Universiteit Leiden
The Netherlands

The relevance of impact: bibliometric-enhanced legal information retrieval

Wiggers, G.

Citation
Wiggers, G. (2023, March 8). *The relevance of impact: bibliometric-enhanced legal information retrieval*. *SIKS Dissertation Series*. Retrieved from https://hdl.handle.net/1887/3570499

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3570499

Note: To cite this publication please use the final published version (if applicable).

The Relevance of Impact

bibliometric-enhanced legal information retrieval

Proefschrift

ter verkrijging van

de graad van doctor aan de Universiteit Leiden,

op gezag van rector magnificus prof.dr.ir. H. Bijl,

volgens besluit van het college voor promoties

te verdedigen op woensdag 8 maart 2023

klokke 11.15 uur

door

Gineke Wiggers

geboren te Hilversum

in 1988

Promotor:	Prof.dr.mr. G.J. Zwenne
Co-promotor:	Dr. S. Verberne
Promotiecommissie:	Prof.mr.dr.ir. B.H.M. Custers
	Dr. F. Dechesne
	Prof.dr.ir. A.P. de Vries (Radboud Universiteit, Nijmegen)
	Prof.dr. A. Hanbury (TU Wien, Austria)
	Prof.mr. C.J.J.M. Stolker

Authornote:
Gineke Wiggers is also employed as a business analyst at Legal Intelligence, a Wolters Kluwer product. Legal Intelligence/Wolters Kluwer did not provide any financial contribution to this research, the research has been funded solely by Leiden University.

SIKS Dissertation Series No. 2023-04
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Contents

1 Introduction **1**
1.1 The legal importance of legal information retrieval 6
1.2 A brief introduction to information retrieval (IR) 8
1.3 Features of legal IR 9
1.4 An overview of common evaluation methods in IR 11
1.5 A brief introduction to bibliometrics 13

2 Relevance **17**
2.1 Introduction 18
2.2 Background 21
2.3 Methods 27
2.3.1 Study design 27
2.3.2 Participants 28
2.3.3 Selection of stimuli 29
2.3.4 Extracting and mapping relevance factors 33
2.3.5 Statistical analysis 39
2.4 Results 40
2.4.1 Relevance factors 42
2.4.2 Differences between user sub-groups 46
2.5 Discussion 50
2.5.1 Implications 50
2.5.2 Limitations of the study 51
2.6 Conclusions 52

2.7 Acknowledgements 53

3 Citation Metrics **55**
3.1 Introduction 57
3.2 Literature analysis 58
3.2.1 Citations as a Form of Impact Measurement 59
3.2.2 Citations in Dutch Legal documents 61
3.2.3 'Scholarly' Legal Documents 62
3.2.4 Citations in Information Retrieval 65
3.3 Methods 67
3.3.1 Document sampling criteria 68
3.3.2 Document classification 68
3.3.3 Readership 73
3.4 Results 73
3.4.1 Citations between documents 74
3.4.2 Usage of documents 75
3.5 Discussion 76
3.5.1 Inter-rater agreement 76
3.5.2 Citations between documents 77
3.5.3 Usage of documents 81
3.5.4 Using bibliometrics in legal IR 82
3.6 Conclusion 83
3.7 Acknowledgements 86

4 Evaluation **89**
4.1 Introduction 90
4.2 Legal IR 92
4.2.1 The User 93
4.2.2 The IR Systems 95
4.2.3 The Documents 96
4.2.4 Relevance 97
4.2.5 Small Data 98
4.2.6 Legal Search and Professional Search 99
4.2.7 Summary 100

4.3 Expected limitations to common evaluation methods 101
4.3.1 Test Collections 101
4.3.2 User surveys 103
4.3.3 A/B testing 104
4.4 Empirical assessment of evaluation methods 105
4.4.1 Test collection based on expert relevance judgments 105
4.4.2 Test collection based on implicit feedback 107
4.4.3 Survey for ranking preferences 110
4.4.4 Survey based on the Net Promotor Score 114
4.5 Conclusion 114
4.6 Acknowledgements 116

5 Algorithm 117
5.1 Introduction 118
5.2 Background 120
5.2.1 Citations and usage in bibliometrics 120
5.2.2 Correlation between usage and citations 121
5.2.3 Usage in evaluation 121
5.3 Data analysis 123
5.4 Methods 129
5.4.1 Our proposed bibliometric-enhanced ranking variable 130
5.4.2 Evaluation 134
5.5 Results and analysis 139
5.5.1 Results 139
5.5.2 Statistical analysis (without clicks) 139
5.5.3 Statistical analysis (including clicks) 141
5.5.4 Practical significance 141
5.5.5 Analysis of long sessions 142
5.6 Conclusions 142
5.7 Acknowledgements 143

6 Discussion and conclusions 145
6.1 The answer (to the research question) 149
6.2 What does this answer mean for the future? 150

Bibliography **153**

Summary **171**

Samenvatting (Dutch Summary) **173**

Appendices **177**

A Composition of search results **179**

A.1 Example query 1 179

A.2 Example query 2 180

B Queries and Seed Documents **183**

C Seed Documents and Data **187**

D Baseline and degraded ranking **189**

E Example of Survey Question **193**

F Curriculum Vitae **195**

G Acknowledgements **197**

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.

Chapter 2

Relevance

Exploration of Domain Relevance by Legal Professionals in Information Retrieval Systems

Previously published as: Wiggers, G., Verberne, S., Zwenne, G-J., Loon van, W.S. (2022). Exploration of Domain Relevance by Legal Professionals in Information Retrieval Systems. Legal Information Management 22 (1) , pp. 49-67.

This paper addresses relevance in legal information retrieval (IR). We investigate whether the conceptual framework of relevance in legal IR, as described by Van Opijnen and Santos [131], can be confirmed in practice.

The research is conducted with a user questionnaire in which users of a legal IR system had to choose which of two results they would like to see ranked higher for a query and were asked to provide a reasoning for their choice. To avoid questions with an obvious answer and extract as much information as possible about the reasoning process, the search results were chosen to differ on relevance factors from the literature, where one result scores high on one factor, and the other on another factor. The questionnaire had eleven pairs of search results. A total of 43 legal professionals participated: 14 legal information specialists, 6 legal scholars and 23 legal practitioners.

The results confirms the existence of domain relevance as described in the theoretical framework by Van Opijnen and Santos [131]. Based on the factors mentioned by the respondents, we can conclude that document type, recency, level of depth, legal hierarchy, authority, usability and whether a document is annotated are factors of domain relevance that are largely independent of the task context.

We also investigated whether different sub-groups of users of legal IR systems (legal information specialists who are searching for others, legal scholars, and legal practitioners) differ in terms of the factors they consider in judging the relevance of legal documents outside of a task context. Using a PERMANOVA we found no significant difference in the factors reported by these groups. At this moment there is no reason to treat these sub-groups differently in legal IR systems.

2.1 Introduction

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." [2]. In everyday language, it is used to describe the effectiveness of information in a given context [110, p. 203]. In information retrieval (IR), the theory of

relevance has several dimensions, including algorithmic relevance, topical relevance, cognitive relevance, situational relevance, motivational relevance, and, in particular for legal information retrieval, bibliographic relevance [131].

The practice however, is that legal IR systems rely heavily on matching the text of the query with the text of the documents (algorithmic and topical relevance).[1] As Barry [12] points out, this may lead to poor user satisfaction.

Park [96] suggests that users of (legal) IR systems have implicit criteria for the relevance/value judgments about documents presented to them. This is supported by anecdotal evidence from employees of Legal Intelligence, one of two large legal content integration and information retrieval systems in the Netherlands. Users of the Legal Intelligence system have reported a preference for documents with certain characteristics over others, for example a preference for recent case law over older case law, case law from higher courts over case law from lower courts; sources which are considered authoritative (government publications) over blogs or news items, well-known authors over lesser-know authors, and/or the official version (case law or law) over reprints.

Van Opijnen and Santos describe a form of relevance that they call *domain relevance*, as "relevance of information objects within the legal domain itself (and hence not to 'work task or problem at hand')" [131, p. 71]. They relate domain relevance to the *socio-cognitive relevance* as defined by Cosijn and Ingwersen: "Socio-cognitive relevance is measured in terms of the relation between the situation, work task or problem at hand in a given socio-cultural context and the information objects, as perceived by one or several cognitive agents." [34, p. 541]. Until now, the implicit criteria as mentioned by [96], the domain relevance by [131], and anecdotal evidence from the Legal Intelligence users have not been connected to each other or studied systematically with users of legal IR systems.

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

Cosijn and Ingwersen state that "It is interesting to note that some central aspects of socio-cognitive relevance are tangible." [34, p. 541]. In this paper, we present the result of a user study investigating these tangible document characteristics (Saracevic's manifestations of relevance [110]) of domain relevance in legal IR, bridging that gap.

We focus on domain relevance because situational relevance and cognitive relevance are difficult to incorporate in legal IR systems, as the task and cognitive state of the users is usually not known. But if domain relevance is indeed shared between users, manifestations thereof may be used by legal IR systems to improve the relevance of their ranking, as suggested in the work of [32].

The First International Workshop on Professional Search[2] describes professional search as: "professional search takes place in the work context, by specialists, and using specialist sources, often with controlled vocabularies." [132] This definition covers people from multiple domains, including librarians, scholars, lawyers, and other knowledge work professions. In the context of users of Legal IR systems this includes professions such as lawyers, judges, government employees, legal information specialists and legal scholars. Experts in IR rather than law, such as professional support lawyers and information specialists, may, by nature of their role, have a different perspective of relevance. A professional support lawyer or information specialist retrieves information for a third person rather than themselves and will likely focus on completeness of the information, not being aware of the cognitive state of the client. A legal professional however, searching for themselves will likely focus only on information that is new to them. From this different perspective, they may have a different perception of relevance [109, p. 341].

In this paper we will distinguish three sub-groups of users: (a) legal information specialists, (b) legal scholars, and (c) legal practitioners. When referring to the overarching group of users, we will use the term legal professionals. Next to determining shared relevance factors between individual users, we will investigate whether these different user sub-groups show

[2]Held at SIGIR 2018, the report of which is available as Verberne et al. [132]

agreement on domain relevance factors in the context of judging search results.

We conducted the study with users of the Legal Intelligence[3] system, following Park, who argued that it is important to test with real users of the information retrieval system [96, p. 322]. We address the following research 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?

In answering these questions, this paper's contributions compared to previous work are: (1) we conducted a user study with 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; (2) using a statistical analysis we demonstrate 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.

2.2 Background

Our research is done in the context of the theory of relevance as described by Saracevic [109, 110]. He defines four research issues regarding relevance: the nature, manifestations, behavior and effects of relevance. Our research focuses on the manifestations of relevance; the different ways in which relevance manifests itself to users in legal IR systems. Saracevic also calls this 'clue research', 'uncovering and classifying attributes or criteria that users concentrate on while making relevance inferences' [110, p. 12]. These clues are described as attributes, criteria or factors of relevance, depending on the author. In the context of this paper, we will use the term 'factors'.

[3] `https://www.legalintelligence.com`

Saracevic [110] describes five types or spheres of relevance in which the manifestations can be grouped: algorithmic relevance, topical relevance, cognitive relevance, situational relevance and motivational/affective relevance. Van Opijnen and Santos [131] apply these spheres of relevance to the legal domain and developed a schema with six spheres of relevance, shown in Figure 2.1.

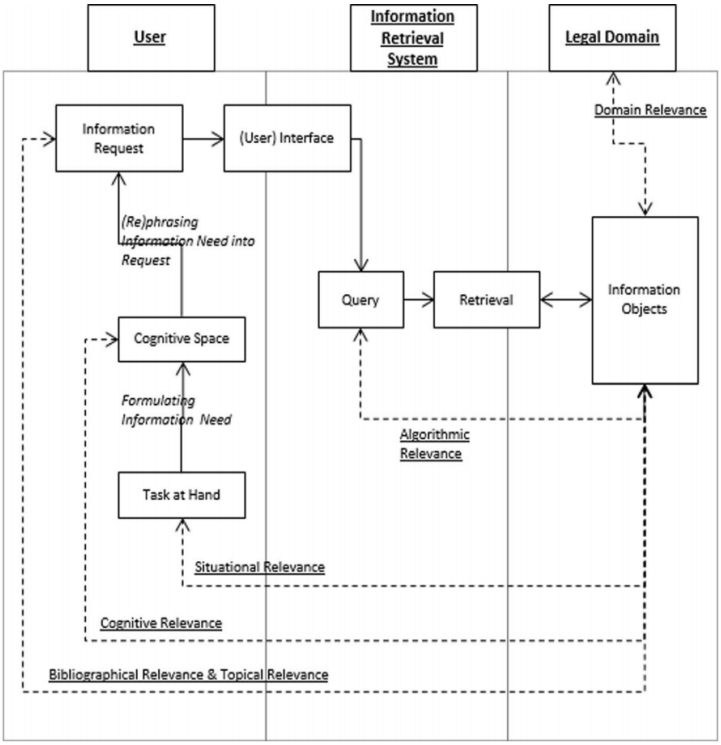

Figure 2.1: Relevance schema of [131]

Algorithmic relevance, sometimes called systemic relevance, is the degree to which the terms in the result match the terms in the query [131, p. 70]. This dimension focuses solely on the relation between the query and

the document, not on the user.

Topical relevance, which has also been referred to as 'aboutness' [25], moves beyond the query and the document, to the topic. An example would be that a query containing the term 'trust' in legal IR refers to a type of agreement involving three parties, rather than a belief that something or someone is good, and returning documents accordingly.

Van Opijnen and Santos [131] add bibliographical relevance to the list of Saracevic. Bibliographical relevance is the relation between the document searched for, and the document retrieved, also described as 'isness'. This is especially relevant for the retrieval of known documents. What makes legal IR interesting, is that it is not only about finding information about something, but that there can be an important legal difference between two documents holding the same information. For example, the officially published version of a law in the United States Statutes at Large or the Official Journal of the European Union, and a reprint of that same law [131, p. 70-71 and 76-77].

Van Opijnen and Santos [131], following Cosijn and Ingwersen [34] do not include Saracevic's affective relevance in their schema, but add domain relevance. This is chosen as their representation of Cosijn and Ingwersen's socio-cognitive relevance, who describe it as 'Socio-cognitive relevance is measured in terms of the relation between the situation, work task or problem at hand in a given socio-cultural context and the information objects, as perceived by one or several cognitive agents. It encompasses the system, a group of individual users or agents, and the socio-organisational environment.' [34, p. 541] Van Opijnen and Santos takes this to mean 'relevance of information objects within the legal domain itself (and hence not to 'work task or problem at hand')' [131, p. 71]. They describe this as 'the relation between the legal crowd and information objects' with 'legal importance' as criterion for success [131, p. 73].

Cognitive relevance focuses on the relation between the cognitive state and information need of the user and the document. It is unique to the user and the specific point in time, as it encompasses factors like informativeness, quality and novelty [113, p. 760].

Situational relevance as described by Saracevic is "...relevance to a par-

ticular individual's situation – but to the situation as he sees it, not as others see it, nor as it really is" [139, 109, p. 335]. [131] describe it as the relation between the documents and the work task of the user. Situational relevance plays a role in domain specific search, because relevance in specific domains depends on the expertise and context of the searcher [60].

Our research explores which document characteristics (Saracevic's manifestations of relevance [110]) reflect domain relevance when judging the relevance of search results in legal IR. This is inspired by the work of Cool et al. [32], and focuses on factors of relevance that are 'representable and usable in the support of information interaction/retrieval' [32], meaning that these factors can be used by the developers of legal IR systems to improve their ranking. It is also inspired by the work of Toms et al. [128] with regards to relating relevance factors to spheres of relevance. It is likely that a relevance judgment based on document representations (in our research title and publisher curated summary) differs from the relevance judgment upon reading the entire document [109, p. 340]. Because the document representation is what the user bases their judgement whether to open the document on, our research focuses on the document representation, and we use the term 'perceived' relevance.[4]

Van Opijnen and Santos [131] divide domain relevance into two subgroups: (a) the legal importance of classes of information objects, and (b) the legal importance of individual information objects. An example of the legal importance of classes of information objects are the prevalence of the constitution over other types of laws, and verdicts from the supreme court which have more legal authority than verdicts from lower courts. The legal importance of individual information objects is more difficult to describe in manifestations of relevance, but can for example be established through citation analysis.

Prior research has identified relevance factors, manifestations of the spheres of relevance. Rieh and Belkin [106] addressed the user's perception of quality and authority as relevance factors. They identified seven different factors of information quality: source, content, format, presen-

[4]Barry [12, p. 152] uses the term predicted relevance.

tation, currency, accuracy, and speed of loading, and two different levels of source authority: individual and institutional. Savolainen [112] found in an exploratory study that specificity, topicality, familiarity, and variety were the four most mentioned factors in user-formulated relevance judgments, but there was a high number of individual factors mentioned by the participants. This research has been expanded upon by Taylor et al. [125, 123, 124].

Previous research on factors of relevance in the context of professional search has been done by, amongst others, Cuadra and Katter [39] and Rees and Schultz [103], who examined judgements by expert reviewers. Barry [12] expands on this research by inviting users (rather than expert reviewers) to submit a request for unknown or unfamiliar information[5], for which she retrieved documents. The results lists were presented to the participants, who were asked to review whether they would or would not pursue the documents contained in the list. The study was done using an open-ended interview technique, to generate a complete overview of relevance factors. Barry identified 23 categories of relevance factors, grouped into seven classes [12]:

- the information content of documents: depth/scope, objective accuracy/validity, tangibility, effectiveness, clarity, recency;
- the sources of documents: source quality, source reputation/visibility;
- the document as a physical entity: obtainability, cost;
- other information or sources within the environment: consensus, external verification, availability within the environment, personal availability;
- the user's situation: time constraints, relationship with author;
- the user's belief and preferences: subjective accuracy/validity, affectiveness;

[5]Thereby excluding known item retrieval.

- the user's previous experience and background: background/experience, ability to understand, content novelty, source novelty, stimulus document novelty.

Barry [12] distinguishes between 'tangible characteristics of documents', subjective qualities and situational factors.

Schamber, in Barry and Schamber [13], conducted structured time-line interviews with users searching for weather information. Schamber identified 22 categories of relevance factors, grouped into ten classes:

- accuracy;
- currency: time frame
- specificity: summary/interpretation, variety/volume;
- geographic proximity;
- reliability: expertise, directly observed, source confidence, consistency;
- accessibility: availability, usability, affordability;
- verifiability: source agreement;
- clarity: verbal clarity, visual clarity;
- dynamism: interactivity, tracking/projection, zooming;
- presentation quality: human quality, nonweather information, permanence, presentation preference, entertainment value, choice of format.

As opposed to the work of Barry and Schamber, the aim of our research is not to generate a complete overview of relevance factors that may possibly be considered, but to determine whether it is possible to identify relevance factors that can be classified as domain relevance, requiring a level of agreement between different (groups of) users to establish the legal importance/wisdom of the legal crowd as described by Van Opijnen and Santos [131]. Furthermore, we investigate relevance in the legal domain, as opposed to general academic search and weather information.

2.3 Methods

2.3.1 Study design

Previous studies addressing relevance factors conducted user observation studies with a thinking-aloud protocol or interviews, or a combination of both [12, 112, 106, 60]. This research focuses on domain relevance. Based on the definition of [131] this encompasses two aspects: (1) a level of agreement between users, the 'legal crowd', and (2) it is not related to the task or problem at hand. An observation study with information needs submitted by the respondents is likely to also trigger responses related to situational and cognitive relevance. This is not desired in our case. And since observation studies are time consuming and therefore difficult to conduct with legal professionals, we decided to conduct a questionnaire.

The choice to use actual users rather than domain experts such as grad students was influenced by Park [96]. A preliminary pilot questionnaire suggested that the target audience is not likely to complete a questionnaire that takes more than 12 minutes, because lawyers often bill per 6 minutes, and are unwilling to spend more than two billable units on a questionnaire. To ensure maximum response, we aimed to keep the questionnaire under 12 minutes.

Forced choice/relative relevance judgments In the questionnaire respondents are shown an example query and two search results and forced to make a choice between two options; a relative relevance judgment by indicating which of the two results they would like to see ranked higher than the other. We chose a method of forced choice/relative relevance scoring because research by [109] shows that the less a person knows about the subject, the more results they will mark as relevant (cognitive relevance), and that relative scoring (thereby limiting the effect of cognitive relevance) leads to more consistent results across respondents of different backgrounds than individual document scorings.

2.3.2 Participants

All users of the Legal Intelligence system were able to fill in the questionnaire. The questionnaire was made available online, so that respondents could fill it in at a moment convenient for them, to ensure maximum response. It was distributed to the national government and large law firms through their information specialists, and to all other users by a newsletter and a LinkedIn post. The questionnaire was brought to the attention of acquaintances who work in the legal field via email. We aimed for 50 responses, distributed over the different affiliation types, law area specialisms, and roles. The number of participants is more than in previous qualitative studies in professional search, see for example Schamber et al. [113] with 30 respondents, Barry [12] with 18 respondents, and Park [96] with 10 respondents. For our questionnaire type analysis (rather than the interviews of Schamber, Barry and Park), it is a feasible number.

Structure The questionnaire consisted of three parts. The first part covered general questions regarding the legal field the respondent is active in, their function profile, and their level of expertise.

For each of the next two parts of the questionnaire, the respondents were shown an example search query. The respondents were first asked to indicate what information need they think the user is trying to fulfill by issuing this query. It is expected that because we use example queries rather than the users' own information needs, respondents will focus on relevance factors from the algorithmic, bibliographical, topical and domain relevance spheres, rather than the cognitive and situational relevance spheres. This is in accordance with the aim of our research of finding relevance factors related to domain relevance. With this question we aimed to determine the extent to which cognitive and situational relevance played a role in the mind of the respondent.

Research has shown that the primitive/intuitive definition of relevance prevails when respondents are confronted with questions regarding relevance judgment [110]. For that reason, no formal definition of relevance was given in the questionnaire. In the introduction of the questionnaire

some examples of factors were given[6]. To avoid leading the respondents, and to encourage respondents to consider all aspects from both results, these examples were not repeated alongside the questions.

2.3.3 Selection of stimuli

We manually selected two example queries from the query logs of the Legal Intelligence search engine (see Appendix A). The example queries are shown to provide the context for the search results, and are broadly recognizable, so that all respondents will have an understanding of algorithmic, topical and bibliographical relevance of the search results (document representation) in relation to the query. To exclude query bias, all respondents are shown the same two queries.

The query along with two related search results, were shown as images from document representations as they are shown in the actual legal IR system. The interface of the pairwise choices is illustrated in Figure 2.2.

[6]Translated the examples read: 'This could be because the title or summary seems more relevant, the result comes from an authoritative source, the publication date of the document, or because it is a document type where you expect to find the answer to the query.'

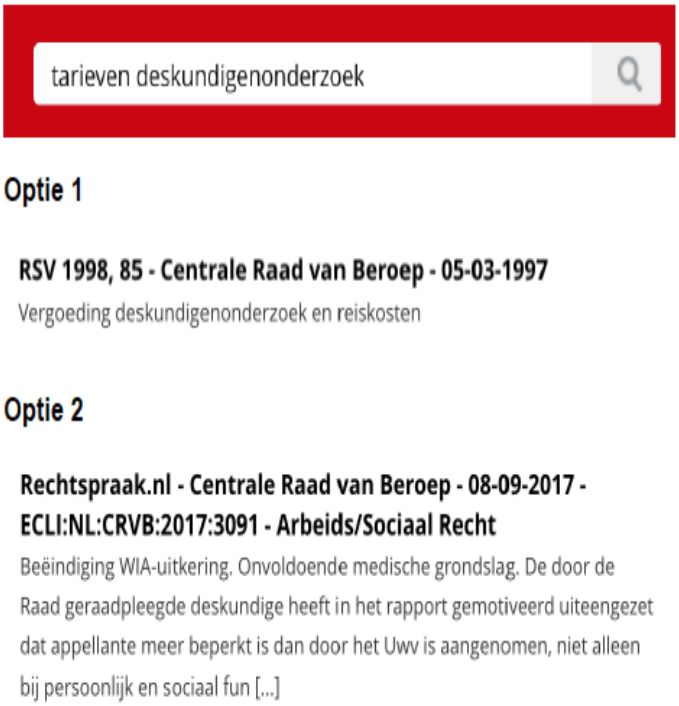

Figure 2.2: A screenshot of the questionnaire. The example query is shown in the query field on top and the two search results (choices) are listed as 'optie 1' and 'optie 2' below the query.

Selection of search results

To make sure the forced choices/relative relevance judgments for the example queries encouraged critical thinking, we chose search results from the actual results list of the Legal Intelligence system. We chose two search results that differ on two of the relevance factors that are mentioned in the literature as relevance factors for (Legal) IR (see Section 2.2) that can be discerned from the information provided in the search result (document

representation). We chose two search results where one has a higher score on the first factor and a lower on the second, and the other has a lower score on the first factor and a higher score on the second. The other factors are kept as similar as possible, given that the examples have to be chosen from actual search results. For the creation of these options, see Appendix A.

By ensuring that the two search results differ on two factors while the other factors are as similar as possible, we aimed to avoid creating an 'obvious' choice and encourage users to describe their reasoning process.

