Chapter 1

Introduction

1.1 Information retrieval systems

An information retrieval system (IR system) is a form of computer system which stores and makes available information, typically in the form of discrete documents.[3] Together, this forms a document collection (or corpus).

As the name suggests, the point of using IR systems is to retrieve information. In order to be useful, the system must not only be able to store a collection of documents, but also to provide individual documents to an end user. Since any interesting collection will contain far more documents than is practical for the end user to read, it must be able to provide those documents that the user wants to read. This is not always easy, as the user may not know which documents are wanted, or even if such documents exists at all.

The user of an IR system uses it to satisfy a need for some information, usually in order to perform some task. In the following, we use the term information need for this. The IR system must provide some way for the user to express that need. The main information retrieval paradigm in such systems is to provide a search form where the user can enter some sort of query, typically in the form of a few keywords, and then to return a result set of documents that matches the query.

Other paradigms include faceted browsing interfaces where the user can navigate through all available documents by restricting different facets of documents (i.e. documents of a certain type, by a certain author, or published a certain year), or by automatically constructing lists of related documents for a particular context.

Regardless of the paradigm used, at some point the user is presented with a result set of documents intended to satisfy his/hers information need. This

[3] That is, documents having clearly defined boundaries, as opposed to a more general database system, in which many small pieces of data can be combined and aggregated to form dynamic documents or views that contains a subset of the entire database content.

list will by necessity be sorted or *ranked* by some specific criteria. Alphabetic or chronological order are common ways of ranking result sets, but the most interesting way is to order them by relevance.[4] For large result sets (or large document collections), relevance ranking may be essential in order to make the system user friendly enough to be satisfactory.[5]

This kind of ranking requires a model of how to estimate relevance.[6] Such an algorithmic relevance model contains rules for what constitutes relevance in the relationship between information and a need for information, and how to measure aspects of this relevance, such as its strength or probability.

The perfect relevance model should be able to select exactly those documents that the user needs to read in order to satisfy the information need, while not selecting any documents that do not contribute to this satisfaction. Unfortunately, no relevance model can be perfect in practice. One indication of this is that that even experts disagree on what is or isn't relevant.[7] This makes it impossible to construct the definitive, objectively correct, list of documents that are relevant for a certain information need. Any such list (often called a *gold standard*) will to some extent bear traces of one or more relevance judges' opinions.

The fact that experts disagree on whether a particular piece of information is relevant for a particular information need suggests that relevance is a subjective notion. If relevance truly is subjective, it spells trouble for any attempt to build an algorithmic relevance model. Such a model contains, as previously stated, *rules* for what constitutes relevance. These rules are general and are applied equally for all users. If user A expresses a information need in the same way as user B, an IR system has no choice but to create the same list of presumably relevant documents in the same order. An IR system cannot deal with a subjective notion of relevance.[8]

However, the concept of relevance can be interpreted in many ways. Not

[4]There are scenarios where ranking in general and relevance models in particular is not needed. The most common is when the user seeks a specific document and knows some identifying property of it, such as the case number for a legal case. These scenarios are not considered in this thesis, but are generally solved by functions known as exact match or identity functions. See further sec. 3.1.1.

[5]Graham Greenleaf: Jon Bing and the History of Computerised Legal Research - Some Missing Links, in: Et tilbakeblikk på fremtiden, Oslo 2004, p. 64.

[6]Less obviously, it requires a model of how to express the user's information needs, i.e. in the form of a query

[7]See e.g. Jon Bing/Trygve Harvold: Legal Decisions and Information Systems, Oslo 1977, p. 40 or Daniel P. Dabney: The Curse of Thamus: An analysis of Full-Text Legal Document Retrieval, in: Law Library Journal 78.5 (1986), pp. 5–40, here p. 15. See also Cuadra and Katter (1967) for research about factors that cause disagreement between judges of relevance

[8]Actually, this is a simplification – techniques like manual query expansion, relevance feedback, click-tracking and background collection of information about the user (such as the location of the user) can be used to inform the system of the users preferences, and so inject subjective knowledge into the ranking process. But these techniques merely change the weighting of terms and parameters, they do not replace the relevance model as such.

all of these interpretations are subjective. [9] At least sometimes, relevance is *not* in the eye of the beholder.[10]

