# Part II

# A legal relevance ranking function

# Chapter 4

# Previous work

The legal information retrieval field has a rich history of ideas concerning relevance and different ways of assessing it in automated ways. In this overview, only those ideas that have been put to test in actual systems are mentioned. More theoretical ideas are mentioned in section 2.4 and 3.2.4. The below list is by no means exhaustive.

Attempt has been done to put different studies under suitable headings. This thesis falls under the last category, "Citation analysis"

## 4.1 Automated concept indexing and classification for retrieval

In traditional full text information retrieval, a set of terms are extracted from the text of the document. The documents are indexed under these terms, and the total set of terms in the systems can be massive (as many as there are unique words in the corpus). This process can be augmented by assigning a similar but smaller set of terms based not on the wording, but the concepts and meaning of the document by the process of classification. Having documents assigned to a set of smaller, carefully crafted taxonomy of terms can help tremendously in crafting high-precision queries, and can increase recall as well. But this classification is very costly to do by hand, and human classifiers are often inconsistent in their decision. Automating this process can lower costs as well as increasing retrieval performance.

**SCALIR:**[262] Traditional full-text retrieval uses only *surface qualities* of the legal text. As the language in law is lexically and semantically am-

[262] A shorter description of primarily technical aspects in D. E. Rose/R. K. Belew: Legal Information Retrieval: A Hybrid Approach, in: Proceedings of the second international on conference on artificial intelligence and law, 1989, pp. 138–146, a longer report including legal theory aspects in Rose/Belew: A connectionist and symbolic hybrid for improving legal research (see n. 40). The Ph.D. dissertation describing the system is reviewed in Jon Bing: Book Reviews: A Symbolic and Connectionist Approach to Legal Information Retrieval, in: Information Processing & Managment 31.6 (1995), pp. 903–910

biguous, open-ended and dynamic, traditional retrieval usage patterns give worse-than-expected recall.[263] In the SCALIR system, an AI-derived approach was used, combining symbolic AI, using explicit knowledge and links, such as references to cases and statues as well as term usage, with connectionist AI, using neural networks and subsymbolic manipulation. This created a layered network of nodes, where each node could be a term, a case, or a statue section. Nodes were connected by differently weighted links, forming a neural network to identify relevant cases for a query, and providing the opportunity to improve the network using user feedback.

**Jurisconsulto:**[264] A system for retrieving cases in Brazilian penal court. The system is built around traditional information retrieval using the vector space model. The core of the invention is automated metadata extraction through a controlled vocabulary (semi-automatically created) and a thesaurus (manually created from the vocabulary). Each document is represented with a key/value-list representing different properties of the case (such as defendant, crime type, damages, etc) and a search is done by entering metadata in the form of a similar key/value-list. Traditional "nearest neighbor"-matching selects the most relevant cases.

**Concept-based ranking:**[265] Automatic construction of a thesaurus from a corpus of 500 000 documents. The thesaurus contains information about broader and narrower terms, as well as synonyms and related terms. The thesaurus is constructed using a belief network,[266] then using this to rank results in a more sophisticated way than classic query expansion. The process improves precision of results, not just recall.

**Austrian Supreme administrative court:**[267] Experiments in automatic classification and clustering of documents using vector space models and the clustering algorithm k-means. A new clustering method (keyword clustering), based on the behavior of legal professionals when they classify tax law cases, was developed. It is based on pre-fixed sets of keywords

[263] Blair/Maron: An Evaluation of Retrieval Effectiveness for a Full-text Document Retrieval System (see n. 37).

[264] Tania C. D'Agostini Bueno et al.: JurisConsulto: Retrieval in Jurisprudential Text Bases using Juridical Terminology, in: Proceedings of the ICAIL-99 conference, 1999.

[265] Maria L. Silveira/Berthier Ribeiro-Neto: Concept-based ranking: a case study in the juridical domain, in: Information Processing & Management 40 (2004), pp. 791–805.

