Chapter 5

A prototype of a legal relevance function

As stated in the introduction, the hypothesis put forth in the introduction have been tested against a document collection consisting of the case history of the European Court of Justice. Following standard IR evaluation methodology (in the Cranfield paradigm tradition), a series of information needs (sometimes described as *Topics*) have been specified, and a set of relevance judgments (the *gold standard*) have been created.[283] This is tested against two different IR systems; one (the *baseline*) being built on best-of-breed general probabilistic retrieval algorithms, and the other built on the set of algorithms specified in sec. 3.3. The results are evaluated using the mean average precision (MAP) metric.

5.1 Corpus, citation network and information needs

5.1.1 Document collection

The document collection was fetched from the English EURLEX database at August 1st, 2011, and consists of 14 327 decisions from the European Court of Justice (ECJ) and the General Court (EGC). Each case has a basic set of metadata such as its identifier, date of decision, and similar properties. More importantly, it contains 254 698 links to cited statutory law (the founding treaties, directives and similar) and ECJ case law. Of these, 107 473 are links to other ECJ cases. These constitute the citation network. The links to statutory law provides a useful mechanism for restricting the network to just those cases that directly refer to e.g. a particular treaty article.

The document collection was fetched from the English EURLEX database at August 1st, 2011, and consists of 14 327 decisions from the European Court of Justice (ECJ) and the General Court (EGC). Each case has a basic

[283] For more on IR evaluation methodology, see sec. 3.1.3

set of metadata such as its identifier, date of decision, and similar properties. More importantly, it contains 254 698 links to cited statutory law (the founding treaties, directives and similar) and ECJ case law. Of these, 107 473 are links to other ECJ cases. These constitute the citation network. The links to statutory law provides a useful mechanism for restricting the network to just those cases that directly refer to e.g. a particular treaty article.

The Treaty of Lisbon resulted in substantial changes to the EU treaties, including changing the name of "The Treaty establishing the European Community" (TEC) to "Treaty on the Functioning of the European Union" (TFEU), and a complete re-numbering of the articles. A similar renumbering occurred with the Amsterdam treaty. TFEU contains 385 articles, each mainly dealing with a single concept.

Older case law does not refer to the articles as numbered in the Lisbon treaty, but instead using the numbering established in the older Amsterdam, Maastricht or Rome treaties. A table of equivalencies between these treaties was constructed, so that old case law that references e.g. article 30 in the original Rome treaty was considered in the same way as a newer case that referenced the substantially same article 34 in the Lisbon treaty.

5.1.2 Information needs

Each information need used in evaluation is specified as a need to find the most illuminating legal cases (the *landmark cases*) for understanding the topic of the article in question. Constructing a gold standard set of judgments for each article would require a substantial amount of work for little benefit. The ten most significant articles, based on case law citation frequency, were selected.

For the baseline system, the text of each article is used to construct a query consisting of a set of terms. These terms are selected by treating each article as a document, the treaty in full as a collection, doing basic stemming on both sets of terms,[284] and then selecting the top 5 terms with highest TF-IDF value. The theory behind this is that these terms will be the most significant for expressing the topic of the article. A better baseline could be constructed by manually creating a set of query terms.

For the prototype, the query is based on the subset of the citation network that cites each individual article. Only such cases, and all such cases, that directly cite a particular article were part of the result set for that article (including equivalencies as described above). They were then ranked using either the entire citation network or a restricted network (only including cases that cited the same or an equivalent article).

[284] Stemming is the process of reducing different inflections of a word to a common basic form or "stem", see Karlgren: Information Retrieval: Statistics and Linguistics (see n. 102), sec. 3.2

5.1.3 Document structure

Both TFEU and the cases are available in electronic form from EURLEX. The consolidated version of TFEU, meaning that changes to the original treaties that were made by the adaption of the Treaty of Lisbon have been incorporated into the text, was used.[285]

As available from the EURLEX service, both the treaty and the cases lack *semantic structure*. In order to analyze the citation patterns, we need at least some semantic information. For the treaty itself, we need to be able to tell which articles it contains, and the text of each article. For each legal case, we need at least its CELEX number, the case number, the date and a list of other cases (identified by CELEX numbers) that the current case cites.

Since the treaty and in particular the legal case collection are both so vast, it would not have been practical to do this work manually. Instead, a program capable of automatically downloading and analyzing both treaty and cases was developed. This program recreates the content of both treaty and cases with added semantic information in the form of a XHTML+RDFa document.

