

On the Concept of Relevance in Legal Information Retrieval

Marc van Opijnen¹ and Cristiana Santos²

Abstract. The concept of ‘relevance’ is crucial to legal information retrieval, but because of its intuitive understanding it goes undefined too easily and unexplored too often. We discuss a conceptual framework on relevance within legal information retrieval, based on a typology of five relevance dimensions used within general information retrieval science, but tailored to the specific features of legal information. We come forward with several suggestions to improve the design and performance of legal information retrieval systems.

1 INTRODUCTION

Legal Information Retrieval (LIR) has always been a research topic within Artificial Intelligence & Law (‘AI & Law’): in ‘A History of AI & Law in 50 papers’ [1] seven of those 50 papers have a relation to LIR. For the legal user though much research seems to be only remotely relevant for solving their daily problems in information seeking. The underrepresentation of legal practitioners within the AI & Law community might offer an explanation: “A lawyer has always the huge text body and his degree of mastery of a special topic in mind. For a computer scientist, a high-level formalisation with many ways of using and reformulating it is the aim.”³ Not surprisingly, LIR has been approached within AI & Law primarily with a focus on conceptualization of legal information, while for daily legal work that might not always be the most effective approach.

Meanwhile, due to the advancements of the information era and the Open Data movement the number of legal documents published online is growing exponentially, but accessibility and searchability have not kept pace with this growth rate. Poorly written or relatively unimportant court decisions are available at the click of the mouse, exposing the comforting myth that all results with the same juristic status are equal. An overload of information (particularly if of low-quality) carries the risk of undermining knowledge acquisition possibilities and even access to justice.

Apart from the problems with the quantities, also the qualitative complexities of legal search cannot easily be underestimated. Legal work is an intertwined combination of research, drafting, negotiation and argumentation. To limit the role of LIR within daily legal practice to just finding the court decisions relevant to the case at hand underestimates the complexities of the law and legal research. Any legal information retrieval system built without

sufficient knowledge, not just of the actual legal information needs but also of the ‘juristic mind’, is apt to fail.

To aid researchers and system designers in designing or developing LIR applications it might be an interesting exercise to approach LIR more explicitly as a subtype of Information Retrieval (IR) instead of (merely) a topic within AI & Law. Since ‘relevance’ is the basic notion in IR, it could be a useful starting point for analysing the specificities of LIR. In this paper we develop a conceptual framework and come forward with suggestions for improvements in LIR.

In section 2 we define ‘Legal Information Retrieval’ by, on the one hand, distinguishing it from Legal Expert Systems and, on the other hand, describing the characteristics that justify its classification as a specific subtype of IR. In section 3 we discuss the concept of relevance in LIR, guided by a topology of five different ‘dimensions’ of relevance. In section 4 we will draw some conclusions and make suggestions for future work.

2 LEGAL INFORMATION RETRIEVAL

2.1 Inference Versus Querying

In a variety of ways information technology is working its way into the legal domain and even endangering the livelihood of its inhabitants.[2] Out of all these different systems we highlight two types of information systems: Legal Expert Systems (LES) and Legal Information Retrieval (LIR), on the hand with a view to articulate the particularities of LIR systems and on the other hand to underline the need – at least for many years to come – of LIR for the legal profession. The main aspects of LES and LIR are listed in table 1.

In research interesting cross-fertilisation experiments started a long way back [3] and many of the recent developments within the legal semantic web (as summarized in e.g. [4]) are also of importance for LIR, but it is highly unlikely that the two types will completely merge. LIR starts where LES isn’t able to provide an answer. And notwithstanding the improvements AI & Law brings to LES, there will always be questions left and relevant documents to be discovered, since the lack of any final scheme is inherent to the legal domain.

¹ Publications Office of the Netherlands (UBR|KOOP), email: marc.opijnen@koop.overheid.nl.

² University of Barcelona (IDT-UAB) and University of Luxembourg, email: cristiana.teixeirasantos@gmail.com

³ E. Schweighofer in [1, par. 2.4]

Table 1. A comparison between Legal Expert Systems (LES) and Legal Information Retrieval (LIR).

Aspect	LES	LIR
Goal	Establish a legal position on specific case	Provide relevant legal information
Input	Facts	Request
Content	Legal rules encoding the domain expertise	Documents
Method	Inference	Querying
Output	Decision, advice, forecast.	Set of documents
Preferred use	Answering 'happy flow' questions within a specific and limited domain	Finding information objects in huge repositories
Advantage	Can provide straightforward answers	Unlimited content, input and output
Drawback	What has not been modelled, cannot be answered	User always has to read, interpret and decide for himself
Basic notion	Uncertainty	Relevance

2.2 Characteristics of Legal Information

A variety of specific features justify – and compel – the positioning of Legal Information Retrieval as a specific subtype of Information Retrieval [5]. On describing these features we will briefly elucidate some shortcomings of general IR in meeting the needs arising from the legal domain.