There are a number of standard models of relevance in use in IR today. These models are *general* in that they are not dependent on any particular kinds of documents. Instead, they are based on general properties of documents and collections, such as how frequently a given term occurs in a document as compared to the document collection as a whole. [11]

1.2 Legal information, norms and meta-norms

IR systems have been used for legal research since their inception.[12] This is not surprising, given that law might be the most information intensive professional occupation there is.[13]

All IR systems have some form of boundary for what information it stores and makes available. For some systems this boundary is very wide, incorporating a great number of documents (such as the boundary "all documents that can be found on the open web" used by web search engines), for some it incorporates only a few selected texts. A common boundary for IR systems that stores and makes available information intended for legal research (which we, from now on, will call a legal IR system)[14] is "documents that can be used to find out what the law is" (most often with the qualifier "in a particular jurisdiction or set of jurisdictions", as few legal IR systems cover documents from all the worlds jurisdictions).

These documents generally encompass statutes and other generally binding legislation, case law that are considered to have precedence, preparatory works for enacted legislation and jurisprudential literature such as treatises and law review articles. We call the contents of these documents *legal information*, and the sources for such information *legal sources*. We call the rules or norms that can be found by interpreting this legal information *substantive norms*.[15]

[9] Particularly, the rule of law requires that legal decisions are not subjective. See further 1.3

[10] In chapter 2, we will explore the notion of relevance from different aspects.

[11] These are described in detail in 3.1.1

[12] Some of the earliest initiatives, dating back to the late 1950's, have been described in Jon Bing: Handbook of Legal Information Retrieval, Oslo 1984, p. 61

[13] Richard E. Susskind: The future of law : facing the challenges of information technology, Oxford 1998, p. 79 See also K. Tamsin Maxwell/Burkhart Schaefer: Concept and Context in Legal Information Retrieval, in: Legal Knowledge and Information Systems - Jurix 2008: The Twenty-First Annual Conference, 2008, who state "There is possibly more textual data for law than any other domain".

[14] A synonym for Legal IR often used is the term "Computer Aided Legal Research" (CALR), but as this thesis is more concerned with the efficient retrieval of information rather than the larger process of research, the term CALR SYSTEM will not be used.

[15] c.f. Peter Wahlgren: Automation of Legal Reasoning - A Study on Artifical Intelligaence and Law, Deventer-Boston / Stockholm 1992, p. 173, "The doctrine of legal sources

There are many kinds of information that relate to law and substantive norms that do not qualify as legal information. For example, a newspaper editorial which describes problems with existing statutory law, and which contains suggestions for changing it, is not generally considered legal information. But if the exact same problem description and suggestion for change is included in a preparatory work, it might be classified as legal information. The key difference is that legal information has *authority* in that it affects the content and/or the interpretation of the law.

The way authority works in the practice of law means that legal information is used slightly different than information in other fields. A statute is not merely a description of what the law is[16] – it constitutes the law.[17]

The authority of texts can often be traced through references or citations. The status of a text can be altered through events occurring after the production of it. One obvious example is that the content of a statute can be modified by another, more recent statute. When a new statute amends an older statute, it does so by explicitly referencing the older statute and stating the changes that the amendment entails. Similarly, a legal rule which first is formulated in one legal case can be made more precise, stronger, weaker or even abolished through later events, both by statutory law and later legal cases.[18] When a subsequent legal case uses a legal rule formulated in an earlier legal case, it usually cites the earlier case.[19]

Citations between different types of text generally flow in a single direction (i.e. a legal case may cite a statute, but a statute never cites a legal case). In most cases, citations also follow a temporal flow (i.e. a newer document can cite an older document, but not the other way round).

Over time, these citations create a citation network between different documents (or parts of documents). This is an important characteristic of legal information that can reveal much about how it is used by different actors in legal processes. Similar networks can be found in other types of information, such as scientific publishing[20] and the general web.[21]

Legal information is just text. The law is really based on norms, evidence of which can be found in such text. The task of finding out the law is

is basically a theory about substantive law"

[16]Here the difference between the textual expression of a legal norm that can be found in a statute and that legal norm itself is disregarded, but see 3.2.3 for more about this distinction.

[17]Howard Turtle: Text Retrieval in the Legal World, in: Artificial Intelligence and Law 3 (1995), pp. 5–54, here p. 7.

[18]To what extent earlier case law can be modified by later case law is affected by whether the legal tradition in question adheres to the *stare decisis* doctrine.