Relevance factors

The factors used for the selection of the search results are selected from the literature (see Section 2.2). We focus on factors of relevance that manifest themselves in the document representation (as shown on the result page), since the questionnaire does not include the document itself. To avoid bias by leading the respondents to answers demonstrating the existence of domain relevance and allow factors of algorithmic, topical and bibliographical relevance to be considered, these selection factors are not limited to expected factors of domain relevance, but encompass a broader scope of factors of relevance. In the setup of the questionnaire each possible relevance factor occurs multiple times (see Appendix A). The factors used were:

- Recency [12, p. 156]: it has been suggested that recent case law is more relevant than older case law (<2 years; 2 – 10 years; >10 years old), though recency can also be related to the specific period the case played in [131, p. 80], in which case it would be situational relevance; Schamber's time frame factor [13];

- Legal hierarchy/importance [131, p. 68]: case law from higher courts carries more weight than case law from lower courts (supreme court; courts of appeal; courts of first instance);

- Presence of annotation[7]: annotated case law (providing context for the case) is more relevant than case law that is not annotated. Related to Schamber's factor of summary/interpretation [13];

- Source authority[8] [12, p. 156]: sources that are considered authoritative are preferred over other sources (government documents, leading publications; mid-range publications; blogs);

- Author authority[9] [12, p. 155-156]: documents written by well-known authors are considered more authoritative than other documents;

- Bibliographical relevance [131, p. 71]: the official version (case law or law) is more relevant than reprints;

- Title relevance: results with search term in the title or summary are considered more relevant than results with the search term not in the title/summary (the visibility of algorithmic and topical relevance for respondents);

- Document type[10]: document types that pertain to the perceived information need are considered more relevant than other document types (depending on perceived information need expressed in the query as interpreted by respondent). Related to Schamber's presentation quality, especially the underlying factors of presentation preference and choice of format [13].

The respondents were not informed for which relevance factors the paired results were chosen. The chosen factors were not mentioned explicitly in the questionnaire, to avoid leading respondents.

Where authoritative sources or authors are used in the examples, it was attempted to show sources and authors that are so generally known that

[7] Mentioned by users to Legal Intelligence employees.

[8] Also described as source quality.

[9] Also described as relationship with author and source reputation/visibility

[10] Van Opijnen and Santos [131, p. 68] mention the large diversity in document types in legal information retrieval.

respondents from other legal fields will likely recognize these names from their legal education, or can estimate it by the academic title of the author. It is assumed that the other factors used in the examples, such as whether a case is annotated, are valid for all legal fields.

Though the factors chosen to base the examples on are prominent in literature, they are by no means an exhaustive list. Nor do they have to be, since they are used as a tool to select good examples that encourage the respondent to think and describe their thought process. Respondents are given a free text field to give their own motivation for their choice.

Because of the time limit discovered during the pilot questionnaire, the number of queries is limited to two and it was not possible to show all possible combinations of factors. Each participant saw eleven pairs of search results spread over the two example queries. Because of the expected sample size, we presented all users with the same search results, to ensure enough data per question. Since the purpose of our research is to gather qualitative information to understand the factors that influence the perception of (domain) relevance, and the factors are inputted into the questions only to avoid an 'obvious' choice, the fact that not all combinations were tested does not limit the outcome of the research.

Likewise, because the research focuses on the factors that influence the choice, rather than the choice itself, there was no benefit in presenting the questions in a different order to different users.

2.3.4 Extracting and mapping relevance factors

Respondents could give a free text explanation for each of the forced choices. Often, these explanations contained one or more relevance factors, or a statement indicating the respondent had no preference. We manually aggregated and linked the free text explanations to the most similar relevance factors found in literature (see Section 2.2), which include the factors used to select the options (see Section 2.3.3), but also other factors mentioned by users.

Examples of the mapping include:

- Recency: 'Newer' or 'Appears out of date';

- Legal hierarchy: 'Supreme court higher than appeals courts';
- Annotated: 'Annotated case law is always relevant';
- Authority: 'Hartkamp is a well-known author' or 'If a verdict is reprinted in a journal it says something about the importance';
- Bibliographical relevance: 'Source instead of derivative';
- Title relevance: 'Doesn't show anything about the possible contents of the document';
- Document type: 'It's the law!' or 'explanatory memorandum not first thing to look at';
- Level of depth: 'General piece' or 'Good broad starting point';
- Law area (topic): 'Because it is civil law';
- Usability: 'More relevant information quickly' or 'Convenient source';
- Document length: 'Reports are often very long' or 'option 2 would take more time to read';

Some answers did not contain relevance factors. Either because none were given (e.g. 'Duh') or because the answer was too vague to extract relevance factors (e.g. 'More relevant'). It was also possible for a single answer to mention multiple factors. An example is the response 'Option 1 because it comes from a higher court. From option 2 the annotation is interesting.'[11] In four instances the respondent indicated 'same answer as before'. In those instances we looked at the response from the previous example and noted the same factor(s) as for the previous examples.

In the explanations, users regularly referenced authority, without mentioning whether they meant the authority of the source or the author.[12]

[11] translated from: 'optie 1 omdat dit hogere rechtspraak is. Bij optie 2 is de noot interessant.'

[12] Of the 31 times authority was mentioned, 15 mentions appear to be related to source authority, 9 mentions appear to be related to the authority of the author, and 7 mentions provide no context as to whether the authority refers to the source or author.

This is related to Schamber's reliability factor, which covers both author and source [13]. Like Schamber we grouped the authority arguments. This lead to a total of 11 main relevance factors which were mentioned at least once.

After the mapping, we counted for each participant the number of times each of the 11 relevance factors were mentioned. This counting was performed across all answers, since the individual questions were not of interest. This way we obtained a vector of 11 relevance factor frequencies for each participant.

Relevance factors are often grouped into types [12, 13, 125]. Barry [13] creates groups for factors pertaining to the information content, the user's background, the user's beliefs and preferences, other information in the environment, the sources of the document, the physical entity of the document, and the user's situation. Barry and Schamber [13] also show an example of grouping into accuracy, currency, specificity, geographical proximity, reliability, accessibility, verifiability, clarity, dynamism, and presentation quality. These groups show similarities with the spheres of relevance, but are not the same.[13] To be able to see to what extent domain relevance can be demonstrated, and which relevance factors (manifestations) are related to it, we manually relate the found relevance factors to the six spheres of relevance mentioned in Section 2.2.

Based on Van Opijnen and Santos [131], two conditions need to be met before relevance factors can be mapped to domain relevance (whether on information class level or individual document level): (1) there is a level of agreement between users, the 'legal crowd', and (2) the relevance factor is not related to the task or problem at hand as described in Section 2.2. If

[13]Most notable is the absence of a group for factors of algorithmic or topical relevance. The research of Schamber, as described in Barry and Schamber [13], extracted relevance judgments from documents that were pursued after viewing the search result, meaning the algorithmic and topical relevance can be inferred (otherwise the respondent would not have pursued the document). In the research of Barry [12] however, responses were also noted for documents that were crossed out as 'would not pursue'. But also in this research the focus was on the identification of relevance 'criteria beyond topical appropriateness' [12, p. 149].

respondents do not indicate a situational context for the example queries used in the questionnaire, the mapping of the relevance factors found to spheres of relevance can be done with the assumption that the method with which the data is gathered implies that factors mentioned are not related to task context. Similarly, since the query shown was an example rather than a query from the user themselves, the user is asked to take a step back from what they already know (their personal cognitive state). This means that factors like 'novelty' (whether the information is new to the user or not), which would be a factor of cognitive relevance, is less likely, as the user is not considering the information in relation to themselves, but to a hypothetical other user. Because the user is not relating their answers to themselves, but to this hypothetical other user, factors that would normally be grouped under cognitive relevance become a factor of domain relevance.

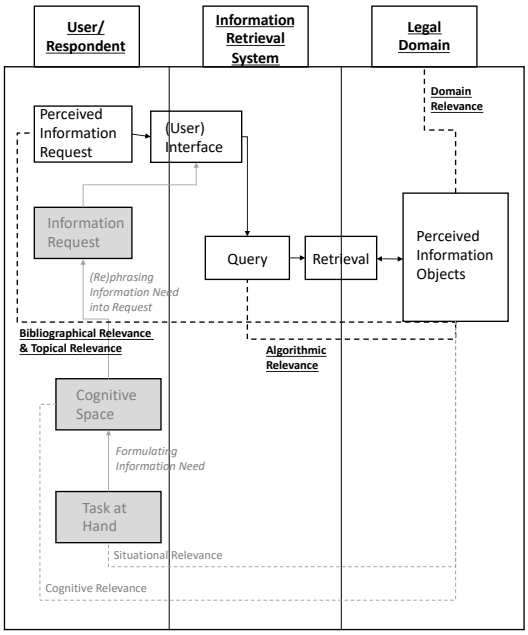

Figure 2.3: An adaptation to the relevance schema of Van Opijnen and Santos [131] to reflect the method using a perceived information request and the lack of cognitive and situational relevance.

The mapping of the relevance factors is then done based on the schema shown in Figure 2.3 using the following sequential steps:

- If the relevance factor is about the (perceived) 'computational relationship between a query and information representation' [131, p. 70]:
  - Then the relevance sphere is (perceived) algorithmic relevance.

- Else if the relevance factor is about the (perceived) 'relationship between the 'topic' (concept, subject) of a request and the information objects' [131, p. 70]:
  - Then the relevance sphere is topical relevance.
- Else if the relevance factor is about the 'relationship between a request and the bibliographic closeness of the information objects' [131, p. 71] (document 'isness'):
  - Then the relevance sphere is bibliographical relevance.
- Else:
  - The relevance sphere is domain relevance.

For factors mapped to domain relevance according to the steps above, we will report additional information: (a) whether it constitutes domain relevance on the document class level or on the individual document level [131], and (b) whether it would have been classified under situational, cognitive or domain relevance had the user given this answer in relation to a personal task. For this additional information we will use the following definitions:

- If the relevance factor is about 'the relation between the information needs of a user and the information objects' [131, p. 71]:
  - Then the relevance sphere is cognitive relevance.
- If the relevance factor is about 'the relationship between the problem or task of the user and the information objects' [131, p. 71]:
  - Then the relevance sphere is situational relevance.
- If the relevance factor is about 'the relevance of information objects within the legal domain itself' [131, p. 71]:
  - Then the relevance sphere is domain relevance.

Using these grouped and mapped factors, in the format of counts in the vector of 11 relevance factor frequencies for each participant, we conduct a statistical analysis.

2.3.5 Statistical analysis

To test whether sub-groups of users of legal IR systems differ on the factors of relevance they consider, we first grouped every respondent who reported their function or job title as legal information specialist, librarian, professional support lawyer or a function in knowledge management as legal information specialists. Next, those who reported their job title as scholar were grouped as legal scholars. All other job titles were grouped as legal practitioners.

We then calculated inter- and intra-group dissimilarity, and performed a permutational multivariate analysis of variance (PERMANOVA) [7, 8]. Classical MANOVA assumes multivariate normality, which is unrealistic given that our data consists of relevance factor frequencies. PERMANOVA is a semiparametric alternative to MANOVA that does not assume multivariate normality [8]. Briefly, the PERMANOVA procedure with one predictor variable is as follows: (1) Calculate a suitable measure of dissimilarity for each pair of respondents. (2) Calculate sums of squares attributed to differences among the groups (SS_A), and differences within each group (SS_W). (3) Calculate the pseudo-F statistic $F = (SS_A/SS_W) \cdot [(N-g)/(g-1)]$, where N is the sample size and g the number of groups. (4) Perform a permutation test to obtain a p-value. PERMANOVA thus compares dissimilarities between individuals in different groups (SS_A) with dissimilarities between individuals in the same group (SS_W); if the ratio between the two quantities is sufficiently large, the null hypothesis of no difference between the groups will be rejected. The approach for situations with more than one predictor variable is similar, with sums of squares attributed to each variable; for a detailed description of the PERMANOVA method we refer to [8]. As a suitable dissimilarity measure we used the cosine dissimilarity (one minus the cosine similarity), which is commonly used for judging the dissimilarity of documents when they are represented as word frequency

vectors [114]. Cosine dissimilarity also allows for respondents who habitually provide more factors per explanation than others. It considers the relative frequency of mentions of a factor for that respondent, not the absolute value. In the same way as it normalizes for document length when measuring text similarity, it compensates for the different amounts of factors respondents provide per question. PERMANOVA was performed in R [101], using package `vegan` [93].

2.4 Results

A total of 43 respondents completed the questionnaire. The respondents came from a range of areas of legal expertise, function types, organization types and years of work experience, as shown in Table 2.1. There were 11 query-answer pairs, leading to $(43*11 =)$ 473 choices made. In 28 instances (6%), the respondent indicated they had no preference for one option or the other. In 90 instances (19%), there was no (clear) explanation.

Respondents were asked to indicate what information they thought the user was trying to find with this query. This question was asked for two reasons: (1) to verify that the query is broadly recognisable, and (2) to verify that users interpret the example query without situational context, and do not for example imagine a situational context in their mind. Though users sometimes interpreted the query to be aimed at a specific information type (e.g. a law [14] or case law [15]), none of them described a situational context (e.g. They want to hire an expert witness for their case and want to know how expensive that will be, or I have contract that I want to get out of with retroactive effect and I want to know how to do that and what the consequences will be.). This is also reflected in the fact that out of the 86 responses given to these interpretation questions, the word 'I' was only used twice, which also affirms our assumption that situational and cognitive

[14] 'Regelgeving, bijvoorbeeld een Staatscourant.' translated: 'Regulation, for example the Government Gazette'

[15] 'Naar rechtspraak over de vernietiging van een overeenkomst met terugwerkende kracht' translated: 'case law on the cancellation of an agreement with retroactive effect'

Law area	Number of respondents
Labour law	8
Intellectual property law	7
Multiple law areas (e.g. information specialists)	5
Criminal law	3
IT/Privacy law	3
Tax law	3
Administrative law	2
Bankruptcy law	2
Contract law	2
Environmental law	2
Competition law	1
Family law	1
Financial law	1
Tort/Liability law	1
Transport law	1
Other	1
Group	**Number of respondents**
Lawyers/legal practitioners	20
Legal information specialists (including librarians and professional support lawyers)	14
Legal scholars	6
Students	2
Management	1
Affiliation	**Number of respondents**
Law firm	28
Government (including courts and local government)	9
University	5
Legal department commercial organization	1
Work experience	**Number of respondents**
0-5 years	10
6-10 years	4
11-20 years	19
More than 20 years	10

Table 2.1: Breakdown of respondents

relevance can be precluded, as discussed in Section 2.3.4.

The respondents often mention relevance factors when selecting a search result that were not part of the factors for which the corresponding search results were chosen. It therefore seems that the factors behind the selection of the examples were not so obvious in the presented questions that they lead respondents in their answers. On average, respondents are split 31:12 over the choices. The highest agreement reached for a choice was a division of 40:3, and the lowest agreement was 20:23. This suggests that the method used to select the search results invited critical thinking.

The relevance factors mentioned in the motivations of the respondents are listed, analyzed and discussed in the next subsection.

2.4.1 Relevance factors

All mentioned factors are listed in Table 2.2. As described in Section 2.3.4, and based on the responses given by the respondents to the query interpretation question, when mapping relevance factors to spheres of relevance, we exclude mapping to situational and cognitive relevance. Because respondents were not asked to assign a weight to the factors in the outcome of their choice or whether it was the determining factor, the raw count of the factor does not indicate its importance, only how often it was mentioned.

Aside from the factors mentioned used to select the examples, which are described in the Section 2.3, respondents mentioned four factors in their considerations of which documents they wish to see ranked higher: (1) the level of depth or detail of a document; described by Barry [12] as depth/scope and by Schamber as specificity [13]; (2) the law area of the document, as determined through the title, source or author; (3) the usability of the document [13], described in the factors of Barry [12] as effectiveness; (4) the length of the document, related to what Barry [12] describes as time constraints, and Schamber as variety/volume [13].

Factor	Times mentioned	Associated Sphere of Relevance
Title relevance	154	(perceived) Algorithmic Relevance
Document type	68	Domain Relevance
Recency	56	Domain Relevance
Level of depth *	59	Domain Relevance
Legal hierarchy	42	Domain Relevance
Law area (topic) *	31	(perceived) Topical Relevance
Authority/credibility (total)	31	Domain Relevance
Usability *	15	Domain Relevance
Bibliographical relevance	12	Bibliographical Relevance
Annotated	7	Domain Relevance
Document length *	2	Domain Relevance

Table 2.2: Relevance factors sorted by number of mentions in the free text field. An asterisk (*) indicates that the factor was not one of the factors used to select example results (listed in Section 2.3) but added by participants

Title Relevance

The Title relevance is a factor of perceived algorithmic relevance, because it covers the (perceived) relationship between the query and the document representation [131, p. 71]. Though all results shown were actual results returned by the Legal Intelligence system, and therefore deemed to be algorithmicly relevant, users mentioned the presence or absence of the query terms in the snippet as factor to prefer one result over the other. We therefore call this 'perceived' algorithmic relevance. In the work of Schamber [13] this might be considered a factor of 'presentation quality'.

It is not surprising that the perceived algorithmic relevance is the most often named factor. Park [96] and Saracevic [111] also describe this as a major factor in the perception of relevance in relation to professional search. This is likely because snippets play a role in understanding the algorithmic relevance of search results [88].

Document Type

Document type constitutes a factor of domain relevance on the document class level. In a different context than this research, it would likely constitute a factor of cognitive relevance, since it deals with the relationship between the information need of the user and the information objects [131, p. 71]. Though we have excluded that respondents perceived cognitive relevance in the example query, this focus on information type lingers, as is demonstrated by the fact that the factor of document type was already visible in the question regarding the interpretation of the query. Responses include references to a law or case law.

Recency

Recency is a factor of domain relevance, which can be argued to be both on document class and individual document level. While anecdotal evidence suggests that in general newer documents are considered more relevant than older documents, this may differ if the legal professional is dealing with a case from the past. In a different context than this research this would be a factor of either cognitive relevance (newer information is more likely to be novel for the user and thus more likely to solve an information need of the user) or situational relevance (if they are working on a case from a particular period recency becomes a factor dealing with the relationship between the task of the user and the information object) [131, p. 71].

Level of Depth

The level of depth or detail of a document is a factor of domain relevance on the individual document level. In a different context than this research this factor would be considered to relate to cognitive relevance. Depending on the familiarity of the user on the subject, they will be looking for a high-level document (introduction to a subject they are not yet familiar with), or a very detailed document. It therefore deals with the relationship between the information need of the user and the information object [131, p. 71].

Legal Hierarchy

The legal hierarchy is a factor of domain relevance on the document class level. Given the legal status represented by this factor (e.g. in case of appeals), this factor would also be mapped to domain relevance in situations other than this questionnaire.

Law Area (Topic)

The factor of Law Area (topic) is a factor of (perceived) topical relevance. Respondents indicate that they are only interested in results that relate to a specific law area.[16] It therefore deals with the relationship between the (perceived) topic of the request and the information objects[131, p. 70], and suggests that the respondent considers topical relevance to be delimited by law area.

Authority/Credibility

The factor of Authority/Credibility is a factor of domain relevance on the individual document level. In a context other than this questionnaire it would most likely be related to the sphere of situational relevance. It is often considered in relation to the persuasiveness/citability of the document, and would therefore likely be related to the work task of the user [131, p. 71] rather than their cognitive state.

Usability

The factor of Usability is a factor of domain relevance on the individual document level. In other instances than this research it would most likely be related to situational relevance, as usability relates to the underlying motivation for information retrieval [131]. It shows overlap with citability.

[16] Translated: "The second option because it is civil law", "Civil, not fiscal"

Bibliographical Relevance

The factor of bibliographical relevance is related to the sphere of bibliographical relevance, as it concerns the 'isness' of the document. Or, as described by Van Opijnen and Santos [131, p. 71] 'the relationship between a request and the bibliographic closeness of the information objects'.

Annotated

The factor of whether a document is annotated or not, is a factor of domain relevance on the document class level. Annotations provide context for the reader, and this is considered preferred in general (though individual annotations may be considered irrelevant because of a (perceived) lack of quality). In a different context the factor of annotated would be considered a factor of cognitive relevance, as it regards the relationship between the information need of the user and the information object [131, p. 71].

Document Length

The factor of document length is a factor of domain relevance, likely on the document class level. In other situations this factor would likely be mapped to situational relevance, as it relates to the task of the user and the amount of available time and completeness required [131, p. 71]. However, the mentioning of this factor in the questionnaire, even when there is no task and therefore no time constraint (note that the users did not have access to the document, only to the snippet shown on the result page), suggests that legal professionals prefer not to read very long reports in general.[17].

2.4.2 Differences between user sub-groups

Table 2.3 shows the total frequencies of each factor, aggregated per subgroup. At a first glance there appear to be small differences in the factors mentioned between legal practitioners, legal scholars, and legal information

[17]Translated: 'reports are often very long', original:'zijn rapporten vaak erg lang'

specialists. Practitioners mentioned the length of the document as a factor, which the information specialists and scholars did not. The factor of usability is also named relatively more often by practitioners than other sub-groups. The group of information specialists appear to mention the level of depth less than other sub-groups. Whereas the group of legal scholars mention legal hierarchy and authority less than the other sub-groups.

Relevance factor	Information Specialists	Scholars	Practitioners
Title relevance	45	28	81
Document type	19	10	39
Level of depth	10	14	35
Recency	17	7	32
Law area (topic)	5	8	18
Legal hierarchy	19	3	20
Authority/credibility	12	3	12
Usability	1	4	10
Bibliographical relevance	5	1	6
Annotated	2	2	5
Document length	0	0	2

Table 2.3: Relevance factors sorted by the number of mentions in the free text field, according to function type

To further analyze the differences between the groups, we visualize the differences between respondents at an individual level using the cosine dissimilarities between them. The dissimilarities are visualized in two dimensions in Figure 2.4. It can be observed that the different groups are not well-separated in the two dimensional space, and the observed distances between the group centroids are small compared to the observed distances within each group.

To test whether the differences between legal information specialists, legal scholars, and legal practitioners are statistically significant, we performed a PERMANOVA. We included two predictor variables in our anal-

ysis. The primary predictor variable of interest was whether someone is a legal professional, scholar, or an information specialist. Years of working experience (0-5 years, 6-10 years, 11-20 years, 21+ years) was added as an additional variable to correct for possible existing differences in years of working experience between the user subgroups in our sample. One respondent was excluded from the analysis because they did not provide an explanation for any of the questions, leading to a total sample size of 42 (23 legal practitioners and 13 information specialists, and 6 scholars). All permutation tests were performed using 10,000 permutations.

PERMANOVA is somewhat sensitive to heterogeneity of multivariate dispersions, meaning that significant results may be caused by different variation within each group, rather than differences between the groups [8, 93]. Therefore, we first performed a permutation test for homogeneity of multivariate dispersions [93] for the user subgroups (pseudo $F = 1.43$, $p = 0.248$) and for years of working experience (pseudo $F = 0.076$, $p = 0.976$). Neither test was significant, thus providing no evidence of different variation within each group.

The PERMANOVA results can be observed in Table 2.4. Note that the test for the interaction is conditioned on the main effects, and each tests for a main effect is conditioned on the other main effect. The interaction effect was not significant ($p = 0.892$). The difference the user subgroups was not significant ($p = 0.243$), nor was the main effect of years of working experience ($p = 0.344$).

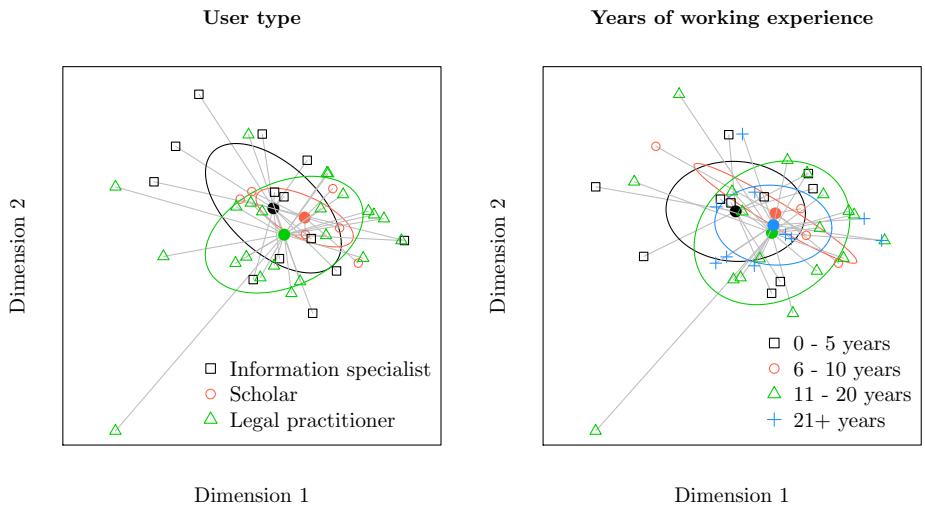

Figure 2.4: The dissimilarities between respondents visualized in two dimensions using principle coordinates analysis (PCoA). Left: Labeled as legal information specialist, legal scholar or legal practitioner. Right: Labeled by years of working experience. The connected dot in the center of each group represents the group centroid; the ellipses represent one standard deviation around each centroid.

Source	df	SS	pseudo F	p
User type	2	0.233	1.476	0.243
Years of working experience	3	0.290	1.225	0.344
Interaction	6	0.236	0.453	0.892
Residual	30	2.606		
Total	41	3.357		

Table 2.4: PERMANOVA results.

2.5 Discussion

2.5.1 Implications

Relevance factors

Legal IR systems appear to focus on algorithmic and topical relevance, while the results of this research show that users (when judging search results outside of the context of a particular worktask) have agreement on other manifestations of relevance that are visible in document representations and could be used in ranking algorithms. Such as the factors of recency, whether a document is annotated or not, legal hierarchy and bibliographical relevance, that can easily be recognised by IR systems to enhance their ranking. Incorporating the lessons learned from this research could enhance the user experience.

It is interesting to note that document type is the second most mentioned consideration for the respondent's relevance choices. This suggests that when users of legal IR systems are searching for something, they know what type of document they are likely to find the information in. Similarly, the level of depth respondents are looking for (fourth most reported argument) influences what document types they open. These factors appear to be related to the sphere of cognitive relevance (the relation between the information need and the document), making it more challenging to incorporate these factors in ranking algorithms.

The mentioning of document length by two respondents separate of each other, in regards to a different question, and in the absence of a situational task, suggests that legal professionals might prefer not to read very long reports. Though the number of respondents is too small to reach strong conclusions, it is interesting, as anecdotal evidence suggests that this might be different in certain situations (e.g. when trying to bury the opposing party in work or when that particular document is very pertinent to a certain task) suggesting that while there is some general consensus on the preference of shorter documents, situational relevance might be stronger than domain relevance.

