The Task: Distinguishing Tasks and Sessions in Legal Information Retrieval

Gineke Wiggers*
Guido Zuccon
g.wiggers@law.leidenuniv.nl
g.zuccon@uq.edu.au
The University of Queensland, School of Information Technology and Electrical Engineering
Brisbane, St. Lucia, Queensland, Australia

ABSTRACT

Legal information retrieval (IR) is a form of professional search often associated with high recall. Information seeking in this context can consist of a single query with no clicks (known as updating behaviour), a literature review where a complex boolean query crafted over several iterations is performed and all documents returned are inspected, or a seeking task spanning days or weeks, consisting of multiple queries interleaved with other tasks. Analysis of query logs is paramount to the improvement of current legal IR systems, and in particular of the system we are associated with, the Dutch Legal Intelligence IR system. This analysis however requires the ability to automatically identify which queries of a user are related to the same search goal — or in other words, related to the same search task. The current practice of defining sessions — a set of user interactions with the IR system with no more than 30 minutes between user actions — and equating a session to representing a search task, might prove ineffective given the characteristics of this user group.

In this paper we provide an initial analysis of a sub-set of the query log from the Dutch Legal Intelligence IR system, comprising of 970 queries issued by 10 users within the space of 1 year. From this query log, we used the 30-minutes heuristic to define sessions, and extract 126 sessions, ranging from 1 to 71 sessions per user. We then independently annotate the query log to manually identify search tasks: this activity leads to the identification of 55 tasks, ranging from 1 to 21 tasks per user. In doing this, we highlight how the currently employed heuristic is not adequate to extract search queries from a user that are related to the same search task. We also show why tasks are more informative than sessions with regards to legal information retrieval. We further describe the potential of using characteristics such as Levenshtein distance, common words and string matching for automated task classification.

CCS CONCEPTS

• **Information systems** → **Expert search**; **Task models**; **Query log analysis**.

KEYWORDS

Information Retrieval, Tasks, Professional Search, User Query Data

ACM Reference Format:
Gineke Wiggers and Guido Zuccon. 2022. The Task: Distinguishing Tasks and Sessions in Legal Information Retrieval. In *Australasian Document Computing Symposium (ADCS '22), December 15–16, 2022, Adelaide, Australia; Author Preprint*. ACM, New York, NY, USA, 4 pages. https://doi.org/TBA

1 INTRODUCTION

Legal Information Retrieval is a form of professional search [19, 21] often associated with the requirement for high (or "total") recall [3, 16, 17]. Only a hand-full of data-sets are publicly available for experimentation [1, 11, 12, 14, 15][1] and even less user data is available about this task. Of the data available, most is from common law jurisdictions where English is (one of) the primary language(s) spoken. Because of the different emphasis on case law between commmon law and civil law jurisdictions[2], the research done using these sets does not easily translate to legal IR systems from civil law jurisdictions.

Information seeking in legal IR can mainly consist of the following three patterns: (1) a single query with no clicks, known as updating behaviour, (2) a literature review where a complex boolean query crafted over several iterations is performed and all documents returned are inspected, or (3) a seeking task spanning days or weeks, consisting of multiple queries interleaved with other tasks. Van der Burg [5] investigated queries from the Legal Intelligence system, the largest legal IR system in the Netherlands, and 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 for Web Search [4]. 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 [5].

Building upon the work of Van der Burg [5], in this paper we discuss the use of queries to define tasks in legal IR. A *task* is

*This paper was written while Gineke Wiggers was conducting a research visit at the University of Queensland.

ADCS '22, December 15–16, 2022, Adelaide, Australia; Author Preprint

ACM ISBN TBA.
https://doi.org/TBA

[1] https://trec.nist.gov/data/legal.html
[2] In common law jurisdictions the law is created by judges through case law. In civil law jurisdictions, law is created by legislative bodies and codified in legal codes (laws), where case law may be used as an interpretative aid [20].

defined by Liao et al [13] as the user interactions in an IR system that relate to one topic. Defined tasks in IR can be used for a variety of purposes, such as recommending sub-tasks [7] or queries, and measuring user satisfaction. A task can consist of multiple queries and query trails. A *query trail* is defined as a query and all behaviour (e.g., filtering or clicks) following that query, ending when a new query is conducted. The end of a task is marked when the user searches for a different topic.

