



Universiteit
Leiden
The Netherlands

The relevance of impact: bibliometric-enhanced legal information retrieval

Wiggers, G.

Citation

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

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3570499>

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

Chapter 2

Relevance

Exploration of Domain Relevance by Legal Professionals in Information Retrieval Systems

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

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

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

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

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

2.1 Introduction

Relevance, in the broadest sense, is a term used to describe “Connection with the subject or point at issue; relation to the matter in hand.” [2]. In everyday language, it is used to describe the effectiveness of information in a given context [110, p. 203]. In information retrieval (IR), the theory of

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

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

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

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

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

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

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

The First International Workshop on Professional Search² describes professional search as: “professional search takes place in the work context, by specialists, and using specialist sources, often with controlled vocabularies.” [132] This definition covers people from multiple domains, including librarians, scholars, lawyers, and other knowledge work professions. In the context of users of Legal IR systems this includes professions such as lawyers, judges, government employees, legal information specialists and legal scholars. Experts in IR rather than law, such as professional support lawyers and information specialists, may, by nature of their role, have a different perspective of relevance. A professional support lawyer or information specialist retrieves information for a third person rather than themselves and will likely focus on completeness of the information, not being aware of the cognitive state of the client. A legal professional however, searching for themselves will likely focus only on information that is new to them. From this different perspective, they may have a different perception of relevance [109, p. 341].

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

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

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

We conducted the study with users of the Legal Intelligence³ system, following Park, who argued that it is important to test with real users of the information retrieval system [96, p. 322]. We address the following research questions: (1) To what extent can we demonstrate the existence and factors of domain relevance in the context of judgment of search results (document representation) in legal IR systems? (2) To what extent do legal information retrieval specialists, legal scholars and legal practitioners show agreement on relevance factors outside of a task context?

In answering these questions, this paper’s contributions compared to previous work are: (1) we conducted a user study with professional users of a legal IR system to research the validity of the concept of domain relevance as described by Van Opijnen and Santos [131] in practice and its possible applications in legal IR systems; (2) using a statistical analysis we demonstrate that sub-groups of users of legal IR systems (legal information specialists, legal scholars and legal professionals) show agreement on relevance factors outside of a task context, when judging search results (document representations) in legal IR.

2.2 Background

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

³<https://www.legalintelligence.com>

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

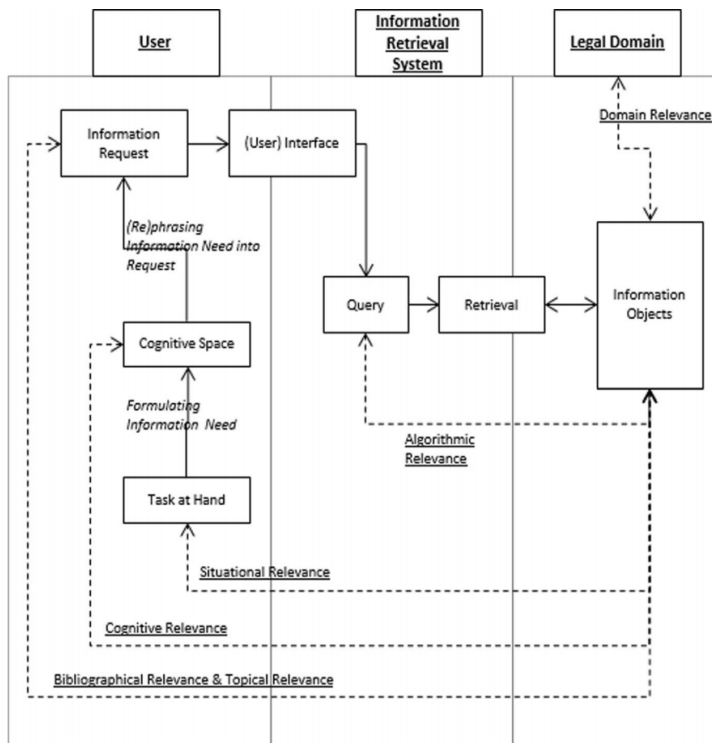


Figure 2.1: Relevance schema of [131]

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

the document, not on the user.