Though not an aim of this research, it is interesting to note that the

most reported relevance factor, whether the word is in the title or summary of the result, suggests that simple changes in the user interface might already improve the perception of the quality of the ranking for users, without actually changing the ranking itself. An IR system will only return documents that are algorithmically relevant – in the sense of containing query terms – but the results suggest that the respondents find it challenging to perceive the relevance of a document if they do not see the search terms in the title or the summary. By showing snippets (where the section of the document where the query terms are found is shown), rather than publisher curated summaries as is currently the case in the system and examples used for our research, users will be able to see the query terms in context. This type of document representation will likely enable users to better estimate the relevance of the document.

User subgroups

In this study the observed differences between legal information specialists, legal scholars, and legal practitioners, in terms of the factors they consider in judging the relevance of legal documents, were not found to be statistically significant. At this moment there is no reason to treat these sub-groups differently in legal IR systems.

2.5.2 Limitations of the study

Our research focuses on relevance factors that are visible as document characteristics. The chosen method excludes situational relevance.[18] The relevance factors found are therefore not an exhaustive list of relevance factors for the legal domain. Because the method focuses on a forced choice between two options, diversity of document types in the ranking is not reflected in the research. Similarly factors related to obtainability are not

[18] It appears that for legal hierarchy, the relevance of the court is not only determined by the level of hierarchy, but also by the question at which court the case for which the research is being done will be decided, related to Schamber's geographic proximity factor [13]. Aspect like this context of legal hierarchy are not made visible by this research.

covered.

Our research is conducted with Dutch legal professionals using Dutch legal examples. Despite the jurisdictional limitation of this research, this research confirms the cross-jurisdictional framework of Van Opijnen and Santos [131]. Given the national nature of the legal domain, it is interesting to conduct further research in other countries to determine whether other legal jurisdictions may provide further insights into the factors related to domain relevance.

In our statistical analysis we investigated the difference between information specialists, legal scholars, and legal practitioners, corrected for pre-existing differences in number of years of working experience. It is possible that the sub-groups of users in our sample differ on other characteristics. For example, we know that the respondents come from a wide variety of law areas (Table 2.1). Due to our modest sample size it was not feasible to include this in the analysis as well.

2.6 Conclusions

With regards to research question 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?

Based on Van Opijnen and Santos [131] domain relevance requires two aspects: (1) a level of agreement between users, the 'legal crowd', and (2) it is not related to the task or problem at hand. Since respondents do not indicate a situational context for the example queries used in the questionnaire, the method with which the data is gathered means that factors mentioned are not related to task context, satisfying the second requirement. The first requirement is satisfied by the number of respondents mentioning the same factors. Based on the factors mentioned by the respondents, we can conclude that document type, recency, level of depth, legal hierarchy, authority, usability and whether a document is annotated are factors of domain relevance when outside of a task context.

The results confirm the existence of domain relevance as described in

the theoretical framework by Van Opijnen and Santos [131], and the factors related to domain relevance confirm the anecdotal evidence given by Legal Intelligence users.

With regards to research question 2: To what extent do legal information retrieval specialists, legal scholars and legal practitioners show agreement on relevance factors outside of a task context? Despite small differences in reported factors, we did not find evidence to conclude that legal information specialists, legal scholars, and legal practitioners differ significantly in terms of the factors they consider in judging the relevance of legal documents outside of a task context.

At this moment there is no reason to treat these sub-groups differently in legal IR systems. In the near future we will use these findings, in particular the factors of domain relevance on document class level that can be established through document representations, to extend our research into improvements for ranking algorithms in legal information retrieval systems.

2.7 Acknowledgements

The authors wish to thank the employees of Legal Intelligence, in particular dr. T.E. de Greef and mr. S. Beaufort, for their cooperation in this research and the distribution of the questionnaire to the users of the system.

Chapter 3

Citation Metrics

Citation Metrics for Legal Information Retrieval: scholars and practitioners intertwined?

Previously published as: Wiggers, G., Verberne, S., Zwenne, G-J. (2022). Citation Metrics for Legal Information Retrieval: scholars and practitioners intertwined? Legal Information Management 22 (2), pp. 88-103.
Changes have been made to sections 3.2.3, 3.3.2 and 3.5.1.

This paper examines citations in legal documents in the context of bibliometric-enhanced legal information retrieval. It is suggested that users of legal information retrieval systems wish to see both scholarly and non-scholarly information, and legal information retrieval systems are developed to be used by both scholarly and non-scholarly users. Since the use of citations in building arguments plays an important role in the legal domain, bibliometric information (such as citations) is an instrument to enhance legal information retrieval systems. This paper examines, through literature and data analysis, whether a bibliometric-enhanced ranking for legal information retrieval should consider both scholarly and non-scholarly publications, and whether this ranking could serve both user groups, or whether a distinction needs to be made.

Our literature analysis suggests that for legal documents, there is no strict separation between scholarly and non-scholarly documents. There is no clear mark by which the two groups can be separated, and in as far as a distinction can be made, literature shows that both scholars and practitioners (non-scholars) use both types.

We perform a data analysis to analyze this finding for legal information retrieval in practice, using citation and usage data from a legal search engine in the Netherlands. We first create a method to classify legal documents as either scholarly or non-scholarly based on criteria found in the literature. We then semi-automatically analyze a set of seed documents and register by what (type of) documents they are cited. This resulted in a set of 52 cited (seed) documents and 3086 citing documents. Based on the affiliation of users of the search engine, we analyzed the relation between user group and document type.

Our data analysis confirms the literature analysis and shows much cross-citations between scholarly and non-scholarly documents. In addition, we find that scholarly users often open non-scholarly documents and vice versa. Our results suggest that for use in legal information retrieval systems citations in legal documents measure part of a broad scope of impact, or relevance, on the entire legal field. This means that for bibliometric-enhanced ranking in legal information retrieval, both scholarly and non-scholarly documents should be considered. The disregard by both scholarly and non-

scholarly users of the distinction between scholarly and non-scholarly publications also suggests that the affiliation of the user is not likely a suitable factor to differentiate rankings on. The data in combination with literature suggests that a differentiation on user intent might be more suitable.

3.1 Introduction

Bibliometric-enhanced information retrieval (IR) aims to improve IR by using bibliometrics, for example citation metrics. Citation metrics are often associated with the notion of scientific impact; the impact of scholarly publications on other scholars. However, legal bibliometrics, and thereby legal bibliometric-enhanced IR, differs from other research domains in two manners: (1) its strong national ties [19] and (2) the often strong interconnection between research and practice, especially in civil law jurisdictions. In the Dutch legal domain this is demonstrated by the use of Dutch language in legal scholarly output, and by the lack of formal distinction between legal scholarly and practitioner (hereafter called non-scholarly) documents. This lack of distinction suggests that users expect both scholarly and non-scholarly documents to be included in legal IR systems. In turn, the developers of these systems aim to serve both scholars and practitioners as customers.

The ultimate aim of our research is to achieve bibliometric enhancement in legal IR systems; to improve the effectiveness of legal search by using citation metrics as a factor of (impact) relevance in ranking algorithms. But before we can implement such a bibliometric based relevance factor, we have to determine whether both user groups have a common understanding of impact relevance. It is important to know whether both user groups can be served using the same bibliometric-enhanced ranking function, or whether each group requires their own function. In order to determine this, we have to understand the meaning of citations in the legal domain.

In this paper we first discuss the literature addressing citations, with a focus on citations in Dutch legal documents. Next we perform a data analysis, for which we create a rule-based, semi-automatic classification

method to classify a set of 52 seed documents – 10 legal cases and 42 journal articles – into scholarly and non-scholarly publications based on document type, publisher reported intended audience, and author affiliations. We further analyze the 3086 documents that cite our seed documents: for each seed document, we register by what (type of) documents they are cited, using the same classification method as used for the cited documents. In our discussion, we link the data analysis to the literature analysis, and conclude with suggestions for using citation metrics in bibliometric-enhanced ranking for legal IR.

The following research questions are addressed in this paper:

1. 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?

2. Does a quantitative data analysis of citations in, and usage of, legal documents support the findings from the literature?[1]

In answering these questions, we distinguish between the implementation of bibliometric-enhanced ranking in legal IR (should the bibliometric-enhanced ranking function consider citations from both scholarly and non-scholarly documents) and the consequences of the implementation choice (given the implementation, would this bibliometric-enhanced ranking function serve both scholarly and non-scholarly users).

The contributions of this paper are twofold: first, we examine the meaning of citation metrics in legal documents using literature and quantitative data analysis. Second, we show, using literature and data analysis, a possible approach for bibliometric-enhanced ranking for legal information retrieval.

3.2 Literature analysis

For the literature analysis, we start by reviewing the general practice of using citations as a form of impact measurement. Next, we compare Dutch

[1] Research questions 3 and 4 in this thesis.

legal citation practices to this general practice. This is followed by a section on the debate on the classification of certain legal documents as 'scholarly', to highlight the highly intertwined legal publishing culture in the Netherlands. We conclude our literature analysis with a section on the use of citations in IR, our intended use case.

3.2.1 Citations as a Form of Impact Measurement

Cronin [37] tells us that the first written form of disseminating scholarly knowledge was the letter; two learned people would write to each other to discuss their thoughts and research. Some of these letters were copied by intermediaries for broader distribution. The networks of learned people would sometimes get together, and from this the learned societies grew. In time, these learned societies established journals as a more structured form of communication. From these journals systems like peer review and citations were developed, to ensure the quality of the content and acknowledge the work of others.

The use of citations as a proxy for impact was introduced by Eugene Garfield. Garfield stated that "Since authors refer to previous material to support, illustrate, or elaborate on a particular point, the act of citing is an expression of the importance of the material. The total number of such expressions is about the most objective measure there is of the material's importance to current research." [48, p. 23]. De Bellis, referring to the work of Merton, stated that: "Citing, specifically, is the same as peer-reviewing, just on a smaller scale. Hence bibliographic citations are atomic components of the cognitive and reward system of scientific communication." [41, 86]. De Bellis also stated that "Being cited by other authors is not simply a matter of intellectual lineage. When the score gets high, it is likely that the cited document is exercising an impact on citing sources" [41, p. 32]. And that: "This forward-pushing potential, in turn, is the hallmark of scientific quality" [41, p. 32]. Another description of the meaning of citation measurements comes from Kurtz and Henneken: "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, p. 696]. What these authors have in common is that they consider the total number of citations a proxy for the impact of the document *on other scholarly documents and scholars.*[2]

Beel and Gipp state that "a citation measures impact but not quality in general." [16, p. 440]. Garfield, though a proponent of using citations as a form of impact or 'significance' [47, p. 473] measurement, does note that: "citation frequency reflects a journal's value and the use made of it, but 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 widely read journal that he states readers read to keep up to date.

The question whether citations in the humanities behave like citations in the hard sciences (i.e. provide insight into the impact on other scholars) has been a topic of interest in the past decade. Bonaccorsi et al. have shown that distribution of citations of articles in the social sciences and humanities is similar to the distribution of citations in the hard sciences [19]. Hicks discusses different document types that play an important role in the social sciences and humanities that may not be covered by a citation index [58]: books, national literature and non-scholarly literature. The need to include books in citation indexes has been discussed by Giménez-Toledo et al. [51]. Zuccala and Cornacchia [144] have conducted research as to the methodology by which to include books and the challenges therein.

[2]Work by e.g. Teufel [127] narrows this down by looking at the words surrounding the citation, to see whether the author cites in a positive or negative manner, but this falls out of the scope of this paper.

3.2.2 Citations in Dutch Legal documents

The topic of legal research and the Dutch legal publishing culture has been extensively described by Stolker [120]. Stolker notes that the legal publishing culture has a strong tradition in book publishing. Even though law journal articles are becoming more important, a perspective of legal documents is not complete without considering books, confirming the statements from Hicks and Giménez-Toledo et al. [58, 51, 144]. Stolker further argues that because law is a national research topic, and documents are often bound, by topic and language, to a national audience, an analysis of such documents should be done on a national level.

Snel [117] states that there are three main reasons for citing in scholarly legal documents: to provide context for the research, to legitimize statements made in the research, and to allow others to check the quality of the research. His article is aimed at scholars, and contains advice for writing sound scholarly articles. But he mentions non-scholarly documents as possible sources of reference [117, p. 255]. To provide societal context for the legal research, he writes that authors may refer to newspaper articles. To legitimize their statements, they may refer to law articles, and legal cases [117, p. 256]. To help navigate readers to more information on the topic, they may refer to overview articles or legal handbooks [116, p. 167-168]. This demonstrates that citing non-scholarly sources is accepted practice in scholarly legal articles.

Van Opijnen [94] and Winkels and colleagues [143, 142, 140, 141] have applied citation analysis to Dutch law and case law, but did not include legal literature, such as journal articles and books. Wirt Soetenhorst presented a proof of concept of a Dutch legal literature citation index in 2017 [118], incorporating all legal articles, making no distinction between scholarly and non-scholarly legal articles. However, a literature search has not returned any information that this citation index has been completed. In his book, Stolker[120] cites several sources [130, 108] that are critical of citation metrics as a form of impact measurement for legal documents, which might explain why a legal citation index has not been created up until this point. However, he focuses exclusively on impact measurement for research

evaluation systems, not for use in IR systems. This in contrast to Garfield [47], who originally focuses on applications in library management and the creating of reading lists for scholars and students. Use for research evaluation is mentioned, but does not appear to be Garfield's original focus.

An example where this distinction – measuring for research evaluation or measuring for IR – becomes visible is document type of the cited and/or citing document. While research evaluation may take the effort and quality into consideration, regardless of the form of the document, citation indexes for IR in the hard sciences (like Garfield's original science citation index) only consider the impact on other scientific articles, as the collection the citation index is used for is limited to those scientific articles.

3.2.3 'Scholarly' Legal Documents

There is debate in the Dutch legal domain about whether a distinction can be made between scholarly and non-scholarly legal documents. Stolker describes three types of legal journals: "journals primarily focusing on the scholarly debate; journals merely focusing on dissemination (notes/annotations and short commentaries); and journals – probably the majority – doing both." [120, p. 257]. Stolker further indicates that law journals, unlike journals in the hard sciences, often do not have external peer review, but are reviewed by the editorial board. The members of this editorial board may be scholars, but may also be practitioners [67]. The Dutch legal journals are also not classified in A-, B- and C- journals, as is done in economics [119, p. 32] and other fields[3]. This means that, unlike in the hard sciences, there is no immediate mark which indicates which documents are scholarly and which are not that can be used for bibliometric research.

Research by Snel [116] further shows that legal scholars are not always explicit about their methodology and their choice of sources. This means that many publications do not contain a methodology section, and so this feature cannot be used to distinguish between scholarly and non-scholarly

[3]In Italy it appears that legal journals are distinguished between A class and other classes, see Bonaccorsi et al. citeBonaccorsi. The classification of journals into categories is mentioned to have been conducted by experts.

publications. Snel interviewed a panel of law professors, who indicated that certain approaches are so common, that they do not have to be made explicit. Examples named are using legal historical or grammatical reasoning to interpret laws, using only case law from the supreme court[4], using the snowball approach to gather literature (rather than describing which database/IR system is used and which queries), and not explaining why non-controversial interpretations from other sources are followed. Only when deviating from one of these standard approaches, the scholar will have to make their methodology explicit.

Krans [74], in his article on the scholarly status of the annotation, indicates that for research evaluation purposes most universities classify an annotation to case law as a practitioner oriented document ('vakpublicatie'). He argues that this is not necessarily so, and that it should be judged based on the content, not the form. The president of the Dutch Supreme Court, Maarten Feteris, divides annotations in four types: (1) summarizations, (2) affirming annotations, (3) annotations that reach a different conclusion based on the same facts, and (4) annotations that shed light on arguments, points of view or consequences that the court did not consider to the full extent and that could lead to a different conclusion [40]. An article by Damen [40] shows anecdotal evidence that annotations can influence courts in later decisions. Krans uses this anecdotal evidence to argue that because of the potentially high quality and impact of annotations, the content could be scholarly [74]. Systematically this could be achieved through a reversal of the burden of evidence, where the author has to demonstrate the quality of the work in order to claim scholarly status [134].

From a research evaluation point of view the impact on judges of this fourth group of annotations could be a valid reason to classify these annotations as scholarly, as the work and quality put into it will not differ much from a journal article. However, if the determining criterion is the aim of furthering of the body of knowledge – the impact on scholars and scholarly documents – the argument that they impact judges and other cases is less persuasive. Judges write case law not for the purpose of furthering the

[4]Thereby not considering case law from lower courts.

scholarly debate, but as side product of the judiciary branch of government.

From a historical point of view it is also interesting to consider how responses to journal articles ('Reacties') should be classified. In the Dutch legal field, it is not uncommon for scholars to write a short response to a journal article of a peer[5], in a form which is similar to the historical copied and distributed correspondence described by Cronin [37]. While such a response would constitute dissemination of knowledge and participation in the scholarly debate, the short nature of these responses, often focusing only on a single point from the original article, means it is not usually on the same level of skill and effort as a full journal article.

Snel [116] agrees that there is a lack of guidelines for what constitutes (good) academic legal doctrinal research. Because of this it is hard to make a clear distinction between scholarly legal documents and non-scholarly legal documents. Snel [116] suggests scholars to look at the content, the reputation of the author, the journal/publisher and the incoming citations when determining the reliability of a document. Citing Van Gestel and Vranken, who in turn base themselves on a VSNU[6] report [134], Snel further indicates three factors to take into consideration: (1) originality, (2) thoroughness and (3) profundity. Originality in this context means that the document has to add something to the current body of knowledge and/or further the academic debate. Profundity is taken to mean "the extent to which the publication should provide a comprehensive answer to the research question through reliance on relevant sources"

The difficulty in separating scholarly and non-scholarly legal publications demonstrates the intertwined nature of legal scholarship and legal practice. Suggesting that impact, as measured through citations, should consider citations from both scholarly and non-scholarly documents. It also suggests that the different contexts for citing as described by Snel's [117]: context, legitimisation of claims and reproducability/quality control, may be more indicative of different information needs and corresponding

[5]See, for example the journals Ars Aequi and Nederlands Juristenblad

[6]VSNU refers to the Association of Cooperating Universities in the Netherlands (Vereniging van Samenwerkende Nederlandse Universiteiten), currently called Universities of the Netherlands (Universiteiten van Nederland).

adjusted rankings than the division between scholarly and non-scholarly legal professionals.

3.2.4 Citations in Information Retrieval

Legal IR has a number of characteristics that distinguish it from other IR domains. One of those aspects is that the same legal IR systems are used by practitioners (lawyers and legal professionals) and scholars. Legal IR systems are therefor both professional and academic search systems. Stolker states that "For the massive number of research results available via the Internet today, researchers need some guidance on both the content and the quality." [120, p. 243]. In IR, this is referred to as the concept of relevance, which consists of multiple forms or spheres of relevance, of which topical relevance is one [109]. Another characteristic that distinguishes legal IR is a form of relevance called bibliographic relevance, where there is a legal difference between the official government sanctioned version of a document and a reprint of the same document [131, 138]. Impact, as measured through bibliometrics, can also be seen as a form of relevance.

An example of using citations as ranking criterion in academic search, including potential negative effects, is the work by Beel and Gipp [16]. They investigated the role of citations in Google Scholar and found that citations have a significant influence on the ranking, though more so for title searches than for other searches [16, p. 442,444]. It appears that since their research, Google has slightly adapted the algorithm to also include how recently the article has last been cited.[7] This is most likely done to mitigate the Matthew effect, where highly cited documents, which are likely older to have been able to generate such a high number of citations, remain at the top at the expense of newer documents. By displaying these highly cited documents at the top, they are more likely to be cited, creating a self-reinforcing effect. Beel and Gipp [16] named this Matthew effect as one of two main points of criticism for using citations in ranking algorithms in their paper.

[7] `https://scholar.google.com/intl/en-US/scholar/about.html`

Use of citations in legal IR systems can be seen in, for example, the American legal IR system Westlaw[8]. As Jackson and Al-Kofahi [61] indicate though, the more factors like citations play a role in ranking, the harder it is for a user to understand why certain results appear in certain positions. This appears to be one of the reasons why Dutch legal IR systems have focused on thesauri and synonyms to improve their systems[9], rather than more complex to explain methods such as Page-Rank.

Furthermore the scale of the Dutch jurisdiction, and thereby the size of Dutch legal IR companies and the datasets they have available to them, do not compare to the US and Westlaw. And Westlaw's techniques cannot simply be copied to other jurisdictions, because of the large difference between common law jurisdictions (like the US and the UK) who focus mainly on case law, and civil law jurisdictions (like the Netherlands and most continental European countries), who focus on legal codes, with case law as an interpretative tool [143].

The above mentioned literature shows that if the aim is to use impact as a factor for legal IR systems, bibliometrics from scholarly and non-scholarly publications should be taken together because (1) scholars cite non-scholarly sources, and non-scholarly sources (e.g. case law) cite scholarly sources, meaning an assessment of impact is incomplete without considering citations from all documents. (2) There is debate on whether distinguishing between scholarly and non-scholarly legal documents is even possible, and on what grounds it could/should be. When users themselves cannot reach agreement on which citations are and aren't a measure of impact for them, it is prohibitively difficult to make this distinction in legal IR systems. Since the collections of legal IR systems contain both scholarly and non-scholarly documents, bibliometric data from both types of documents is available, and can be taken together, to measure a broader form of impact than the scholar-on-scholar impact of traditional citation measures such as those proposed by Garfield [48].

[8] `http://lscontent.westlaw.com/images/content/L-355700_West-Search-brochure.pdf`

[9] `https://clin28.cls.ru.nl/#abstract-36`

Thus, the answer to the implementation question from the introduction is that citations from all document types should be considered, and that these citations measure not only scholar-on-scholar or practitioner-on-practitioner impact, but a broader form of impact on the legal domain as a whole. Therefore, literature does not appear to give an indication that the bibliometric-enhanced ranking for legal IR should be differentiated.

3.3 Methods

To validate the literature, we create a method to distinguish between scholarly and non-scholarly documents – based on guidance from the literature –, to analyse (1) what types of documents cite each other, and (2) what types of users use what types of documents. Our method is motivated by the discussion in Section 3.2.3, which showed that a generalized distinction is necessary to allow us to quantify the interaction between practitioners and 'scholarly' publications and vice-versa, to determine whether a bibliometric-enhanced ranking algorithm could serve both user groups, or whether separate algorithms need to be developed.

For this research, we used data from the Legal Intelligence system. Legal Intelligence is one of two large commercial legal IR systems in the Netherlands, covering all government publications and legal publishers. We collected 52 seed documents from the year 2014 – 10 legal cases and 42 journal articles – and the documents that cite them. For both the seed and citing documents, we extract from the logs what type of document it is (e.g. journal article, case law), the source, the title, the name(s) of the author(s) and what the usage of the document is. Along with assessing whether scholarly and non-scholarly documents cite each others, we analyse which types of documents have usage from users affiliated with a university and users affiliated with other types of organizations.

All document types in the Legal Intelligence system are included in our citation analysis, including blogs and newspaper articles, since we want to validate whether the literature is correct in that Dutch legal scholars cite non-scholarly documents and vice versa.

3.3.1 Document sampling criteria

Bornmann et al. [20, p. 214], citing Boyack [21], have remarked that the distribution of citation counts over documents is skewed.[10] This means that a large portion of documents receive no citations, whereas a small number of documents receive a large number of citations [87]. For that reason, a random selection of documents would not be informative for our study, because the majority of randomly selected document has no or very few citations. We selected the documents for our analysis as follows.

We chose seed documents from the year 2014 because of the time it takes for documents in the social sciences to gather citations [108]. The citing documents were from the period 2014 to August 2019, the most up-to-date data available at the time of the research.

To be able to analyze what types of documents cite, we needed to select documents from 2014 that were likely to have been cited. Based on the assumption that documents that are sought often are also read often, and documents are often read before being cited, we used the 2015 query logs from the Legal Intelligence system. We sorted the queries by frequency of occurrence. We manually went through this list and looked at all queries that are clearly related to a case (journal identifier, ECLI number or party/case name) or journal article (journal identifier or title (more than one word)) published in 2014. Case law and journal articles from other years were skipped, to avoid a citation bias based on time since publication. The documents selected are the first documents in the query list that meet these criteria. The documents selected are shown in Table B in Appendix B.

3.3.2 Document classification

For each of the seed documents, we searched our citation index [136] for documents citing it based on the unique document identifier. These citing documents are not only journal articles, but all documents in the Legal Intelligence system. This includes books, as indicated important by Stolker

[10] See also Bonaccorsi et al. [19]

[120]. This resulted in 3086 citing documents.[11] For these citing documents we also retrieved the source, the title, the name(s) of the author(s), the document type and the usage.[12]

Our first step is to attempt to categorize these documents into scholarly and non-scholarly documents. To determine the category of documents, we consider three cumulative factors:

1. The intent criterion [37]: the document is written with the intent to further the body of knowledge and/or foster academic debate;

2. Related to this is the originality criterion [134, 130]: the document is not merely repetitive or descriptive, but adds interpretation or recommendations;

3. The thoroughness and profundity criteria [134, 130]:

   (a) The document is based on more than one source;

   (b) The document has proper references.

Because of the size of the data-set, it was not possible to manually assess each document. Based on our three categorization factors we looked for proxy factors in the data and settled on document type and author affiliation, further explained below. These two proxy factors are cumulative to ensure the least possible false positives in scholarly documents. To limit the manual work required, we only assess author affiliations of documents that do not have a non-scholarly document type; meaning they have either a scholarly document type or the document type alone is inconclusive as to whether the document is scholarly or non-scholarly and also has to be assessed manually.

[11] Books are indexed per chapter. This means that if multiple chapters cite the same seed document, each chapter is treated as a separate document for the purpose of this analysis.

[12] Information about the citing documents can be found at `https://github.com/G-Wiggers/Citation-Metrics-for-Legal-Information-Retrieval-Systems`.

1. **Document type** We used the type of a document to assess the intent of the document, as well as the originality and thoroughness and profundity criteria.