[19]For very established legal rules, it is less common for such cites to be explicit as the rule, over time, is considered to be a general principle of and not particularly tied to the first case. This process is similar to the phenomenon "obliteration through incorporation", further described at page 30.

[20]Such citation networks is studied in the field of bibliometrics. See 2.3.2.

[21]Link analysis algorithms have been a field of research for around 15 years, see 3.1.2

essentially the task of finding out what legal norm(s) exists that are applicable to the facts of the case. Extracting norms from text is a central part of the legal reasoning process.[22]

Jurisprudence, the philosophy of law, includes norms about how law is practiced. This includes norms about how the law may and must be interpreted in different circumstances, but also about what constitutes the law – from which documents knowledge about the law may be sought. These norms are not formally specified e.g. in the form of statutes. We call this type of norms meta-norms.[23]

Meta-norms that determine which documents that can be considered legal information are sometimes referred (i.e. which sources are legal sources) to as the doctrine of legal sources. The doctrine of legal sources varies between different legal traditions and can be used to select which information should be included in legal information systems.[24]

If a particular document is not of a type that a court (or an arbitrator, or any other legal decision-maker) may consider when attempting to find out what the law is, it *cannot* be legally relevant. If it is, it *may* be relevant, but this is in itself only a neccesary condition for relevance, not a sufficient condition for the same.[25] The doctrine of legal sources thus acts as a filter for what constitutes legal information.[26] There are also other meta-norms for e.g. determining the relative importance of different norms in the event that they conflict.[27]

These meta-norms affect the legal relevance of legal information. By analyzing them, jurisprudence-based model of relevance can be formulated.[28] From this model, a relevance ranking algorithm can be constructed. A search engine that uses this algorithm can perform better than one based on traditional, general, models of relevance. The question of what "better" means in this context is the central theme of this thesis. A tentative definition, perhaps somewhat circular, is "more aligned with jurisprudential meta-norms concerning the doctrine of legal sources and conflict of norms".

1.3 Motivation

Why is relevance, and particularly ranking of search results by relevance, important? The practice of law seems to be in a perpetual state of informa-

[22] Further explained in 3.2.3

[23] Bing: Handbook of Legal Information Retrieval (see n. 12), p. 10.

[24] Peter Wahlgren: Law and Information Technology - Swedish Views, in: 2002, chap. The Quest for Law. Legal Sources via IT, p. 210.

[25] This is a reformulation of Bing's definition of legal relevance, further described in 2.4.3.

[26] Or, in the mathematical sense of the word, a *predicate*, i.e. a function that determines if a particular element is part of a set.

[27] See further sec. 2.4.1

[28] This is done in 2.4

tion overload.[29] In particular, the law of all member states in the European Union consists of two, sometimes radically different, often conflicting, sets of norms – national law and community law.

The rule of law requires that laws must be, amongst other things, publically declared and that its application is predictable. In order to have a predictable application of law, all practitioners must have access to the same set of documents that contain knowledge about the law, and they must be able to read and comprehend those documents that affect the question at hand. To invoke Ronald Dworkin's Hercules metaphor, predictable application of the law in all cases, including the so-called "hard cases",[30] require that judges (or other applicators of law) be as omniscient as Hercules the judge.[31] Not very many human judges can match the cognitive capacity of Hercules.[32] The result is unpredictable application of the law. Therefore it can be argued that improving legal information systems is worthwhile in order to ensure the rule of law.[33]

Furthermore, the rule of law requires that the application of law, and thus reasoning based on legal information, be predictable. Therefore, relevance in a legal context cannot be any more subjective than a judge can, or else the rule of law is not upheld.[34]

Traditional legal information retrieval is based on full text retrieval.[35] This model of retrieval (further explained in 3.1) is known to have problems, especially for typical law research scenarios where recall is more important

[29]See e.g. Louis O. Kelso: Does The Law Need a Technological Revolution?, in: Rocky Mountain Law Review 18 (1946), pp. 378–392, here p. 378, or Bing's description of the 1970's "information crisis of the law" in Jon Bing: The Policies Of Legal Information Services: A Perspective Of Three Decades, in: Lee Bygrave (ed.): Yulex 2003, Oslo 2003, pp. 37–58

