The relevance of impact: bibliometric-enhanced legal information retrieval

Wiggers, G.

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.

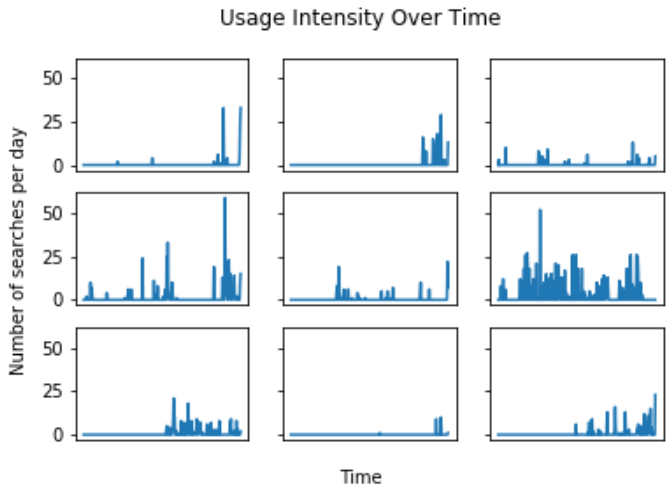

Figure 4.1: The number of searches per day for nine users over 10 months.

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.

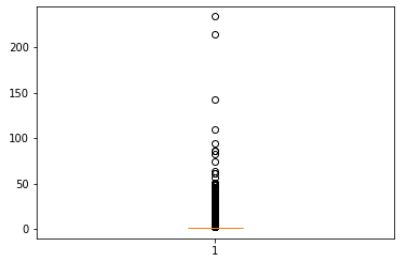

Figure 4.2: Distribution of queries and number of users conducting them

	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.