2. **Author affiliation** To aid in the assessment of the intent criterion, we considered the affiliation of the author.

Documents are classified as scholarly or non-scholarly based on these two cumulative criteria.

To automate as much of the classification as possible, we developed a Python script using the proxy factors and a set of rules to determine for each of the documents whether it is classified as scholarly or non-scholarly (intended for practitioners). This process is visualized in Figure 3.1.

Classification based on document type

- If the document is a government document or case law, then it is classified as non-scholarly. Because these documents are created as a byproduct of the practice of the legislature, the executive, and the judiciary; they are not written for the advancement of scholarship and fail the intent criterion. This means, for example, that our 10 case law seed documents are all classified as non-scholarly because they are byproducts of the judiciary.

- If the document is a news article or notification of publication (short summaries with references to new books or case law), then it is classified as non-scholarly. These documents fail the intent and originality criteria.

- If the document is an annotation to case law, then it is classified as non-scholarly. Though debatable, because the theory above shows that there is a subgroup of annotations that may be considered scholarly based on quality and originality, most of these documents are not written with the intent to further scholarship but to provide interpretation of a legal decision. For this reason, they are likely to fail the intent criterion.

- If the document is a dissertation, then the intent is considered to be the advancement of scholarship and it is classified as directed towards scholars.

- If the document is a journal article or book, we add manual steps (marked in blue in Figure 3.1. Journal articles and books can have many possible intentions. Therefor, for journals and books we checked the (self-reported) publisher information to find out whether the journal or book in its entirety (on source level) was directed more towards scholars or non-scholars. Every time we encountered a new source to check, we added the outcome to a list, to allow automatic classification of other documents from the same source. We classified a document as non-scholarly if the title or description mentioned things like 'practical information' or 'for practice'. If the publisher information mentioned only scholarly use, it is classified as directed towards scholars. If the publisher information mentioned both, we continued to the next step. If the publisher information mentioned nothing, we considered the source to be non-scholarly.

- If the publisher information of a journal states that it has both scholarly and other articles, we analyzed all documents from that journal in our dataset individually. If the document is an announcement or similar document, then it fails the originality and intent criteria and is considered non-scholarly. If it is an article, we check whether it analyses several cases and/or literature and uses proper references. If it meets these thoroughness and profundity criteria we consider it a scholarly article. In case of uncertainty, the documents are categorized as non-scholarly.

For documents for which the document type is inconclusive, we manually assess the last two steps in the classification schema (marked in blue in figure 3.1; whether the document covers multiple documents, and whether there are sufficient references). This manual last step is done by two legal professionals. To assess the reliability of the manual part of the classification, we calculate the inter-rater agreement in terms of Cohen's κ [28].

Classification of authors

If a document has a scholarly document type, we analyze the affiliation(s) of the author(s) as follows:

- We check if a document had author information. Not all documents (e.g. journal articles) have author information.
- If author information is available, we retrieve the affiliation of the authors primarily from the author information in the document.
- If the author affiliation was not provided in the document, a Google search is conducted and all affiliations mentioned on the first page of the search results are considered.
- If the affiliation is to the government, the intent of the author is not considered to be the furthering of scholarship, since that is not the main goal of the government. This despite the high/scholarly level of quality of some of these documents.
- If an author has multiple affiliations and one of the affiliations is a university, we classify the document as scholarly.
- If a document has more than one author, we classify the document as scholarly if at least one of the authors is affiliated with a university.

Final document classification based on document type and authors

For both the seed documents and the citing documents, we consider a document to be scholarly when the classification based on document type is scholarly and at least one of the authors is affiliated to a university as analyzed in the author classification. These cumulative conditions were chosen to ensure the least possible false positives in scholarly documents. This is chosen since our aim is to attempt to separate between the purely scientific impact of documents, as measured by citation indexes in the hard sciences, and broader impact on the (practitioners in the) legal field.

3.3.3 Readership

To analyze the reading behavior of scholarly and non-scholarly users (reading scholarly and non-scholarly documents), we separate the document usage by scholarly users (all users affiliated with a university) and the usage by non-scholars (all users not affiliated to a university). To do this, we use the organization ID available in the Legal Intelligence data. This organization ID determines the subscription access for users affiliated to that organization. We received a list of organization IDs associated with universities. We first queried the usage by all users with an organization ID associated with a university (this data includes students as the position of the user in the organization is not included in the data), followed by users affiliated to other organization types (such as government, courts, law firms and corporations).

The usage is measured by the number of times the document is opened (clicks), where the same user can use a document on multiple occasions. The data only reflects online usage in the Legal Intelligence system.[13]

The group of users affiliated to a university is roughly 28% of total users, and is therefore smaller than the group of users not affiliated to a university. It is possible that an author affiliated with a law firm writes a scholarly article, so that a click from a user not affiliated to a university could in fact represent use in a scholarly manner. Especially if the user has multiple affiliations. However, it is not possible to determine the reader's intent from the data. For that reason, clicks from organizations other than universities are considered to be for other purposes than writing scholarly articles.

3.4 Results

The number of documents that underwent the manual last two steps of the classification (marked in blue in figure 3.1) is 311 out of the total 3138

[13] It is possible that users have alternative methods to access information, for example through paper versions of books and journals. We have no reason to assume that this would apply more to one group than to the other.

(3086+52). This means that 90% could be classified automatically and 10% needed manual classification. Of the 311 documents, 253 were assessed by both assessors. 58 documents were assessed by only 1 assessor because the second assessor experienced 'page not found' or 'insufficient access rights' errors.[14] Cohen's κ, calculated on 253 documents, is 0.58. This indicates moderate agreement in the application of the classification schema for the most difficult to classify documents. For further analyses, we used the classification of rater 1 in cases were both raters disagreed.

Table C in Appendix C shows the detailed results of the classification of the seed documents. It also quantifies the usage and citations for each seed document. The columns Scholarly citations and Non-Scholarly (N-S) citations show the classification of the citing documents according to the rules described in Section 3.3.2. For our analysis we show the usage by users affiliated to a university (shown as Usage Schol.) and the usage by users affiliated to other organization types (shown as Usage N-S).[15]

3.4.1 Citations between documents

To analyze the extent to which documents classified as scholarly and non-scholarly cite each-other, we counted the aggregated numbers of citations between scholarly and non-scholarly documents. Table 3.1 shows the summary of these counts. As expected based on the general theory of citation metrics, using a χ^2 test, we found that there is a significant relationship between the two variables ($\chi^2(1, N = 253) = 22.8, p =< .0001$): citations to scholarly seed papers are more likely to come from scholarly papers than

[14] 55 errors were access rights errors, 3 were 'page not found errors'. Of these 58 documents, 22 documents were books, 14 were articles, 13 were case law reprints in student collections, 4 were notifications/summaries, and 5 documents were other types. 48 were classified as non-scholarly, 10 were classified as scholarly.

[15] The difference in usage between documents could in part be explained by the access rights system of the IR system. Though the IR system works the same for every user, only results from publications the user has a subscription to are shown in the results list. All government documents are freely accessible to all users, as well as open access documents. It appears that certain journals have a higher subscription rate than others, and that digital availability of books is limited to a small share of the user group.

from non-scholarly papers. Note that this test has an expected frequency of cross-citations based on the data, and the table indicates that the categories are far from exclusive in their respective citations: citations from scholarly to non-scholarly documents make up 92% of the total number of citations from documents classified as scholarly (138/(12+138)).

It is also important to note here that there is a strong class imbalance on the data: out of the 52 seed documents, 13 documents were classified as scholarly articles based on the criteria in Section 3.3.2. This is why the χ^2 test is important, even though this test presupposes citations between the two groups of documents exist. The same holds for the distribution of citations over documents, which is highly skewed, as expected based on literature [20, 21, 19, 87]. 1155 of the non-scholarly to non-scholarly citations come from 1 seed document, document 14281373. 14 documents (8 documents classified as non-scholarly oriented, 6 documents classified as scholarly oriented), did not receive any citations.

	Scholarly seed	Non-scholarly seed
Scholarly citing	12	138
Non-Scholarly citing	59	2877

Table 3.1: Results: aggregated citations counts. The columns show the classification of the seed documents, the rows the classification of the citing documents.

3.4.2 Usage of documents

To analyze the usage of both classes of documents by users of the Legal Intelligence system, we counted the aggregated numbers of usage between the types of users and the types of documents. Table 3.2 shows the seed documents and the usage thereof subdivided into users affiliated to a university, and users affiliated to other organization types. Similar to the citation data, using a χ^2 test, we found that there is a significant relationship between

	Scholarly seed	Non-scholarly seed
Scholars	1062	3290
Non-Scholars	560	2577

Table 3.2: Results: usage. The columns show the classification of the seed documents, the rows the classification of the usage based on the company identifier linked to the user account.

the two variables ($\chi^2(1, N = 3086) = 46.1, p =< .0001$): a scholarly paper is more likely to be accessed by a scholar than by a non-scholar. Again, this test has an expected frequency of cross-usage based on the data, and it appears to be quite common for a non-scholar to read a scholarly paper: of all the papers accessed by non-scholars, 18% are scholarly (560/(560+2577). And it is also common for scholars to read non-scholarly documents: 76% of the documents accessed by scholars are non-scholarly (3290/(1062+3290)).

3.5 Discussion

The above describes a method to distinguish between scholarly and non-scholarly documents and the results thereof. In this section we will briefly discuss the documents that the two manual classifiers did not agree on, followed by (1) an analysis of what types of documents cite each other, and (2) an analysis of the usage data of these documents, and compare these results with the literature. We conclude this section by discussing the implications of these results on the creation of a bibliometric-enhanced legal IR ranking algorithm.

3.5.1 Inter-rater agreement

With an inter-rater agreement of $\kappa = 0.58$, we find that there is moderate agreement between the two raters. Although we judge this as satisfactory considering that these 311 documents were the most difficult documents to classify, we analyzed the differences in classification between rater 1 and

rater 2 in more detail. We noticed three things. First, the debate about the classification of annotations to case law (see Section 3.2.3) is reflected in the results. Rater 1 strictly adhered to the classification scheme and classified all instances of annotated case law that occurred in the manual classification (e.g. because the publisher information did not identify the document as annotated case law) as non-scholarly. Rater 2 however looked at the content of the annotations, and classified 11 of them as scholarly, with a note stating that the quality of these annotations was such that they could have been published as articles. Second, rater 1 classified 7 documents that were a response to a previously published article as non-scholarly, because of the short length of these documents. Rater 2 classified these as scholarly, with as motivation that they contribute to the scholarly debate. Third, rater 1 classified 3 documents that were reports of conferences of legal experts as non-scholarly, whereas rater 2 classified these as scholarly, again referring to their contribution to the scholarly debate.

The examples mentioned above show that rater 1 focused only on the classification of the work, whereas rater 2 focused on the quality of the work. Rater 2 appeared to have used a reversal of the burden of proof similar to that mentioned by the VSNU [134].

3.5.2 Citations between documents

The analysis of the classified documents – 52 cited (seed) documents and 3086 citing documents – shows a significant relationship ($\chi^2(1, N = 253) = 22.8, p =< .0001$) of scholars citing scholars, in a setting where cross-citation is expected. This level of cross-citation the χ^2 test expects from the data shows that scholarly articles also cite non-scholarly oriented documents, as well as the other way around. When we look at Table 3.1, the largest group by far is non-scholarly documents citing other non-scholarly documents. This is partly caused by document 14281373, a legal case and therefore non-scholarly document, which has 1155 citing non-scholarly documents (See Table C in Appendix C).

However, as mentioned in Section 3.4.1, the dataset is unbalanced, meaning that the group of seed documents classified as scholarly is much

smaller than the group classified as non-scholarly. Furthermore, our chosen classification method has strict criteria before a document can be classified as scholarly to avoid false positives, which may result in false negatives, creating further imbalance.

We see that case law documents[16] are widely read and cited by both non-scholars and scholars. NJB, which is a journal aimed at both legal practitioners (non-scholars) and scholars, also receives citations from both groups.[17] It is interesting to see that the journal Arbeidsrecht, which is marketed as a journal for practitioners (non-scholars) receives no citations from either group in this dataset.[18]

Document 13627420 attracts a lot of response articles. The article was published in the journal for private law, notaries and registration[19], which according to the website of the publisher contains both scholarly articles and non-scholarly oriented articles.[20] The author information in the article indicates that the author, mr. R.J. Abendroth [3], is affiliated to a law firm, with no mention of an affiliation to a university. The article is about the order of securities on a good. It received 60 citations, of which 3 are a direct chain of responses. After the original article, prof.mr. F.E.J. Beekhoven van den Boezem (scholar) writes a direct response ('Reactie') in document 15442271. Abendroth (practitioner) responds to this in document 15442265. In 16492944, mr. K.J. Krzeminski (practitioner) responds to both authors. Though this is just one example, it demonstrates an interaction between non-scholars and scholars. It also shows that the informal letter or 'Reactie', which from a research evaluation point of view may not be equal to a journal article in terms of time investment and academic rigour (as pointed out by rater 1 in section 3.5.1), from a dissemination of knowledge point of view may have just as much impact in the legal debate

[16] The documents in Table C below the line.

[17] See document ids 14151738, 12987162, 13330606, 12926733, 14177758, 13235698 and 13580788 in Table C

[18] See document id's 13002758, 14124128, 12987652, 14124136, 22171998, 13241348, 12882340 and 12660424 in Table C

[19] 'Weekblad voor Privaatrecht, Notariaat en Registratie'

[20] `https://www.sdu.nl/shop/weekblad-voor-privaatrecht-notariaat-en-registratie-abonnement.html`

(as pointed out by rater 2 in section 3.5.1).

The work of Snel [117], as discussed in section 3.2.2 shows us multiple reasons why scholarly articles may cite non-scholarly documents, and vice versa. This theoretical research, as shown in Figure 3.2, may explain the cross-citations found in the data. When looking at these reasons, we notice that Snel is not just referring to the use of non-scholarly documents in scholarly documents, or vice versa, but also to the use of one document type in another type of document. Snel [117] suggests that a highly cited case could signify that a novel problem was solved (e.g. the first case that dealt with the question whether a digital item is a good in the sense of property law), or that the court veered from a previous ruling. The high number of citations in articles could mean that the case has sparked a legal debate, and has thereby contributed to the furthering of knowledge (intent criterion). This is an example of a non-scholarly work influencing a scholarly work.

A citation from a journal article in a reference work could signify that the article has a lasting impact, for example because it has a novel contribution to legal scholarship (intent criterion) and is of high quality (thoroughness and profundity criteria). Though the reasons for citing as shown by Snel [117] differ, they are all indications of relevance for the legal domain as a whole.

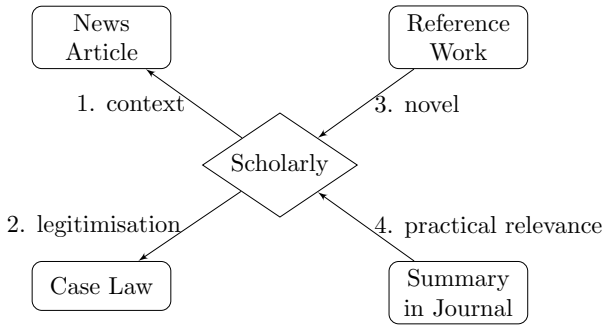

Figure 3.2: Citations in legal documents. 1. Scholarly articles may cite news to give context [117]. 2. Scholarly articles may cite case law to legitimize their claim [116]. If the case is cited often, this may indicate that the court decided a novel problem, or veered from a previous ruling. 3. If a reference work cites a scholarly article, this may indicate that the article had a novel contribution and was of high quality. 4. If the scholarly article is cited in summary in a journal, this may indicate that the article is also relevant for practitioners.

If we were to consider only the impact of scholarly documents on scholarly documents (upper left in Table 3.1), as in citation metrics in the hard sciences, we would miss part of the impact that the scholarly documents have (bottom left in Table 3.1), as well as the impact of non-scholarly documents on the scholarly documents (upper right in Table 3.1). Therefore, for bibliometric-enhanced legal information retrieval, a citation index which does not also look at non-scholarly oriented documents, both what they are cited by and what they cite has an incomplete picture of the legal field. This system of cross-citations also suggests that citations reflect not scholarly impact like in the hard sciences, but part of a broader scope of

impact, or relevance, for the entire legal field.

Given the sizeable number of documents without citation information, as discussed in Section 3.3.1 and visible in Table C, we also looked at usage data from the two user groups, to see whether that shows similar patterns, and whether it could potentially be useful to fill in the gaps in the data for use in legal IR.

3.5.3 Usage of documents

We see in Table 3.2 that even though the group of users affiliated to a university is smaller, their usage is higher than that of the group not affiliated to a university. Our results indicate that there is a relation between usage of scholarly documents and legal scholars, but that legal scholars also read documents classified as non-scholarly, and legal practitioners read documents classified as scholarly.

This finding is supported by the literature (e.g. Snel [117]). An example supporting this is document 12702866. This document is annotated case law and is therefore classified as non-scholarly. But the data shows high usage from users affiliated with a university and a relatively high number of citations by scholars when compared to the other documents in this research. Upon analyzing the document, it appears to be a seven page analysis by Prof.mr. T. Kooijmans [72] on the legal concept of recklessness. This document appears to be one of the annotations that Krans [74] argues should be classified as scholarly due to the high quality, an argument also mentioned by rater 2 (see section 3.5.1). For research evaluation this might be a strong argument to categorize these annotations as scholarly, but that has to be weighed against the intent criterion (see section 3.3.2).

It should be noted that it was not possible in this data set to distinguish between students and employees of the universities, meaning students were classified as scholars. This raises the question whether law students behave more like scholars (contributing to the upper left of Table 3.2) or like practitioners (contributing to the upper right of Table 3.2) in their legal information seeking and needs[21]. It is possible that the high number of usage

[21] The investigation of the search behaviour of students versus legal scholars falls outside

of non-scholarly publications by users affiliated to a university (upper right of Table 3.2 is partly caused by this lack of distinction. This does not, however, explain the high number of usage of scholarly documents by users not affiliated to a university (bottom left of Table 3.2). We therefor consider that while further research is required to determine whether law students have information needs and information seeking behaviour like scholars or like legal practitioners to refine these results, this does not negate the observation that there is usage of non-scholars of scholarly information and usage of non-scholarly information by users affiliated to universities.

It is interesting to note that for case law, in particular documents 14281373 and 12827114, the number of citations for the document is higher than the number of times the document has been opened. We propose two possible explanations: (1) the document has been reprinted in case law journals. Users have read the case in one of these reprint forms, but have decided to cite this version. Both documents are the official government reported versions[22] and referenced by the European Case Law Identifier (ECLI). (2) Users access these cases outside of the IR system, for example by going directly to the government website publishing the cases or from a print subscription. Whatever the cause may be, this discrepancy between the usage and citations suggests that the usage data does not provide a complete picture of the readership of a document.

3.5.4 Using bibliometrics in legal IR

The data shows that legal scholars and legal professionals use the same legal IR systems and (at least to some extent) the same documents. This suggests that creating a separation between scholarly and non-scholarly documents in legal IR systems may not be useful for the users. Legal professionals open the most useful or relevant documents for their information need, regardless of the document type and/or the intended audience of the publisher. The bi-directionality of the disregard of scholarly and non-scholarly users to the distinction between scholarly and non-scholarly publications also suggests

the scope of the data set and thereby of this paper

[22] As published on the government website www.rechtspraak.nl

that the affiliation of the user is not likely a suitable factor to differentiate rankings on. The data, in combination with the work of Snel [117] suggests that a differentiation on user intent might be more suitable.

The consequence of using citations from all document types in a citation index is that we move from a pure scholarly citation index and the theory behind that, so that the meaning of a citation might also differ. The citation data in this research shows that when we look solely at scholarly impact (scholarly to scholarly citations), we miss part of the picture of the total impact the document has. Similarly, we miss the impact non-scholarly documents have on the scholarly debate. When we combine data from all documents for bibliometric-enhanced legal IR, we are looking at impact on the legal field as a whole rather than solely scholarly impact. To measure this broader impact, citations alone may not provide enough information, since not all documents are cited, and for those that are cited, we only capture impact on authors (scholarly and non-scholarly). We are therefore looking at a part of the impact on the entire legal field.

To enrich the view on this impact on the legal field as a whole a combination with usage metrics appears to be an obvious combination, though it has to be kept in mind that the usage data of legal IR systems may not offer a complete view of usage of legal information, as shown in section 3.5.3. It will however, further fill in the picture of the impact of a document on the legal field as a whole. When implementing usage into a bibliometric-enhanced ranking, usage from both users affiliated to a university and users affiliated to other organization types should be considered, to reflect this impact on the legal field as a whole.

3.6 Conclusion

In this paper we addressed two research questions to try and determine how bibliometric-enhanced ranking can be introduced in legal information retrieval:

(1) Does the literature suggest the use of one bibliometric-enhanced ranking function in legal IR, or should there be sep-

arate bibliometric-enhanced ranking functions for legal scholars and legal practitioners?

The literature discussed in Section 3.2 shows that if the aim is to use impact as a factor for legal IR systems, bibliometrics from both scholarly and non-scholarly publications should be taken together because (1) legal scholarly articles use non-scholarly documents to support their claim, and in turn are mentioned in non-scholarly documents, meaning an assessment of impact is incomplete without considering citations from all documents; (2) There is debate on whether distinguishing between scholarly and non-scholarly legal documents is even possible, and on what grounds it could/should be. When users themselves cannot reach agreement on which citations are and aren't a measure of impact for them, it is prohibitively difficult to make this distinction in legal IR systems. Since both scholars and legal professionals access the same sources and use the same legal IR systems, bibliometric data from both users groups is available, and can be taken together.

Thus, when using citations from both scholarly and non-scholarly publications, we measure a broader form of impact than the scholar-on-scholar impact of traditional citation measures such as those proposed by Garfield [48], and will measure part of a broader form of impact on the legal domain as a whole.

(2) Does a quantitative data analysis of citations in legal documents support the findings from the literature? To validate the findings of our literature analysis we created a classification schema for scholarly and non-scholarly documents based on three cumulative criteria: intent, originality, and thoroughness and profundity. Most of the documents were classified based on rules, the 311 remaining documents were classified with manual steps.

The results of the citation analysis on 52 seed documents and 3086 citing documents confirm that scholarly articles cite non-scholarly documents and vice versa. The usage data shows that users affiliated to a university use both scholarly and non-scholarly documents, as well as users affiliated to other organization types. This is in line with our findings from the literature, and suggests that citations in legal documents do not measure

impact on scholarly documents and scholars in the same way as in the hard sciences, but quantify part of a broader scope of impact, or relevance, for the entire legal field.

This disregard by both scholarly and non-scholarly users of the distinction between scholarly and non-scholarly publications, and especially the fact that this occurs in both directions, also suggests that the affiliation of the user is not likely a suitable factor to differentiate rankings on. The data in combination with literature suggests that a differentiation on user intent might be more suitable. Further research focusing on differentiating queries into the user intents defined by Snel [117] will show whether a differentiated bibliometric-enhanced boost on these grounds will be possible.

Because of the modest sample size used in this research, the results of this paper cannot be extrapolated to all documents in the Dutch legal domain. Because of the national characteristics of legal domains, this example from the Netherlands can also not be extrapolated to other countries. The results do, however, provide the valuable insight that the theory and methods for impact measurement from the hard sciences cannot simply be copied to use as metric for impact in legal IR systems.

When creating a citation index that included both scholarly and non-scholarly documents, and using this for biblometric-enhanced rankings for both scholars and practitioners alike, we encountered some missing data, since a substantial number of documents are never cited, and since citations only capture impact on authors. For documents that are never cited, the illusion could exist that they have had no impact on the field even though, like the Scientific American example, they may have had a different form of impact. For documents that have been cited, we have data on the impact they have had on other authors, but not on non-author users, meaning that we may be missing part of the picture of the total impact the document has had. We therefore suggest to combine citation metrics with usage metrics. Future research will focus on the correlation between citations and usage, and possibilities to combine these two metrics into an overarching view of the impact of legal documents on the legal community as a whole for use in bibliometric-enhanced legal IR systems.

3.7 Acknowledgements

The authors thank the employees of Legal Intelligence, in particular dr. T.E. de Greef and P. van Boxtel for their cooperation in this research.

The authors acknowledge prof. W.H. van Boom who, in the spirit of scholarly debate, send us his inaugural lecture on annotations after reading an earlier version of this paper.

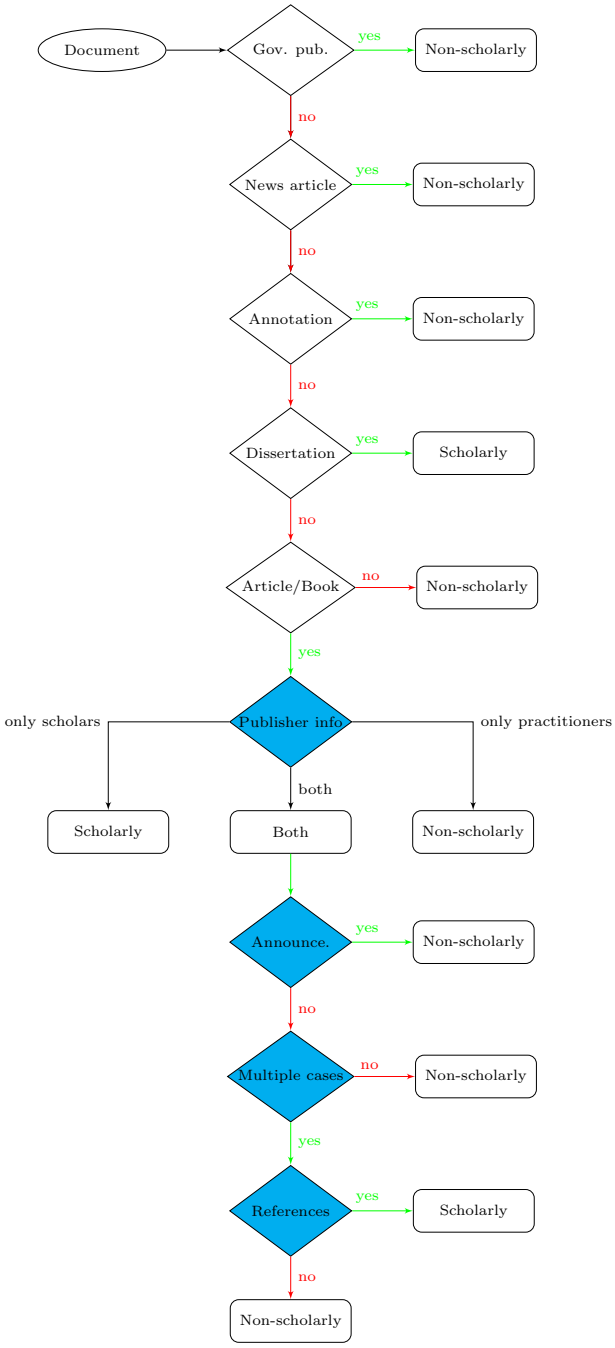

Figure 3.1: Flowchart of classification schema

Chapter 4

Evaluation

High Recall, Small Data: The Challenges of Within-System Evaluation in a Live Legal Search System

Under review as: Wiggers, G., Verberne, S., de Vries, A., van der Burg, R. (2022). High Recall, Small Data: The Challenges of Within-System Evaluation in a Live Legal Search System.