1. *Volume.* Although in the age of 'big data' the longstanding impressive volumes of legal materials have been surpassed by e.g. telecommunications and social media data, viewed upon from an information retrieval perspective the volume of legal materials is still impressive. This holds true for public repositories (like case law databases) as well as for private repositories (e.g. case files within law firms or courts).
2. *Document size.* Compared to other domains, legal documents tend to be quite long. Although metadata and summaries are often added, access to (and searchability of) the full documents is of paramount importance.
3. *Structure.* Legal documents have very specific (internal) structures, which often also are of substantive relevance. Although standards for structuring legal documents are emerging [6], many legal documents do not have any (computer readable) structure at all.
4. *Heterogeneity of document types.* In the legal sphere a variety of document types exist which are hardly seen in other domains. Apart from the obvious legislation and court decisions, one can think of Parliamentary documents, contracts, loose-leaf commentaries, case-law notes a.s.o.
5. *Self-contained documents.* Contrary to many other domains, documents in the legal domain are not just 'about' the domain, they actually contain the domain itself and hence they have specific authority, depending on the type of document. A statute is not merely a description of what the law is, it constitutes the law itself [5]. Notwithstanding the notion that in a bibliographical sense a document is only a manifestation of an abstract work [7], for information retrieval purposes the object to be retrieved embodies the object itself.
6. *Legal hierarchy.* The legal domain itself defines a hierarchical organization with regard to the type of documents and its authority. Formal hierarchies depend on the specific jurisdiction or domain, and factual hierarchies often also depend on interpretation, e.g. the general rule *lex specialis derogat legi generalis* requires a decision on its applicability in a specific situation.
7. *Temporal aspects.* Within the incessant flow of legislative processes, legislative texts and amendments follow one another and may overlap. Recurrent challenges stem from tracing the history of a specific legal document by searching the temporal axis of its force and efficacy [8] and by retrieving the applicable law in respect to the timeframes covered by the events subject to regulation [9].
8. *Importance of citations.* In most other scientific domains citation indexes exist for academic papers. In the legal domain, citations are a more integral part of text and argumentation: "*Legal communication has two principal components: words and citations*" [10, p. 1453]. Citations can be internal (cross-references), linking one normative provision to another normative provision in the same document or normative provisions to recitals [11]. Citations can also be external, linking e.g. a court decision to a normative provision, a normative document to another normative document, or an academic work to a Parliamentary report. Citations can be explicit or implicit and they can express a whole variety of different relationships: they can be instrumental (or 'formal') – e.g. a court of appeal referring to the appealed first instance decision – or of a purely substantive nature, but having distinct intensions.
9. *Legal terminology.* Legal terminology has a rich and very specific vocabulary, characterized by ambiguity, polysemy and definitions that are very precise and vague at the same time.
10. *Audience.* Legal information is queried by a wide variety of audiences. Laymen with different levels of legal knowledge and jurists with completely different professions (e.g. scholars, judges, lawyers, notaries or legal aid workers) have completely different information retrieval needs.
11. *Personal data.* Many legal documents contain personal data. Apart from the consequences for the publication of e.g. court decisions, it also weighs on LIR, since the juristic memory is often built on names of persons and places.
12. *Multilingualism and multi-jurisdictionality.* In many (scientific) domains English is the pivotal language, and in the legal domain the same goes for common law jurisdictions. Civil law jurisdictions though have a variety of languages; language and jurisdiction have such a strong relationship that translated documents can only be a derivative of the original. As a result European or international information retrieval poses its own problems.

Strongly related to these specific characteristics of legal information, 'legal search' differs substantially from 'non legal search' [12, 13], e.g. with regard to history tracking.

3 RELEVANCE WITHIN LEGAL SEARCH

3.1 Nature of Relevance in LIR

The science of Information Retrieval is basically about ‘Relevance’: how to retrieve the most relevant documents from – in principle – an unlimited set? Before any methodology or system for retrieval can be developed or discussed, the concept of ‘relevance’ has to be examined. This seems to be a trivial undertaking since this concept has a tendency to be immediately understood by everybody. A thorough understanding though is of the utmost importance for the effectiveness of LIR systems, and hence it needs continuous consideration. The foundations of a conceptual framework can be adopted from general IR science. Saracevic defined ‘relevance’ as: ‘*pertaining to the matter at hand*’ [14], or, more extended: ‘*As a cognitive notion relevance involves an interactive, dynamic establishment of a relation by inference, with intentions toward a context.*’ From this definition it follows that relevance has a contextual dependency since it is measured in comparison to the ‘matter at hand’. From this definition it also follows that relevance is a comparative concept or a measure (irrelevant, weakly relevant, very relevant), which by using a specific threshold can be turned into a binary value and hence a property (relevant or not). Because of its dynamic establishment relevance may change over time and it involves some kind of selection.[15]

3.2 Dimensions of Relevance in LIR

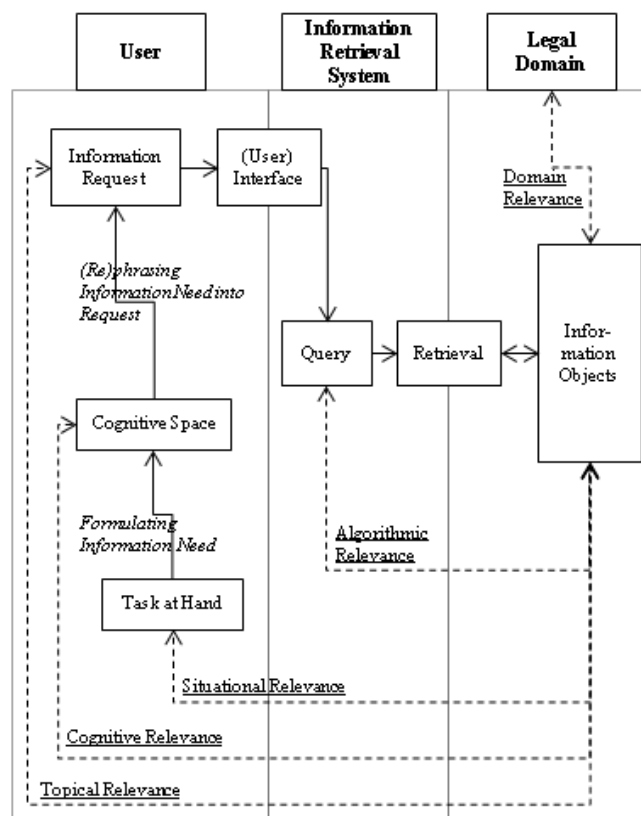
To understand the concept of relevance it is important to disambiguate the various ‘manifestations’ as understood by Saracevic [15], although we prefer to use Cosijn’s label ‘dimensions’ [16]. We discuss them here in brief, and elaborate them in the following sub-sections. The role of these dimensions in the interplay between user, information retrieval system and legal domain is depicted in figure 1.⁴