A complex task might take more than one session to complete. *Sessions* have been commonly defined in a heuristic manner as a user interacting with an information retrieval (IR) system with no more than 30 minutes between user actions [8]. Alternatively, a session might contain several interleaved tasks. Jones and Klinkner [9] have demonstrated for web-search that tasks are a better representation of the user experience than sessions for the purpose of evaluation.

In this paper we highlight how the currently employed definition of 30 minute sessions is not adequate in legal information retrieval to extract search queries from a user that are related to the same search task. We also show why tasks are more informative than sessions with regards to legal information retrieval. We further describe the potential of using characteristics such as Levenshtein distance, common words and string matching for automated task classification. Using session information to learn about the different types of search activities that users in legal IR systems perform, gives us more insight into the requirements of these systems, and allows us to move beyond the "total recall" ideal of legal IR and towards the "research reality" described by Geist [6]. This is of particular value for session-based evaluation metrics, such as the session Discounted Cumulative Gain (sDCG) metric proposed by Järvelin et al. [10].

2 DATA-SET CREATION

We took all user queries and corresponding user actions (queries and query trails) from all users affiliated with Leiden University in the Netherlands from the Legal Intelligence system for the academic year 2021/2022 (defined as September 1st 2021 until August 31 2022). We identified 5,027 unique users, issuing a total of 272,877 queries, with a mean of 54 queries and a median of 14 queries per user[3].

The set of users encompasses both students and employees of the university. We expect the dataset to contain both examples of multiple tasks in a day (e.g., a student answering multiple questions/cases in preparation for a class) and examples of one task spanning multiple days or weeks (e.g., research for a PhD thesis, or a legal consultancy case for an academic).

We used this data to create a ground truth of tasks in this data set. Using random sampling, we retrieved all available data from 1 user at a time, and checked whether the data contained more than one query. If it contained more than one query, it was accepted as a sample. In this manner we selected 10 users for further analysis.

For these 10 users we grouped the data by user id, followed by date, and then time, ascending. This allowed us to group the data using the default rule of 30 minutes between actions to create sessions. Using this method we defined 126 sessions, ranging from 1 to 71 sessions per user.

[3]Ranging from 0 to 1,726 with a standard deviation of 117 queries.

We created an overview of all queries conducted by a user, sorted by date/time ascending, but with the date/time and session masked. Masking was done to avoid the annotator basing their labelling on the date/time or session. Using our domain knowledge we labelled all queries based on the perceived task, starting with number 1, and creating a new task label when a change in topic is perceived [13]. When in doubt as to the intention of the query, the search engine result page (SERP) the user was exposed to was reproduced and used as an interpretative aid. If reproducing the SERP also proved inconclusive, the query was marked as inconclusive; 18 queries were marked as such.

The queries were labelled by one annotator. Given the high expertise involved in the creation of the queries on the user side, it is expected that if there are more annotators, each with the legal background to aid in the interpretation of the queries, there would be a fair to moderate agreement between the annotators. It is expected that any difference would be mainly caused by understanding different aspects as belonging to the same case (e.g., being underage and self-defense are both types of defenses for criminal responsibility) that could be resolved through discussion between the annotators.

Tasks can be sequential, but can also be interleaved, so the annotator could return to earlier labels. In this manner we annotated the actions of 10 users into tasks. This resulted in a set of 55 defined tasks, ranging from 1 to 21 tasks per user.

An example of two queries considered to deal with the same topic, and thereby labelled as belonging to the same task is "Wet Werk en Zekerheid" ("Work and Security Act") and "ontslag op staande voet 20 april 2012" ("instant dismissal 20 april 2012"). An example of two queries considered to be different topics, and thereby not in the same task is "noodweer en noodweerexces" ("self-defense and excessive self-defense") and "bestanddeel in vereniging" ("part of an association").

3 RELATIONSHIP BETWEEN SESSIONS AND TASKS

Given the expertise and time investment required to create labelled task data, and the valuable insights that such data could provide for search result diversification or task based evaluation methods, we try to determine which features can be used to automatically classify queries to tasks. We do this by training a Support Vector Machine (SVM) on our labelled data and analysing the weights assigned to the features. SVMs are very suitable for classification tasks like these, where there are two classes (the same task or not the same task). Though the weights could also have been determined using a regression analysis, with the end goal of automatically classifying queries into tasks in mind, an SVM seemed preferable. This has as additional advantage that it allows us to contrast our results with those of Liao et al. [13].