This paper addresses the limitations of common ranking evaluation methods for legal information retrieval (IR). We show these limitations with log data from a live legal search system and two user studies.

We provide an overview of aspects of legal IR, and the implications of these aspects for the expected limitations of common evaluation methods: test collections based on explicit and implicit feedback, user surveys, and A/B testing. Next, we empirically demonstrate the limitations of common evaluation methods using data from a live, commercial, legal search engine.

We specifically focus on methods for monitoring the effectiveness of (continuous) changes to document ranking by a single IR system over time.

We show how the combination of characteristics in legal IR systems and limited user data provides unique challenges that cause each common evaluation method to be sub-optimal.

In our future work we will therefore focus on less common evaluation methods, such as cost-based evaluation models.

4.1 Introduction

In the legal domain, the amount of information available digitally is increasing rapidly. Legal scholars and professionals have to navigate this information to find the case law and articles relevant for them. They often do this under the time pressure of having to account for every minute spend on a case. A study by LexisNexis showed that attorneys spend approximately 15 hours in a week seeking case law [77]. Legal information retrieval (IR) systems exist to help legal professionals navigate this information overload to find relevant information in the most efficient way. In order to do this, legal IR systems are continuously improving their retrieval and ranking algorithms. Evaluation of these systems is important from a commercial and academic point of view; however, in practice this is not always conducted in a consistent manner.

That evaluation of legal IR is not always conducted in a consistent manner was shown by Conrad and Zeleznikow in their work on the use of evaluation methods in articles on legal IR in the ICAIL proceedings

[30] and the journal Artificial Intelligence and Law [31]. They find that "there may remain some cause for concern insofar as a scientific research community that champions Artificial Intelligence for the benefit of the legal domain may still have as many as a fifth of its empirical conference works presenting no performance evaluation at all." [31, p. 185]. Aside from this one fifth missing evaluation at all, their results show that 46% of the papers use gold data created by domain experts as evaluation method and a further 22% use manual assessment by grad students or research assistants. Conrad and Zeleznikow argue that if the research community in AI and law wishes to remain relevant to legal practitioners, they have to develop methods to show the value of their work [31]. This would mean including evaluation in every paper, and perhaps moving towards evaluation involving end users.

In this paper we show that evaluating legal IR systems is not only lacking for certain research settings, but that the challenges causing this missing evaluation also occurs for live legal IR systems. We describe evaluation challenges and limitations based on the literature about legal IR and demonstrate why the common evaluation approaches do not work for live professional search systems. We do so using data from a live legal IR system and two user studies. We focus on within-system evaluation of changes in ranking algorithms. This applies to situations where a change to the algorithm is made that affects the ranking of the documents but not the number of documents retrieved to allow scholars and developers to assess the effect of the change in the ranking algorithm. We address the following research questions:

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

2. What are the limitations of common evaluation methods and metrics for evaluating ranking changes in live professional IR systems?[1]

The data for our work is provided by Legal Intelligence, one of the largest legal content aggregators and legal IR systems in the Netherlands.

[1] Research questions 5 and 6 in this thesis.

The contribution of this paper is to demonstrate why common ranking evaluation methods cannot be applied to live professional search systems. We do this by (1) providing insight in the characteristics of legal IR in practice that make the task different from common ranking evaluation tasks; (2) describing the limitations to common evaluation methods to be expected based on these characteristics; and (3) showing, using data from a live legal search engine, the limitations of common ranking evaluation methods.

To define which evaluation methods are common, we based ourselves on the classic textbook from Manning et al. [82]. We assess the following evaluation methods for our problem: (a) a test collection based on information needs and relevance judgments by domain experts, (b) a test collection based on implicit feedback from clickthrough/log analysis, (c) user satisfaction studies (in particular surveys), and (d) A/B testing.

In Section 4.2 we conduct a literature analysis to answer research question 1. In Section 4.3 we discuss expected limitations of common evaluation methods and metrics for live professional IR systems. In Section 4.4 we demonstrate, using data from our legal search engine, the found limitations of these methods. Based on the information from the literature analysis (research question 1) and the data we will conclude in Section 4.5 by answering research question 2.

4.2 Legal IR

To understand why common ranking evaluation methods are not suitable for legal IR systems, we need to have a clear picture of the characteristics of these systems and their users. This section starts with the description of the characteristics of legal IR, its users and its documents, contrasting its properties with these of Web search where possible. It also relates legal IR to professional IR in general, to further specify the characteristics of legal IR.

4.2.1 The User

The classical image of a legal professional is a lawyer who (1) works under high time pressure and (2) cannot afford to miss information that might be relevant in court. The time pressure for lawyers (and other legal professionals) often stems from the billing system, where every hour or even minute dedicated to a case has to be accounted for. This is often tracked using specific software.[2]

At the same time, legal professionals cannot afford to miss any important information. Their professional reputation would be damaged if the opposing party has information they have missed. Konstan et al. [71] analyzed the cost-benefit values for different user groups, and show that for legal users, missing an item that turns out to be valuable has a very high negative impact. In contrast, 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. This is in line with the conclusion by Bock [18] that the main focus in legal IR should lie on high recall.[3] Manning et al. [82, p. 156] even go as far as to say that paralegals will tolerate fairly low precision results to obtain this high recall.

Geist observes in [50] 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'[4]: "Simply put, it is in a legal dispute first of all important to know more than the opposing lawyer(s) and not to fulfill abstract ideals of completeness".

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 depends on the user (e.g. a novice versus a senior lawyer, or a general practice lawyer versus a highly spe-

[2] e.g. [90].
[3] See also Mart [83].
[4] 'Vollständigkeit(sideal) und Recherche-Realität' [50, p. 158], translation by author.

cialised lawyer) and the case at hand. Geist [50] argues that only a good relevance ranking can provide users with both high recall (completeness) and high precision (most relevant results first).

A secondary effect of the time pressure of legal professionals is that the gathering of explicit feedback (asking users or judges to evaluate search results) is prohibitively expensive for developers of legal IR systems. This leads scholars and developers to use feedback from graduate students.

A practise shown very often by legal professionals [81] and much less in Web search [68, 126] is *updating*. Updating behaviour refers to gaining understanding about the current importance or status of a particular document [81]. It could be regarded as a type of known-item retrieval: the user is aware of the existence of a document, or a state of a document, and needs to know if their knowledge is still current and up-to-date. An example of this is monitoring for changes of legal documents like amendments to laws to verify if something is still the accepted interpretation of the law. This updating behaviour is mostly done in a direct way by querying for the particular document or indirectly by means of an automatic citator service [81].

Van der Burg [42] found that of all queries investigated, 25% is inferred, or assumed known-item search and 75% are other searches. This frequency of known-item searches lies close to the 20% navigational queries found by Broder [23]. Van der Burg describes that the queries in the assumed known-item set are on average shorter than those in the remainder set, and that the clicks related to the assumed known-item set are more often on the highest ranked documents [42].

Another characteristic of legal professionals is that they wish to have control in their search [121]. Mart [83] describes the ranking algorithms of two leading American legal IR systems, Westlaw and Lexis. She explains that companies treat their ranking algorithms as trade secrets, and are therefore reluctant to discuss them in detail, but based on the information she gathered from various sources, it appears that Westlaw considers "...commercial user document interaction history" [83, p. 400][97] in their ranking, something that is common in Web search. Lexis on the other hand states: "This is not a popularity algorithm! Our algorithms provide

you with more control over your research...” [78]. This need for control makes the user requirements for legal IR systems different from those of Web search engines like Google.

4.2.2 The IR Systems

What most legal IR systems have in common, with the exception of a small number of commercial IR systems, is that they limit themselves to one jurisdiction. This limited scope distinguishes legal IR from Web search. When looking in more detail, legal IR systems can be divided into two broad groups, based on their owners: (1) governments and (2) publishers [50].

Governments, in their role as legislative and judiciary branch [89], create laws and case law. These are often published on government websites with an IR system build into it.[5] These systems are often limited to one information type, either law or case law, and in federal government structures often further delimited to federal law/case law or state law.

Publishers create commercial legal IR systems to make their publications more accessible to legal professionals on subscription basis.[6] These

[5]e.g. `https://www.govinfo.gov/app/collection/STATUTE` for US Statutes at Large and `https://www.loc.gov/collections/united-states-reports/` for selected US Reports, `https://www.gesetze-im-internet.de/index.html` for German laws and `https://www.bundesverfassungsgericht.de/DE/Homepage/homepage_node.html` for case law from the German Bundesverfassungsgericht, `https://www.ris.bka.gv.at/` for Austrian law and case law, and `https://wetten.overheid.nl/zoeken` and `https://www.rechtspraak.nl/` for Dutch law and case law.

[6]Westlaw, an American legal IR system active in many countries is owned by ThomsonReuters, see `www.westlaw.com`. LexisNexis, another US based system operating in many countries is owned by the RELX group, formerly known as Reed Elsevier, see `www.lexisnexis.com`. In Austria [50], there is RDB owned by publisher Manz (`www.rdb.at`), LexisNexis Austria (`www.lexisnexis.at`), and Linde Digital owned by Linde Publishers (`https://www.lindedigital.at/`). Exception to the rule appears to be RIDA created and maintained by prof. Jahnel, see `http://www.rida.at/Wer-entwickelt-RIDA.321.0.html`. In the Netherlands there is Legal Intelligence owned by publisher Wolters Kluwer (`https://www.wolterskluwer.nl/shop/serie/legal-intelligence/Legal-Intelligence/`), and Rechtsorde owned by publisher Sdu (`https://www.sdu.nl/juridisch/producten-diensten/rechtsorde`), who in

commercial legal IR systems usually deal with multiple documents types. Systems like Westlaw, LexisNexis, and the Legal Intelligence system that we work with in this research, include not only laws and case law, but also legal journals, books, government reports and newspaper items.

4.2.3 The Documents

When looking at legal IR systems with diverse document types, the large deviation in length of the documents in the index is often the most notable feature. Lengths may vary between a government report (161 pages)[7] and a newspaper item (57 words)[8].[9] There is also a difference in genre, varying from the structured form of legal codes and case law, to the free form of blog posts and newspaper items.

The scope of the collection of a legal IR system is smaller than in Web search, and pre-determined by the owner of the IR system. As mentioned above the collection is often limited to one legal jurisdiction. Documents included in the collection of a legal IR system are all from sources that are considered to be relevant to legal professionals. This restricted scope reduces noise, especially when dealing with homonyms. The word 'trust' for example in a legal context has a specific meaning [49]. To distinguish between the meaning of terms in ordinary speech and 'legalese', law dictionaries are created, the most famous being Black's Law Dictionary [49]. By reducing the scope of the collection of the legal IR system to documents relevant to legal professionals, a search for 'trust' by a legal professional will result in documents regarding this topic, rather than results about the company Trust and the character quality one might find in Web search[10].

turn is part of publishing company Lefebvre Sarrut (`https://www.lefebvre-sarrut.eu/en/by-your-side/`). In Germany, there is Juris, owned in part by the German state and in part by Sdu (`https://www.juris.de/jportal/nav/juris_2015/unternehmen_2/ueber_juris/ueber_juris.jsp`) and thus by Lefebvre Sarrut, and Beck Online owned by C.H. Beck publishers (`https://beck-online.beck.de`).

[7] DocumentID 34474736.

[8] DocumentID 34582268.

[9] Note that books are often indexed by chapter or paragraph.

[10] Incognito Google search conducted on October 30th 2020.

A further narrowing of the scope of the collection comes from journals/sources with a subscription model. Where the government or a university is likely to purchase a blanket subscription to journals from all law areas, a niche law firm will likely subscribe to a limited amount of journals relevant to their work to limit expenses.[11] Because of the difference in amount of documents accessible for each user, the same query will generate a different set of results for the lawyer than for the scholar.

When looking at the structure of the documents, 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. Though legal professionals have multiple methods to cite a document (e.g. party names, case number, journal reprint reference number), the various references can be mapped using regular expressions to provide an overview of the relations between documents. It appears though, that this information is not always used to the fullest extent possible [50]. This in contrast to websearch, where PageRank has become the standard [82].

4.2.4 Relevance

IR, including legal IR, has as aim to aid users to find relevant information. For legal IR, this notion of relevance can be described by the following relevance factors, as identified in prior work [138]: title relevance, document type, recency [121], level of depth, legal hierarchy, law area (topic), authority (credibility), bibliographical relevance, source authority, usability, whether the document is annotated, and the length of the document. These relevance factors are similar to those in other fields, as demonstrated by the work of Barry and Schamber [12, 13]. Van Opijnen and Santos [131] established that legal professionals tend to agree strongly on factors like authority, legal hierarchy and whether the document is annotated. While these factors are usually grouped under 'cognitive' or 'situational relevance' and thereby considered to be specific to the user or task, because of the

[11] Though this depends on the price models used by the publishers, who sometimes price packages of content in such a way that a package deal with more content is cheaper than subscribing to only the journals needed.

general agreement between users in the legal domain on these factors, Van Opijnen en Santos [131] group these as 'domain relevance'.

The importance of recency has motivated the use of so-called 'recency boosts' in rankings in legal IR. This has two functions. It is used to be able to show the most up to date information, but it is also a way to ensure that appeal decisions, which are from a higher court but by definition also more recent, are shown above the decision in first instance.[12] Legal IR systems are aware of this, and boost newer documents to the top of the results list. Because of this, and because of the large amount of documents published, the top of a results list for a given query may be completely different from month to month.

4.2.5 Small Data

Because of the time pressure users are under, and the associated labor costs, as mentioned in section 4.2.1, it is often not possible for developers of legal IR systems to obtain large quantities of explicit feedback or relevance judgments. The use of implicit feedback collected in the course of normal search activities [69] is also limited, because legal IR systems are often bound to a particular jurisdiction. This means that the number of users in a system is limited to the legal professionals within that country. In the

[12]The legal importance of recency in legal IR systems is hinted at in the case of *GC, AF, BH and ED against Commission nationale de l'informatique et des libertés (CNIL)*, where the Court of Justice of the European Union made clear that search engines (Google in the case at hand) need to ensure that the search results reflect the current status of a case: "Having regard to the above considerations, the answer to Question 4 is that the provisions of Directive 95/46 must be interpreted as meaning that ... second, the operator of a search engine is required to accede to a request for de-referencing relating to links to web pages displaying such information, where the information relates to an earlier stage of the legal proceedings in question and, having regard to the progress of the proceedings, no longer corresponds to the current situation, in so far as it is established in the verification of the reasons of substantial public interest referred to in Article 8(4) of Directive 95/46 that, in the light of all the circumstances of the case, the data subject's fundamental rights guaranteed by Articles 7 and 8 of the Charter override the rights of potentially interested internet users protected by Article 11 of the Charter.", Court of Justice of the European Union case ECLI:EU:C:2019:773.

case of the Netherlands, the largest legal IR system has between 75 000 and 100 000 users. The amount of usage data available is therefor much lower than in IR systems for generic Web search.

This smaller dataset due to the size of the audience is narrowed even further when we consider that legal IR systems are not used daily. When we add to this the high attention to recency, and the changing results lists this creates as mentioned in Section 4.2.4, as well as the differences in subscriptions, few users have seen the same results lists or query-results pairs. This means the data available for implicit feedback analysis is also limited.

4.2.6 Legal Search and Professional Search

Legal IR is a form of professional search, and shares many characteristics with it, as well as with other types of domain specific search. Understanding these similarities and differences might provide insight into suitable evaluation methods. The First International Workshop on Professional Search[13] describes professional search as: "professional search takes place in the work context, by specialists, and using specialist sources, often with controlled vocabularies." [132]. It covers people from multiple domains, including librarians, scientists, lawyers, and other knowledge worker professions. They describe six characteristics: (1) a restricted scope and domain. Users do not wish to retrieve information from all possible sources, but only from within their domain (e.g. legal, medical). (2) Not all sources are equally accessible; subscriptions are required to access some sources. This means that two professionals with different subscriptions will retrieve different result sets. (3) the use of multiple systems; (4) a tolerance for low precision; professionals create lengthy queries and often take time to refine them. (5) the need for users to be in control: "explaining the predominance of Boolean search in, e.g., prior art search and systematic review." [132] (6) the use of controlled vocabularies.

When applying these six characteristics to legal IR, we notice that (1)

[13]Held at SIGIR 2018.

the restricted scope and (2) subscription access are indeed characteristics of legal IR, as shown in Section 4.2.3. Characteristic (3), the use of multiple systems, may vary from jurisdiction to jurisdiction. In countries like the United States and the Netherlands systems like Westlaw, LexisNexis and Legal Intelligence provide content integration as well as IR functionalities. Geist [50] however describes that in Austria licensing issues have caused situations where legal IR systems include summaries of publications from other publishers in their index, but users must use the print version or change IR systems to be able to access the full-text of these documents.

As described in Section 4.2.1, the (4) tolerance to low precision is described by Manning et al. [82, p. 156] to include legal IR, but debated by Geist [50]. This is often related to (5) the need for control. Two well-known high recall tasks, often conducted using boolean queries for reproducability, are systemic review tasks (academic[14]/medical search) and prior art search (patent search). However, several professional search domains, such as medical search and legal search, include instances of these high recall tasks, aside more applied search behaviours. The legal domain for example has a citation culture where legal scholarly articles may cite publications from legal practice [137]. The last characteristic, (6) the use of controlled vocabulary, is demonstrated by the existence of law dictionaries and has been discussed in Section 4.2.3.

4.2.7 Summary

Legal IR has several characteristics that challenge common evaluation methods: (1) The cost of missing results is high, but the tolerance to low precision results drops under time pressure. This means that early-precision metrics are not sufficient; lower-ranked documents also have to be considered in evaluation. (2) Explicit relevance judgements are expensive to gather. (3) Because the field of legal research is highly specific, the user group and number of user interactions is limited. (4) Different users see different results in their results list, based on the journals/sources they are

[14]For the purpose of this paper we will consider the search for scholarly information – academic search – part of professional search.

subscribed to, and thus have access to. This limits the use of implicit feedback models further. (5) Recency is considered very important, and plays a large role in the ranking algorithm. Because of this, and the high frequency with which new documents are published (and boosted in the ranking algorithms of legal IR systems), the top of the results list is highly dynamic, meaning that static evaluation methods are difficult to use for live systems.

4.3 Expected limitations to common evaluation methods

All IR systems share the same aim: user satisfaction [82]. This comprises multiple components, including speed, user interface[15], and satisfaction with the results returned. The satisfaction with the results returned depends on the number of relevant results returned, and the order in which the results are returned. This research focuses on evaluation methods comparing two different versions of a ranking algorithm, in particular the following four common methods: (a) a test collection based on information needs and relevance judgments by domain experts, (b) a test collection based on implicit feedback, (c) user surveys, and (d) A/B testing. In the following subsections we discuss each of these in relation to our problem: the evaluation of a live legal search engine.

4.3.1 Test Collections

A common method of evaluation is test collections [63], such as the TREC collections [54]. An example for the legal domain is the test collection created by Locke and Zuccon [80]. An initiative for benchmarking in legal IR is the Competition on Legal Information Extraction/Entailment (COLIEE), active since 2014 [102].[16] COLIEE's specific focus is on case law, using Canadian test collections.

[15] For the importance of snippets in Legal IR, see Wiggers et al. [138].

[16] `https://sites.ualberta.ca/~rabelo/COLIEE2021/`

Conducting evaluations on these public test collections is less informative for legal search systems that cover non-English language civil law jurisdictions, as the content in the actual collection will be in the language of the jurisdiction, and the focus of the user may be more on legal statutes and less on case law. The evaluation of such a system on an English language test collection with a limited task (e.g. retrieving only case law or e-discovery) will provide little information on the performance of the system when used in daily legal practice in the home jurisdiction. In addition, case retrieval tasks such as the one in COLIEE are document-to-document tasks, where the query is a case law document, as opposed to a keyword query. Most commercial professional search engines, including ours, use keyword queries.

Hawking [57] suggests that a test collection for professional search (in his situation enterprise search) should be created specifically for the company in order to be a suitable evaluation method. The set will have to be tailored to the company because of the highly specialized content used in the system.

Conrad and Zeleznikow [31] mention that relevance assessments are often created by some sort of domain expert, for example grad students or research assistants. However, as Cole and Kuhlthau [29] have shown, there is a difference between what an early career legal professional classifies as relevant, and what a senior legal professional classifies as relevant, in line with the notion of cognitive relevance of Saracevic [109]. This is also the reason why relevance assessments are usually gathered from multiple assessors. In the case of legal professionals, that would require relevance judgments of not only junior but also (more expensive) senior legal professionals, as well as participation from scholars and the judiciary.

As stated by Voorhees [133], for many evaluation metrics used in test collections, all documents in the results list need to be judged. When this needs to be done by multiple assessors, and requires the inclusion of high level experts as described above, this becomes prohibitively complex and expensive.

An alternative to using test collections with expert judgments is the use of implicit feedback. In Section 4.4 we will assess the value of test collections based on explicit or implicit feedback for the evaluation of a live

professional search engine.

4.3.2 User surveys

Asking a user directly whether they are satisfied provides valuable information. However, the research of Blair and Maron [17] suggests that there is likely to be a mismatch between the recall the users think they have achieved and the recall calculated based on random samples of documents in the collection. In their research with legal professionals the average calculated recall was 20 percent, whereas the legal professionals questioned believed they were at 75 percent recall or higher.

Furthermore, as suggested by Turpin and Hersh [129], a ranking that scores higher on system oriented metrics does not always score higher using user oriented evaluation metrics. Literature suggests this to be especially true when the difference between the rankings is small and not at the extreme ends of performance (e.g. both are not extremely poor systems or extremely good systems) [115]. Users can adapt their search strategies to achieve similar levels of results for different levels of quality systems [6], for example by refining their queries [129]. This might be a limitation for use as an evaluation method for professional search systems, as a commercial system is unlikely to be an extremely poor system, and a change to the ranking algorithm is unlikely to create drastic changes such as a complete reversal of the ranking.

For commercial websites and webservices, measuring user satisfaction is often done through Reichheld's Net Promotor Score [105], a very short survey that measures user satisfaction. The appeal of the Net Promotor Score (NPS) as compared to other types of surveys is that the shortness makes for a higher response rate.

It should be noted that Reichheld shows that the NPS score has a lower correlation with sales where the purchase decision is not made by the individual user, but by company management, such as computer systems [105, p. 6]. It is therefore important to carefully consider the framing of the question in a manner that corresponds with the information desired.

In Section 4.4 we fill assess the value of two types of user surveys – a

ranking preferences survey and an NPS survey – for our problem.

4.3.3 A/B testing

For large scale systems like Google, the evaluation is often done with live user-oriented evaluation methods in the form of an A/B test [122]. A/B testing is a between-group design that usually consists of (1) randomly splitting the users into two representative groups, a test group and a control group, and (2) presenting the test group a feature (whether in the interface or in the ranking algorithm) while keeping the control group on the current version of the system [122]. The two groups are then compared on variables such as user engagement.

The legal domain has both users that search for themselves and users (e.g. paralegals) that search for others. In conversations with management of the Legal Intelligence system we found that customers expect the system to return the same results for all users. This so that the work of the paralegal or intern can be replicated and checked. Therefore, in the legal domain, it is commercially not acceptable to differentiate between users from the same organization. When trying to split the user group on organizational level, we found that due to the many firms who specialize in one area of the law, it is difficult to create two groups that are both representative. There is also commercial pressure to provide the latest (and thereby believed to be best) version of the system to all customers. For these commercial reasons it is not possible to divide the entire customer base of a live system into two groups, whether on user or on organisation level. This appears to be a blocking factor for using A/B testing in practice.

This means that we have three evaluation methods left (test collections based on expert judgments, test collections based on implicit feedback, and user surveys), which we will apply and empirically assess for our problem based on data from the search engine and user studies.

4.4 Empirical assessment of evaluation methods

In this section we show, supported by descriptive statistics of data from the search engine and two user studies, the implications of applying common evaluation methods to a live professional search engine: (a) a test collection based on expert relevance judgments (Section 4.4.1), (b) a test collection based on implicit feedback (Section 4.4.2), and (c) two surveys: a survey measuring users' preferences for rankings (Section 4.4.3), and a survey based on the Net Promotor Score (Section 4.4.4). For each method we discuss the suitability and limitations of the method for legal IR in practice, with a focus on monitoring the effectiveness of changes to a single legal IR system over time.

4.4.1 Test collection based on expert relevance judgments

In the case of Legal Intelligence, an early precision (or shallow pool) golden standard, or golden data set, internally known as the 'golden answer set', is available. This data set contains queries and their 'golden answers'; documents that are expected to be the top ranked results. This set of queries and their corresponding golden answers has been created by editors of legal journals, who are domain experts in their law area. The set contains 194 queries with for each query between 1 and 17 golden answers. The collection has been built by sampling from queries conducted by domain experts in the past, eliciting the results they would have liked to have seen in top positions. This set is subdivided into case law (51 queries), literature (51 queries), legal codes (46 queries) and legal commentary (46 queries).