1. *Algorithmic or system relevance.* A computational relationship between a query and information objects, based on matching or a similarity between them. Traditionally models have been described within the context of full-text search, e.g. being Boolean, probabilistic, vector-space a.s.o. Natural language processing is also perceived to be within algorithmic relevance, although in our view it covers also those processes which do not take place during the actual querying, but are intended to improve algorithmic relevance at a later stage. Examples are pre-processing of documents, automatic classification a.s.o. Unlike all other relevance dimensions that can be observed and assessed without a computer, algorithmic relevance cannot: it is system-dependent.
2. *Topical relevance.* The relationship between the ‘topic’ (concept, subject) of a query and the information objects retrieved about that topic. A topicality relation is assumed to be an objective property, independent of any particular user. ‘Aboutness’ is the traditional distinctive criterion. The topics of the information objects might be hand-coded or computed, e.g. by classification algorithms. The self-containment feature

of legal information adds ‘isness’ – as a sibling of ‘aboutness’ – to topical relevance in LIR.

3. *Cognitive relevance or pertinence.* Concerns the relation between the information needs of a user and the information objects. Unlike algorithmic and topical relevance, cognitive relevance is user-dependent, with criteria like informativeness, preferences, correspondence and novelty as measuring elements.

Figure 1. Interplay between user, information retrieval system and legal domain.



4. *Situational relevance or utility.* Defined as the relationship between the problem or task of the user and the information objects in the system. Also this dimension of relevance is dependent on the specific user, but unlike the cognitive relevance it does not focus on the request as formulated, but on the underlying motivation for starting the information retrieval process. Inferred criteria for situational relevance are the usefulness for decision-making, appropriateness in problem solving and reduction of uncertainty.
5. *Domain relevance.* As the fifth dimension Saracevic [14] used ‘Motivational or affective relevance’, but in a critical assessment Cosijn and Ingwersen [16] replaced this dimension by ‘socio-cognitive relevance’, which “[I]s 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.” Given the specific features of legal information as well as for reasons of modelling, we define this dimension as the relevance of information objects within the legal domain itself

⁴ Inspired by [17].

(and hence not to ‘work task or problem at hand’). For convenience we label it ‘domain relevance’.

Already here it should be observed that relevance dimensions easily overlap and intermingle: *“The effectiveness of IR depends on the effectiveness of the interplay and adaptation of various relevance manifestations, organized in a system of relevancies.”*[14, p. 216] In the design of IR systems it is hence of the utmost importance to distinguish between its various dimensions and to pay specific attention to each of them, in the user interface, the retrieval engine and the document collection. It will definitely improve the user’s perception of the system’s performance on retrieving the most relevant information. This perception – or ‘criterion for success’ – depends on the relevance dimension(s) invoked. These criteria are, together with the nature of the respective dimensions, summarized in table 2.

Table 2. Dimensions of Relevance Compared

Dimension	Describes a relation between	Criterion for ‘success’
Algorithmic relevance	Query and information objects	Comparative effectiveness in inferring relevance
Topical relevance	Topic or bibliographical object expressed in the request and topic or bibliographical object covered by information objects	Isness / aboutness
Cognitive relevance	Information needs of the user and information objects	Cognitive correspondence, novelty, information quality, authoritativeness, informativeness
Situational relevance	Situation / task at hand	Usefulness in decision-making and problem-solving
Domain relevance	Opinion of the legal crowd and information objects	Legal importance / authority

In the following subsections we will elaborate the five relevance dimensions in LIR and discuss how these dimensions may help to classify past and current spectrum of approaches and how it might help bridging the conceptual gap between lawyers and informaticians.

3.2.1 Algorithmic Relevance

Algorithmic relevance concerns the computational core of information retrieval. As expressed in figure 1 it is the relation between the information objects and the query; this ‘query’ is to be understood as the computer processable translation of the request as entered in the user interface or any other intermediary component. Algorithmic relevance is about the capability of the engine to retrieve a given set of information objects (the ‘gold standard’) that should be retrieved with a given query (measured in ‘recall’) with a minimum of false positives (measured in ‘precision’).

