The relevance of impact: bibliometric-enhanced legal information retrieval

Wiggers, G.

Chapter 1

Introduction

"Not everything that can be counted counts, and not everything that counts can be counted." attributed to Albert Einstein and William Bruce Cameron

Information retrieval (IR) can be defined as "...finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)."[82, p. 1]. Legal professionals spend up to a third of their time doing research [77]. During this research legal information retrieval is a form of intelligent assistance (IA); the system helps its users find information that is relevant[1] for them. This intelligent assistance is important because "...the number of legal documents published online is growing exponentially, but accessibility and searchability have not kept pace with this growth rate." [131]. This can be described as the proverbial needle in a haystack situation.

The quality of IR systems is often described as a trade-off between precision and recall. *Precision* is the fraction of the returned documents that are actually relevant out of the total number of documents shown to the user. *Recall* is the fraction of relevant results that are retrieved, out of the total number of relevant results in the collection [82]. If a system focuses on recall, it is likely to show more documents to the user, and thereby lower the precision. But if the system focuses on precision, it is likely to show fewer results to the user, and thereby lower the recall. Up until now, it was assumed that legal IR should focus on recall [82, 71]. However, Geist [50] describes the precision and recall trade-off as the 'completeness ideal', where legal professionals require full recall, and the 'research reality' where legal professionals do not have the time to go through irrelevant documents, and therefore require high precision. Geist [50] concludes that the solution for the perceived trade-off between high recall and high precision (where one can be improved only at the expense of the other) lies in improving the ranking. The ranking, the order in which results are displayed, allows legal IR systems to return the full recall of documents, whilst optimizing the order of the results according to precision. This way, users find the (expected) most relevant documents first, and if they are satisfied enough (due to the research reality), they can stop. If their task at hand requires full recall however, they still have access to all recalled documents.

This thesis is about improving legal information retrieval. More specifi-

[1] For more information on relevance, see Chapter 2

cally about how we can use bibliometrics to improve ranking algorithms for legal information retrieval in the Netherlands. Improving in this research means presenting relevant documents higher in the results list. The main research question is: **How can bibliometrics improve common ranking algorithms in legal information retrieval?** This is subdivided into the following sub-questions:

1. To what extent can we demonstrate the existence and factors of domain relevance in the context of judgment of search results (document representation) in legal IR systems?

2. To what extent do legal information retrieval specialists, legal scholars and legal practitioners show agreement on relevance factors outside of a task context?

3. Does the literature suggest the use of one bibliometric-enhanced ranking function in legal IR, or should there be separate bibliometric-enhanced ranking functions for legal scholars and legal practitioners?

4. Does quantitative data analysis of citations in, and usage of, legal documents support the findings from the literature?

5. What are the characteristics of legal IR that influence the choice of ranking evaluation methods and metrics?

6. What are the limitations of common evaluation methods and metrics for evaluating ranking changes in live professional IR systems?

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

8. To what extent are usage and citations correlated?

9. Can bibliometrics improve common ranking algorithms in legal information retrieval?

To answer the main research question chapter 2 discusses the concept of relevance in legal IR and addresses sub-questions 1 and 2. The answers to these sub-questions build the theoretical framework of what relevance means to legal professionals and which factors they use to assess this. This framework is needed to understand which documents in the results list are relevant documents in the eyes of the user, as well as which aspects of the documents we should focus on in the ranking algorithm to ensure that those documents are returned higher in the results list.

Chapter 3 covers the meaning of citations in legal documents, and covers sub-questions 3 and 4. This chapter builds on chapter 2 and lays the foundation for using bibliometrics to improve ranking algorithms in legal IR. It is the theoretical underpinning for the ranking algorithm proposed in this thesis.

Chapter 4 explains the challenges of common evaluation methods for IR when applying them to the legal domain, and answers sub-questions 5 and 6. In answering these sub-questions, we show the challenges encountered while determining what an appropriate evaluation measure and metric is to evaluate changes to the ranking algorithm of a legal IR system. This is required to be able to demonstrate the bibliometric-enhanced ranking algorithm proposed in this thesis is indeed an improvement to the existing ranking algorithm.