Because this data set focuses on early precision through golden answers (results expected on top positions), it does not contain relevance judgments for all results returned. This requires less relevance judgments, and is therefore cheaper to make. This is, however, also the most important limitation of this method. Because the set is only limited to only a small number of relevance judgments, this tool cannot be used to assess the ranking algorithm for high recall scenarios. The use of this set is limited to 'research reality' scenarios as described by Geist [50] where the focus is on early

precision.

Further limitations include the age of the set. The set was created in 2018, meaning that newer, perhaps more relevant results, have not been included. Regularly updating this data set is time intensive, and therefore expensive. In practice, the problems with the age of the judgements are circumvented by using a document collection with publication dates up until 2018, and pretending it is early 2019 to ensure that date boosts are functioning correctly. While this method allows developers an easy way to compare two versions of a ranking, this clearly does not reflect the reality that the top of the results list is highly dynamic. This limitation exists for all test collections, but is more prominent when using the method for the evaluation of continuing updates to a single system.

An early precision golden data set does not provide information that can be used to infer pairwise preferences: document A is expected above B, but when B is also marked as relevant, that cannot be taken to mean that either A or B in isolation does not provide sufficient information for the information need behind this query, as that was not considered when creating this test collection. A further limitation is the subscription model used for legal publications. The document marked most relevant for the query may be outside the subscription of the user. If no alternative document has been marked as 'second best', the golden standard set does not reflect the user experience of users who are not subscribed to the publication this document appeared in.

Because of these limitations, the golden standard set is only suitable for developers to conduct sanity checks when developing a new ranking algorithm, taking into account that the results only reflect early precision use cases, not high recall use cases. An updated test collection with relevance assessments done by multiple users including senior legal professionals is too expensive. Test collections are therefore not a viable method to evaluate changes made to the ranking algorithm of legal IR systems.

4.4.2 Test collection based on implicit feedback

Implicit feedback appears promising because, unlike the test collection mentioned above, it does not require a time investment from the users or domain experts and is usually readily available in legal IR systems. As it is collected during the normal work process of the user, the data is always up-to-date.

Implicit relevance judgments can be used to infer relevance from (user) interactions. In the Netherlands, legal scholar Van Opijnen [94] studied implicit feedback as signal for the relevance of case law. This work focused mainly on (re)publication as signal rather than user interaction.

Addressing the interactions of users with the search engine, Oard and Kim [92] have created a framework that describes the different types of user behaviour that could be monitored for implicit feedback. Methods that have been proposed to assign relevance scores to documents include Click Through Rate (CTR) and pairwise inference (see e.g. Joachims et al. [64], further expanded on by Chuklin et al. [27] and Agrawal et al. [5]).

The implicit feedback data that we use contains the clicks registered in the logs of the Legal Intelligence system, with a pseudonomized user ID, the document ID, the position of the document in the ranking, the text of the query and a datestamp.

The search engine result page of Legal Intelligence contains links to 20 documents. When a user scrolls to the bottom of the results page, a further 20 results will be loaded, if available. Each document is described by a publisher curated abstract that consists of the title of the document and varying amounts of meta-data. When a user clicks on a result, they will be directed towards the full article on the platform of the publisher of the article. Because the user is outside the Legal Intelligence system while reading the article, and is able to click through to other articles while on the publisher platform, reading time is not logged in the Legal Intelligence logs; we only use clicks as the signal of (implicit) relevance.

The amount of data per user To explore the data available to a commercial legal IR system, and the limitations it causes, we looked at the patterns of user interactions per user. To measure the activity of users of

legal IR systems, and how much data they generate per person in their day to day activities, we selected the nine users who conducted the most recent queries reported in the logs. For these nine users, we tracked the number of queries in the Legal Intelligence system from the first of January 2020 to the 20th of October 2020. Figure 4.1 shows the usage patterns of these nine users. Though the average number of queries varies between users, all users show periods of more intense research and periods of less intense research. This means that of the total user group, only a part is active on a given day.

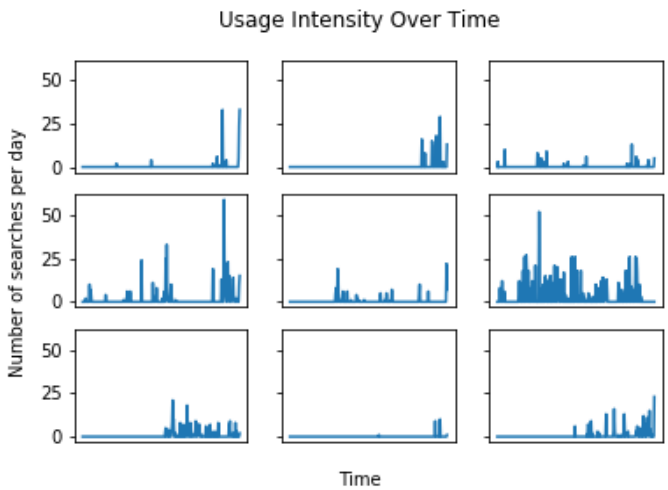

Figure 4.1: The number of searches per day for nine users over 10 months.

Queries are usually unique We looked at the number of queries that have been issued by multiple users within one month. We zoom in on a period of one month (October 2020), because of the highly dynamic top of the results list, as discussed in Section 4.2.4.

To create implicit feedback models, whether through click-through rates or pairwise inferences, we need queries that are conducted by multiple users.

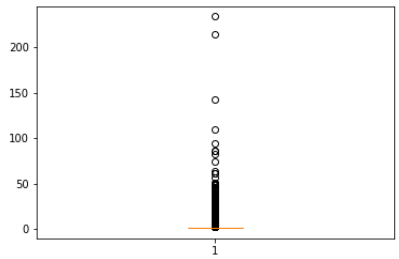

Figure 4.2: Distribution of queries and number of users conducting them

	No. of users issuing the same query
mean	1.16
std	1.25
min	1.00
25%	1.00
50%	1.00
75%	1.00
max	234

Table 4.1: Distribution of queries and number of users conducting them

In general we need enough data to rate the entire results list, or the @k results specified in the evaluation metric. But in legal IR we also need enough data to compensate for the fact that the users may have seen different results list due to differences in subscription or new documents being added to the collection. Different result lists mean users have seen different pairs of results and generate different pairwise inferences. As shown in Figure 4.2 and Table 4.1 the majority of queries is unique to one user. This is not unexpected as professional search deals with experts. It is not unreasonable to assume that the more expertise a user has on a topic, the more unique the queries become [29, 132].

Queries issued by multiple users When we look at the top 10 queries ordered by number of users that conducted that query, we see in Table 4.2 that these queries are often navigational queries where a user wishes to find and open a particular source, for example a book or journal. We consider this separately from known item retrieval, where the user wants to access a particular document from that source, for example an article or chapter. In Table 4.2 this difference is illustrated by queries on source names ‘lexicon’, ‘tekst en commentaar’, ‘asser’ and ‘wpnr’, and queries for

sepcific documents ‘awb’ and ‘ECLI:NL:HR:2013:BZ2653’. These navigational queries would provide a very one-sided image of legal IR if used in implicit feedback models.

Query	Number of Users
poging tot doodslag (‘attempted homicide’)	234
* (a wildcard query to retrieve all documents)[17]	215
lexicon (source name)	142
tekst en commentaar (source name)	109
onrechtmatige daad (‘tort’)	94
awb (law name)	86
corona (colloquial reference to the SARS-COV-2 virus)	86
asser (source name)	83
ECLI:NL:HR:2013:BZ2653 (case law identifier)	74
wpnr (source name)	64

Table 4.2: Distribution of Queries and Number of Users Conducting Them

This means that using implicit feedback to infer relevance for test collections, even in the case of partially judged results lists, is not viable for a professional search system like Legal Intelligence.

4.4.3 Survey for ranking preferences

To asses surveys as an evaluation method, we created one. The survey was created using compilations of screenshots from the search engine. It shows the query, followed by two images of result lists, as shown in Figure E.1 in Appendix E. Respondents are asked to indicate which ranking they prefer. Respondents also have the option to indicate no preference.

The two rankings used are a baseline ranking (the then current ranking in the legal IR system) and a degraded model, inspired by Smith and Kantor

[17] Users may use this if they wish to navigate using filters rather than a query.

[115]. In our test set up the degraded model was created by removing boost functions from the baseline model of which we know that they are wished by users. Thus, we know that the degraded model differs in a manner relevant to the user. We chose a relative relevance assessment method ("which of the two rankings do you prefer"), since it has been demonstrated that humans can make such relative decisions more reliably [44], and it helps negate the bias of work experience [109].

Survey design As per TREC [54] convention, we aimed at 50 reviewed query/rankings pairs (QRPs), with a minimum of 25 reviewed pairs [133] and a minimum of 3 respondents per pair. The QRPs were divided per law area. Users were asked to indicate the law area they practice in, and were shown QRP's accordingly. This was done to ensure experts in a particular legal domain reviewed only QRPs for which they were able to assess the information need behind the query, and the relevance of the results for the query. We also include general practice queries for which respondents were able to asses the general information needs.

We selected queries that multiple users have issued, from multiple companies, to avoid privacy sensitive queries. This also reduces the risk of noise by accidental clicks. As shown in Section 4.4.2 queries issues by multiple users tend to either be less specific or navigational. If those are used in an evaluation method they will give an incomplete image of the quality of the system, but in the context of testing whether users can agree on a preferred ranking the general nature of the queries may be helpful as it will allow users to understand the information need behind the query. We selected queries from 7 law areas: corporate law, IT law, environmental law, labour law, tax law, criminal law and generic legal practice. Each law area included at least one query for a law article, one query for a law name, and one or more queries for a legal concept. Each set of queries included one query (except the general group, which had two) that was also included in one of the other sets, leading to a total of 55 different queries. With these 55 queries we created 9 QRPs per law area for 7 law areas, for a total of 63 pairs.

Respondents were given 9 QRPs to review, each with two rankings of 10 results, but were able to end the survey earlier. We decided to allow this to ensure the largest number of participants possible.

We inspected the rankings to confirm that they are different. On average 2.4 documents in the top-10 remained in the same position, whilst 7.6 documents changed position. Of these 7.6 documents 1.4 documents moved up, 2.9 moved down, and 3.2 were replaced. However, as Table D.1 in Appendix D shows, in some cases the results list of the degraded model had no documents in common with the results list of the baseline model. To show that the changes in the order of results were relevant, we created a highly simplified implicit feedback model. As shown in Section 4.4.2 we only had generic queries that were done by multiple persons, and for those we had on average a total of 3.7 clicks (from all users combined) in the top 20 to base our nDCG calculation on. We considered a clicked document to be relevant, and an un-clicked document to be neutral. Using this click data we calculated the nDCG@20 under the old and new ranking. This was 2.08 for the old ranking, and 1.96 for the new ranking. While we expected the nDCG to reflect that the degraded model, because we removed boosts added to the system at the request of the users, was less preferable, the score suggests otherwise. However, for the purpose of this survey the question is not which is better, but whether users see a difference, and indicate the same preferences.

The order of the baseline model and the degraded model was alternated. Our hypothesis is that if the survey is an appropriate evaluation tool, users will notice difference between the the two rankings and indicate a preference for one of the rankings.

Users Prefer Baseline	Users Prefer Degraded	Users Tie
29	23	11

Table 4.3: Number of QRPs (total 63) by majority preference (excluding no preference). Users considered tied when number of users indicating choice 1 and 2 is equal (regardless of number of no preferences).

Users Prefer Baseline	Users Prefer Degraded	Users Tie
12	9	42

Table 4.4: Number of QRPs (total 63) by majority preference (including no preference). Users considered tied when number of users indicating no preferences is higher than choice 1 and/or 2.

Results The survey was completed by 77 respondents. Each of the 7 law areas had at least 3 respondents. For our analysis, we selected the majority answer for each of the 63 QRPs. In Table 4.3 we excluded the answers from respondents who indicated that they had no preference; in Table 4.4 we considered the pair also tied when the number of respondents indicating no preference was higher than the number of users indicating option 1 or 2.

To test the significance of these results we conducted a three-way ANOVA. The three factors (independent variables) of the analysis are the ranking, the query, and the law area; the dependent variable is the percentage of respondents choosing the ranking. When we look at the relation between the ranking and the percentage of respondents choosing that ranking, we found an insignificant relation ($p = 0.21$). We also looked at the relation between the query and the choice of the respondents, and the relation between the law area the respondents belong to and their choices. Both of these relations are insignificant ($p = 0.51$ $p = 0.67$ respectively).

Analysis We expected to find a preference from the users for one ranking over the other, as the nDCG scores indicated that the relevant documents had moved, and the change we made to create the degraded model was a boost function introduced at the request of the users and as such is expected to be noticeable by the users. As shown in Table 4.3 and Table 4.4, this was not always the case. This means that a survey of this kind does not elicit enough information to base an evaluation on. We conclude that a ranking preference survey is not a usable evaluation method for our problem.

4.4.4 Survey based on the Net Promotor Score

As a second type of survey, we experimented with the Net Promoter Score (NPS) as described in Section 4.3.2, because of the low user effort required. The NPS data is constantly being collected for commercial purposes, meaning the data is readily available. The NPS measures overall user satisfaction, and does not focus specifically on the ranking. Nevertheless, one would expect that an improvement in the ranking of the search results would also improve the overall user satisfaction.

For our experiment we chose a real live change in the ranking algorithm of the Legal Intelligence system that went live on September 14th 2020. In our situation the NPS score is gathered per month, so we compared August 2020 with October 2020. The NPS question is not always presented to users. To avoid irritating users the question is posed at the most once every six months. Furthermore a user has to be logged in to see the NPS question. As shown in Section 4.4.2, users do not use the system daily, so the user population that is shown the NPS question on a given day is small. Of the users that are shown the NPS question, not all respond. In both months, ten users responded to the NPS survey.

The scores were exactly the same for both months.[18] Like with the other survey, this may be explained by the difference being small, and because of the adaptability of research strategies by users. The combination of the broadness of the measure and the low number of respondents mean that the NPS is not a good approach to assess differences in ranking within a legal IR system, especially for jurisdictions of a modest size.

4.5 Conclusion

Legal professionals are confronted with information overload, and are in need of effective legal IR systems. Though evaluation of these systems is considered important from an academic point of view, in practice this is not always conducted in a consistent manner. In this paper we showed,

[18] Because of commercial interests the exact NPS score cannot be reported in this paper.

using data from a live professional search system, the limitations of common evaluation methods.

The focus of this research is on situations where a change is made to the algorithm that affects the ranking of the documents but not the number of documents retrieved or other changes to the IR system, including the user interface. Its application is therefore limited to within-system comparisons, not between–systems comparisons. The applicability of our work is limited to commercial, medium-sized professional IR systems.

The common evaluation methods were defined as: (a) a test collection based on information needs and relevance judgments by domain experts, (b) a test collection based on implicit feedback from clickthrough/log analysis, (c) user satisfaction studies (in particular surveys), and (d) A/B testing.

As argued in Section 4.3.3, A/B testing is not an option because in the legal domain commercial reasons prohibit different users seeing different results. As shown in Section 4.4.1 test collections based on relevance judgments from domain experts are too expensive to gather and keep up to date. Implicit feedback data is also not suitable for creating test collections, as the available data is too sparse, in particular with regards to queries issued by multiple users, as shown in Section 4.4.2

As shown in Section 4.4.3, surveys are not a suitable evaluation method to evaluate differences in ranking algorithms in legal IR. The survey on ranking preferences in our legal search engine showed inconclusive results. The NPS survey analysis shows that the number of users exposed to the NPS questions and the broad nature of the question make it not suitable.

Given the found limitations, we find that all of the common evaluation methods are sub-optimal for use in evaluating changes to ranking algorithms in live professional information retrieval systems. In our future work we will focus on less common evaluation methods, such as a cost-based evaluation model as described by Järvelin et al. [63].

4.6 Acknowledgements

The authors would like to thank the respondents for their participation in this research. The authors would also like to thank Legal Intelligence, in particular T.E. de Greef and P. van Boxtel, for their cooperation.

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.

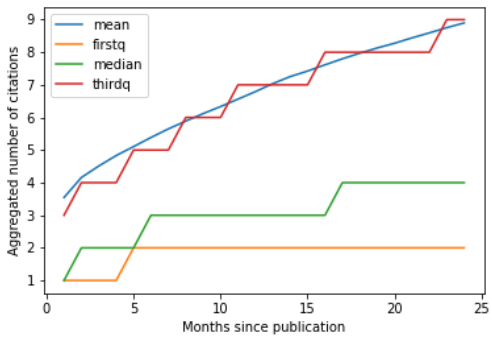

Figure 5.1: Aggregated citations per month after publication

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.

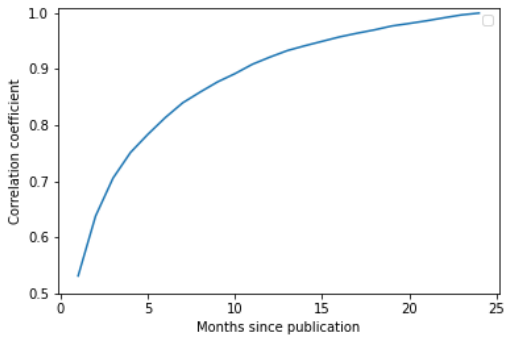

Figure 5.2: Correlation per month of citations up to and including that month with citations after 24 months

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.

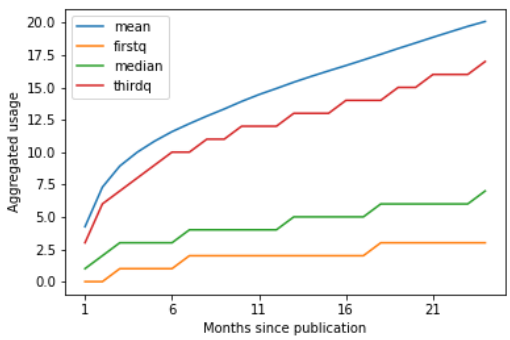

Figure 5.3: Aggregated usage per month after publication

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 $\rho = 0.64$.

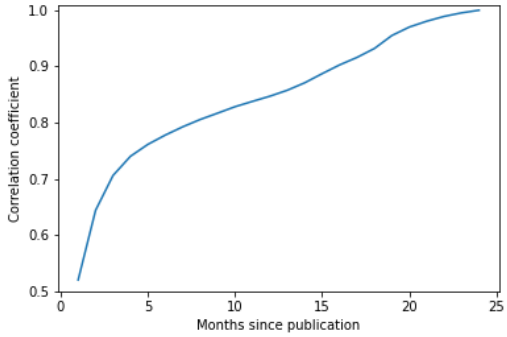

Figure 5.4: Correlation per month of usage up to and including that month with usage after 24 months

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.

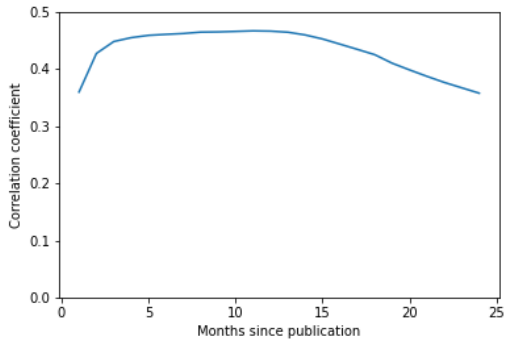

Figure 5.5: Correlation per month of usage up to and including that month with citations after 24 months

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).

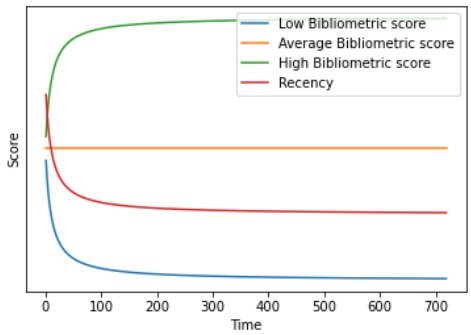

Figure 5.6: A visualisation of the ranking variable for a low, average and high citation/usage score, and the corresponding recency variable

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.

Chapter 6

Discussion and conclusions

"So once you know what the question actually is, you'll know what the answer means." Douglas Adams – The Hitchhiker's Guide to the Galaxy

This research started from the top down, with the idea that it must be possible to improve the ranking of legal IR systems by adding meta-information about the documents, and the availability of a substantial amount of data. Certain that something like this had already been done, a Google search led me[1] to the discovery of altmetrics [100] and the theory behind Google's PageRank algorithm [95]. With this confirmation that a theoretical underpinning for this idea existed, and with sufficient data available, this could have led to immediate implementation in the Legal Intelligence system. Users would have been happy, or not, and the development team would move on to the next project. But the question 'what does it mean', moved it from a mere idea, ready to be implemented in two sprints, to a full PhD project.

It soon became obvious that this research would become interdisciplinary. The scholarly field of IR focuses a lot on state-of-the-art web-search, whilst domain specific (e.g. legal or archaeological) IR applications often still rely heavily on BM25[70], developed in the 1980's and 1990's [65]. That the latest academic developments focus only on web-search is unfortunate, since domain specific IR applications are often very valuable to end-users.

Because of this academic focus on web-search it can be hard to find the right context for domain specific IR research (and to find venues to publish that work). This missing context means an interdisciplinary approach is required which looks at all the steps in the process from foundational theory to application [22]. It requires a combination of domain specific (user) knowledge as well as information science and information retrieval.

Every step towards implementation of the bibliometric-enhanced ranking model led to more questions. A big challenge of interdisciplinary research is that of vocabulary. The vocabulary in bibliometrics and information science is not the same as that of information retrieval, which makes it hard to find relevant literature. In that regard the work of Van Opijnen

[1] Because of the more personal nature of this discussion, and because there are no co-authors for this chapter, this discussion is written in the first-person singular as opposed to the earlier chapters, which were written in the first-person plural.

and Santos [131] became a Rosetta Stone of sorts. Their paper applies the work of Saracevic [110] to legal IR, and thereby not only introduced the concept of bibliographic relevance as a feature of legal IR, but for me also bridged a language gap between information science and computer science. The BIR community[2] helped further bridge this gap and gave a name to the research, whilst the JURIX community[3] helped bridge the gap between the legal domain and computer science.

The question whether legal scholars and legal professionals have the same perception of relevance, and thereby whether one legal IR system can serve both user groups, was the first question answered in this research, in what has become Chapter 2. Using a survey and conducting a PERMANOVA on the answers we found no significant difference in the factors reported by these two user groups. This meant that there is no reason to treat these sub-groups differently in legal IR systems (sub-question 2). The agreement of the respondents on factors of relevance, in a survey setting without situational relevance, also shows the existence of domain relevance as described by Van Opijnen and Santos as 'relevance of information objects within the legal domain ...' [131] (sub-question 1).

The next challenge was the question 'what does a citation mean in the legal domain'? The work of Stolker [120] provided valuable information about the publication culture of Dutch legal professionals. The work of Merton [87] lead to the question what a citation in legal documents represents, a question answered by Snel [116, 117] (sub-question 3). This also introduced the second main challenge of interdisciplinary research: sidetracks. Citation metrics for scholarly evaluation have been discussed in the Dutch legal domain [130, 108], but their use for Dutch legal IR less [94]. The negative light in which they had been discussed for research evaluation may prove to be the reason why this is the case. It was tempting to become part of the debate on the merits of bibliometrics for research evaluation, but the scope of the research had to be limited.

In Chapter 3 we conducted a data analysis, which confirmed the work

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

of Snel and Stolker and showed many cross-citations between scholarly and non-scholarly documents. We also found cross-usage. The literature suggested, and the data showed, a disregard by both scholarly and non-scholarly users of the distinction between scholarly and practitioner-oriented publications. This confirmed to us that the affiliation of the user (legal scholar or legal practitioner) is not a suitable factor to differentiate rankings on (sub-question 4). It also provided us with the theoretical insight that citations in legal documents measure part of a broad scope of impact, or relevance, on the entire legal field. We say *part* of a broad scope of impact, because for documents that are never cited, the illusion could exist that they have had no impact on the field even though they may have had a different form of impact. We therefore suggest to combine citation metrics with usage metrics.

The third question, on the interdisciplinary sphere of academia and industry, was 'what does this mean in practice', or how to implement this. The first example I found of how to implement usage and citation counts into a live IR system was the work of Kurtz and Henneken [75]. This work led to the work of the CWTS [135] on different ways to implement citation counts, from raw counts to normalized counts, and why some methods are preferred over others.

The most prominent example of 'what does it mean in practice' has been and still is the question of evaluation. The aim of BIR is to improve IR systems. But what is 'better', and how to measure it? Järvelin [63] stated that to understand what an effective method of evaluation for a (legal) IR system is, we need to understand the theoretical background (sub-question 5). Azzopardi and colleagues [9, 11] have developed a framework to create user models to aid in this. But we ran into many practical problems trying to evaluate a live domain specific IR system, as demonstrated in Chapter 4 (sub-question 6).

In the end the works of Järvelin's [63] and Azzopardi and colleagues [9] inspired us to create a cost based model for evaluation, as discussed and implemented in Chapter 5. Chapter 5 also describes the other practical questions asked in the implementation process and how we found the answers. We discovered, through the work of Geist [50], that the completeness

ideal assumed by the IR community to exist in legal IR, is offset by the research reality of legal practice. We also found that citations appear, and are a reliable predictor of future citations, much earlier than thought [108] (sub-question 7). We also confirmed the correlation between usage and citations found by, amongst others, Brody et al. [24] (sub-question 8), and were thereby able to confirm our theory from Chapter 3 that usage data can be used along citation data to represent different flavors of impact. Applying a linear mixed model (LMM) to data from this user model before and after the introduction of our Bibliometric-enhanced ranking algorithm, we found a reduction of cost for the user of 2 to 3% for situations other than known-item retrieval (sub-question 9).