From our conceptual perspective the type of query as well as the type of retrieval framework is not relevant, but given the legal

information features of volume, document size and lack of structure, textual search has for long had the focus. In the early days Boolean search was the core of any legal retrieval system, and it is still an indispensable element in most LIR systems today. In a Boolean system both the user request and the documents are regarded as a set of terms, and the system will return documents where the terms in the query are present. Boolean searches often result in the retrieval of a large number of documents. In addition, they provide little or no assistance to the user in formulating or refining a query and they lack domain expertise that could improve the search outcome. Relevance performance was improved by using models as the *vector space model* [18] and *TF-IDF* (term frequency – inverse document frequency). Nevertheless, recall is often below acceptable levels because the design of full-text retrieval systems: *“(I)s based on the assumption that it is a simple matter for users to foresee the exact words and phrases that will be used in the documents they will find useful, and only in those documents.”* [19]. Ambiguity, synonymy and complexity of legal expressions contribute substantially to this problem.[20] Natural language processing (NLP) is gaining popularity as an addition to or alternative to pure text-based search.[21]

Apart from text-based search also other types of calculations can be considered within ‘algorithmic relevance’, like the use of ontologies as higher level knowledge models [4, 22] as well as network statistics, especially when used for citation analysis [23, 24].

3.2.2 Topical Relevance

Topical relevance is about the relevancy relation between the topic as formulated in the user request and the topic of the information objects. But before we can discuss this ‘traditional’ topicality within LIR, attention should be drawn to an often overlooked feature of legal information that is of crucial importance for topical relevance: its self-containment. A classic car database contains documents *about* classic cars, not the cars themselves, while a legislation database does contain the legislative texts themselves. Because the same repository might also contain other acts or court decisions citing it or scholarly writings discussing it, we add ‘isness’ to ‘aboutness’ as a separate concept within topical relevance. We will discuss both concepts below.

Isness

In general two types of searching can be distinguished: searching the known and searching the unknown. Searching the known in LIR concerns ‘isness’: finding a specific law, court case, Parliamentary document or scholarly article, generally by keying in some kind of identifier (e.g. a title or a code). Although this might look like a problem of ‘data retrieval’ instead of ‘information retrieval’ [25, par. 1.1.1] and hence a no-brainer [26], in most legal information systems it is still a real brainteaser. A first reason for this is that ‘isness’ is too easily confused with ‘an exact match’ or a specific document while, on the contrary, a whole set of different documents can be correctly retrieved by an isness request: an initial act as published in an Official Journal, as well as a series of consolidated versions, all in different language expressions and/or in different formats. A second reason is that lawyers often refer to the work level, [7] while the search engine is not clever enough to relate the work level to the actual information objects. A third reason is the improper or incomplete pre-parsing of the user

request, resulting in interpreting the request as text query instead of understanding it as an identifier for a (series of) information objects.

This can e.g. be illustrated in EUR-Lex: using the quick search field for searching by document number ('*Regulation (EEC) No 1408/71*'), often used formatting variants ('*Reg. 1408/71*') or aliases ('*Services directive*', '*Dublin Regulation*') does not render the documents which are identified by these labels on top of the result list, and when using the advanced search – where one has to split the document number into a 'year' part and a 'number' part – the non-specialist user is probably puzzled where to put which digits for 'Directive 96/95/EC', 'Regulation 98/2014' or 'Regulation 2015/2016'.⁵

To improve topical relevance it is important to understand that a user of legal information retrieval systems generally prefers – if possible – isness over aboutness. To achieve such improvement isness should always rank higher in a result list than aboutness, or even better: be labelled as such. Secondly, the capabilities of the system should be improved as to recognize requests that refer to isness, including all the many ways in which isness can be expressed, such as: different types of identifiers for the same thing, many different formatting styles even for one type of identifier, the use of 'global aliases' like 'Bolkestein Directive' or 'General Data Protection Regulation'. Reference parsers have been developed for detecting citations in the documents themselves [27] (see below, section 3.2.5), but can also be used for parsing user requests.

Aboutness

While 'isness' relates to searching the known, 'aboutness' relates to searching the unknown: one is not searching for a specific document (or work), but for information or knowledge about something.

Using free text search and mapping the searched terms to the terms indexed from the information objects renders poor results since legal concepts can be expressed in a variety of ways, while completely different concepts can textually be quite similar.

Adding head notes and keywords from taxonomies or thesauri has been a long tradition within the legal information industry. Although aboutness is assumed to be an objective property, independent of any particular user, manual indexing is inherently subjective, and even the same indexer may sort the same document under different terms depending on which context the document is presented in [28]. "*Manual indexing is only as good as the ability of the indexer to anticipate questions to which the indexed document might be found relevant. It is limited by the quality of its thesaurus. It is necessarily pre-coordinated and is thus also limited in its depth. Finally, like any human enterprise, it is not always done as well as it might be.*" [20, p. 14] Semi-automated classification using ontologies [29] is gaining popularity, but automatic classification turns out not to perform better than human indexing. [30] For huge public databases manual tagging is hardly an option. And a general drawback of such systems is the mandatory use in the user interface of the classification scheme. This forces the user to limit or to reformulate his request to align it with the available classification system. A problem that can only be

solved by the time-consuming and tedious task of "*Using a combination of automated and manual techniques, [constructing] a list of concepts and variations for expressing a concept.*"⁶ This requires in-depth legal knowledge, analysis of search engine log files and continuous maintenance.

Search in multilingual legal repositories – e.g. in the ECLI Search Engine on the European e-Justice portal⁷ – poses specific problems: the terms used in the request do not only have to be translated into the language of the information objects, but also into the specific legal terminology of the jurisdiction the information objects are about. Sufficient solutions have not yet been developed. EuroVoc⁸ is a large multilingual vocabulary; although it is used for tagging in the EUR-Lex database, it is too much policy-oriented and too less legal to be of practical use for LIR. Aligning legal vocabularies of different legal systems and/or languages has proven to be quite difficult [31]. Within the Legivoc project various national legal vocabularies have been mapped [32], but it needs more elaboration to be of practical use.