[30]A "hard case" can be characterized as a legal question for which there is no determinable rule. Raymond Wacks: Philosophy of Law - A Very Shor Introduction, Oxford 2006, p. 42

[31]For more on the metaphor of Hercules, see Ronald Dworkin: Taking rights seriously, 1977, p. 105

[32]Over half a century ago, Kelso wrote "[T]he absurdity of the presumption that [judges] know *the law*, that is, that they have surveyed and have become informed on all legal data bearing on the point at hand, increases by the hour". Kelso: Does The Law Need a Technological Revolution? (See n. 29), p. 380

[33]The concept of "the rule of law" can be characterized as a ideological set of values regarding the role and responsibilities of law in society. Eckhoff has described it through the five elements "predictability, fair and just process, objectivity in application of norms and discretion, the principle of equality and democratic control". See Bing/Harvold: Legal Decisions and Information Systems (see n. 7), pp. 225

[34]Bing: Handbook of Legal Information Retrieval (see n. 12), p. 198: "In legal decision-making, as in other formal situations, it is therefore not appropriate to regard relevance as entirely subjective".

[35]Christine Kirchberger: Paper and stone: How technology has not changed the retrieval of legal information, yet, in: Proceedings of the BILETA 2007 Annual Conference, 2007, p. 3.

than precision, and the corpus is very large.[36] This model often yields poor recall since it's based on the assumption that users are able to foresee what words and phrases will be used in useful documents.[37] 25 years of further research in advanced weighting techniques and document vector distances hasn't changed the simple fact that there usually is not enough information in the free text query itself to construct useful results from.[38]

The main problem with full text retrieval is that it is hard to guess what kinds of words may be present in those documents that the user may find relevant. The user of a legal information system is typically interested in concepts, not words.[39] If the user is looking for cases concerning the legality of videocassette recorders (VCRs) and uses the term "VCR" as a search query, he/she will miss documents that uses other terms for the same thing.[40] Ways around this problem include automatic query expansion or conceptor-based strategies.[41]

These methods to increase the number of search results have their own drawbacks. Instead of missing relevant documents we get too many non-relevant documents. We experience information overload when it gets impractical to examine large amounts of documents in order to find the few ones with actual relevance.

There exists alternatives, such as systems based on knowledge engineering or case-based reasoning techniques, but these require either manual or automatic indexing. Indexing covers activities such as tagging or classifying texts (or parts thereof) so that the meaning of the text (or a restricted subset thereof) is available in a machine-readable way.[42]

Manual indexing (also known as "intellectual indexing") is too expensive to be practical for large document collections, particularly since the process can yield inconsistent results as two different human indexers may classify the same document differently, or the same indexer may classify the same document differently at different points in time.[43]

[36] The terms "recall" and "precision" are defined in 3.1.3.

[37] David C. Blair/M. E. Maron: An Evaluation of Retrieval Effectiveness for a Full-text Document Retrieval System, in: Communications of the ACM 28.3 (Mar. 1985), pp. 289–299, here p. 295.

[38] Trygve Harvold: Is searching the best way to retrieve legal documents?, in: Lov & Data 98 (2009), pp. 22–26.

[39] Marie-Francine Moens: Innovative techniques for legal text retrieval, in: Artificial Intelligence and Law 9 (2001), pp. 29–57, here p. 29.

[40] The well-known Sony v Universal case (464 U.S. 417), which established that manufacturers of such devices could not be held liable for contributory copyright infringement, does not use this term but instead the older "video tape recorder". This problem is described in detail in Daniel E. Rose/Richard K. Belew: A connectionist and symbolic hybrid for improving legal research, in: International Journal of Man-Machine Studies 35 (1991), pp. 1–33, here p. 6.

[41] Conceptors are described in sec. 3.1.1

[42] Moens: Innovative techniques for legal text retrieval (see n. 39), pp. 35.

[43] Blair/Maron: An Evaluation of Retrieval Effectiveness for a Full-text Document Re-

Automatic indexing require natural language processing[44] and, depending on which semantic fidelity one needs, automated reasoning, which is a very hard problem to solve in a general manner. Furthermore, attempts to augment the authentic text of legal sources with semantic metadata introduce an interpretation of the legal sources, which gets hidden in the system unbeknownst to the user, even if the users' interpretation may differ from the systems.

