The relevance of impact: bibliometric-enhanced legal information retrieval

Wiggers, G.

Chapter 5

Algorithm

Bibliometric-enhanced Legal Information Retrieval: combining usage and citations as flavors of impact relevance

Under review as: Wiggers, G., Verberne, S., Loon van, W.S., Zwenne, G-J. (2022). Bibliometric-enhanced Legal Information Retrieval: combining usage and citations as flavors of impact relevance.

Bibliometric-enhanced information retrieval uses bibliometrics (e.g. citations) to improve ranking algorithms. Using a data-driven approach, this paper 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-enhanced ranking 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 research session of legal professionals by 2 to 3% on average for use cases other than known-item retrieval or updating behaviour. Given the high hourly tariff of legal professionals and the limited time they can spend on research, this is expected to lead to increased user satisfaction, especially for users with extremely long search sessions.

5.1 Introduction

It is often thought that in legal IR, the focus should be on high recall (see e.g. [18, 83, 82]). However, Geist [50] 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'[1].

The 'completeness ideal' suggests that legal professionals do not stop their research until they have achieved full recall. 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

[1] 'Vollständigkeit(sideal) und Recherche-Realität' [50, p. 158], translation by authors.

specialised lawyer) and the case at hand. Geist [50] argues that only a good relevance ranking can provide users with both high recall and high precision.

Legal Information Retrieval (IR) systems still rely heavily on algorithmic and topical relevance[2], the occurrence of the query term in the result returned. This does not encompass all aspects of relevance for the user, as described by Saracevic [110], Van Opijnen and Santos [131], and Wiggers et al. [138]. As Barry [12] points out, this may lead to poor user satisfaction.

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. The use of citations and statistical methods to analyse the impact of books, articles and other publications is commonly referred to as bibliometrics. Usage of documents (clicks in the search engine) could be an additional source of information for measuring impact on readers [56], and thereby constitute another aspect of relevance [99]. 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.

This paper covers the analysis of usage and citation data in a legal IR system and the process of balancing the indicators to create a bibliometric-enhanced ranking variable, as well as balancing this variable with other existing variables in the ranking algorithm, such as a term-frequency based variable. The term 'ranking variable' therefore refers to one factor in the relevance ranking, whereas the term 'ranking algorithm' refers to the whole model for relevance ranking. In this research 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.

This paper addresses the following research question: **can bibliometrics improve common ranking algorithms in legal information retrieval**?[3] The contributions of this paper are threefold: (1) we show that

[2]As discussed by Mart [83] the algorithms of commercial legal IR systems are trade secrets, but her work and information obtained from Lexis [78] and the system used in our previous research [138], Legal Intelligence, indicate that algorithmic and topical relevance are still the main focus.

[3]Research question 9 in this thesis.

bibliometrics can be seen as a manifestation of impact relevance; (2) we show that ranking algorithms in legal IR can be improved using bibliometrics; (3) we show, in a data-driven manner, how such a bibliometric-enhanced ranking variable can be created; and (4) we set an example of cost-based evaluation of live, domain specific search engines.

5.2 Background

From an IR perspective, Oard and Kim [92] have created a framework that describes the different types of user behaviour that could be monitored for implicit feedback on the relevance of documents. They have subdivided the behaviours 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. [55], expanded upon by Erdt et al. [43] from a bibliometric perspective, created a framework for user interactions with research objects (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).

This research focuses on the two metrics that are most readily available in legal IR systems: clicks (part of the examine category from Oard and Kim, and part of the accessing category from Haustein et al.), and citations (part of the reference category from Oard and Kim and part of the applying category from Haustein et al.).

5.2.1 Citations and usage in bibliometrics

The use of citations as a proxy for impact was introduced by Garfield [48]. 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." [75]. Piwowar [99] describes citations and usage as different flavors of impact.

As Kousha and Thelwall [73] 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 [120]. For this reason, books are included in legal IR systems and will be included in this research.

5.2.2 Correlation between usage and citations

For the above reasons, we aim to combine metrics for document usage and citations. Because some readers are also authors, a correlation between usage and citations counts is expected. Priem et al. [100], in the early stages of what they described as 'altmetrics', 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 [98] analyzed 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. [24], 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 [56, 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 [107] therefore propose the term 'influmetrics'.

5.2.3 Usage in evaluation

Next to using clicks as a sign of impact in bibliometrics, clicks are used as implicit feedback of relevance for the evaluation of IR systems [92] (the

examine behaviour category on the object level). Cooper and Chen [33] describe how multiple reasons exist for clicking on an article, but all have an implicit assumption of relevance to the user. For that reason, implicit feedback, in the form of clicks or other user interactions, is not an absolute relevance judgment, but is a good approximation of the perception of the relevance of the item for that particular user at that point in time.