Chapter 5 combines the information learned in the previous chapters and shows the new bibliometric-enhanced ranking algorithm – sub-questions 7 and 8 – and the evaluation thereof, sub-question 9. When we show, with the answer to sub-question 9, that a bibliometric-enhancement can indeed improve the ranking algorithm in a live legal IR system, we have all the pieces to answer the main research question in chapter 6. Here we answer the question *how* bibliometrics can improve common ranking algorithms in legal information retrieval; the sum of all the sub-questions. Chapter 6 also discusses the importance of interdisciplinary research in domain specific IR, and why the IR community should make more space for such research.

Contribution

The contribution of this thesis is the application of bibliometrics and IR (BIR) to the legal domain. The application of bibliometrics in IR was already envisioned by Garfield [48], has been applied by e.g. Kurtz and Henneken [75], and has led to the creation of regular BIR workshops at IR conferences[2]. Its application to the legal domain, and the thorough description thereof from information science, bibliometrics, and information retrieval perspective, is how this thesis contributes to the field.

This thorough description starts with a user study involving professional users of a legal IR system to research the validity of the concept of domain relevance as described by Van Opijnen and Santos [131] in practice and its possible applications in legal IR systems. This is followed by the statistical analysis which demonstrates that sub-groups of users of legal IR systems (legal information specialists, legal scholars and legal professionals) show agreement on relevance factors outside of a task context, when judging search results (document representations) in legal IR. And that there is no indication, at this point, to differentiate the ranking for subgroups of users based on their role or affiliation.

The contribution also includes the insight, through literature and quantative data analysis, that bibliometrics can be seen as a manifestation of impact relevance. It shows that citations in legal documents represent part of a broader form of impact on the legal domain as a whole, and should be used alongside usage data to also see the impact on non-authors.

In the evaluation process we provide insight in the characteristics of legal IR in practice that make the evaluation of a live legal IR system different from common ranking evaluation tasks, and describe the limitations to common evaluation methods based on literature and data. We create an example of cost-based evaluation of live, domain specific search engines.

These contributions culminate in the demonstration, in a data-driven manner, how such a bibliometric-enhanced ranking variable can be created, and that ranking algorithms in legal IR can be improved using bibliometrics.

[2] See `https://sites.google.com/view/bir-ws/home`

This research is conducted with data from Legal Intelligence, the largest legal IR system in the Netherlands.

Reading guide

Bibliometric-enhanced information retrieval, when applied to legal IR systems, encounters the challenge that the legal domain is often overlooked when it comes to bibliometrics, and that domain specific IR is often overlooked in the field of IR.

This introduction is therefore split in multiple sections. Section 1.1 describes the legal importance of (good) legal IR. Section 1.2 contains a brief introduction to information retrieval, to familiarize legal professionals with the terminology used in IR. Section 1.3 provides an overview of characteristic features of legal IR as compared to other domain specific fields of IR research, to aid IR professionals in understanding the particularities of legal IR. Section 1.4 is an overview of common evaluation methods in (non-domain specific fields of IR research), to provide a background for non-IR readers for Chapter 4. Section 1.5 gives a brief introduction to bibliometrics.

1.1 The legal importance of legal information retrieval

Aside from practical assistance for legal professionals, legal IR has legal importance in and of itself. It can be argued that IR systems that contain too many documents in an unstructured manner, make it as hard to access information as when it has not been published (online): "Consideration should be given to avoiding complicating the use of the system by the accumulation of a growing amount of obsolete information." [35]. "An overload of information (particularly if of low-quality) carries the risk of undermining knowledge acquisition possibilities and even access to justice." [131]

In the Netherlands, the government has determined that case law has

to be made public and accessible [1, p. 4]. Van Opijnen [94] describes three types of access: make public ('openbaarheid'), provide ('verstrekking') and make accessible ('toegangelijkheid'). Whether or not a document is *made public* is a legal status. A public document can be unfindable, but a not public (leaked) document can be widely available. *Providing* sees to the practical nature of public documents; and can be either on request only, or being made widely available through publication on the internet. Making a document *accessible* has to do with the ease with which an interested party can get the information in the document. This is taken to include findability [1, p. 7-8].