6.1 The answer (to the research question)

The research question of this thesis is **How can bibliometrics improve common ranking algorithms in legal information retrieval?** Combining the answers of the sub-questions above, we can conclude that a bibliometric-enhanced ranking feature needs to take into account both usage and citations (two flavors of impact relevance), and needs to increase in influence as the reliability of the data grows (in combination with a recency feature that gives new documents the benefit of the doubt and decreases at the same rate as the bibliometric feature increases). With such a bibliometric-enhanced ranking feature we can reduce the cost required from legal professionals (whether practitioner, scholar or legal information professional) to find enough information for their information need.

The contribution of this thesis lies not only in the answer as a whole, but in the steps taken to reach this conclusion:

1. that there is no reason at this point to differentiate the ranking for sub-groups of users of legal IR systems based on their role or affiliation;

2. that bibliometrics can be seen as a manifestation of impact relevance and 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;

3. how common evaluation methods are limited by the characteristics of legal IR when used for the evaluation of live domain specific search engines, and how a cost-based evaluation can be used instead;

4. a clear step-by-step description how such a bibliometric-enhanced ranking variable can be created, and that ranking algorithms in legal IR can indeed be improved using bibliometrics.

6.2 What does this answer mean for the future?

The research in this thesis has raised even more questions for me, and possibly for others. For example: to what extent the perception of relevance differs from snippets as opposed to full documents, what the optimal level of detail is to use in normalization of bibliometrics, and how it is possible that legal documents get cited so quickly in published documents.

Future work should focus on the development of domain specific, live evaluation models, so that non-academic developers can adequately evaluate their system. Aside from the benefits for their own system, this will allow them to find venues to publish their work more easily. In a similar way that companies like Google, Yahoo and Microsoft contribute a lot to the scientific community around web-search by producing research output and datasets, we need companies to further the academic debate about domain specific IR. But in order to participate in the academic debate, these companies need to use evaluation methods that the scientific community agrees on.

It is in the interest of the users, and the legal profession as a whole, that the legal IR systems implement these evaluation methods, and cooperate with researchers to improve the systems. By providing insight into the completeness ideal and research reality, users themselves can also contribute to the improvement of these systems.

Azzopardi and colleagues [9] have developed the C/W/L framework (pronounced 'cool'), which (based on user data) can be used to predict user interactions with the search results based on their position in the results list, and by extension can also be used to evaluate IR systems. A similar model, tailored to small-scale live IR systems, would be very suitable, since it can be altered to suit the characteristics of the users of the domain. With such a system changes in domain specific IR systems can be reported in a uniform manner, and easily be interpreted by the wider IR community. This will force open the door for the IR community to pay more attention to domain specific IR.

A unified and interpretable evaluation method like C/W/L will hopefully also remove roadblocks for publication of such work. Currently IR journals consider the sample sizes of domain specific IR small, and a barrier to publication. A recognised evaluation method may remedy this. Similarly, work in the BIR community is often published in a special issue of the journal of Scientometrics[4], but does not have an obvious outlet in the IR community. This makes it harder for researchers in BIR to reach the IR community. These challenges and limitations of interdisciplinary work are often not recognised by journals when making their publication decisions while they should be.

Within universities, interdisciplinary research should be further normalized. Not just through the creation of interdisciplinary research groups, which appears to be a growing development, but also through the facilitating of introductory skill-based courses for (senior) researchers and assistance with navigating publishing interdisciplinary work. Often, the skills associated with a field or discipline are taught throughout content-heavy courses. But for researchers from other disciplines, this means following courses of which the majority of information is irrelevant, or independent (online) study where they do not benefit from the knowledge and skills of their colleagues. By creating modular, skill-based courses for researchers (e.g. programming or descriptive statistics), they are able to select those (LEGO) building blocks that they need, in an environment which best suits

[4] e.g. `https://sites.google.com/view/scientometrics-si2019-bir`

their needs.[5]

Creating a learning environment where researchers can learn new research methods with peers might also encourage researchers to use a wider range of research methods (e.g. a law scholar might start using quantitative research methods next to their normative research works). Using interdisciplinary skills in their domain (a transfer learning of sorts) may also lead to the construction of novel ideas. With the added benefit that increased contact between researchers from different faculties may lead to more interdisciplinary collaboration.

[5]This research was part of the interdisciplinary data science research program. I started this research with limited knowledge of statistics, so my own PhD trajectory is a good example that it can be difficult to find such courses. The options available were several bachelor courses with the skills weaved into domain specific knowledge, or online learning.

Bibliography

[1] *Kamerstukken II* 1996/97, **20644, nr. 30**

[2] Oxford English Dictionary. Oxford University Press, Oxford, United Kingdom (2019), `https://www.oed.com/view/Entry/161891?redirectedFrom=relevance#eid`

[3] Abendroth, R.: Rangwisseling pandrecht door eigenlijke achterstelling. Weekblad voor Privaatrecht, Notariaat en Registratie **7029**, 756–762 (2014)

[4] Abrol, M., Latarche, N., Mahadevan, U., Mao, J., Mukherjee, R., Raghavan, P., ..., Zhang, G.: Navigating large-scale semi-structured data in business portals. VLDB **1**, 663–666 (2001)

[5] Agrawal, R., Halverson, A., Kenthapadi, K., Mishra, N., Tsaparas, P.: Generating labels from clicks. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining. pp. 172–181 (2009)

[6] Allan, J., Carterette, B., Lewis, J.: When will information retrieval be 'good enough'? In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 433–440 (2005)

[7] Anderson, M.J.: A new method for non-parametric multivariate analysis of variance. Austral ecology **26**(1), 32–46 (2001)

[8] Anderson, M.: Permutational Multivariate Analysis of Variance (PERMANOVA), pp. 1–15 (2017). https://doi.org/10.1002/9781118445112.stat07841

[9] Azzopardi, L., Moffat, A., Thomas, P., Zuccon, G.: User models, metrics and measures of search: a tutorial on the c/w/l evaluation framework. In: Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. pp. 347–348 (2021)

[10] Azzopardi, L.: Building cost-benefit models of information interactions. In: Proceedings of the 2017 Conference on Human Information Interaction and Retrieval. pp. 425–428 (2017)

[11] Azzopardi, L., Zuccon, G.: Two scrolls or one click: A cost model for browsing search results. In: Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016. pp. 696–702 (2016)

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

[13] Barry, C., Schamber, L.: Users' criteria for relevance evaluation: a cross-situational comparison. Information Processing and Management **34**(2/3), 219–236 (1998)

[14] Baskaya, F., Keskustalo, H., Järvelin, K.: 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 (2012)

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

[16] Beel, J., Gipp, B.: Google scholar's ranking algorithm: the impact of citation counts (an empirical study). In: Third International Conference on Research Challenges in Information Science (RCIS). pp. 439–446 (2009)

[17] Blair, D.C., Maron, M.E.: An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM **28**(3), 289–299 (1985)

[18] Bock, A.: Gütezeichen als Qualitätsaussage im digitalen Informationsmarkt: dargestellt am Beispiel elektronischer Rechtsdatenbanken. S. Toeche-Mittler (2000)

[19] Bonaccorsi, A., Daraio, C., Fantoni, S., Folli, V., Leonetti, M., Ruocco, G.: Do social sciences and humanities behave like life and hard sciences? Scientometrics **112**(1), 607–653 (2017)

[20] Bornmann, L., Bowman, B., Bauer, J., Marx, W., Schier, H., Palzenberg, M.: Bibliometric Standards for Evaluating Research Institutes in the Natural Sciences, pp. 201–224. Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact, MIT Press (2014)

[21] Boyack, K.: Mapping knowledge domains: characterizing pnas. In: Proceedings of the National Academy of Sciences of the United States of America. vol. 101, pp. 5192–5199 (2004)

[22] Brandsen, A., K Lambers, K., Verberne, S., Wansleeben, M.: User requirement solicitation for an information retrieval system applied to dutch grey literature in the archaeology domain. Journal of Computer Applications in Archaeology **2**(1), 21–30 (2019)

[23] Broder, A.: A taxonomy of web search. ACM SIGIR forum **36**, 3–10 (2002)

[24] Brody, T., Harnad, S., Carr, L.: Earlier web usage statistics as predictors of later citation impact. Journal of the American Society for Information Science and Technology **57**(8), 1060–1072 (2006)

[25] Bruza, P.D., Huibers, T.W.: A study of aboutness in information retrieval. Artificial Intelligence Review **10**(5-6), 381–407 (1996)

[26] Buckley, C.: Trec eval ir evaluation package (2004), `https://github.com/usnistgov/trec_eval`

[27] Chuklin, A., Markov, I., de Rijke, M.: Click models for web search. Synthesis Lectures on Information Concepts, Retrieval, and Services **7**(3), 1–115 (2015)

[28] Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)

[29] Cole, C., Kuhlthau, C.: Information and information seeking of novice versus expert lawyers: how experts add value. The New Review of Information Behaviour Research **1**, 103–115 (2000)

[30] Conrad, J.G., Zeleznikow, J.: The significance of evaluation in ai and law: a case study re-examining icail proceedings. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law. pp. 186–191 (2013)

[31] Conrad, J.G., Zeleznikow, J.: The role of evaluation in ai and law: an examination of its different forms in the ai and law journal. In: Proceedings of the 15th international conference on artificial intelligence and law. pp. 181–186 (2015)

[32] Cool, C., Belkin, N., Frieder, O., Kantor, P.: Characteristics of text affecting relevance judgments. National online meeting **14**, 77 (1993)

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

[34] Cosijn, E., Ingwersen, P.: Dimensions of relevance. Information Processing and Management **36**, 533–550 (2000)

[35] Council of Europe, Committee of Ministers: Recommendation r(95)11, concerning the selection, processing, presentation and archiving of court decisions in legal information retrieval systems, appendix I, art. III-5

[36] Council of Europe, Committee of Ministers: Recommendation rec(2002)13, on the publication and dissemination in the member states of the text of the european convention on human rights and of the case-law of the european court of human rights, explanatory memorandum, under 19

[37] Cronin, B.: Scholars and Scripts, Spoors and Scores, pp. 3–21. Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact, MIT Press (2014)

[38] Cronin, B., Sugimoto, C. (eds.): Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact. MIT Press, Cambridge, Massachusetts (2014)

[39] Cuadra, C.A., Katter, R.V.: Experimental studies of relevance judgments: Final report. Vol. I: Project summary. (NSF Report No. TM-3520/001/00). System Development Corp, Santa Monica, CA, United States (1967)

[40] Damen, L.: De invloed van bestuursrechtelijke annotaties op de rechtspraak. Nederlands Tijdschrift voor Bestuursrecht **35**, 280–281 (2016)

[41] De Bellis, N.: History and Evolution of (Biblio)Metrics, pp. 23–44. Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact, MIT Press (2014)

[42] Burg van der, R.: A query log analysis in the context of Legal Information Retrieval. Master thesis, Radboud University, Nijmegen, the Netherlands (2020)

[43] Erdt, M., Nagarajan, A., Sin, S., Theng, Y.: Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media. Scientometrics **109**(2), 1117–1166 (2016)

[44] Fox, S., Karnawat, K., Mydland, M., Dumais, S., White, T.: Evaluating implicit measures to improve the search experiences. In: SIGIR03 Workshop on Implicit Measures of User Interests and Preferences (2003)

[45] Free Access to Law Movement: Declaration on Free Access to Law (2002), `http://fatlm.org/declaration/`

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

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

[48] Garfield, G.: Citation Indexing: its theory and application in science, technology, and humanities. John Wiley & Sons, Inc., New York, NY (1979)

[49] Garner, B. (ed.): Black's Law Dictionary. 11 edn. (2019)

[50] Geist, A.C.J.: Rechtsdatenbanken und Relevanzsortierung. Doctoral dissertation, uniwien, Vienna, Austria (2016)

[51] Giménez-Toledo, E., Mañana Rodríguez, J., Engels, T.C., Ingwersen, P., Pölönen, J., Sivertsen, G., ..., Zuccala, A.A.: Taking scholarly books into account: current developments in five european countries. Scientometrics **107**, 685–699 (2016)

[52] Grossman, M.R., Cormack, G.V.: Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. Rich. JL & Tech. **17**, 1 (2010)

[53] Guo, F., Liu, C., Kannan, A., Minka, T., Taylor, M., Wang, Y.M., Faloutsos, C.: Click chain model in web search. In: Proceedings of the 18th international conference on World wide web. pp. 11–20 (2009)

[54] Harman, D.K.: The trec test collections. (2005), `https://trec.nist.gov/`

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

[56] Haustein, S.: Readership Metrics, pp. 327–344. Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact, MIT Press (2014)

[57] Hawking, D.: Challenges in enterprise search. ADC **4**, 15–24 (2004)

[58] Hicks, D.: he four literatures of social science, pp. 473–496. Handbook of quantitative science and technology research, Springer, Dordrecht, The Netherlands (2004)

[59] Hook, P.: Evaluating the work of judges, pp. 345–364. Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact, MIT Press (2014)

[60] Ingwersen, P., Järvelin, K.: Information retrieval in context: Irix. ACM SIGIR forum **39**(2), 31–39 (2005)

[61] Jackson, P., Al-Kofahi, K.: Human Expertise and Artificial Intelligence in Legal Search. Strukturierung der Juristischen Semantik, Weblaw, Bern, Switzerland (2011)

[62] Jansen, B.J., Spink, A., Kathuria, V.: How to define searching sessions on web search engines. In: Nasraoui, O., Spiliopoulou, M., Srivastava, J., Mobasher, B., Masand, B. (eds.) Advances in Web Mining and Web Usage Analysis. pp. 92–109. Springer, Berlin, Heidelberg, Germany (2007)

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

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

[65] Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments. In: Information Processing and Management. pp. 779–840 (2000)

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

[67] Kaltenbrunner, W., Rijcke de, S.: Quantifying "output' for evaluation: Administrative knowledge politics and changing epistemic cultures in dutch law faculties. Science and Public Policy **44**, 284–293 (2017)

[68] Kellar, M., Watters, C., Inkpen, K.M.: An exploration of web-based monitoring: implications for design. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 377–386 (2007)

[69] Kelly, D., Teevan, J.: Implicit feedback for inferring user preference: A bibliography. ACM SIGIR Forum **37**(2), 18–28 (2003)

[70] Kelly, D.: The future of IR (2018)

[71] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., Riedl, J.: Grouplens: Applying collaborative filtering to usenet news. Communications of the ACM **40**(3), 77–87 (1997)

[72] Kooijmans, T.: De roekeloze automobilist. Ars Aequi Maandblad **118**, 118–124 (2014)

[73] Kousha, K., Thelwall, M.: Web impact metrics for research assessment. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact p. 289 (2014)

[74] Krans, B.: Vorm of inhoud: de annotatie als wetenschappelijke publicatie? Ars Aequi Maandblad **237**, 237–239 (2017)

[75] Kurtz, M., Henneken, E.: Measuring metrics - a 40-year longitudinal cross-validation of citations, downloads, and peer review in astrophysics. Journal of the Association for Information Science and Technology **68**, 695–708 (2017)

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

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

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

[79] Lindijer, V.: De goede procesorde: een onderzoek naar de betekenis van de goede procesorde als normatief begrip in het burgerlijk procesrecht. Burgerlijk Proces & Praktijk nr. IV, Wolters Kluwer (2006)

[80] Locke, D., Zuccon, G.: A test collection for evaluating legal case law search. In: Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR). pp. 1261–1264 (2018)

[81] Makri, S., Blandford, A., Cox, A.: Investigating the information-seeking behaviour of academic lawyers: From ellis's model to design. Information Processing & Management **44**, 613–634 (2008)

[82] Manning, C., Schütze, H., Raghavan, P.: Introduction to information retrieval. Cambridge university press, Cambridge, United Kingdom (2008)

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

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

[85] McGregor, M., Azzopardi, L., Halvey, M.: 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 (2021)

[86] Merton, R.: The normative structure of science, pp. 267–278. The sociology of science: theoretical and empirical investigations (1973)

[87] Merton, R.K.: The matthew effect in science, ii: Cumulative advantage and the symbolism of intellectual property. isis **79**(4), 606–623 (1988)

[88] Mi, S., Jiang, J.: Understanding the interpretability of search result summaries. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 989–992 (2019)

[89] Montesquieu, C.: L'esprit des lois. A. Belin (1817)

[90] Nederlandse Orde van Advocaten: Modelkantoorhandboek. The Hague, The Netherlands (2020), `https://www.advocatenorde.nl/kantoorhandboek`

[91] Newton, I.: Letter to Robert Hooke (1675)

[92] Oard, D.W., Kim, J.: 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 (2001)

[93] Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H.: vegan: Community ecology package (2019), `https://CRAN.R-project.org/package=vegan`, r package version 2.5-5

[94] Opijnen van, M.: Op en in het web: Hoe de toegankelijkheid van rechterlijke uitspraken kan worden verbeterd. Boom Juridische uitgevers, The Hague, The Netherlands (2014)

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

[96] Park, T.: The nature of relevance in information retrieval: an empirical study. Library Quarterly **63**(3), 318–351 (1993)

[97] Peoples, L.F.: Testing the limits of westlawnext. Legal Reference Services Quarterly **31**, 125–149 (2012), `https://ssrn.com/abstract=1910766`

[98] Perneger, T.: Relation between online "hit counts" and subsequent citations: prospective study of research papers in the bmj. British Medical Journal **329**, 546–547 (2004)

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

[100] Priem, J., Taraborelli, D., Groth, P., Neylon, C.: Altmetrics: A manifesto (2011), `https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1187&context=scholcom`

[101] R Core Team: R: A language and environment for statistical computing (2017), `https://www.R-project.org/`

[102] Rabelo, J., Kim, M.Y., Goebel, R., Yoshioka, M., Kano, Y., Satoh, K.: Coliee 2020: Methods for legal document retrieval and entailment. In: JSAI International Symposium on Artificial Intelligence. pp. 196–210 (2020)

[103] Rees, A., Schultz, D.G.: A field experimental approach to the study of relevance assessments in relation to document searching. Vol. I. Final report (NSF Contract No. C-423). Case Western Reserve University, Cleveland, OH, United States (1967)

[104] Rehn, C., Kronman, U., Wadskog, D.: Bibliometric Indicators – Definitions and Usage at Karolinska Institutet. Karolinska Institutet University Library, Stockholm, Sweden (2014)

[105] Reichheld, F.: The one number you need to grow. Harvard business review **81.12**, 46–55 (2003)

[106] Rieh, S.Y., Belkin, N.J.: Understanding judgment of information quality and cognitive authority in the www. In: Proceedings of the 61st annual meeting of the American society for information science. pp. 279–289 (1998)

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

[108] Royal Netherlands Academy of Arts and Sciences (KNAW): 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, Amsterdam, The Netherlands (2005)

[109] Saracevic, T.: 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 (1975)

[110] Saracevic, T.: Relevance reconsidered, information science: Integration in perspectives. In: Proceedings of the Second Conference on Conceptions of Library and Information Science. pp. 201–218 (1996)

[111] Saracevic, T.: Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part III: Behaviour and Effects of Relevance. Journal of the American Society for Information Science and Technology **58**(13), 2126–2144 (2007)

[112] Savolainen, R., Kari, J.: User-defined relevance criteria in web searching. Journal of Documentation **62**(6), 685–707 (2006)

[113] Schamber, L., Eisenberg, M., Nilan, M.: A re-examination of relevance: toward a dynamic, situational definition. Information Processing and Management **26**(6), 755–776 (1990)

[114] Singhal, A.: Modern information retrieval: A brief overview. IEEE Data Eng. Bull. **24**(4), 35–43 (2001)

[115] Smith, C.L., Kantor, P.B.: User adaptation: good results from poor systems. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 147–154 (2008)

[116] Snel, M.: 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, The Hague, The Netherlands (2016)

[117] Snel, M.: Hoera, een lijstje! over bronvermelden. Ars Aequi **3**, 254–260 (2018)

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

[119] Stolker, C.: Een discipline in transitie: Rechtswetenschappelijk onderzoek na de commissie koers. Recht en Methode in onderzoek en onderwijs **1**, 13–43 (2011)

[120] Stolker, C.: Rethinking the Law School: Education, research, outreach and governance. Cambridge University Press, Cambridge, United Kingdom (2015)

[121] T., R.R., Chamberlain, J., Azzopardi, L.: Information retrieval in the workplace: A comparison of professional search practices. Information Processing & Management **54**, 1042–1057 (2018)

[122] Tang, D., Agarwal, A., O'Brien, D., Meyer, M.: Overlapping experiment infrastructure: More, better, faster experimentation. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 17–26 (2010)

[123] Taylor, A.: User relevance criteria choices and the information search process. Information Processing and Management **48**, 136–153 (2012)

[124] Taylor, A.: Examination of work task and criteria choices for the relevance judgment process. Journal of Documentation **69**(4), 523–544 (2013)

[125] Taylor, A., Cool, C., Belkin, N., Amadio, W.: Relationships between categories of relevance criteria and stage in task completion. Information Processing and Management **43**, 1071–1084 (2007)

[126] Teevan, J., Dumais, S.T., Liebling, D.J., Hughes, R.L.: Changing how people view changes on the web. In: Proceedings of the 22nd annual ACM symposium on User interface software and technology. pp. 237–246 (2009)

[127] Teufel, S.: Argumentative zoning for improved citation indexing. Computing attitude and affect in text: Theory and Applications. Springer, Dordrecht, The Netherlands (2006)

[128] Toms, E.G., O'Brien, H.L., Kopak, R., Freund, L.: Searching for relevance in the relevance of search. In: International Conference on Conceptions of Library and Information Sciences. pp. 59–78. Springer, Berlin, Heidelberg, Germany (2005)

[129] Turpin, A., Hersh, W.: Why batch and user evaluations do not give the same results. In: Proceedings of the TwentyFourth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. pp. 225–231 (2001)

[130] Van Gestel, R., Vranken, J.: Assessing legal research: Sense and nonsense of peer versus bibliometrics and the need for a european approach. German Law Journal **12**, 901–929 (2011)

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

[132] Verberne, S., He, J., Kruschwitz, U., Wiggers, G., Larsen, B., Russell-Rose, T., de Vries, A.P.: First international workshop on professional search. ACM SIGIR forum **52**(2), 153–162 (2019)

[133] Voorhees, E.M.: The philosophy of information retrieval evaluation. In: Workshop of the Cross-language Evaluation Forum for European Languages. pp. 355–370. Springer (2001)

[134] VSNU: Oordelen over rechten. rapport commissie voorbereiding onderzoeksbeoordelingrechtsgeleerdheid (2005)

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

[136] Wiggers, G., Lamers, W.: Shepard's citations revisited - citation metrics for dutch legal information retrieval. In: Proceedings of 17th International Conference of Scientometrics and Informetrics (2019)

[137] Wiggers, G., Verberne, S., Zwenne, G.J.: Citation metrics for legal information retrieval: scholars and practitioners intertwined? Legal Information Management **22**, 88–103 (2022)

[138] Wiggers, G., Verberne, S., Zwenne, G.J., Loon van, W.: Exploration of domain relevance by legal professionals in information retrieval systems. Legal Information Management **22**, 49–67 (2022)

[139] Wilson, P.: Situational relevance. Information Storage and Retrieval **9**(8), 457–471 (1973)

[140] Winkels, R., Boer, A., Plantevin, I.: Creating context networks in dutch legislation. In: Proceedings of JURIX. pp. 155–164 (2013)

[141] Winkels, R., Boer, A., Vredebregt, B., Someren van, A.: Towards a legal recommender system. In: Proceedings of JURIX. vol. 271, pp. 169–178 (2014)

[142] Winkels, R., Ruyter de, J.: 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 (2011)

[143] Winkels, R., Ruyter de, J., Kroese, H.: Determining authority of dutch case law. Legal Knowledge and Information Systems **235**, 103–112 (2011)

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

Summary

Legal professionals spend up to a third of their time doing research. During this research legal information retrieval (IR) helps users find information that is relevant for them. These legal IR systems are important because the number of legal documents published online is growing exponentially.

This research addresses the question: how can bibliometrics improve common ranking algorithms in legal information retrieval?

Chapter 2 focuses on the users of legal IR systems. Users were surveyed to determine whether legal practitioners (searching for themselves) and information professionals (searching for others) have the same perception of relevance. This was done by comparing the factors of relevance they consider then evaluating search results. We found no reason to distinguish between these user groups. With regards to the distinction between legal scholars and legal practitioners, it was determined in Chapter 3 that the usage and citations between scholarly and non-scholarly publications show no reason to create separate rankings users based on their affiliation.

Chapter 3 regards the documents in the legal IR system. The citation and usage analysis provided the theoretical insight that citations in legal documents measure part of a broad scope of impact, or relevance, on the entire legal field. Using this information a bibliometric-enhanced ranking variable was created.

There are several challenges to evaluating a live domain specific IR system. Chapter 4 deals with these challenges and why common evaluation methods in IR are not applicable. In the end, in Chapter 5, a cost based model is used for evaluation, which shows a reduction of cost for the user.

Combining all this information this thesis shows that a bibliometric-enhanced ranking feature that takes into account both usage and citations (two flavors of impact relevance), and increases in influence as the reliability of the data grows (in combination with a recency feature that gives new documents the benefit of the doubt and decreases at the same rate as the bibliometric feature increases), can reduce the cost required from legal professionals (whether practitioner, scholar or legal information professional) to find the point of satisfaction in the completeness ideal/research reality trade-off.

Samenvatting

Juridische professionals besteden tot een derde van hun tijd aan onderzoek. Gedurende dit onderzoek helpen juridische zoeksystemen gebruikers de informatie vinden die voor hun relevant is. Deze juridische zoeksystemen zijn belangrijk omdat het aantal juridische documenten dat online beschikbaar is exponentieel groeit.

Dit onderzoek behandelt de vraag: hoe kan bibliometrie de ranking algoritmes van juridische zoeksystemen verbeteren?