3.1 Analysis Methodology

Inspired by the work of Liao et al. [13] we experimented with the following methods to compare the relationship between two queries:

- time: the difference, in seconds, between the two queries;
- Levenshtein distance: the Levenshtein distance between the two queries;

- Levenshtein distance (2): the Levenshtein distance between the two queries after removing stopwords from the queries;
- common terms: the average rate of common terms between the two queries (the number of common terms divided by the number of terms in the longest query);
- common terms (2): the average rate of common terms between the two queries after removing stopwords from the queries;
- common terms (3): the average rate of common terms between the two queries, where words are considered equal if one contains the string of the other (e.g. 'narrow' and 'narrower' are considered equal);
- common characters: the rate of common characters calculated from the left;
- common characters (2): the rate of common characters calculated from the right;
- common sub-string: the rate of the longest common sub-string;
- contains: whether one query contains the other.

We made pairs of all queries for each user. This led to a total of 139,703 pairs. We calculated the above features for each pair. To calculate the Levenshtein distance, the NLTK package was used [2]. The list of stopwords used was retrieved from https://snowballstem.org/algorithms/dutch/stop.txt. The features were not scaled. A linear[4]SVM was then trained using the scikit-learn package [18]; a SVM classifier was trained with session labels (i.e. a query pair has label 1 if they belong to the same session, and 0 vice-versa) and a separate classifier was trained with task labels. The classifiers were trained so as to then examine the feature weights and compare weights across the two classifiers as an indication of feature importance in the session vs. task data. In doing so, we followed the analysis method by Liao et al. [13].

3.2 Results Analysis

We first analyse common statistics of the data-set we created; these can be seen in Table 1[5]. We note that there is a difference between the mean and median of queries per session and task. This suggests a long-tail distribution, as expected based on the work of Van der Burg [5]. Table 1 further shows that tasks in legal IR usually involve multiple queries, and often (roughly 50% of the tasks) take more than one session. The task with the highest number of sessions involved 35 sessions. The sessions with the highest number of tasks involved 5 tasks (3 occurrences). Though there is overlap between sessions and tasks, the task information provides more insight into the information behaviour of legal professionals, by differentiating between the short known-item or updating tasks and the longer information seeking tasks. This can also be inferred from the larger difference between median and mean number of queries in tasks than in sessions.

[4]We chose a linear SVM to be able to compare our results to those of Liao et al. [13].
[5]When comparing these results to the work of Liao et al. [13] note that Liao et al. grouped sessions based on total session time, whereas this paper defines sessions based on the time between actions.
[6]N.B. when comparing these results to the work of Liao et al. [13] note that Liao et al. grouped sessions based on total session time, whereas this paper defines sessions based on the time between actions.

Table 1: Analysis of data-set statistics per session and per task.

Feature	Value
Mean queries per session	11.22
Median queries per session	9.60
Mean queries per task	13.12
Median queries per task	7.75
Mean tasks per session	1.55
Mean sessions per task	3.44
% of multi-task sessions	35%
% single-query tasks	16%
% of single-query sessions	14%

Table 2: Features weights across the two SVM classifiers we trained on the labelled data (session vs. task). For reference, in the second column, we also report the original weights identified by Liao et al. on their data [13]

Feature	Weight Liao et al.	Weight session	Weight task
Time	-0.1121	-0.0000	-0.0000
Levenshtein distance	0.0106	0.0275	0.0233
Levenshtein distance (2)	-0.1951	-0.0361	-0.0162
6 Common terms	-0.2870	0.4124	0.0237
Common terms (2)	1.2058	-0.4397	0.0512
Common terms (3)	0.5292	0.0150	-0.0035
Common characters	1.6318	0.0489	0.0856
Common characters (2)	0.4014	-0.0053	0.1446
Common sub-string	0.4941	-0.1204	0.3487
Contains	0.6361	0.0246	0.1796

Next we present the analysis of the feature weights obtained when training an SVM on the labelled data. As expected based on the work of Liao et al. [13], Table 2 shows that the query based features have more weight than the time-based feature: this occurs both when we train the classifier for time-based sessions, and when we train it for the tasks. The weights assigned to the features differ remarkably when comparing sessions and tasks given that both SVMs are trained on the same query pairs.

4 CONCLUSIONS

In this paper we set out to explore the difference between tasks and sessions in legal information retrieval. We did this by annotating queries into topically bounded tasks and investigating which weights a SVM classifier would assign to different features.