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

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

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

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

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

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

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

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

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

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

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

Previous research on factors of relevance in the context of professional search has been done by, amongst others, Cuadra and Katter [39] and Rees and Schultz [103], who examined judgements by expert reviewers. Barry [12] expands on this research by inviting users (rather than expert reviewers) to submit a request for unknown or unfamiliar information⁵, for which she retrieved documents. The results lists were presented to the participants, who were asked to review whether they would or would not pursue the documents contained in the list. The study was done using an open-ended interview technique, to generate a complete overview of relevance factors. Barry identified 23 categories of relevance factors, grouped into seven classes [12]:

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

⁵Thereby excluding known item retrieval.

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

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

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

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

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

2.3 Methods

2.3.1 Study design

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

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

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

2.3.2 Participants

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

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

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

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

some examples of factors were given⁶. To avoid leading the respondents, and to encourage respondents to consider all aspects from both results, these examples were not repeated alongside the questions.

2.3.3 Selection of stimuli

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

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

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



The screenshot shows a red rectangular box containing a search interface. At the top is a white search bar with the text 'tarieven deskundigenonderzoek' and a magnifying glass icon to its right. Below the search bar, the results are organized into two sections, 'Optie 1' and 'Optie 2', each with a bold title and a descriptive paragraph.

Optie 1

RSV 1998, 85 - Centrale Raad van Beroep - 05-03-1997
 Vergoeding deskundigenonderzoek en reiskosten

Optie 2

**Rechtspraak.nl - Centrale Raad van Beroep - 08-09-2017 -
 ECLI:NL:CRVB:2017:3091 - Arbeids/Sociaal Recht**
 Beëindiging WIA-uitkering. Onvoldoende medische grondslag. De door de Raad geraadpleegde deskundige heeft in het rapport gemotiveerd uiteengezet dat appellante meer beperkt is dan door het Uvw is aangenomen, niet alleen bij persoonlijk en sociaal fun [...]

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

Selection of search results

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

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

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

Relevance factors

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

- Recency [12, p. 156]: it has been suggested that recent case law is more relevant than older case law (<2 years; 2 – 10 years; >10 years old), though recency can also be related to the specific period the case played in [131, p. 80], in which case it would be situational relevance; Schamber’s time frame factor [13];
- Legal hierarchy/importance [131, p. 68]: case law from higher courts carries more weight than case law from lower courts (supreme court; courts of appeal; courts of first instance);

- Presence of annotation⁷: annotated case law (providing context for the case) is more relevant than case law that is not annotated. Related to Schamber's factor of summary/interpretation [13];
- Source authority⁸ [12, p. 156]: sources that are considered authoritative are preferred over other sources (government documents, leading publications; mid-range publications; blogs);
- Author authority⁹ [12, p. 155-156]: documents written by well-known authors are considered more authoritative than other documents;
- Bibliographical relevance [131, p. 71]: the official version (case law or law) is more relevant than reprints;
- Title relevance: results with search term in the title or summary are considered more relevant than results with the search term not in the title/summary (the visibility of algorithmic and topical relevance for respondents);
- Document type¹⁰: document types that pertain to the perceived information need are considered more relevant than other document types (depending on perceived information need expressed in the query as interpreted by respondent). Related to Schamber's presentation quality, especially the underlying factors of presentation preference and choice of format [13].

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

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

⁷Mentioned by users to Legal Intelligence employees.

⁸Also described as source quality.

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

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

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

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

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

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

2.3.4 Extracting and mapping relevance factors

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

Examples of the mapping include:

- Recency: ‘Newer’ or ‘Appears out of date’;

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

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

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

¹¹translated from: ‘optie 1 omdat dit hogere rechtspraak is. Bij optie 2 is de noot interessant.’