Joachims et al. [64] assume that search engine users scan lists from top to bottom in a exhaustive fashion (the 'cascade model'). This assumption is adopted by later user interaction models, such as the commonly used Click Chain Model [53].

Baskaya et al. [14] researched search behaviour for 60, 90 and 120 second 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 behaviour 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 behaviour clues. Järvelin et al. [66] developed the DCG 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 [63] further 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 [63] 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.

Azzopardi et al. [10, 11] have used such cost based models to determine the effectiveness of changes to the user interface. Maxwell [84] 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). As shown by the work of Geist [50] we can assume that a legal professional will not stop searching until they reach a point in the 'research reality' [50] trade-off where they are satisfied enough to stop, given that their professional reputation is on the line.

McGregor et al. [85] differentiate between load, effort and cost. Load is taken to refer to the total amount of resources used to complete the task, internal and external. Effort represents the internal resources spent (e.g. cognitive effort), while cost represents the external resources spend (e.g. time or money). Cost can be measured in time-orientated cost or interaction orientated/count based costs.

5.3 Data analysis

In this section we discuss the data analysis that preceded the creation of the bibliometric-enhanced 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?[4]

The KNAW, the Koninklijke Nederlandse Akademie van Wetenschappen[5], has indicated that it can take up to two years for documents in the humanities to gather sufficient citations for research evaluation [108]. For

[4]Research questions 7 and 8 in this thesis.

[5]the Royal Netherlands Academy of Arts and Sciences

this reason, we decided to use documents from the Legal Intelligence system from the first half of 2017 for our analysis.[6]

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 in their main text. Using the document identifier, we extract the usage data (clicks) from the search engine logs.

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

Citation data

After accumulating all citations (excluding self-citations), we see that 235,609 documents have received citations. 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.[7] However, based on citations in other fields, it is also to be expected that a large number of documents does not generate citations.[8] 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.

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

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

[8] See, for example Brody et al. [24]

Analysis

To analyse 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 5.1 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.

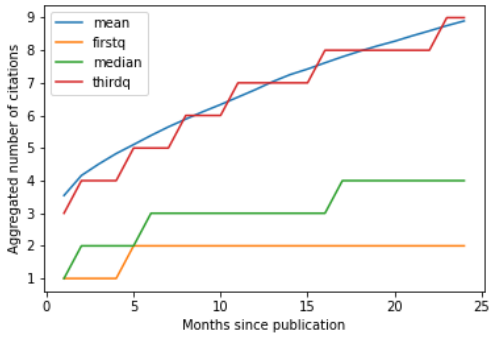

Figure 5.1: Aggregated citations per month after publication

Figure 5.1 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 [24, 20], which states that citation counts often show long-tail distributions.

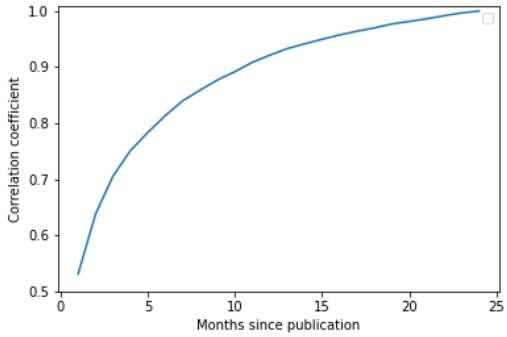

Figure 5.2: Correlation per month of citations up to and including that month with citations after 24 months

Figure 5.2 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 two months is valuable. It is also possible to update the data regularly[9], so increases in citation counts can

[9] e.g. monthly

be incorporated as they occur.

2. To what extent are usage and citations correlated?

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 5.3. This is again caused by a long-tail distribution, and is seen throughout the 24 months.

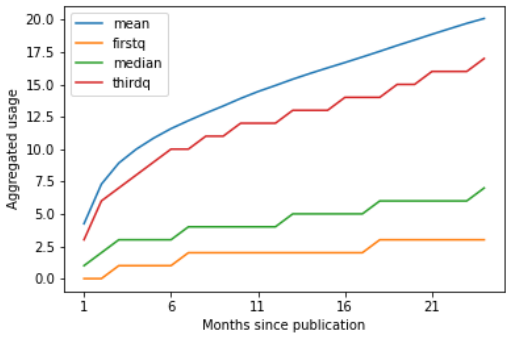

Figure 5.3: Aggregated usage per month after publication

Figure 5.4 shows a Spearman correlation between usage after 1 month and usage after 24 months of $\rho = 0.52$. The Spearman correlation between usage after two months and usage after 24 months is $\rho = 0.64$.

