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 that is 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 tasks. This analysis however requires the ability to automatically identify which queries of a user are related to the same search 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 common law and civil law jurisdictions, 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.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

ACDS ’22, December 15–16, 2022, Adelaide, Australia; Author Preprint
© 2022 Copyright held by the owner/author(s).
ACM ISBN TBA
https://doi.org/TBA

CCS CONCEPTS

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

KEYWORDS

Information Retrieval, Tasks, Professional Search, User Query Data

ACM Reference Format:

1 INTRODUCTION

Legal Information Retrieval is a form of professional search 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] 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 common law and civil law jurisdictions, 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

1https://trec.nist.gov/data/legal.html
2In 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].
We identified 5,027 unique users, issuing a total of 272,877 queries. We took all user queries and corresponding user actions (queries with a mean of 54 queries and a median of 14 queries per user (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.

The set of users encompasses both students and employees of the university. We expect the dataset to contain both examples of tasks, ranging from 1 to 21 tasks per user.

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. 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 tasks considered to deal with the same topic, 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;
We first analyse common statistics of the data-set we created; these was used [2]. The list of stopwords used was retrieved from https://snowballstem.org/algorithms/dutch/stop.txt. The features were versa) and a separate classifier was trained with task labels. The task based query grouping provides more meaningful query groups than in sessions.

The data-set statistics per session and per task.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean queries per session</td>
<td>11.22</td>
</tr>
<tr>
<td>Median queries per session</td>
<td>9.60</td>
</tr>
<tr>
<td>Mean queries per task</td>
<td>13.12</td>
</tr>
<tr>
<td>Median queries per task</td>
<td>7.75</td>
</tr>
<tr>
<td>Mean tasks per session</td>
<td>1.55</td>
</tr>
<tr>
<td>Mean sessions per task</td>
<td>3.44</td>
</tr>
<tr>
<td>% of multi-task sessions</td>
<td>35%</td>
</tr>
<tr>
<td>% single-query tasks</td>
<td>16%</td>
</tr>
<tr>
<td>% of single-query sessions</td>
<td>14%</td>
</tr>
</tbody>
</table>

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].

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

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