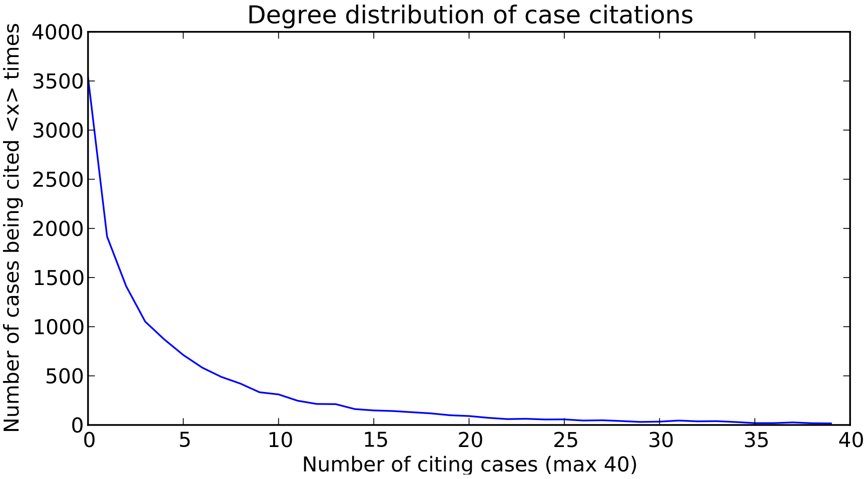
This automatic processing is further described in appendix A.

5.1.4 Citation network properties

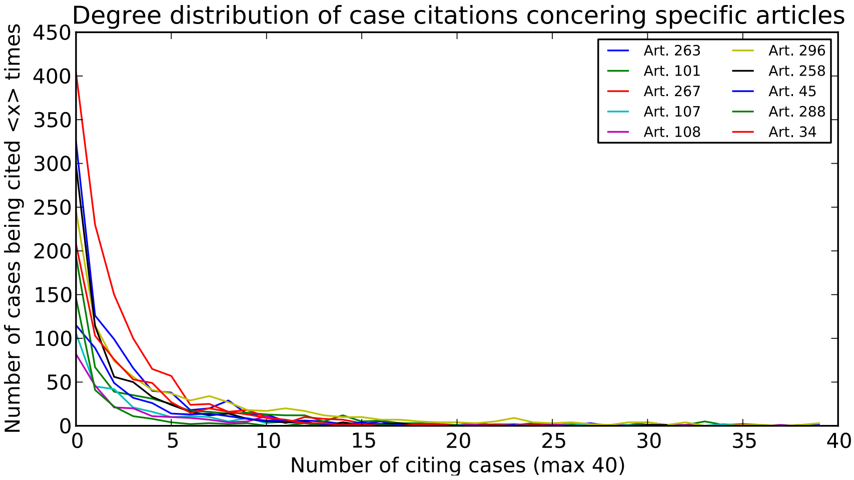
Before describing the prototype and the evaluation, it may be interesting to examine a few key aspects of the citation network that the corpus forms. These aspects can be visualized as graphs.

In the first graph, we examine the degree distribution of the entire network. We can observe that it exhibits scale-free network properties – more cases gets cited zero times than one, more cases get cited once than twice, and so on. The graphs x-axis is cut off at maximum 40 citations. This is because beyond that point there are very few cases with that number of inbound citations, but a few outliers have a massive number of citations (around 300). Setting a cut-off point helps focusing on the interesting parts of the graph.

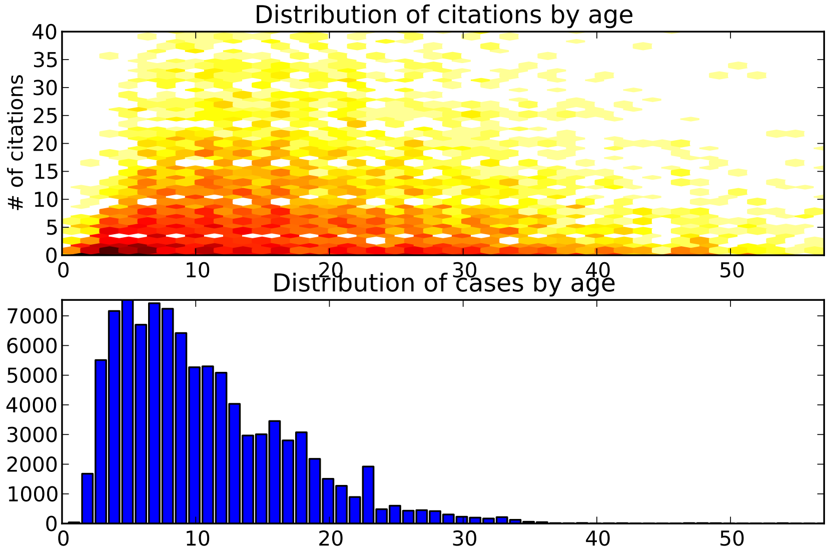
[285] As of the time of writing at the url *http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2008:115:0001:01:EN:HTML*



In the second graph, we examine the degree distributions of citation network that are restricted to just citations from cases that also cite a specific TFEU article (this is the same network type known as "unrestricted" in the algorithm configuration below). Again, we can see that the degree distribution in general mirror the degree distribution of the entire network, with some small individual differences.