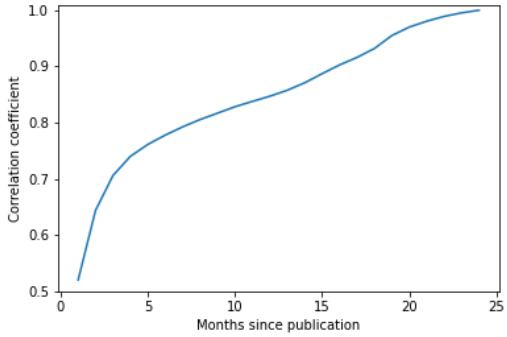

Figure 5.4: Correlation per month of usage up to and including that month with usage after 24 months

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 5.3). 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.

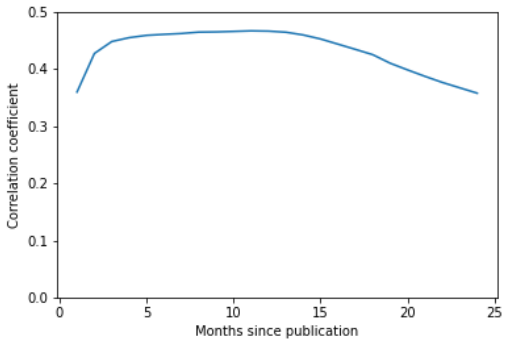

Figure 5.5: Correlation per month of usage up to and including that month with citations after 24 months

The development of the correlation between usage and citations is as expected. Brody et al. [24] found that the increase of the correlation between usage and citations is not linear with time, but reaches it's highest point after about 6-7 months. In their paper Brody et al. [24] indicate that after these 6 months the correlation increases by a small amount. The decline in the correlation in Figure 5.5 can be explained as the usage no longer grows much whilst the citations do, leading to a lower correlation between the two.

As indicated by Haustein [56], medium positive correlations (in this research between $\rho = 0.52$ and $\rho = 0.64$), show that citations and usage measure different flavors of impact.

5.4 Methods

In this paper, we propose a bibliometric-enhanced 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.

5.4.1 Our proposed bibliometric-enhanced 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. 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.

Normalization

The normalization of the raw citation and usage counts of the publications is based on the NCS score of the CWTS [135] and the work of Rehn et al. [104] on the normalization of citations. This 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. 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. We decided to apply the same normalization to the

usage counts.

This normalization is achieved by dividing the number of clicks/citations of the document ($\text{citations}_\text{d}$) 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 ($\text{citations}_\text{a}$):

$$W_d = citations_d / \ citations_a \tag{5.1}$$

Our normalization differs 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 [87] 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 ranking functions (such as TF-IDF). A log based normalisation risks that outliers exceed the maximum, in our data even a log10 scale exceeded the chosen maximum of 2 for certain extreme outliers. These would then have to be capped anyway.

The bibliometric-enhanced 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 can not yet provide a reliable prediction of the impact the document will have, and increases in influence as the data 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:

$$I_d = c + ((\beta - (s/(t_d + \alpha))) * (W_d - 1)). \tag{5.2}$$

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 5.4.1), and subtract 1, to get a score ranging from -1 to 1.[10] 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 be capped off at a maximum below the TF-IDF or BM25 score, to prevent this variable (representing

[10]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.

the impact form of relevance) from overruling other variables (representing other forms of relevance).

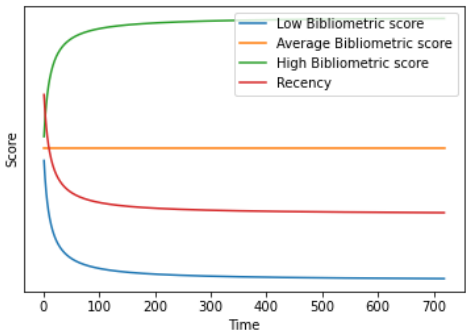

Figure 5.6: A visualisation of the ranking variable for a low, average and high citation/usage score, and the corresponding recency variable

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:

$$R_d = c_2 + (s/(t_d + \alpha)). \tag{5.3}$$

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

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

part is the same as in the bibliometric boost. 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 as 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.

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.

