Theng, Y. L. (2016). Altmetrics: An analysis of the state-of-the-art in measuring research impact on social media. *Scientometrics*, *109*(2), 1117–1166.

Fuhr, N. (2018). Some common mistakes in IR evaluation, and how they can be avoided. *ACM SIGIR Forum*, *51*(3), 32–41.

Garfield, G. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471–479.

Garfield, G. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*. John Wiley & Sons.

Geist, A. C. J. (2016). *Rechtsdatenbanken und Relevanzsortierung* (Doctoral dissertation). uniwien.

Giménez-Toledo, E., Mañana Rodríguez, J., Engels, T. C., Ingwersen, P., Pölönen, J., Sivertsen, G., Verleysen, F. T., & Zuccala, A. A. (2016). Taking scholarly books into account: Current developments in five European countries. *Scientometrics*, *107*(2016), 685–699.

Haustein, S., Bowman, T. D., & Costas, R. (2016). *Interpreting 'Altmetrics': Viewing acts on social media through the lens of citation and social theories* (pp. 372–406). De Gruyter.

Haustein, S. (2014). Readership metrics. In *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact* (Vol. 2014, p. 327). MIT Press.

Hersh, W. (1994). Relevance and retrieval evaluation: Perspectives from medicine. *Journal of the American Society for Information Science*, *45*(3), 201–206.

Hicks, D. (2004). *he four literatures of social science* (pp. 473–496). Springer.

Jansen, B. J., Spink, A., & Kathuria, V. (2007). How to define searching sessions on web search engines. In O. Nasraoui, M. Spiliopoulou, J. Srivastava, B. Mobasher, & B. Masand (Eds.), *Advances in web mining and web usage analysis* (pp. 92–109). Springer Berlin Heidelberg.

Järvelin, K. (2009). Explaining user performance in information retrieval: Challenges to IR evaluation. In *Conference on the Theory of Information Retrieval* (pp. 289–296).

Järvelin, K., Price, S. L., Delcambre, L. M., & Nielsen, M. L. (2008). Discounted cumulated gain based evaluation of multiple-query IR sessions. In *European Conference on Information Retrieval* (pp. 4–15).

Joachims, T., Granka, L. A., Pan, B., Hembrooke, H., & Gay, G. (2017). Accurately interpreting clickthrough data as implicit feedback. *ACM SIGIR Forum*, *51*(1), 4–11.

KNAW. (2005). *Judging research on its merits—An advisory report by the Council for the Humanities and the Social Sciences council*. Royal Netherlands Academy of Arts and Sciences.

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). GroupLens: Applying collaborative filtering to Usenet News. *Communications of the ACM*, *40*(3), 77–87.

Kousha, K., & Thelwall, M. (2014). Web impact metrics for research assessment. In *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact* (Vol. 2014, p. 289). MIT Press.

Kurtz, M. J., & Henneken, E. A. (2017). 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*(2017), 695–708.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lastres, S. A. (2013). *Rebooting legal research in a digital age*. https://www.lexisnexis.com/documents/pdf/20130806061418_large.pdf

LexisNexis. (2013). LexisNexisLawSchools. *Understanding the technology and search algorithm behind Lexis Advance*. https://www.youtube.com/watch?v=bxJzfYLwXYQ&feature=youtu.be

Makri, S., Blandford, A., & Cox, A. L. (2008). Investigating the information-seeking behaviour of academic lawyers: From Ellis's model to design. *Information Processing & Management*, *44*(2), 613–634. https://doi.org/10.1016/j.ipm.2007.05.001

Maleki, A. (2022). OCLC library holdings: Assessing availability of academic books in libraries in print and electronic compared to citations and altmetrics. *Scientometrics*, *127*(2022), 991–1020. https://doi.org/10.1007/s11192-021-04220-6

Manning, C. D., Schütze, H., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mart, S. N. (2017). The algorithm as a human artifact: Implications for legal [re]search. *Law Library Journal*, *109*(2017), 387.

Maxwell, D. M. (2019). *Modelling search and stopping in interactive information retrieval* (Doctoral dissertation). University of Glasgow.

McGregor, M., Azzopardi, L., & Halvey, M. (2021). Untangling cost, effort, and load in information seeking and retrieval. In *Proceedings of the 2021 conference on human information interaction and retrieval* (pp. 151–161). MIT Press.

Merton, R. K. (1988). The Matthew effect in science, II: Cumulative advantage and the symbolism of intellectual property. *Isis*, *79*(4), 606–623.

Oard, D. W., & Kim, J. (2001). Modeling information content using observable behavior. In *Proceedings of the 64th annual meeting of the American Society for Information Science and Technology* (pp. 38–45). https://www.learntechlib.org/p/92951/.