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

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

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

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

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

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

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

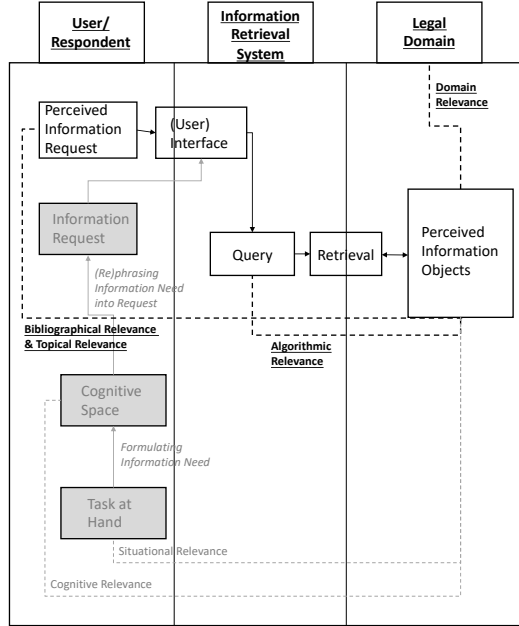


Figure 2.3: An adaptation to the relevance schema of Van Opijnen and Santos [131] to reflect the method using a perceived information request and the lack of cognitive and situational relevance.

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

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

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

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

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

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

2.3.5 Statistical analysis

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

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

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

2.4 Results

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

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

¹⁴‘Regelgeving, bijvoorbeeld een Staatscourant.’ translated: ‘Regulation, for example the Government Gazette’

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

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

Table 2.1: Breakdown of respondents

relevance can be precluded, as discussed in Section 2.3.4.

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

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

2.4.1 Relevance factors

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

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

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

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

Title Relevance

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

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

Document Type

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

Recency

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

Level of Depth

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

Legal Hierarchy

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

Law Area (Topic)

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

Authority/Credibility

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

Usability

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

¹⁶Translated: “The second option because it is civil law”, “Civil, not fiscal”

Bibliographical Relevance

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

Annotated

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

Document Length

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

2.4.2 Differences between user sub-groups

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

¹⁷Translated: ‘reports are often very long’, original: ‘zijn rapporten vaak erg lang’

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

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

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

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

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

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

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

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

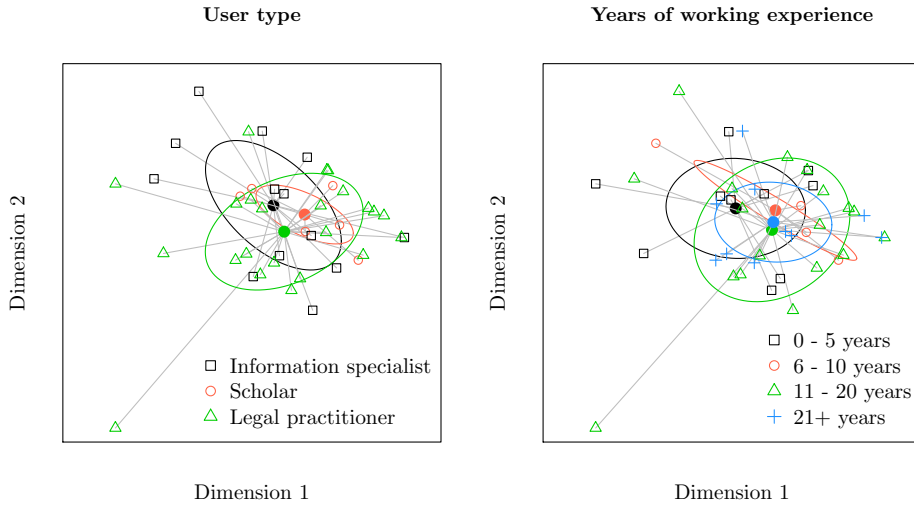


Figure 2.4: The dissimilarities between respondents visualized in two dimensions using principle coordinates analysis (PCoA). Left: Labeled as legal information specialist, legal scholar or legal practitioner. Right: Labeled by years of working experience. The connected dot in the center of each group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

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

Table 2.4: PERMANOVA results.

2.5 Discussion

2.5.1 Implications

Relevance factors

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

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

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

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

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

User subgroups

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

2.5.2 Limitations of the study

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

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

covered.

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

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

2.6 Conclusions

With regards to research question 1: to what extent can we demonstrate the existence and factors of domain relevance in the context of judgment of search results (document representation) in legal IR systems?

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

The results confirm the existence of domain relevance as described in

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

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

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

2.7 Acknowledgements

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