Meanwhile, developers of LIR systems should consider whether the investment is worth the effort: surveys have shown that classification systems are not very popular among users [33], contrary to searches by relationship [34]. Many topics in law, at least in the juristic mindset and information seeking behaviour, have a strong connection to other legal documents. Typical requests may refer to a search for (everything) about a specific paragraph of law or court decision. In such requests these information objects represent a specific legal concept, but the only reason lawyers rephrase it might be related to the fact that the search engine cannot cope with their actual request. For well-known acts and codes such aboutness information is structured in treatises or loose-leaf encyclopaedias, but they are optimized for browsing, not for search. Since such works do not cover the whole legal domain, performing searches on citations might in principle be the obvious choice.

In common law countries citators are very popular for such 'topical citation search', like LawCite.org in the public domain and Shepard's in the private domain. The latter is based on manual tagging and also contains qualifications of these relations. In continental Europe the importance of search by citation – as a type of aboutness – needs more attention from search providers. In EUR-Lex, HUDOC and various national legislative databases, relations between documents are tagged and searchable/browsable, but especially in national case law databases citation search is extremely difficult. A first reason is that judges have lousy citations habits: research showed that only 36% of cited EU acts was in conformity with the prescribed citation style, the other citations were made with a wide range of other styles [35]. Comparable problems appear when searching for case law citations, where additional complexity is added by the fact that one decision can be cited by many different identifiers. [36] Also, slashes, commas and hyphens are essential elements of legal identifiers, but are out-of-the-box interpreted by search engines as specific search instructions (like 'near' for '/' or 'not' for '-'). Manual tagging for large scale public databases is undoable, so reference parsers have to be developed [27]; as indicated in section

⁵ The year is 96, 2014 and 2015 respectively. If the citation the user has at its disposal is correct and if he is knowledgeable about EU document numbering he can solve the problem, but often citations are incomplete or poorly formatted [27]. In a directive the year comes first, in a regulation the number comes first. But from 1 January 2015 onwards the year comes first in all acts <eur-lex.europa.eu/content/tools/elaw/OA0614022END.pdf>.

⁶ P. Zhang, key-note speech on ICAIL 2015 Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts. <www.lrdc.pitt.edu/ashley/icail2015nlp/>.

⁷ <https://e-justice.europa.eu/content_ecli_search_engine-430-en.do>.

⁸ <eurovoc.europa.eu>.

3.2.1 they can be used both for recognizing the citations in the information objects as well as in the request.

3.2.3 Cognitive Relevance

Cognitive relevance concerns the extent to which an information object matches the cognitive information needs of the user. Unlike algorithmic and topical relevance, this dimension is of a subjective nature: do the retrieved documents fit to the user's state of knowledge? Is he offered the temporal version that matches his information need? Are there any metadata regarding the information objects retrieved he should be aware of?

Since this dimension is of subjective nature, the cognitive relevance performance of a LIR system depends, broadly speaking, on the continuum between on the one hand, system features to tailor the search experience to personal needs, and on the other hand, the ability of the system to explicitly or implicitly understand the information needs of each individual user. An example of the former is time travelling: jurists often need to know the delta between the temporal version T of an act and version T+1. Up until recently many legislative databases were only able to serve version T and T+1 in parallel, without actually showing the delta. By offering such functionality,⁹ a user is served in his personal information needs, although the information retrieved is the same for all users.

On the other end of the continuum we find systems for personalized search, using filters to recommend information objects that are deemed relevant for a specific user at a specific stage in his information collecting process. Within such 'recommender systems' two types of filtering can be distinguished. 'Collaborative filtering' recommends documents by making use of the user's past search behaviour and/or that of a peer group. 'Content-based filtering' uses shared features of the document at hand and other documents, based on e.g. topical resemblance, having comparable metadata or closeness in a citation network. Recommender mechanisms can be used to limit the number of documents retrieved (e.g. because the system knows user A is only interested in tax law and not in criminal law) or to increase the number of documents: by offering 'more like this' buttons or navigable citation graphs users can be supported in serendipitous information discovery.[37] Being tailored to the individual need of the user, recommender system can also be used for pro-active search: notification systems informing a user about information objects that have been added to the repository and might be of interest for him. Within legal information retrieval recommender systems have not had too much attention yet. [38]

3.2.4 Situational Relevance

While cognitive relevance is associated with search task execution, situational relevance pertains to work task execution; the relevance of documents is measured by their usefulness for the task at hand, e.g. decision-making or problem-solving.[17] It should be noted that the system is not asked to solve the problem itself – then it would be a legal expert system, not a LIR system.

Situational relevance in legal information retrieval comes close to – but should not be confused with – 'legal relevance', which usually means that information is relevant to a proposition when it

affects, positively or negatively, the probability that the proposition is true [39, p. 148].¹⁰

The difference between 'legal relevance' and situational relevance can be understood with the help of the following definition by Jon Bing:

A legal source is relevant if:

1. *The argument of the user would have been different if the user did not have any knowledge of the source, i.e. at least one argument must be derived from the source; or*
2. *legal meta-norms require that the user considers whether the source belongs to category (1); or*
3. *the user himself deems it appropriate to consider whether the source belongs to category (1).[41]*

In this definition (1) pertains to the strict notion of 'legal relevance', while situational relevance in legal information retrieval also covers (2) and (3).