Hoofdstuk 2 gaat over de gebruikers van juridische zoeksystemen. Gebruikers zijn bevraagd om te bepalen of juridische professionals (die voor zichzelf zoeken) en informatiespecialisten (die voor een ander zoeken) hetzelfde beeld hebben van relevantie. Dit is onderzocht door de factoren van relevantie die zij benoemen wanneer ze een zoekresultaat beoordelen te vergelijken. Uit deze analyse blijkt dat er geen reden is om onderscheid te maken tussen deze groepen gebruikers. Met betrekking tot het onderscheid tussen wetenschappers en juridische professionals, toont het onderzoek in Hoofdstuk 3 aan dat het gebruik van, en de citaties tussen, wetenschappelijke en niet-wetenschappelijke publicaties geen reden geeft om aparte ranking algoritmes te maken voor gebruikers op basis van hun affiliatie.

Hoofdstuk 3 kijkt naar de documenten in juridische zoeksystemen. De citatie- en gebruiks-analyse geeft het theoretische inzicht dat citaties in juridische documenten deel van een brede impact, of relevantie, op het gehele juridische vakgebied weergeven. Het door bibliometrie verrijkte ranking algoritme is op basis van dit inzicht gemaakt.

Er zijn verschillende uitdaging met betrekking tot het evalueren van een live domein-specifiek zoeksysteem. Hoofdstuk 4 laat zien wat deze uitdagingen zijn, en waarom standaard evaluatiemethoden uit het onderzoeksveld van zoektechnologie niet bruikbaar zijn. In Hoofdstuk 5 wordt daarom een evaluatiemodel gebruikt op basis van de tijd die het gebruikers kost om informatie te vinden. Dit model toont een afname van de kosten van de gebruiker voor een met bibliometrie verrijkt ranking algoritme in vergelijking met een ranking algoritme zonder bibliometrie.

In zijn geheel toont dit werk aan dat een met bibliometrie verrijkte ranking dat zowel gebruik en citaties meeneemt (twee smaken van impact), en dat in gewicht toeneemt naar mate de data betrouwbaarder wordt (in

combinatie met een recentheid boost dat nieuwe documenten het voordeel van de twijfel geeft en afneemt in hetzelfde tempo als het gewicht van de bibliometrie toeneemt), de kosten van gebruikers van juridische zoeksystemen (zij het juridische professional, wetenschapper of informatiespecialist) vermindert, om een punt van verzadiging te bereiken in het compleetheid-sideaal/onderzoekswerkelijkheid compromis.

Appendices

Appendix A

Composition of search results

Documents and positions as retrieved from Leiden Law School data set.

For the hierarchy level of the court, the newness of the document, and the authority of the source a three-point scale is used. For all other factors the presence or absence of the factor in the result is shown.

A.1 Example query 1

Query: 'Tarieven deskundigenonderzoek' Translated: 'Fees expert witnesses'

Total number of results: 4945

Question 1:
Option 1: Recency (1), Legal hierarchy (3)
Option 2: Recency (2), Legal hierarchy (2)

Question 2:
Option 1: Title relevance (y), Recency (3)
Option 2: Title relevance (n), Recency (1)

Question 3:
Option 1: Legal hierarchy (1), Bibliographical relevance (n)
Option 2: Legal hierarchy (2), Bibliographical relevance (y)

Question 4:
Option 1: Source authority (1, Asser, T&C), Title relevance (n)
Option 2: Source authority (3, blog, news), Title relevance (y)

Question 5:
Option 1: Source authority (1, Asser, T&C), Document type (n)
Option 2: Source authority (2, mid), Document type (y)

Question 6:
Option 1: Source authority (2, mid), Authority author (n)
Option 2: Source authority (3, blog, news), Authority author (y)

Question 7:
Option 1: Document type (y), Bibliographical relevance (n)
Option 2: Document type (n), Bibliographical relevance (y)

A.2 Example query 2

Query: 'vernietiging overeenkomst terugwerkende kracht' Translated: 'voidable contract retroactive effect'

Total number of result: 1325

Question 1:
Option 1: Recency (2), Annotated (n)
Option 2: Recency (3), Annotated (y)

Question 2:
Option 1: Source authority (2, mid), Authority author (n)

Option 2: Source authority (3, blog, news), Authority author (y)

Question 3:
Option 1: Legal hierarchy (1), Annotated (n)
Option 2: Legal hierarchy (3), Annotated (y)

Question 4:
Option 1: Authority author (y), Recency (3)
Option 2: Authority author (n), Recency (1)

Appendix B

Queries and Seed Documents

Table B.1: Queries that we selected to sample documents for our dataset, and the IDs of the corresponding documents.

Query	DocID
Cancun	12923916
ECLI:NL:HR:2014:948	12981736
JAR 2014/298	14290648
Zalco	12871782
Berzona	13580788
ECLI:NL:HR:2014:3351	14223358
ECLI:NL:HR:2014:3077	14145097
Coface/Intergamma	12827114
NJ 2014/268	13238467
NJ 2014/62	12701453
Bescheidenheid en moed	14151738
Informatieverstrekking aan derden in het licht van goed werkgeverschap: is zwijgen de norm?	13002758
Preventieve hechtenis in Veen	12987162
de andere kant van de ZSM-medaille	13330606
TRA 2014/75	13800385

TAP 2014/1	12654375
hoe verder met de klachtplicht	12538900
Klaarheid over het Clearing House	14003488
We zijn geen padvinders: een verkennend onderzoek naar de criminele carriëres van leden van een procent motorclubs	14121997
schikken in het nieuwe ontslagrecht	13697909
Enkele aspecten van cao-recht	12654376
Het nieuwe jeugdstelsel en de jeugdbescherming	14013961
wat is er mis met een vrijspraak	14331724
Houdt de WWZ voldoende rekening met de contractuele grondslag van het cao-recht?	14124128
WFR 2014/1067	13705093
WFR 2014/1168	13835404
Waarheidsvinding in de jeugdzorg	12926733
NJB 2014/2056	14177758
De roekeloze automobilist	12702866
TRA 2014/76	13800386
Daar gaan we weer? Het concurrentiebeding revisited	13211319
De Vrijgestelde beleggingsinstelling	14309602
NJB 2014/1139	13235698
is de staat aansprakelijk voor klimaatverandering	12685430
de procedure na cassatie en verwijzing	14165599
het geheim van raadkamer	12987652
ArbeidsRecht 2014/53	14124136
Curator en overwaardearrangement	22171998
NTB 2014/3	12707423
Naar een vervanging van de unus-testisregel van artikel 342 Sv	13241348
Partneralimentatie in de praktijk: is maatwerk mogelijk?	14226701
Rangwisseling pandrecht door eigenlijke achterstelling	13627420
TFO 2014/134.1	13400193
Arbeidsrecht 2014/21	12882340
Arbeidsrecht en onderwijs	12660424

Cessie- en verpandingsverboden: nieuw arrest, nieuwe problemen	13361780
De civielrechtelijke aansprakelijkheid voor schade veroorzaakt door een autonome auto	14111819
FIP 2014/360	14340903
Het doel van garanties bij bedrijfsovernames: informatie of risico	13570943
TAP 2014/4	12654373
WFR 2014/1384	14154576
heeft het bw een politieke kleur	12658261

Appendix C

Seed Documents and Data

Table C.1: Results: the usage and citations for the 52 analyzed documents. 'N-S' refers to non-scholarly/non-scholars and 'Schol.' refers to scholarly/scholars. The first 42 lines are journal articles, the 10 lines below are legal cases.

DocumentID	Classification on Document Type	Schol. Affiliations	Final Classification	Schol. Citations	N-S Citations	Usage Schol.	Usage N-S[1]
14151738	Non-scholarly	0	Non-scholarly	1	5	9	24
12987162	Scholarly	1	Scholarly	1	2	138	17
13330606	Scholarly	1	Scholarly	1	8	13	20
13800385	Non-scholarly	0	Non-scholarly	0	20	60	91
12654375	Non-scholarly	0	Non-scholarly	9	136	164	73
12538900	Scholarly	0	Non-scholarly	4	38	58	50
12654376	Non-scholarly	0	Non-scholarly	0	5	101	114
14013961	Non-scholarly	0	Non-scholarly	0	26	89	23
13705093	Non-scholarly	0	Non-scholarly	0	19	2	17
13835404	Non-scholarly	0	Non-scholarly	0	3	34	6
12926733	Scholarly	1	Scholarly	1	13	77	13
14177758	Scholarly	1	Scholarly	1	15	116	88
12702866	Non-scholarly	1	Non-scholarly	8	16	484	45
13800386	Non-scholarly	0	Non-scholarly	0	16	33	64
13211319	Non-scholarly	0	Non-scholarly	0	14	241	168
14309602	Non-scholarly	0	Non-scholarly	0	49	75	37
13235698	Non-scholarly	0	Non-scholarly	0	2	18	10
12685430	Scholarly	0	Non-scholarly	7	18	49	17

14165599	Non-scholarly	0	Non-scholarly	0	14	4	5
12707423	Scholarly	1	Scholarly	5	12	159	49
14226701	Non-scholarly	0	Non-scholarly	0	2	8	13
13627420	Scholarly	0	Non-scholarly	9	51	64	283
13400193	Non-scholarly	0	Non-scholarly	2	57	63	118
13361780	Scholarly	1	Scholarly	2	8	48	35
14340903	Non-scholarly	0	Non-scholarly	1	13	34	18
12654373	Non-scholarly	0	Non-scholarly	1	13	54	32
14154576	Non-scholarly	0	Non-scholarly	0	11	62	5
12658261	Scholarly	1	Scholarly	1	1	12	5
13002758	Non-scholarly	0	Non-scholarly	0	0	165	72
14003488	Scholarly	1	Scholarly	0	0	28	9
14121997	Scholarly	1	Scholarly	0	0	41	94
13697909	Scholarly	1	Scholarly	0	0	67	88
14331724	Scholarly	2	Scholarly	0	0	164	16
14124128	Non-scholarly	0	Non-scholarly	0	0	66	69
12987652	Non-scholarly	0	Non-scholarly	0	0	3	26
14124136	Non-scholarly	0	Non-scholarly	0	0	73	295
22171998	Non-scholarly	0	Non-scholarly	0	0	16	11
13241348	Scholarly	1	Scholarly	0	0	157	31
12882340	Non-scholarly	0	Non-scholarly	0	0	64	20
12660424	Non-scholarly	0	Non-scholarly	0	0	86	36
14111819	Non-scholarly	0	Non-scholarly	0	0	108	18
13570943	Scholarly	1	Scholarly	0	0	42	95
12981736	Non-scholarly	0	Non-scholarly	9	134	143	42
14290648	Non-scholarly	0	Non-scholarly	1	65	99	48
12871782	Non-scholarly	0	Non-scholarly	1	49	20	108
13580788	Non-scholarly	0	Non-scholarly	2	10	18	7
14223358	Non-scholarly	0	Non-scholarly	0	17	4	7
14145097	Non-scholarly	0	Non-scholarly	3	179	48	28
12827114	Non-scholarly	0	Non-scholarly	60	554	177	164
13238467	Non-scholarly	0	Non-scholarly	11	158	276	34
12701453	Non-scholarly	0	Non-scholarly	5	28	54	17
14281373	Non-scholarly	0	Non-scholarly	4	1155	164	362

Appendix D

Baseline and degraded ranking

This table shows the the difference between the baseline ranking and the degraded ranking. Each row in the table represents one query. Each ranking shown was 10 documents. Each number in the column corresponds with the number of documents for that query that was ranked higher or lower in the degraded model, documents that were not present in the degraded model but replaced by another document, and the number of documents that remained in the same position.

QueryID	Moved Up	Moved Down	Document Replaced	Same Position
1	1	2	2	5
2	3	4	1	2
3	1	6	2	1
4	1	4	3	2
5	0	4	6	0
6	0	2	8	0
7	0	2	6	2
8	2	2	1	5
9	3	4	1	2
10	2	3	3	2

11	2	3	4	1
12	4	2	2	2
13	2	3	4	1
14	1	1	1	7
15	1	4	4	1
16	1	2	5	2
17	0	5	4	1
18	1	4	3	2
19	1	3	1	5
20	3	3	1	3
21	1	3	2	4
22	3	3	1	3
23	4	3	1	2
24	0	1	4	5
25	1	1	2	6
26	1	2	3	4
27	0	7	2	1
28	2	1	2	5
29	1	4	3	2
30	3	2	3	2
31	3	2	3	2
32	0	3	3	4
33	3	4	3	0
34	1	4	2	3
35	2	3	3	2
36	0	7	3	0
37	2	0	2	6
38	0	0	10	0
39	2	1	3	4
40	1	6	3	0
41	2	3	4	1
42	0	6	4	0
43	0	6	2	2
44	0	2	6	2

45	0	0	10	0
46	1	0	9	0
47	3	3	2	2
48	3	0	2	5
49	0	5	3	2
50	2	1	1	6
51	0	2	2	6
52	5	1	3	1
53	3	2	2	3
54	1	3	5	1
55	2	5	2	1

Table D.1: The number of the documents from the top-10 that changed position in the degraded ranking as compared to the baseline ranking.

Appendix E

Example of Survey Question

ECLI:NL:GHAMS:2019:1693 - Gerechtshof Amsterdam - 10-5-2019 - Hoger beroep - Strafrecht

Overtreding gebiedsverbod. Vrijspraakverweer; onvoldoende duidelijk welke straten onder het verbod vielen.

Gst. 2018/125 - Met noot - Rooij van, M. e.a. - Raad van State - 28-3-2018

Toepasselijkheid onschuldpresumptie gebiedsverbod art. 172a Gemeentewet. (Amsterdam)

ECLI:NL:RVS:2018:1043 **JOM 2018/319** **VIND Bestuursrecht 2018**

Vindplaatsen

AB 2019/141 - Met noot - Brouwer, J.G. e.a. - Rechtbank Midden-Nederland - 24-5-2018

Doodsbedreiging ernstige verstoring openbare orde; gebiedsverbod rondom woonwagen.

ECLI:NL:RBMNE:2018:2287 **VIND Bestuursrecht 2018**

Vindplaatsen

ECLI:NL:RBLIM:2018:9320 - Rechtbank Limburg - 3-10-2018

Vovo hangende bezwaar. Verweerder heeft eerder een gebiedsverbod aan verzoeker opgelegd op grond van artikel 172a van de Gemeentewet. De voorzieningenrechter heeft het vorige besluit geschorst, omdat dit een deugdelijke motivering ontbeerde. Hierna h [...]

ECLI:NL:RBLIM:2018:8273 - Rechtbank Limburg - 31-8-2018 - Voorlopige voorziening - Bestuursrecht

Vovo hangende bezwaar. Verweerder heeft verzoeker een gebiedsverbod op grond van artikel 172a van de Gemeentewet opgelegd voor twee gebieden in Heerlen, omdat hij structureel overlast heeft veroorzaakt en de openbare orde heeft verstoord. Het oplegge [...]

Vindplaatsen

ECLI:NL:RBGEL:2018:4774 - Rechtbank Gelderland - 24-10-2018 - Eerste aanleg - enkelvoudig - Civiel recht

Verbintenissenrecht. Onrechtmatige overheidsdaad. Nalaten adequaat optreden tegen strafbare feiten waarvan eisers slachtoffer zijn geworden en die hen tot verhuizing hebben gedwongen. Veiligheidsgarantie. De vorderingen worden afgewezen.

AR 2018/5887 **JA 2019/6** **PS Updates 2018/947**

Vindplaatsen

ECLI:NL:RVS:2018:2112 - Raad van State - 27-6-2018 - Hoger beroep - Bestuursrecht

Bij besluit van 13 april 2016 heeft de burgemeester aan [appellant] een gebiedsverbod opgelegd voor de duur van 3 maanden.

VIND Bestuursrecht 2018

Vindplaatsen

ECLI:NL:GHAMS:2018:3267 - Gerechtshof Amsterdam - 6-9-2018 - Hoger beroep - Strafrecht

Opzettelijk handelen in strijd met het in artikel 2 onder C van de Opiumwet gegeven verbod.diefstal door twee of meer verenigde personen. opzettelijk niet voldoen aan gegeven bevel.

Gst. 2012/67 - Met noot - Berkouwer, E.C. e.a. - Rechtbank Amsterdam - 3-4-2012

Zorgvuldige dossieropbouw is essentieel voor de onderbouwing van een art. 172a Gemw-bevel en de rechterlijke toets.

AB 2012/174 **ECLI:NL:RBAMS:2012:BW1140** **VIND Bestuursrecht 2013**

Vindplaatsen

ECLI:NL:GHAMS:2018:3386 - Gerechtshof Amsterdam - 2-7-2018 - Hoger beroep - Strafrecht

Overtreding gebiedsverboden, wederspannigheid en diefstal. Geldige dagvaarding. Hulpverleningskader dat in een andere strafzaak via bijzonder voorwaarden is geïndiceerd voortzetten.

Appendix F

Curriculum Vitae

Gineke Wiggers was born in Hilversum, the Netherlands, in 1988. From 2006 to 2009 she studied european law with a minor in Dutch law at Maastricht University, and in 2011 she obtained a master's degree in Dutch law from the same university. Her master's thesis was entitled The Potential Influence of the Google Book Settlement on Dutch Copyright Holders.

After graduating from university she applied for a PhD, and learned how not to write a research proposal. She has worked as a legal assistant, publishing assistant, customer service agent and business analyst. From her position as business analyst at Legal Intelligence she came up with a new and improved research proposal and submitted it for a PhD in law and data science at Leiden Law School, Leiden University. She started her PhD in 2017, conducting research at eLaw and the data science research program, whilst she continued to work at Legal Intelligence.

Her work has been published in Legal Information Management, proceedings of the European Conference on Information Retrieval, the BIR workshops and the Nederlands Juristenblad.

Appendix G

Acknowledgements

I would like to thank my supervisors, Gerrit-Jan Zwenne and Suzan Verberne, for their support. Aside from all their valuable feedback, they have provided me with two insights that have helped me get this done: (1) the best theses are the ones that are done, and (2) the end product of a PhD is a researcher, not a thesis.

I want to thank the eLaw family for supporting me. Everyone at eLaw is incredibly kind and motivated, making it an amazing place to work. I especially want to thank Pieter for providing me with good coffee and Jenneke for all the moral support.

The Data Science Research Program is an amazing group of PhD students. I feel truly blessed to have been able to work alongside a group of such smart, talented and fun people. I have thoroughly enjoyed the foosball, pub-quizzes, nachos and murder mystery parties. I have learned more about archeology, deep-vein thrombosis, sign language, whiskey and t-SNE maps than I ever imagined. I have also learned programming in Python, and thanks to Wouter now know a little about statistics. Of which I, as an LL.M., am quite proud.

Roel van den Burg has been of tremendous value in this research. He helped me with the second survey and wrote an incredible master thesis about legal IR. I would also like to thank Arjen de Vries for his valuable

insights. Alan Hanbury, my ECIR doctoral consortium mentor helped me realise that I needed to go back to the basics with regards to evaluation methods. I want to thank Marc van Opijnen for providing feedback on my very first paper. The BIR, Jurix and legal AIIA communities have been welcoming and inspiring. I look forward to in-person meetings again.

This thesis would not have been possible without the support and data of Legal Intelligence. I particularly want to thank Laurens and Roderick for encouraging to apply, Pieter and Tjerk for their help through the process, and Sjoerd and Stephanie for their endless cheerleading.

Wouter van Loon not only helped me with the statistics in this research, but has also been an endless supporter. For me, he is the best thing I gained from this PhD. I look forward to 'walking with dr. Wouter and dr. Wiggers'.

Last, but certainly not least, I want to thank my family and friends. For their encouragement and support, but also for occasionally reminding me how lucky I am to be able to do this. Your reminders that I am loved, regardless of my title, mean the world to me. This page does not have enough space to tell you how grateful I am to have you in my life.

This thesis is dedicated to my grandmothers Grietje Wiggers-Supèr and Bertie de Hartog-Flinterman. I stand not only on the shoulders of giants [91], but have also been upheld by the strongest women I know.

SIKS dissertation series

Overview of theses in the SIKS dissertation series since 2016

2016	01	Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
	02	Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
	03	Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
	04	Laurens Rietveld (VU), Publishing and Consuming Linked Data
	05	Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
	06	Michel Wilson (TUD), Robust scheduling in an uncertain environment
	07	Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
	08	Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
	09	Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
	10	George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
	11	Anne Schuth (UVA), Search Engines that Learn from Their Users
	12	Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
	13	Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
	14	Ravi Khadka (UU), Revisiting Legacy Software System Modernization
	15	Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
	16	Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
	17	Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
	18	Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
	19	Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UVA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UVA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (UvT), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks

	48	Tanja Buttler (TUD), Collecting Lessons Learned
	49	Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
	50	Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains
2017	01	Jan-Jaap Oerlemans (UL), Investigating Cybercrime
	02	Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
	03	Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
	04	Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
	05	Mahdieh Shadi (UVA), Collaboration Behavior
	06	Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
	07	Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
	08	Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
	09	Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
	10	Robby van Delden (UT), (Steering) Interactive Play Behavior
	11	Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
	12	Sander Leemans (TUE), Robust Process Mining with Guarantees
	13	Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
	14	Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
	15	Peter Berck (RUN), Memory-Based Text Correction
	16	Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
	17	Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
	18	Ridho Reinanda (UVA), Entity Associations for Search
	19	Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
	20	Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
	21	Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
	22	Sara Magliacane (VU), Logics for causal inference under uncertainty
	23	David Graus (UVA), Entities of Interest — Discovery in Digital Traces
	24	Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
	25	Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
	26	Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
	27	Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

	28	John Klein (VU), Architecture Practices for Complex Contexts
	29	Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
	30	Wilma Latuny (UvT), The Power of Facial Expressions
	31	Ben Ruijl (UL), Advances in computational methods for QFT calculations
	32	Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
	33	Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
	34	Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
	35	Martine de Vos (VU), Interpreting natural science spreadsheets
	36	Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
	37	Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
	38	Alex Kayal (TUD), Normative Social Applications
	39	Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
	40	Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
	41	Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
	42	Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
	43	Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
	44	Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
	45	Bas Testerink (UU), Decentralized Runtime Norm Enforcement
	46	Jan Schneider (OU), Sensor-based Learning Support
	47	Jie Yang (TUD), Crowd Knowledge Creation Acceleration
	48	Angel Suarez (OU), Collaborative inquiry-based learning
2018	01	Han van der Aa (VUA), Comparing and Aligning Process Representations
	02	Felix Mannhardt (TUE), Multi-perspective Process Mining
	03	Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
	04	Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
	05	Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
	06	Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
	07	Jieting Luo (UU), A formal account of opportunism in multi-agent systems
	08	Rick Smetsers (RUN), Advances in Model Learning for Software Systems
	09	Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

	10	Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
	11	Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
	12	Xixi Lu (TUE), Using behavioral context in process mining
	13	Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
	14	Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
	15	Naser Davarzani (UM), Biomarker discovery in heart failure
	16	Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
	17	Jianpeng Zhang (TUE), On Graph Sample Clustering
	18	Henriette Nakad (UL), De Notaris en Private Rechtspraak
	19	Minh Duc Pham (VUA), Emergent relational schemas for RDF
	20	Manxia Liu (RUN), Time and Bayesian Networks
	21	Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
	22	Eric Fernandes de Mello Araujo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
	23	Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
	24	Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
	25	Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
	26	Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
	27	Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis
	28	Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
	29	Yu Gu (UVT), Emotion Recognition from Mandarin Speech
	30	Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web
2019	01	Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
	02	Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
	03	Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
	04	Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
	05	Sebastiaan van Zelst (TUE), Process Mining with Streaming Data
	06	Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
	07	Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
	08	Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
	09	Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VU), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35 Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming
36 Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills
37 Jian Fang (TUD), Database Acceleration on FPGAs

	38	Akos Kadar (OUN), Learning visually grounded and multilingual representations
2020	01	Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
	02	Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
	03	Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
	04	Maarten van Gompel (RUN), Context as Linguistic Bridges
	05	Yulong Pei (TUE), On local and global structure mining
	06	Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
	07	Wim van der Vegt (OUN), Towards a software architecture for reusable game components
	08	Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
	09	Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
	10	Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining
	11	Sepideh Mesbah (TUD), Semantic-Enhanced Training Data Augmentation-Methods for Long-Tail Entity Recognition Models
	12	Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
	13	Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
	14	Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
	15	Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
	16	Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
	17	Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
	18	Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
	19	Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
	20	Albert Hankel (VU), Embedding Green ICT Maturity in Organisations
	21	Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be
	22	Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
	23	Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
	24	Lenin da Nobrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
	25	Xin Du (TUE), The Uncertainty in Exceptional Model Mining
	26	Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization

	27	Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
	28	Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
	29	Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
	30	Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst
	31	Gongjin Lan (VU), Learning better – From Baby to Better
	32	Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising
	33	Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation
	34	Anna Bon (MU), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
	35	Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production
2021	01	Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
	02	Rijk Mercuur (TUD), Simulating Human Routines:Integrating Social Practice Theory in Agent-Based Models
	03	Seyyed Hadi Hashemi (UVA), Modeling Users Interacting with Smart Devices
	04	Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
	05	Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
	06	Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
	07	Armel Lefebvre (UU), Research data management for open science
	08	Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
	09	Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
	10	Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
	11	Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
	12	Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
	13	Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
	14	Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
	15	Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
	16	Esam A.H. Ghaleb (UM), BIMODAL EMOTION RECOGNITION FROM AUDIO-VISUAL CUES
	17	Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
	18	Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

	19	Roberto Verdecchia (VU), Architectural Technical Debt: Identification and Management
	20	Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
	21	Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
	22	Sihang Qiu (TUD), Conversational Crowdsourcing
	23	Hugo Manuel Proença (LIACS), Robust rules for prediction and description
	24	Kaijie Zhu (TUE), On Efficient Temporal Subgraph Query Processing
	25	Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
	26	Benno Kruit (CWI & VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
	27	Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
	28	Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs
2022	1	Judith van Stegeren (UT), Flavor text generation for role-playing video games
	2	Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
	3	Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
	4	Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
	5	Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
	6	Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
	7	Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
	8	Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
	9	Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
	10	Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
	11	Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
	12	Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
	13	Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
	14	Michiel Overeem (UU), Evolution of Low-Code Platforms
	15	Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
	16	Pieter Gijsbers (TU/e), Systems for AutoML Research
	17	Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification

18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
25 Anna L.D. Latour (LU), Optimal decision-making under constraints and uncertainty
26 Anne Dirkson (LU), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality