



Bibliometric-Enhanced Legal Information Retrieval

Gineke Wiggers^{1,2}(✉) 

¹ Leiden University, Leiden, The Netherlands

g.wiggers@law.leidenuniv.nl

² Legal Intelligence, Alphen aan den Rijn, The Netherlands

Abstract. This research project addresses user-focused ranking in legal information retrieval (IR). It studies the perception of relevance of search results for users of Dutch legal IR systems, the employment of usage and citation variables to improve the ranking of search results (bibliometric-enhanced information retrieval), and user-centred evaluation for ranking improvements. The goal of this project is improve the ranking in legal IR systems. Ultimately, this will help legal professionals find relevant information faster.

Keywords: Legal information retrieval · Ranking · User study

1 Motivation

In the legal domain, the amount of information available digitally is rapidly increasing. Legal scholars and professionals have to navigate this information to find the case law and articles relevant for them. Dutch legal information retrieval systems currently focus primarily on matching queries and documents for their ranking. In order to help users find the most relevant documents, these ranking algorithms should be expanded.

A possible solution is bibliometric-enhanced information retrieval, where citations are measured and used as a proxy for impact in the ranking algorithm. But the legal domain differs from other research domains due to the often strong interconnection between research and practice. In the Dutch legal domain this is demonstrated by the lack of distinction between legal scientific and professional publications. This is one of the reasons why bibliometrics, which in other fields measures impact of scientific publications, has not yet been established within the Dutch legal domain.

This research will cover both the theory (such as relevance factors in legal publications and the meaning of citations in legal publications) and experimentation of applying bibliometrics to legal documents. The result is a new bibliometrics-enhanced ranking algorithm for legal information retrieval systems.

With thanks to Legal Intelligence for providing the data for this research.

© Springer Nature Switzerland AG 2020

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 630–634, 2020.

https://doi.org/10.1007/978-3-030-45442-5_83

2 Research Questions

The main question in this research is: can we use citation and usage (click) metrics to improve ranking in legal IR? This question comprises five sub-questions:

1. What factors influence the perception of relevance of users of legal IR systems?
2. What does a citation signify in legal publications?
3. What is the relation between user interactions with documents (usage) and citations in legal publications?
4. What is the right balance between text-based relevance and user-based relevance in ranking algorithms for legal IR?
5. What is the appropriate user-focused rank evaluation metric for legal IR?

3 Related Work

The notion of using bibliometrics to enhance information retrieval is inspired by the groundwork of Garfield [5] on the theory of citations as a measure of impact, and more recent work of Beel and Gip [4]. The normalization of citations is based on the work of Waltman et al. [18]. Based on their work we have decided on time, field, and document type normalization. The context of Dutch legal publications and their citation culture is provided by Stolker [16] and Snel [15].

This research is inspired by work of Van Opijnen and Santos [10], who translate the theory of Saracevic [13] on the spheres of relevance to the legal domain. The work of Barry and Schamber [2,3] describes the different indicators of relevance as identified by users.

The choice for a user-focused evaluation metric is influenced by the work of Järvelin and Kekäläinen [6]. The decision to use click data as a form of implicit feedback is based on work of Joachims et al. [7,8].

4 Preliminary Work

4.1 What Factors Influence the Perception of Relevance of Users of Legal IR Systems?¹

[17] The goal of our study was to make explicit which document characteristics users consider as factors of relevance when assessing a search result (document representation; such as title and snippet) in legal IR. To achieve this, we conducted a user questionnaire. In this questionnaire we showed users of a legal IR system a query and two search results. The user has to choose which of the two results he/she would like to see ranked higher for the query and is asked to provide a reasoning for his/her choice. The questionnaire had eleven pairs of search results spread over two queries. A total of 43 legal professionals participated in our study.

¹ An earlier version of this paper was presented at the 17th Dutch-Belgian Information Retrieval Workshop.

The identified relevance factors were title relevance, document type, recency, 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. The identified factors confirm previous research, such as the work of Barry and Schamber [2,3]. The identified factors also suggest that there are document characteristics (e.g. authority, legal hierarchy and whether the document is annotated) that are usually grouped under cognitive or situational relevance, and thereby considered to be personal, but that users in the legal domain agree on. The agreement within the field makes that these factors can be grouped under domain relevance as described by Van Opijnen and Santos [10].

4.2 What Does a Citation Signify in Legal Publications?²

[19] In our next paper we examined citations in legal information retrieval. Citation metrics can be a factor of relevance in the ranking algorithms of information retrieval systems. But the challenge in legal bibliometrics, and therefore legal bibliometric-enhanced IR, is that the legal domain differs from other research domains in two manners: (1) its strong national ties and (2) the often strong interconnection between research and practice [16].

First, we contrasted citations in the legal domain to citations in the hard sciences based on the literature on scholarly citations. Second, we applied quantitative analysis of legal documents and citations to test whether the theory described in literature (particularly the distinction between scholarly and practitioners oriented publications) is confirmed by the data.

An analysis of 52 cited (seed) documents and 3196 citing documents showed no strict separation in citations between documents aimed at scholars and documents aimed at practitioners. Our results suggest that citations in legal documents do not measure the impact on scholarly publications and scholars, but measure a broader scope of impact, or relevance, for the legal field.

5 Proposed Methods

5.1 What Is the Relation Between User Interactions with Documents (Usage) and Citations in Legal Publications?

Based on the outcome of the above research question, we wish to create a boost function to create the bibliometric-enhanced ranking algorithm. Because citations in legal documents measure a broader form of impact, it is likely that citations alone do not provide a complete overview of this broader impact, but that other factors, such as usage (through click data), create a more complete overview.

² An earlier version of this paper has been presented at the 8th International Bibliometric-enhanced Information Retrieval Workshop (BIR 2019) at the 41st European Conference on Information Retrieval.

However, the work of Perneger [12] suggests that there might be a correlation between usage and citations. Therefore we took all documents published and added to the Legal Intelligence system (the largest legal IR system in the Netherlands) in February 2017.³ This led to a set of 43,218 documents.

For each document, based on their unique document ID, we collected all click data (usage) until 2019. We then computed the Spearman correlation between the usage and the citations. We chose Spearman correlation because the data, like all citation data, does not have a normal distribution. The results show a Spearman correlation of 0.57 ($p < 0.0001$). This means that there is a moderate positive correlation between citations and usage, which is highly significant.

For the ranking algorithm, this correlation means that two separate boost functions will boost some results too much and others too little. Therefore a single harmonized boost function appears to be a better choice.

5.2 What Is the Right Balance Between Text-Based Relevance and User-Based Relevance in Ranking Algorithms for Legal IR?

Next to balancing citations and usage, the boost function as a whole will also have to be balanced against the current text-based relevance score (TF-IDF, BM25 or similar). Furthermore, the usage (click) and citation data will not be reliable from the moment of publication of the document. These new documents will have to be given the benefit of the doubt, for example through a freshness score. The optimal balance of these variables will be tuned using the evaluation metric.

5.3 What Is the Appropriate User-Focused Rank Evaluation Metric for Legal IR?

One of the challenges for legal information retrieval is finding the correct evaluation metric. This has several causes:

1. The small user group when compared to web search. This makes A/B testing difficult.
2. Differences in results lists. Because of differences in journal subscriptions two users may not see the same results list when they do the same query.
3. The high tariffs of legal experts, which makes a golden answer set prohibitively expensive to create and maintain.

A metric based on an implicit feedback model is of particular interest, as this is affordable and user-focused. We have made a first attempt to create an implicit feedback DCG model, but there is data sparsity. This means it is not possible to find queries for which the entire results list has a relevance judgment. Possible solutions will be sought in the fields of patent search and e-discovery.

³ 2017 was chosen to ensure that the documents have had time to gather citations. February was chosen as January often contains holidays and a lot of review articles over the past year.

References

1. Atzmüller, C., Steiner, P.M.: Experimental vignette studies in survey research. *Methodology* **6**(3), 128–138 (2010)
2. Barry, C.L.: User-defined relevance criteria: an exploratory study. *J. Am. Soc. Inf. Sci.* **45**(3), 149–159 (1994)
3. Barry, C.L., Schamber, L.: Users' criteria for relevance evaluation: a cross-situational comparison. *Inf. Process. Manage.* **34**(2/3), 219–236 (1998)
4. 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. IEEE (2009)
5. Garfield, G.: *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. Wiley, New York (1979)
6. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* **20**(4), 422–446 (2002)
7. Joachims, T.: Optimizing search engines using clickthrough data. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142 (2002)
8. Joachims, T., Granka, L.A., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. *Sigir* **5**, 154–161 (2005)
9. Joachims, T., Swaminathan, A., Schnabel, T.: Unbiased learning-to-rank with biased feedback. In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM 2017)*, pp. 781–789 (2017)
10. van Opijnen, Marc, Santos, Cristiana: On the concept of relevance in legal information retrieval. *Artif. Intell. Law* **25**(1), 65–87 (2017). <https://doi.org/10.1007/s10506-017-9195-8>
11. Park, T.K.: The nature of relevance in information retrieval: an empirical study. *Library Q.* **63**(3), 318–351 (1993)
12. Perneger, T.V.: Relation between online “hit counts” and subsequent citations: prospective study of research papers in the BMJ. *BMJ* **2004**(329), 546 (2004)
13. Saracevic, T.: Relevance: a review of and framework for the thinking on the notion in information science. *J. Am. Soc. Inf. Sci.* **26**, 321–343 (1975)
14. Jones, E., Oliphant, E., Peterson, P., et al.: *SciPy: Open Source Scientific Tools for Python*. <http://www.scipy.org/>. Accessed 22 Oct 2019
15. Snel, M.V.R.: Hoera, een lijstje! Over bronvermelden. *Ars Aequi* **3**, 254–260 (2018)
16. Stolker, C.: *Rethinking the Law School: Education, Research, Outreach and Governance*. Cambridge University Press, Cambridge (2015)
17. Wiggers, G., Verberne, S., Zwenne, G.J.: Exploration of intrinsic relevance judgments by legal professionals in information retrieval systems. In: *Proceedings of the 17th Dutch-Belgian Information Retrieval Workshop*, pp. 5–8 (2018). <https://arxiv.org/abs/1812.04265>
18. Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S., van Raan, A.F.: Towards a new crown indicator: some theoretical considerations. *J. Inform.* **5**(1), 37–47 (2011)
19. Wiggers, G., Verberne, S.: Citation metrics for legal information retrieval systems. In: *Proceedings of the 8th International Workshop on Bibliometric-Enhanced Information Retrieval (BIR) co-located with the 41st European Conference on Information Retrieval (ECIR 2019)*, pp. 39–50 (2019). <http://ceur-ws.org/Vol-2345/>