[266] A belief network, also known as a bayesian network, is a system of nodes representing probability for individual variables, and a set of connections between nodes representing direct influence between variables. See Russel/Norvig: Artificial Intelligence - A Modern Approach (see n. 257), p. 436 and also H. Turtle/W. B. Croft: Inference networks for document retrieval, in: SIGIR '90: Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA 1990, pp. 1–24 for a application of belief networks (in the article called inference network) in information retrieval

[267] Ingo Feinerer/Kurt Hornik: Text Mining of Supreme Administrative Court Jurisdictions, tech. rep. 51, Department of Statistics and Mathematics Wirtschaftsuniversitü Wien, Mar. 2007.

and finds that this yields almost perfect results (compared to legal expert classification).

Experiments in dichotomic clustering (partitioning documents in disjunct sets either dealing or not dealing with a particular regulation) were also performed using both string kernels and support vector machines (SVM). SVM was found to yield better results.

## 4.2 Knowledge engineering and case based retrieval approaches

Knowledge engineering uses techniques more associated with artificial intelligence than information retrieval. The focus is on representing the meaning and concepts featured in the text in such a way that it can be used for automatic reasoning. The system described by Cooper in 1971 in his description of "logical relevance" is a precursor to these approaches.[268]

One important class of applications for knowledge engineering is *expert systems*, which uses large knowledge bases and automatic reasoning (through infering new facts from facts contained in the knowledge base) in order to assist users with answering questions. The knowledge base that these systems operate on is usually constructed with the help of domain experts.[269]

**Knowledge representation model for legal cases:**[270] A database of 150 tort/damages cases where for each case is manually extracted: (I) issues (legal questions) and sub-issues (requisites) are manually extracted, (P) factors favorable for either party (P), and (F) contextual features that are important for the decision but is not directly in favor of either party (such as the roles of the damager and the damages). By calculating a trinary IPF value for each case, a list of relevant cases can be found by ranking on IPF score. The relevance connections matches the explicit legal case cites in the actual text of the cases.

**Comprehensive Legal Ontology:**[271] In computer and information science, a ontology is a formal specification of concepts within a domain, including hierarchies (such as Ordinance *is-a* Statute) and other relations.[272]

[268] See sec. 2.2.4 and Cooper: A definition of relevance for information retrieval (see n. 64)

[269] Richard Susskind: Expert Systems In Law, Oxford 1987, p. 9.

[270] Yiming Zeng et al.: A Knowledge Representation Model for the Intelligent Retrieval of Legal Cases, in: International Journal of Law and Information Technology 15.3 (2006), pp. 299–319.

[271] Erich Schweighofer: Computing Law: From Legal Information Systems to Dynamic Legal Electronic Commentaries, in: Cecilia Magnusson Sjöberg/Peter Wahlgren (eds.): Festskrift till Peter Seipel, 2006, pp. 569–584 and Erich Schweighofer/Doris Liebwald: Advanced lexical ontologies and hybrid knowledge based systems: First steps to a dynamic legal electronic commentary, in: Artificial Intelligence and Law 15.2 (July 2007)

[272] Schweighofer: Computing Law: From Legal Information Systems to Dynamic Legal Electronic Commentaries (see n. 271), p. 572.

A legal ontology is thus a formal specification of concepts within the legal domain. The work on the Comprehensive Legal Ontology (CLO) is an effort to construct a legal ontology that can be used for reasoning about legal questions. This is not yet a working system, only a plan for how such a system could be constructed.

**Semantic events:**[273] Using natural language processing (NLP) to extract "events" (including states and attributions) from legal text. For instance, in the sentence "The defendant replied that no city permit was necessary because the defendant's lands enjoy interjurisdictional immunity", the events "replied", "permit", "enjoy" and "immunity" can be extracted. Based on experiments that were performed on a small selection of Canadian crime cases, hypotheses that similarity between event structure in texts can be used to improve precision and ranking.