In this annotated set we found that roughly 50% of tasks (topically defined) take more than one session, and that tasks involve a median of 7.75 queries (and connected query trails). However, we also find that 16% of tasks involve only one query. We find that using a task based query grouping provides more meaningful query groups than the time based session groups, because tasks allow for the diverse range of search tasks performed in legal IR systems, from the "research reality" [6] of updating behaviour (which might involve

multiple tasks in one session) to total recall sessions (which might be visible in the data as one task performed over multiple days/weeks).

Using heuristics like the 30 minutes intervals is still the common practice for session analysis in legal information retrieval. Our initial work highlights instead the value of classifying user actions into tasks rather than sessions. We plan to extend our exploration into the possibility of automated classification of tasks for legal information retrieval. In future work, we also intend to investigate the inter-annotator agreement when multiple legal professionals group queries into tasks.

Query similarity measures appear promising in our exploration of the weights assigned to features, and we would like to explore whether search results analysis, measuring the overlap in results returned for the query, might be an extension of this. Our ultimate aim is to investigate whether task-based evaluation methods might be an improvement over session-based evaluation methods in the context of legal information retrieval.

Acknowledgment. The authors wish to thank the Stichting Recht & Informatica for funding this research and Legal Intelligence for providing the data used in this research.

REFERENCES

[1] Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. Overview of the FIRE 2019 AILA Track: Artificial Intelligence for Legal Assistance. In *FIRE (Working Notes)*. 1–12.

[2] Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python.* O'Reilly Media Inc., California, United States.

[3] A. Bock. 2000. *Gütezeichen als Qualitätsaussage im digitalen Informationsmarkt: dargestellt am Beispiel elektronischer Rechtsdatenbanken.* S. Toeche-Mittler.

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

[5] R.L. Burg van der. 2020. *A query log analysis in the context of Legal Information Retrieval.* Master thesis, Radboud University, Nijmegen, the Netherlands.

[6] A. C. J. Geist. 2016. *Rechtsdatenbanken und Relevanzsortierung.* Doctoral dissertation, uniwien, Austria, Vienna.

[7] A. Hassan Awadallah, R. W. White, P. Pantel, S. T. Dumais, and Y. M. Wang. 2014. Supporting complex search tasks. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management.* 829–838.

[8] Bernard J. Jansen, Amanda Spink, and Vinish Kathuria. 2007. How to Define Searching Sessions on Web Search Engines. In *Advances in Web Mining and Web Usage Analysis*, Olfa Nasraoui, Myra Spiliopoulou, Jaideep Srivastava, Bamshad Mobasher, and Brij Masand (Eds.). Springer, Berlin, Heidelberg, Germany, 92–109.

[9] R. Jones and K. L. Klinkner. 2008. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management.* 699–708.

[10] K. Järvelin, S. L. Price, L. M. Delcambre, and M. L. Nielsen. 2008. Discounted cumulated gain based evaluation of multiple-query IR sessions. In *European Conference on Information Retrieval.* 4–15.

[11] Marios Koniari, Ioannis Anagnostopoulos, and Yannis Vassiliou. 2017. Evaluation of Diversification Techniques for Legal Information Retrieval. *Algorithms* 10, 1 (2017), 22.

[12] Marios Koniari, Ioannis Anagnostopoulos, and Yannis Vassiliou. 2018. Network Analysis in the Legal Domain: A Complex Model for European Union Legal Sources. *Journal of Complex Networks* 6, 2 (2018), 243–268.

[13] Z. Liao, Y. Song, Y. Huang, L. W. He, and Q. He. 2014. Task trail: An effective segmentation of user search behavior. *IEEE Transactions on Knowledge and Data Engineering* 26, 12 (2014), 3090–3102.

[14] D. Locke and G. Zuccon. 2018. 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).* 1261–1264.

[15] Daniel Locke, Guido Zuccon, and Harrisen Scells. 2017. Automatic Query Generation from Legal Texts for Case Law Retrieval. In *Proceedings of the Asia Information Retrieval Symposium.* 181–193.

[16] C.D. Manning, H. Schütze, and P. Raghavan. 2008. *Introduction to information retrieval.* Cambridge university press, Cambridge, United Kingdom.

[17] S.N. Mart. 2017. The Algorithm as a Human Artifact: Implications for Legal [Re]Search. *Law Library Journal* 109 (2017), 387.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

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

[20] Larissa van den Herik, Ewoud Hondius, and Wim Voermans (Eds.). 2022. *Introduction to Dutch Law* (6th ed.). Wolters Kluwer International, Alphen aan den Rijn, The Netherlands.

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