Probably because of the relative importance of case law in the United States and other common law countries, much LIR research has concentrated on finding the (most) relevant court decisions relating to a case at hand. This can be pursued using a variety of (sometimes combined) technologies, like argumentation mining [42] and natural language processing (NLP) [21].

Navigational features of LIR systems, like memorized search history, storage of relevant documents found, shared folders and customization features do not pertain to situational relevance in an IR sense, unless these data are used for collaborative or content-based recommendations that pertain to the dossier at hand.

3.2.5 Domain Relevance

We defined 'domain relevance' as the relevance of information objects within the legal domain itself. It is independent from an information system and independent from any user request. As can be understood from the previous paragraph we prefer to avoid the term 'legal relevance', but 'legal authority' or 'legal importance' are safe to use as synonyms for 'domain relevance'.

Domain relevance can be applied in LIR systems in different ways. First, it can be used to classify categories of information objects as to their legal weight: a constitution outweighs an ordinary act, which in turns is of more importance than a by-law or ministerial degree. In the same way an opinion of a supreme court can be expected to have more authority than a district court verdict, but it can be superseded by a judgment of the European Court of Human Rights.

Secondly, the concept of domain relevance can be used to classify individual information objects as to their legal authority. Separating the wheat from the chaff has for long been the territory of domain experts: since publication / storage was expensive, and adding documents itself labour-intensive, a selection was made on the input side of any paper or early digital repository. The ease with which information can now be published on the internet has shifted the issue of selection – at least partially – from the input side to the output side: 'selection' has evolved from a publisher's issue into a search issue. Case law publication in the Netherlands could serve as an illustration: the public case law database in the Netherlands¹¹ contains a small percentage (less than 1%) of decided cases, but in fifteen years has accumulated 370.000

⁹ E.g. <wetten.overheid.nl/BWBR0006368/2016-01-01?VergelijkMet=BWBR0006368%3fg%3d2010-02-01>.

¹⁰ Next to this 'logical' or 'probabilistic' definition often also a 'practical' concept is used, meaning 'worth hearing'. [40]

¹¹ <uitspraken.rechtspraak.nl>.

documents. More than 75% of those are not considered important enough to be published in legal magazines.[43]

An example of domain relevance applied at the document level can be observed in the HUDOC database, containing all case law documents produced by the European Court of Human Rights. To aid the user in filtering the nearly 57.000 documents as to their legal authority, four importance levels have been introduced. Except for the highest category, containing all judgments published in the Court Reports, all documents have been tagged manually. Since this importance level is an attribute of each individual document, it can easily be used in combination with other relevance dimensions.

Since manual tagging is labour-intensive, for more massive repositories a computer-aided rating is indispensable. Given the abundant use of citations between court decisions, network analysis is an obvious methodology to assess case law authority [23, 44]. In the 'Model for Automated Rating of Case law' [24] the 'legal crowd' – the domain specialists that rate individual court decisions as to their authority by citing them or not – is extended to legal scholars, while it also uses other variables within regression analysis to predict the odds of a decision rendered today for being cited in the future. It also takes into account changing perceptions over time (see e.g. also [45]). If court decisions are well-structured and citations are made to the paragraph level, importance can be calculated for the sub-document level as well [46]. Comparable techniques can be used for the relevance classification of legislative documents [47] or for a network containing different types of sources [48].

Network analysis is supported by the use of common identifiers, like the European Legislation Identifier,¹² the European Case Law Identifier¹³ [49] and possibly in the future a European Legal Doctrine Identifier (ELDI) [50] or a global standard for legal citations.¹⁴

4 CONCLUSIONS AND FUTURE WORK

Relevance, the basic notion of information retrieval "Is a thoroughly human notion and as all human notions, it is somewhat messy." [15] As upheld in this paper, 'relevance' within legal information retrieval deserves specific attention, due to rapidly growing repositories, the distinct features of legal information objects and the complicated tasks of legal professionals.

Because most LIR systems are designed by retrieval specialists without comprehensive domain knowledge, sometimes assisted by domain specialists with too little knowledge of retrieval technology, users are often disappointed by their relevance performance.

Four main conclusions can be highlighted. First of all, retrieval engineering is focussed too exclusively on algorithmic relevance, but it has been proven sufficiently that without domain specific adaptations every search engine will disappoint legal users. By unravelling the holistic concept of 'relevance' we hope to stimulate a more comprehensive debate on LIR system design. All dimensions of relevance have to be considered explicitly while

designing all components of LIR systems: document pre-processing, (meta)data modelling, query building, retrieval engine and user interface. Within the user interface searching, filtering and browsing should take full advantage of the various relevance dimensions, of course in a way that fits the legal mindset and acknowledging that relevance dimensions are continually interacting in the process of information searching.

Secondly, the 'isness' concept is too often overlooked. Finding (the expressions of) a work is – and not (just) the related works – is an often-used functionality for jurists, but misunderstood by system developers.

Thirdly, domain relevance is also an underdeveloped area. While there is a tendency to publish ever more legal information, especially court decisions, without tagging it as to its juristic value, information overkill will become a serious threat to the accessibility of such databases. Performance on other relevant dimensions will suffer if the problem of domain relevance isn't tackled.

Finally, given the importance of digital information for legal professionals – lawyers easily spend up to fifteen hours per week on search, most of it in electronic resources [34] although the abandonment of paper does not always seem to be a voluntary choice [51] – the gap between LIR systems and user needs is still big. For a full understanding of their search needs just taking stock of their wishes is not going to suffice, since they are not capable of describing the features of a system that does not yet exist. To understand the juristic mindset it is of the utmost importance to follow meticulously their day-to-day retrieval quests. It will for sure reveal interesting insights that can be used to improve the relevance performance of LIR systems.