The debate whether more, or even all, cases should be made public falls outside the scope of this thesis, and can be found in the work of Van Opijnen [94]. The focus of this section is the question whether good legal IR is a prerequisite for access to justice, in the form of accessibility of legal information.

Van Opijnen argues that part of findability, and by extension accessibility, is manageability [94]; large document collections have to be offered in a manner that the user can find the information easily. A parallel can be drawn to the challenge in trial discovery where parties deliver (either to convince or confuse the other) so much documentation that it is difficult to find the relevant information. For those cases, rules have been created (alongside e-Discovery technology [52]). For example, that the party has to indicate the claim the piece of evidence relates to, and the importance thereof [79]. A similar solution is mentioned for the overabundance of case law in legal IR systems in Recommendation R(2002)13, where it is argued with regard to case law from the European Court of Human Rights (ECtHR) that dissemination cannot consist of simply throwing everything on the internet, but that information "has to be assessed and an appropriate commentary added" [36].

However, whilst it might be feasible to add commentary to cases of the ECtHR, adding such information to all government publications of a country for the purpose of improving findability might prove to be prohibitively expensive. For this reason, effective legal IR systems are necessary in order to meet the requirement of manageability, findability, and by extension

accessibility, in order to guarantee access to justice.

As the Free Access Law Movement describes, effective IR encompasses more than technological solutions alone. "...technical assistance, advice and training" may be required to make the information truly easily accessible [45].

1.2 A brief introduction to information retrieval (IR)

Users of an IR system have an information need; they need a piece of information to answer a question or to complete a task. For example: I am working on a legal IR system, and want to know how bibliometrics can be used to enhance the ranking function. This thesis focuses on document retrieval, and therefore assumes that the information need can be satisfied with a written document (as opposed to e.g. an image or video).

The information need is usually expressed by the user as a query; consisting of one or more query terms. In this example the query could be 'bibliometric-enhanced legal information retrieval'. A document is considered potentially relevant by the IR system, and returned as a search result, if it contains the query term(s).

To be able to quickly retrieve all documents that contain a certain query term, the documents are indexed beforehand. For each word in the collection of documents, a list is created with the identifiers of the documents that contain that word (an inverted index). When a user searches for that query term, the IR system knows it has to return the documents in the list for that word. When a user searches for multiple query terms, the IR system can return either the *intersection* (all documents that occur in both lists) if the system uses an AND relation between the query terms, or the *union* (all documents that occur in either list) if the system uses an OR relation between the query terms [82].

To make it easier for the user to find what they are looking for, several steps are taken, both when indexing the documents and when processing a query, to return as many relevant results as possible. These steps often

include (1) making all letters lowercase letters, (2) transforming words into their roots (two possible approaches are stemming and lemmatization), (3) query expansion (for example when the query term is 'IR', also returning documents that contain the phrase 'information retrieval' written in full), and (4) spelling correction [82].

When the IR system has determined which documents need to be returned, it has to decide in which order to return the documents to the user. This is done using a ranking algorithm, a formula that assigns a score to each document. The documents are then returned in order of highest to lowest score. This thesis focuses only on ranking (the order in which results are returned) in legal IR. The retrieval model (the number of results returned) falls outside the scope of this thesis.

Ranking algorithms usually encompass at least the following components: (1) query-document similarity (there are many ways to calculate this, often involving the number of times the word occurs in the document, the number of documents in the collection that contain that word, and the length of the document), (2) the position of the query term(s) in the document (when they occur in the title it is considered more relevant then when they occur in a footnote), (3) the 'freshness' or publication date of the document, (4) the authority of the source (for example through Google's PageRank algorithm), and (5) popularity of the search result. Other possible components are (6) query-term proximity, the closer the query terms are together, the higher the score, and (7) static boosts, where certain points are awarded to documents based on their metadata rather than the query. The weight of each of these components can be determined by the developer of the IR system beforehand, or learned by the system through a learning-to-rank model.

1.3 Features of legal IR

An IR system searches through a 'collection', often (but not always) of documents. This collection can be the entire internet, such as in web-search, or a curated collection, such as in domain specific IR (also called

professional search). Examples of domain specific IR are legal IR, medical IR and academic IR.