A middle road between traditional full text retrieval and knowledge engineering-based retrieval is to use such metadata that are part of the document corpus, yet not part of the authentic text. One such example is the structure of documents included in the corpus. By taking advantage of such divisions into different fields that typically occur in legal information, it is possible to create queries (and interfaces for querying) that result in higher precision searches. [45]

Another example is citation analysis. Extracting citations between cases is a relatively simple process, and they can be a useful indicator of relative importance. Determining the exact semantic implications of what these citations *mean* is decidedly more difficult, though.

1.4 Hypotheses

This thesis presents a relevance ranking system that is *specific* for legal documents, in that its relevance model is based on legal norms. The purpose of designing a ranking system specific to the domain of law is to create more efficient legal IR systems. Efficient in this context means better at facilitating the *communication* between the user (typically a legal professional with a specific problem at hand, but may also be an intermediary such as a paralegal or a legal librarian) and the system. The concept of IR system interaction as a communication process is further explored in 2.1.

As part of the thesis, a basic IR system has been implemented, containing the case history of the European Court of Justice (ECJ) as the document collection and using the text of the Treaty of the Functioning of the European Union (TFEU) as base for the queries.

Three hypotheses are proposed. The system is evaluated to test the third and final of these.

1. The legal method used by legal professionals contains norms for determining the relevance of a legal document with respect to a particular

trieval System (see n. 37), p. 290.

[44]Natural language processing (NLP) is a set of techniques to automatically detect semantic information from written text. See e.g. Daniel Jurafsky/James H. Martin: Speech and language processingn, 2nd ed., 2008

[45]See e.g. Cecilia Magnusson Sjöberg: Critical Factors in Legal Document Management, Stockholm 1998, ch. 12

legal problem. These norms are, for the most part, meta-norms rather than substantive norms. This hypothesis is handled in chapter 2, "The concept of relevance", particularly 2.4.

2. These relevance-determining rules can be *approximately* described in a formal fashion as a ranking function. This hypothesis is handled in chapter 3, "Information retrieval", particularly 3.3.

3. A legal IR system will have better relevance ranking when using such a specific function than when using standard relevance functions. This hypothesis is evaluated in chapter 5, "A prototype of a legal relevance function".

1.5 Method description

The subject matter of this thesis is how legal information is used in the legal reasoning process. In this aspect, the thesis uses traditional jurisprudence method.[46] The concept of relevance is examined from a number of different views (not all of them based in legal theory), and an attempt is made to compile these views, hopefully adding some insight in the process. This method is most pronounced in chapter 2 but also to some extent in chapter 4.

But the thesis also attempts to find measurable aspects of relevance, such that a ranking system can be constructed from them. It proposes a falsifiable hypothesis and attempts to carry out an examination yielding measurable results. These results either falsify or corroborate the hypothesis. In using methods from the natural sciences, but in a legal context, the thesis uses the jurimetrical method. Jurimetrics is concerned with the use of scientific methodology when analyzing legal problems.[47] This method is most pronounced in chapter 5, but the hypothesis proposed is constructed from what is learned in chapter 3.

1.6 Structure of this thesis

In order to design a jurisprudential relevance ranking system, one must attempt to answer the question "What makes something legally relevant?". Chapter 2 deals with general theories of relevance, the notion of relevance in information retrieval and the relationship between legal relevance and the legal method.

[46] This is not as easy to define as legal method in general. An introduction to these difficulties can be found in J.E. Penner: Textbook on Jurisprudence, 4th ed., Oxford 2008, pp. 8

[47] Wahlgren: Automation of Legal Reasoning - A Study on Artifical Intelligaence and Law (see n. 15), pp. 119.

As one of the goals of the thesis is to design and implement an actual working information retrieval system, Chapter 3 introduces the field of information retrieval in general, particular issues for legal IR systems, and examines a number of different models for determining relevance in IR systems. It ends with a tentative definition of a jurisprudential relevance ranking function.

The second, shorter, part of the thesis concerns this relevance function. Chapter 4 describes previous work by examining the history and current state of the art concerning relevance ranking in legal IR systems. Chapter 5 describes the implementation and evaluation of a prototype IR system using the proposed relevance function. Chapter 6 concludes the study and suggests future work.

The second part is complemented by the actual system prototype and two appendixes. The first appendix describes technical aspects of the prototype system including information for acquiring, running and extending it. The second appendix lists the set of gold standard relevance judgments used for evaluating the performance of the system.