REFERENCES

- [1] T. Bench-Capon, M. Araszkievicz, K. Ashley, et al., *A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law*, Artificial Intelligence and Law 20 (2012): 215-319.
- [2] R. Susskind, *Tomorrow's Lawyers: An Introduction To Your Future*, Oxford University Press, 2013.
- [3] E. L. Rissland and J. J. Daniels, *A hybrid CBR-IR approach to legal information retrieval*, *Proceedings of the 5th international conference on Artificial intelligence and law*, ACM, 1995, pp. 52-61.
- [4] P. Casanovas, M. Palmirani, S. Peroni, et al., *Special Issue on the Semantic Web for the Legal Domain Guest Editors' Editorial: The Next Step*, *Semantic Web Journal* 7 (2016): 213-227.
- [5] H. Turtle, *Text Retrieval in the Legal World*, *Artificial Intelligence and Law* 3 (1995): 5-54.
- [6] M. Palmirani, *Legislative XML: Principles and Technical Tools*, ARACNE, Rome, 2012.
- [7] International Federation of Library Associations and Institutions, *Functional Requirements for Bibliographic Records* UBCIM Publications - New Series Vol 19.
- [8] M. Araszkievicz, *Time, Trust and Normative Change. On Certain Sources of Complexity on Judicial Decision-Making*, in P. Casanovas, U. Pagallo, M. Palmirani and G. Sartor, eds., *AI*

¹² Council conclusions inviting the introduction of the European Legislation Identifier (ELI), CELEX: 52012XG1026(01).

¹³ Council conclusions inviting the introduction of the European Case Law Identifier (ECLI) and a minimum set of uniform metadata for case law, CELEX: 52011XG0429(01).

¹⁴ LegalCiteM: <www.oasis-open.org/committees/legalciteM/>.

- Approaches to the Complexity of Legal Systems: AICOL 2013*, Springer, 2014, pp. 100-114.
- [9] M. Palmirani and R. Brighi, *Time Model for Managing the Dynamic of Normative System*, in M. Wimmer, H. Scholl, Å. Grönlund and K. Andersen, eds., *Electronic Government; Lecture Notes in Computer Science*, Springer, Heidelberg, 2006, pp. 207-218.
- [10] F. R. Shapiro, *The Most-Cited Articles from The Yale Law Journal*, *Yale Law Journal* 100 (1991): 1449.
- [11] L. Humphreys, C. Santos, L. d. Caro, et al., *Mapping Recitals to Normative Provisions in EU Legislation to Assist Legal Interpretation*, in A. Rotolo, ed., *Legal Knowledge and Information Systems - JURIX 2015: The Twenty-Eighth Annual Conference*, IOS Press, Amsterdam, 2015, pp. 41-49.
- [12] S. Makri, *A study of lawyers' information behaviour leading to the development of two methods for evaluating electronic resources*, 2008.
- [13] S. Davidson, *Way Beyond Legal Research: Understanding the Research Habits of Legal Scholars*, *Law Library Journal* 102 (2010): 561-579.
- [14] T. Saracevic, *Relevance Reconsidered, Information science: Integration in perspectives. Second Conference on Conceptions of Library and Information Science*, 1996, pp. 201-218.
- [15] T. Saracevic, *Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance*, *Journal of the American Society for Information Science and Technology* 58 (2007): 1915-1933.
- [16] E. Cosijn and P. Ingwersen, *Dimensions of relevance*, *Information Processing and Management* 36 (2000): 533-550.
- [17] E. Cosijn and T. Bothma, *Contexts of relevance for information retrieval system design*, *Proceedings of the 5th international conference on Context: Conceptions of Library and Information Sciences*, Springer-Verlag, 2154833, 2005, pp. 47-58.
- [18] G. Salton, A. Wong and C. S. Yang, *A vector space model for automatic indexing*, *Communications of the ACM* 18 (1975): 613-620.
- [19] D. C. Blair and M. E. Maron, *An evaluation of retrieval effectiveness for a full-text document-retrieval system*, *Communications of the ACM* 28 (1985): 289-299.
- [20] D. P. Dabney, *The Curse of Thamuis: An Analysis of Full-Text Legal Document Retrieval*, *Law Library Journal* 78 (1986): 5-40.
- [21] K. T. Maxwell and B. Schafer, *Concept and Context in Legal Information Retrieval*, in E. Francesconi, G. Sartor and D. Tiscornia, eds., *Legal Knowledge and Information Systems - JURIX 2008: The Twenty-First Annual Conference*, IOS Press, Amsterdam, 2008, pp. 63-72.
- [22] M. Saravanan, B. Ravindran and S. Raman, *Improving legal information retrieval using an ontological framework*, *Artificial Intelligence and Law* 17 (2009): 101-124.
- [23] J. H. Fowler and S. Jeon, *The authority of Supreme Court precedent*, *Social Networks* 30 (2008): 16-30.
- [24] M. van Opijnen, *A Model for Automated Rating of Case Law, Fourteenth International Conference on Artificial Intelligence and Law*, ACM, New York, 2013, pp. 140-149.
- [25] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Pearson Education, Essex, 1999.
- [26] T. Harvold, *Is searching the best way to retrieve legal documents?*, *e-Stockholm '08 Legal Conference*, 2008.
- [27] M. van Opijnen, N. Verwer and J. Meijer, *Beyond the Experiment: the eXtensible Legal Link eXtractor, Workshop on Automated Detection, Extraction and Analysis of Semantic Information in Legal Texts, held in conjunction with the 2015 International Conference on Artificial Intelligence and Law (ICAAIL)*, 2015.
- [28] J. Bing and T. Harvold, *Legal Decisions and Information Systems*, Universitets Forlaget, Oslo, 1977.
- [29] G. Boella, L. Di Caro, L. Humphreys, et al., *Eunomos, a legal document and knowledge management system for the web*, *Semantic Web Journal*.
- [30] S. N. Mart, *The Relevance of Results Generated by Human Indexing and Computer Algorithms: A Study of West's Headnotes and Key Numbers and LexisNexis's Headnotes and Topics*, *Law Library Journal* 102 (2010): 221-249.
- [31] E. Francesconi and G. Peruginelli, *Semantic Interoperability among Thesauri: A Challenge in the Multicultural Legal Domain*, in W. Abramowicz, R. Tolksdorf and K. Węcel, eds., *Business Information Systems Workshops: BIS 2010 International Workshops, Berlin, Germany, May 3-5, 2010*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 280-291.
- [32] H.-J. Vibert, P. Jouvelot and B. Pin, *Legivoc – connecting laws in a changing world*, *Journal of Open Access to Law* 1 (2013).
- [33] L. F. Peoples, *The Death of the Digest and the Pitfalls of Electronic Research: What Is the Modern Legal Researcher to Do?*, *Law Library Journal* 97 (2005): 661-679.
- [34] S. A. Lastres, *Rebooting Legal Research in a Digital Age*, 2015.
- [35] M. van Opijnen, *Searching for References to Secondary EU Legislation*, in S. Tojo, ed., *Fourth International Workshop on Juris-informatics (JURISIN 2010)*, 2010.
- [36] M. van Opijnen, *Canonicalizing Complex Case Law Citations*, in R. Winkels, ed., *Legal Knowledge and Information Systems - JURIX 2010: The Twenty-Third Annual Conference*, IOS Press, Amsterdam, 2010, pp. 97-106.
- [37] E. Toms, *Serendipitous Information Retrieval, DELOS Workshop: Information Seeking, Searching and Querying in Digital Libraries*, 2000.
- [38] R. Winkels, A. Boer, B. Vredebrecht, et al., *Towards a Legal Recommender System 27th International Conference on Legal knowledge and information systems (JURIX 2014)*, IOS Press, 2014, pp. 169-178.
- [39] R. Cross and N. Wilkins, *An Outline of the Law of Evidence*, Butterworths, London, 1964.
- [40] J. Woods, *Relevance in the Law: A Logical Perspective*, in D. M. Gabbay, P. Canivez, S. Rahman and A. Thiercelin, eds., *Approaches to Legal Rationality*, Springer, Dordrecht, 2010.