A limited collection allows for a domain specific vocabulary [132], including domain specific abbreviations. For example, where the term 'trust' in web-search may denote having faith in someone, in legal IR it will possibly relate to one person holding property for another person. In domain specific IR both the user and the documents are likely to use the word in the same manner, which means the user is less likely to be confronted with irrelevant results that use the word in another way.

Another feature of the collection in domain specific search is the use of access rights [132]. Many sources in the collection of professional and academic search systems work with subscriptions, where a user has to pay to be able to access specific content. This means that two users of the same IR system with different access rights will see a different set of results.

A downside of curated collections is that users are sometimes required to access multiple IR systems [132]. Whether this applies to legal IR differs from country to country. Geist [50] suggests that in Austria legal publishers often have their own IR system, meaning users may be required to use multiple systems, similar to other fields of domain specific search. In the Netherlands legal IR systems also function as content integration systems, meaning this is often not required.

In legal IR, there is a large variety of document types and lengths of documents, from a 161 page government report[3] to a 57 word newspaper item[4]. This makes legal IR different from patent search and academic search, where the collection usually consists of documents of roughly the same length. Aside from the document length, documents in a legal collection differ in the level of formal structure. Legal codes and case law often adhere to strict structures, similar to patent search. And legal documents, like prior art search in patents, maintain a high standard of referencing between documents in the collection. Legal blogs and journal articles on the other hand, do not adhere to these strict structures.

[3]DocumentID 34474736 in the Legal Intelligence system.
[4]DocumentID 34582268 in the Legal Intelligence system.

Next to academic and patent search sub-types of domain specific IR, legal search is related to enterprise search, where the collection consists of all the information of that organisation. When we compare legal documents to the documents in enterprise search, legal IR has a very homogeneous collection, because legal documents are often limited to text files or PDF versions thereof. Documents in enterprise search on the other hand can differ from emails to presentation slides to databases [4]. Furthermore, the documents in the collection of a legal IR system are often created by either the government or legal publishers, who make documents search-engine-friendly by, for example, providing metadata and the explicit referencing convention.

A distinguishing feature of legal IR as opposed to other forms of domain specific IR are the national boundaries. Though large academic research platforms such as Web of Science and PubMed cater mainly to English language publications, their audience is global. But legal publications have strong national ties and therefore have a limited scope of applicability and are published in the national language [120]. This means that not only the collection is limited by these national boundaries, but also the audience. Which leads to a limited amount of user data (see Chapter 4).

1.4 An overview of common evaluation methods in IR

Evaluation methods for ranking in IR systems can broadly be classified in two groups, based on the manner in which relevance judgements are obtained: (1) system-oriented evaluation: using test collections of queries and relevance judgments, and (2) user-oriented evaluation: using input from users in evaluation [63]. Conrad and Zeleznikow [31] refer to these two groups as 'comparison with expert judgements' and 'comparison based on human performance'.

The benefit of test collections is that they can be used multiple times. The relevance judgments in test collections are often created by assessors. They are presented with an information need that can be expressed in a

query, and judge whether the returned documents are relevant for the information need [82]. The notion of relevance has been extensively described by Saracevic [110].

Implicit relevance judgments can be obtained through 'click through rates' or pairwise inferences. The Click Through Rate registers whether the user opened a document for the query or not. If they opened it, it is considered relevant, if not, it is considered neutral. The absence of a click is not considered a signal that the result was not relevant, so information about irrelevant results is not available with this method. For pairwise inferences, Joachims et al. [64], further expanded on by Chuklin et al. [27] and Agrawal et al. [5], argue that when a document is clicked, it indicates that the documents above were not considered relevant or did not answer the user's information need fully. Joachims mentions that clicks do not provide an absolute relevance judgment, but that they do provide a relative signal [64].

Both expert relevance judgements and implicit feedback relevance judgements are well-known methods to create test collections. Joachims et al. [64] and Konstan et al. [71] discovered a positive correlation between expert relevance judgements and implicit feedback relevance judgements.