In the third graph, we examine correlations between the age of the case and the number of citations it has received. Each point in the graph is colored accordingly to how many cases of that age have been cited that many times. As could be expected, older cases (10-20 years) receive the most citations. What is perhaps initially more confusing is the relatively low number of highly cited cases that are 30-40 years. Since many of the true landmark cases from ECJ is of that age, one would expect many highly cited cases from that era. One explanation is offered in the lower part of the graph, where it becomes evident that the volume of cases in recent years dwarfs the number of cases that ECJ produced 30-40 years ago.



5.2 Prototype construction

The system was written as a Python application that creates a set of static result set pages.[286] The same system performed the baseline queries (including selecting significant query terms), executing the queries with the prototype algorithm, as well as evaluating the results against the gold standard judgments.

The baseline was created by loading the text of each case into an embedded Whoosh index.[287] This index was searched and ranked using Whoosh's built-in BM25F probabilistic ranking algorithm.

For evaluating the algorithm (in it's different configurations), the metadata from all cases was extracted into a RDF database (the Sesame triple store). This was then queried using SPARQL to yield sets of cases (depending on algorithm configuration) including information of how they cited each other and treaty articles. The information was loaded into a NetworkX graph.[288] The prototype was based on the NetworkX implementations of the InDegree, HITS and PageRank algorithms.

More information about the prototype system, including information on how to download, install and recreate the results, is available in appendix A.

[286] Python is a interactive, interpreted, dynamic computer language available at *http://www.python.org/*.

[287] Whoosh is a Python library for embedded IR systems, available from *https://bitbucket.org/mchaput/whoosh/wiki/Home*.

[288] NetworkX is a Python library for graph constructing and analyzing graphs, available from *http://networkx.lanl.gov/*.

5.3 Evaluation

5.3.1 Set of information needs

Below is the set of information needs used. Each information need is based on an article in the TFEU. Corresponding numbers for the previous TEC is provided, as well as a short description of the subject matter of each article.

Art.	(TEC)	Subject matter
263	230	Provisions governing the institutions (Direct action for annulment)
101	81	Competition (Anticompetitive agreements)
267	234	Provisions governing the institutions (Preliminary rulings)
107	87	Competition (Restrictions on state aid)
108	88	Competition (Restrictions on state aid)
296	253	Provisions governing the institutions (Requirements for legal acts)
258	226	Provisions governing the institutions (Actions against member states)
45	39	Free movement of workers
288	249	Provisions governing the institutions (Adoption of secondary law)
34	28	Customs union (Quantitative import restrictions)

5.3.2 Set of baseline queries

These were the (stemmed) terms with highest TF-IDF value for each article. These terms are used with the OR combinator when constructing a query for that article. The query is run on an index stemmed in the same way, and the result set is ranked using BM25F.

Art.	Terms				
263	vi	legal	produc	brought	intend
101	undertak	concert	categori	practic	share
267	tribun	rais	question	court	pend
107	aid	certain	grant	divis	germani
108	aid	ha	grant	plan	made
296	select	proportion	type	refrain	case
258	matter	opinion	opportun	latter	observ
45	employment	public	entail	embodi	worker
288	bind	entireti	leav	address	but
34	quantit	equival	import	prohibit	restrict

5.3.3 Gold standard relevance judgments

A set of judgments considered to be highly relevant for each article was constructed. The relevance judgments were made based on whether the case was discussed at any substantial length (i.e. more than just a mention) in the standard textbook "EU Law".[289] The theory behind this method of judging was that any case considered illuminating or landmark-like would be featured in such a textbook. All cases not described in any substantial way

[289] Josephine Steiner/Lorna Woods: EU Law, 10th ed., 2009.

were considered not relevant. This includes cases that are only mentioned, not described in any detail. As a rule of thumb, if a case is only mentioned within parentheses, and the mention does not add any substantive information about the reasoning in the case, it is not considered relevant. Note that this method produced wildly differing sizes of relevant judgment sets from to around 80 cases (article 263) to a single case (article 293).

Also note that only ECJ cases are considered. Particularly within the field of competition law, the decisions of the Commission (which may be appealed to CFI and ECJ) is considered a legal source. Such decisions are not considered here.