- [41] J. Bing, *Handbook of Legal Information Retrieval*, Norwegian Research Center for Computers and Law, Oslo, 1991.
- [42] R. Mochales and M.-F. Moens, *Argumentation Mining*, *Artificial Intelligence and Law* 19 (2011): 1-22.
- [43] M. van Opijnen, *Op en in het web. Hoe de toegankelijkheid van rechterlijke uitspraken kan worden verbeterd*, Amsterdam UvA, Den Haag, 2014.
- [44] R. Winkels, J. de Ruyter and H. Kroese, *Determining Authority of Dutch Case Law*, in K. M. Atkinson, ed., *Legal Knowledge and Information Systems. JURIX 2011: The Twenty-Fourth International Conference.*, IOS Press, Amsterdam, 2011, pp. 103-112.
- [45] F. Tarissan and R. Nollez-Goldbach, *Temporal Properties of Legal Decision Networks: A Case Study from the International Criminal Court*, in A. Rotolo, ed., *Legal Knowledge and Information Systems - JURIX 2015: The Twenty-Eighth Annual Conference*, IOS Press, Amsterdam, 2015, pp. 111-120.
- [46] Y. Panagis and U. Šadl, *The Force of EU Case Law: A Multidimensional Study of Case Citations*, in A. Rotolo, ed., *Legal Knowledge and Information Systems - JURIX 2015: The Twenty-Eighth Annual Conference*, IOS Press, Amsterdam, 2015, pp. 71-80.
- [47] P. Mazzega, D. Bourcier and R. Boulet, *The network of French legal codes*, *12th International Conference on Artificial Intelligence and Law*, 2009, pp. 236-239.
- [48] M. Koniaris, I. Anagnostopoulos and Y. Vassiliou, *Network Analysis in the Legal Domain: A complex model for European Union legal sources*, arXiv:1501.05237 [cs.SI]2015).
- [49] M. van Opijnen and A. Ivantchev, *Implementation of ECLI - State of Play*, in A. Rotolo, ed., *Legal Knowledge and Information Systems - JURIX 2015: The Twenty-Eighth Annual Conference*, IOS Press, Amsterdam, 2015, pp. 165-168.
- [50] M. van Opijnen, *The European Legal Semantic Web: Completed Building Blocks and Future Work*, *European Legal Access Conference*, 2012.
- [51] C. C. Kuhlthau and S. L. Tama, *Information Search Process of Lawyers: a Call for 'Just For Me' Information Services*, *journal of Documentation* 57 (2001): 25-43.