Metrics Aside from a method of collecting the relevance judgments, a researcher/developer needs to decide what metric will be applied in the evaluation. We distinguish between two groups of metrics [63]: (1) system-oriented metrics, based on counts of relevant documents, and (2) user-oriented metrics, measuring the effort required by the user to achieve satisfaction.

For the evaluation of rankings, *Mean Average Precision* (MAP) and *normalized Discounted Cumulative Gain* (nDCG) are commonly used system oriented metrics. Most metrics are calculated using a model where points are gained for relevant results and lost for irrelevant results (the precision/recall trade-off). Often metrics require the entire results list to be evaluated, but for measures with a certain cut-off point (e.g. @k measures) the results list only has to be evaluated up to the cut-off point (e.g. doc-

ument k). A commonly used user-oriented metric is *success rate at rank k* [26] (abbreviated as S@k, sometimes as s@n). Another example of a user-oriented metric is a cost/benefit model [63]. McGregor et al. [85] describe that cost could be calculated in different ways; e.g. time or mental effort.

Surveys A completely different approach, that does not use documents and (explicit or implicit) relevance judgements is to present users with a survey to ask whether they are satisfied with the IR system. An example is the Net Promotor Score, or NPS [105]. This is a very short survey to gauge user satisfaction. The question is answered on a zero to ten scale, making it an intuitive question for users. The NPS is popular in commercial settings because the share of promotors (users who give a score of 9 or 10) is found to have a stronger correlation with sales than traditional customer satisfaction surveys [105, p. 4]. Furthermore, the results are very quick to analyse and interpret. If desired, the NPS question can be followed by an optional question for users to explain the reason for their score, bringing it more in line with traditional surveys.

1.5 A brief introduction to bibliometrics

Bibliometrics is defined as: "the application of mathematics and statistical methods to books and other media of communication." [41]. Bibliometrics [38] originates from the quantitative analysis of books, but has expanded to cover all the forms of output of scholars (and practitioners). It is closely related to scientometrics (the measurement of scientific output, e.g. from a person or research institute) and altmetrics (alternative metrics, obtained from e.g. reference managers). Bibliometrics can be used for research evaluation, but also to measure the impact of publications.

Frank Shepard is credited with creating one of the first citation networks, and the best-known legal one; the Legal Case Citator or Shepard's Citations [59]. It is immortalized in LexisNexis' Shepardize function (which shows whether a case has been overturned in appeal). According to Garfield [48], the Legal Case Citator influenced him in the creation of the Science

Citation Index. The Science Citation Index introduced the notion of citation indexing in the context of journal articles, and has been extensively used for impact measurement of journal articles and research evaluation.

Garfield himself acknowledged: "there are undoubtedly highly useful journals that are not cited frequently" [47, p. 476]. "[T]hat does not mean that they are therefore less important or less widely used than journals that are cited more frequently. It merely means that they are written and read primarily for some purpose other than the communication of original research findings." [47, p. 476]. An example he uses is Scientific American, a journal readers read to keep up to date, but tend not to cite. The impact of these sources can not be captured by citation measurement. Haustein [56] describes that though not all readers cite, these non-citing readers might still use the documents in their daily work. This has broadened the application of bibliometrics from citation counts to also include metrics like usage counts.

In its simplest form, bibliometrics can consist of raw counts of citations or usage from other documents. However, this can easily be manipulated by authors clicking on their own paper multiple times, and makes comparison of documents difficult since some fields are larger than others. The CWTS [135] and other organizations [104] developed normalization techniques for raw counts, including the normalized citation score (NCS) score. These methods calculate a citation score normalized for time (based on year of publication), field and document type. In this manner each document is compared to other documents from that time-frame field and document type, and the normalized score shows how it performed compared to its peers.

There is debate about the use of bibliometrics for research evaluation in Dutch law faculties (see e.g. [130, 108]). This discussion falls outside the scope of this thesis. This thesis covers only the use of bibliometrics in IR systems.

To the reader

Each chapter in this thesis has its own background section and conclusion, and can be read as an independent work. The main research question is answered in Chapter 6, which also provides an overview of the steps taken to reach that conclusion. This chapter also provides references to the chapter in which a particular step was taken. This thesis therefore does not have to be read front to back, but can be read non-linearly.