Finally, no attempt to distinguish between different measures of relevance was made. Some cases are clearly of a more landmark-like quality. For example, concerning article 34 (the full text of which reads, "Quantitative restrictions on imports and all measures having equivalent effect shall be prohibited between Member States.", the cases 8/74 (Dassonville), 120/78 (Cassis de Dijon) and the joined cases C-267/91 and C-268/91 (Keck and Mithouard) are of enormous importance and clearly more central to understanding the rules of quantitative restrictions than e.g. case C-67/97 (Ditlev Blume). This distinction is not made in the gold standard judgment set, partly for reasons of evaluation (most evaluation metrics including MAP assume binary relevance judgments) and partly for difficulties drawing the line between "landmark-like" and simply "important" cases.

The full list for all relevant cases for the information needs, with notes, is available in Appendix B.

5.3.4 Results

These are the results of running the 12 variations of the algorithm, together with the baseline algorithm, and evaluating it against the gold standard using the MAP metric:

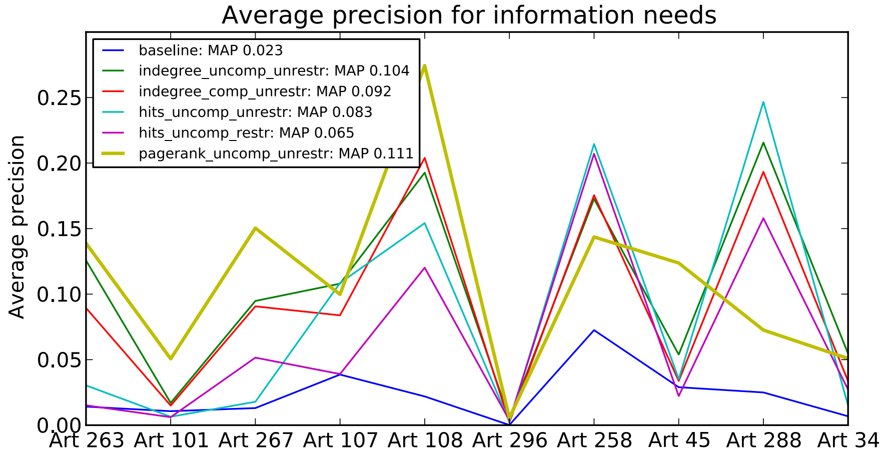
Conf	**263**	**101**	**267**	**107**	**108**	**296**	**258**	**045**	**288**	**034**	**MAP**
base	0.014	0.011	0.013	0.039	0.022	0.000	0.073	0.029	0.025	0.007	0.023
IUU	0.126	0.017	0.095	0.108	0.193	0.005	0.173	0.054	0.216	0.055	0.104
IUR	0.034	0.009	0.050	0.031	0.083	0.005	0.116	0.026	0.094	0.028	0.048
ICU	0.089	0.015	0.091	0.084	0.204	0.002	0.176	0.034	0.193	0.034	0.092
ICR	0.033	0.007	0.038	0.026	0.083	0.003	0.125	0.021	0.095	0.024	0.046
HUU	0.030	0.006	0.018	0.109	0.154	0.003	0.215	0.035	0.247	0.015	0.083
HUR	0.015	0.006	0.052	0.039	0.120	0.004	0.207	0.022	0.158	0.027	0.065
HCU	0.049	0.008	0.032	0.066	0.087	0.005	0.194	0.098	0.046	0.010	0.059
HCR	0.030	0.006	0.055	0.030	0.058	0.011	0.207	0.036	0.073	0.030	0.054
PUU	0.139	0.051	0.150	0.100	0.274	0.005	0.144	0.124	0.073	0.051	0.111
PUR	0.023	0.009	0.062	0.051	0.038	0.050	0.104	0.051	0.066	0.029	0.048
PCU	0.058	0.053	0.051	0.069	0.074	0.001	0.105	0.066	0.028	0.023	0.053
PCR	0.023	0.013	0.046	0.039	0.032	0.020	0.067	0.043	0.039	0.034	0.035

The configurations are named after the three parameters:

- Whether age compensation was used: `comp` (C) or `uncomp` (U).

- The base link analysis algorithm: `indegree` (I), `hits` (H) or `pagerank` (P).
- Whether analysis was performed on a restricted or unrestricted graph: `restr` (R) or `unrestr` (U).

Below is a graph showing the average precision for each information need for the top five performing configurations, as well as for the baseline. In the top right legend, the MAP score for each configuration is found as well.



From this graph, it is clear that link analysis-based ranking outperforms our baseline (probabilistic ranking) for our defined information needs. An anomaly in the above graph is the performance for Art. 108. This can be tracked to the extremely small set of relevant cases in the gold standard (a single case) compared to the large number of cases citing that article (921 cases). In order to get a high AP score in such circumstances, that single relevant case must place very near the top of the ranked results.