5.4.2 Evaluation

For the evaluation of the ranking variable described above, we use a cost model inspired by Järvelin [63] and compare the cost before the introduction of this variable (the intervention) with the cost after the intervention. This model will be limited to cost without gain, as there are no relevance judgements to base gain on. However, as shown in Section 5.2 we can assume that a legal professional will not stop searching until they are satisfied enough to stop. Because of the time pressure legal professionals work under a time-orientated metric, as described by McGregor et al. [85], appears to be the most suitable.

The intervention took place on September 14th 2020 at the close of business day. We took data from the three weeks before the intervention (24th

of August until 14th of September) and three weeks after the intervention (15th of September to 5th of October).

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 minutes between two actions) vary between 1 second and 82,555 seconds (or almost 23 hours), 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 second and 86,125 seconds (or almost 24 hours), 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 minutes [62] 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. [14] use 3 seconds 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 email alert). For those queries, we retrieved the closest login event from the logs, with a maximum of 30 minutes (our chosen boundary for 1 session). This resulted in 144,479 sessions with a median of 14 seconds and a mean of 52.68 seconds.

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 spend on inspecting results is spread evenly over the number of items inspected. This gives us an indication of the time spend inspecting each search result, under the assumption of the cascade model[64]. This gave us a total of 101,711 query-click pairs, with a median inspection time per result of 5 seconds and a mean of 17.22 seconds.

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 second, 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 seconds and a mean of 73.92 seconds.

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 spend 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 seconds and a mean of 73.83 seconds. It is likely that users inspect some of the results before reformulating the query, at a cost of 5 seconds 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 seconds 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 seconds and a mean of 34.60 seconds.

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:

$$Cost = (Q * Tq) + (R * Tr) + (F * Tf) + (I * Ti)), \tag{5.4}$$

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:

$$ExtendedCost = (Q * Tq) + (R * Tr) + (F * Tf) + (I * Ti) + (C * Tc). \tag{5.5}$$

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

$$Cost = (Q * 14) + (R * 18) + (F * 11) + (I * 5), \tag{5.6}$$

and

$$ExtendedCost = (Q * 14) + (R * 18) + (F * 11) + (I * 5) + (C * 24). \tag{5.7}$$

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 behaviour [81], 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.

5.5 Results and analysis

5.5.1 Results

Table 5.1 shows the results of applying the Cost and ExtendedCost formula to the user sessions. Even though, as explained in Section 5.4.2, 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 [50]: according to the completeness ideal, professional users would inspect all results; but in reality many users do not.

5.5.2 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

	Cost		Extended Cost	
	Before	After	Before	After
count	59081.00	66519.00	59081.00	66519.00
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	10788.00	44610.00	84097.00

Table 5.1: Cost per session Before/After

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:

$$\text{log-cost}_{ij} = \alpha + \beta x_{ij} + u_i + e_{ij}, \tag{5.8}$$

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) [101]. Model fitting was performed using `lme4` (version 1.1-27.1) [15]. 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) [76]. Table 5.2 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.

	estimate	SE	df	t	p-value
intercept	4.469	0.005			
effect of intervention	-0.022	0.005	125594.84	-4.644	<0.001

Table 5.2: ANOVA table for the structural part of the model (without clicks).

5.5.3 Statistical analysis (including clicks)

We apply the same model to the data with clicks included. Table 5.3 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.

	estimate	SE	df	t	p-value
intercept	5.327	0.004			
effect of intervention	-0.027	0.005	125593.48	-5.836	<0.001

Table 5.3: ANOVA table for the structural part of the model (including clicks).

5.5.4 Practical significance

To demonstrate the effect of the change on the user, we have reported the estimated geometric mean.[11] 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 seconds for the Cost of a search session (a reduction of 2.2%), and 5 seconds 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 [77]. At a regular hourly tariff of 300 euros for attorneys, a 2 to 3% reduction in search time can have substantial financial impact.

[11] See also Fuhr [46].

5.5.5 Analysis of long sessions

At the extreme end of the long-tail we see user sessions with an Extended Cost of 84,097 seconds (1401 minutes, equals 23.36 hours). It appears unlikely that a user would be conducting research for 23 hours, without pausing for more than 30 minutes. To investigate this particular behaviour, 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 six 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.

5.6 Conclusions

This paper shows the steps required to create an impact relevance variable for use in a bibliometric-enhanced ranking algorithm. The impact relevance 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 research session of legal professionals by 2 to 3% for use cases other than known-item retrieval or updating behaviour. 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 user satisfaction.

5.7 Acknowledgements

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