## 4.3 Legal citation analysis

**Citation vectors:**[274] By observing the citation patterns in English and American legal cases, a number of assumptions about these cases were made. In particular, case A citing case B, and B being decided close in time, by a lowly remote court, probably means that in all contexts where A is relevant, B is also relevant. By modeling each case as a vector, and then comparing vectors, cases can be correlated and clustered.

**SCALIR**: This system was described in sec. 4.1, but it deserves special mention in this section due to the fact that it utilizes explicit legal citations between statutes and cases.

**Semantics-Based Legal Citation Network Viewer:**[275] Observes that two cases can both cite a single older case for reasons of different legal topics. By using NLP techniques on a corpus of US cases, the system can distinguish different topics by analyzing the text around the citation (the reason for citing), and creating separate citation networks for each legal topic. The resulting topical networks are visualized and browsable.

**AustLII:**[276] The Austrialian Legal Information Institute (AustLII) makes available case law for Australian juridictions. The information is available in highly structured form, including citations between cases. In their custom

[273] K. Tamsin Maxwell/Jon Oberlander/Victor Lavrenko: Evaluation of Semantic Events for Legal Case Retrieval, in: Proceedings of the ESAIR '09 conference, 2009.

[274] Tapper: The use of citation vectors for legal information retrieval (see n. 208) and Tapper: An experiment in the use of citation vectors in the area of legal data (see n. 158)

[275] Paul Zhang/Lavanya Koppaka: Semantics-Based Legal Citation Network, in: Proceedings of the ICAIL '07 conference, 2007.

[276] Andrew Mowbray/Philip Chung/Graham Greenleaf: Free-access case law enhancements for Australian law, tech. rep., AustLII, 2008 and Graham Greenleaf/Philip Chung/Andrew Mowbray: The Long Tail(s) of the Law: An exploratory study, in: Law via the Internet 2011 Conference, Hong Kong, Abstract and presentation available from http://www.hklii.hk/conference/programme, 2011

search engine,[277] one of the alternatives for ranking results is by citation frequency (other options are by database, date, relevance or title). The structure of the citation network has further been analyzed in a study.[278]

**Lovdata:**[279] Describes problems with traditional legal IR through search interfaces - if one does not know which particular document to look for, it's hard to find documents since one doesn't know which terms they use. An alternative is to navigate through links. Ends with a summary of statistics for the Lovdata database

**US Supreme Court Authority:**[280] Confirms that the authority score of the HITS algorithm corresponds (and even predict) what lawyers find important, using the citation network of the US Supreme Court history.

**Austrian supreme court citation network:**[281] Observes that the Austrian supreme court citation pattern forms a scale-free network. Uses link analysis to determine the importance of each case, and determines that citation count correlates with relevance.

**Distance measures:**[282] Legal citation networks form, through the time-bound formation of citations, a class of graphs known as acyclic digraphs. For this class of graphs, a distance measure based on the occurrence of "sinks" (nodes that are referenced, but do not themselves reference anything) is defined. This measure is used on the citation network of the US Supreme Court history to find clusters. Compared to traditional graph clustering (or community detection) algorithms defined on general graphs, this method yields a better clustering more correlated with actual legal concepts that occurs in those cases.

[277] Available at *http://www.austlii.edu.au/forms/search1.html*

[278] Greenleaf/Chung/Mowbray: The Long Tail(s) of the Law: An exploratory study (see n. 276).

[279] Harvold: Is searching the best way to retrieve legal documents? (See n. 38).

[280] Fowler/Jeon: The authority of Supreme Court precedent (see n. 217).

[281] Geist: Using Citation Analysis Techniques For Computer-Assisted Legal Research In Continental Jurisdictions (see n. 162).

[282] Michael J. Bommarito II et al.: Distance Measures for Dynamic Citation Networks, in: Physica A 389.19 (2010), pp. 4201–4208.