OED. (2019). *Oxford English Dictionary*. Oxford University Press. https://www.oed.com/view/Entry/161891?redirectedFrom=relevance#eid

Opijnen van, M. (2014). *Op en in het web: Hoe de toegankelijkheid van rechterlijke uitspraken kan worden verbeterd*. Boom Juridische uitgevers.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Perneger, T. V. (2004). Relation between online "hit counts" and subsequent citations: Prospective study of research papers in the BMJ. *Bmj*, *329*(7465), 546–547.

Piwowar, H. (2012). 31 flavors of research impact through altmetrics. *Research Remix*. https://researchremix.wordpress.com/2012/01/31/31-flavours/

Priem, J., & Hemminger, B. H. (2010). Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, *15*(7). https://firstmonday.org/ojs/index.php/fm/article/download/2874/2570

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rehn, C., Kronman, U., & Wadskog, D. (2014). *Bibliometric indicators—Definitions and usage at Karolinska Institutet*. Karolinska Institutet University Library.

Rousseau, R., & Ye, F. Y. (2013). A multi-metric approach for research evaluation. *Chinese Science Bulletin*, *58*(26), 3288–3290.

Saracevic, T. (1975). Relevance: A review of and framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, *1975*, 321–343.

Saracevic, T. (1996). Relevance reconsidered, information science: Integration in perspectives. In *Proceedings of the second conference on conceptions of library and information science* (Copenhagen, Denmark) (pp. 201–218).

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science*, *58*(13), 2126–2144.

Snel, M. V. R. (2016). *Meester(s) over bronnen: een empirische studie naar kwaliteitseisen, gevaren en onderzoekstechnieken die betrekking hebben op het brongebruik in academisch juridisch-dogmatisch onderzoek*. Boom Juridische Uitgevers.

Snel, M. V. R. (2018). Hoera, een lijstje! Over bronvermelden. *Ars Aequi*, *3*(2018), 254–260.

Soetenhorst, W. J. (2017). Een juridische citatie-index: het proof of concept is voorhanden. *Nederlands Juristenblad*, *17*(915), 1184–1186.

Stolker, C. (2015). *Rethinking the law school: Education, research, outreach and governance*. Cambridge University Press.

Tamine-Lechani, L., Boughanem, M., & Daoud, M. (2010). Evaluation of contextual information retrieval effectiveness: Overview of issues and research. *Knowledge and Information Systems*, *24*(1), 1–34.

Thelwall, M. (2020). The pros and cons of the use of altmetrics in research assessment. *Scholarly Assessment Reports*, *2*(1), 2. https://doi.org/10.29024/sar.10

Van Opijnen, M., & Santos, C. (2017). On the concept of relevance in legal information retrieval. *Artificial Intelligence and Law*, *25*(2017), 65–87.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37–47.

Wiggers, G., Verberne, S., & Zwenne, G.-J. (2022). Citation metrics for legal information retrieval: Scholars and practitioners intertwined? *Legal Information Management*, *22*(2), 88–103.

Wiggers, G., Verberne, S., Zwenne, G.-J., & Van Loon, W. S. (2022). Exploration of domain relevance by legal professionals in information retrieval systems. *Legal Information Management*, *22*(1), 49–67.

Wiggers, G., & Verberne, S. (2020). Usage and citation metrics for ranking algorithms in legal information retrieval systems. In *CEUR Workshop proceedings*. (pp. 42–52) CEUR-WS.org.

Winkels, R., Boer, A., & Plantevin, I. (2013). Creating context networks in Dutch legislation. In *Proceedings of JURIX* (pp. 155–164). IOS Press.

Winkels, R., Boer, A., Vredebregt, B., & Someren van, A. (2014). Towards a legal recommender system. In *Proceedings of JURIX* (Vol. 271, pp. 169–178). IOS Press.

Winkels, R., & Ruyter de, J. (2011). Survival of the fittest: network analysis of Dutch supreme court cases. In *Proceedings of the international workshop on AI approaches to the complexity of legal systems* (pp. 106–115). Springer.

Winkels, R., Ruyter de, J., & Kroese, H. (2011). Determining authority of Dutch case law. *Legal Knowledge and Information Systems*, *235*(2011), 103–112.

Zuccala, A., & Cornacchia, R. (2016). Data matching, integration, and interoperability for a metric assessment of monographs. *Scientometrics*, *108*(2016), 465–484.

How to cite this article: Wiggers, G., Verberne, S., van Loon, W., & Zwenne, G.-J. (2023). Bibliometric-enhanced legal information retrieval: Combining usage and citations as flavors of impact relevance. *Journal of the Association for Information Science and Technology*, *74*(8), 1010–1025. https://doi.org/10.1002/asi.24799