Chapter 6

Conclusion and future work

6.1 Analysis of results

Compared to the baseline algorithm, all link analysis ranking algorithms performed better on almost all information needs. This confirms our third hypothesis in sec. 1.4. But the different configurations had very varying performance.

The configuration options for age compensation and citation network restriction was introduced on the theory that they would increase performance by either making more data available (historical progression of a case's authority) or by restricting data for a higher quality dataset (only counting such citations that appear to be on topic). Neither of these options turned out to yield better performance in most cases. This could be because these options were suboptimally implemented (there might be better ways of compensating for age and to restrict an optimal citation network), or it could be because the non-compensated, non-restricted options hit a "sweet spot" in terms of available data.

When using link analysis ranking in web search engines, algorithms that estimate the authority of a node from the authority of linking nodes have a clear advantage, as not every page on the web should be equally trusted. However, at a first glance it would seem that every judgment from the ECJ should be equally trusted in terms of the how much authority one could infer from a citation. If that were the case, one would expect the InDegree-based configurations to perform as well as or better than the more complex HITS and PageRank models.

Yet, the result seem to show that one particular configuration of PageRank is slightly superior to the other configurations for most of the information needs, as well as in MAP score, although close runner-ups include two InDegree-based configurations. The relatively low performance of the HITS based configurations was surprising giving their good performance in

the US Supreme Court study by Fowler and Jeon.[290]

6.2 Analysis of method

The method of this thesis has been a combination of jurisprudence method and jurimetrical methods. The main hurdle has been to combine results from these two methods into a cohesive whole.

The fact that many of the more sophisticated definitions of relevance provide no way of actually assessing it, particularly not in graded terms and in automated ways, means that we cannot use them to build better legal IR systems. Instead, we must use what we can automatically assess, such as popularity. Assuming that this correlates with relevance, ranking on popularity may be a workable way of creating better legal IR systems. Our evaluation shows this to be the case.

The fact that popularity correlates with relevance may not be so strange when considering that citations are make based on individuals' relevance assessments. For the problem domain we have chosen, the information needs are not that unique, and other people, including ECJ judges, have had similar needs, which resulted in the citations. With this interpretation, popularity is not just a correlated value with respect to relevance; it is actually a proxy for it.

This study has dealt with a very limited set of tests, and in particular the methodology for creating the tests and the gold standard is very rudimentary. Some suggestions for improving the methods follow in the next section.

6.3 Suggestions for future work

This is, comparatively, a small study. If time and resources were infinite, it would have been improved in many directions. Some of these directions may instead suggest future work:

- **Better data**: In this study, only case citations from other cases were counted. But one could yield citation data from other sources as well. Jurisprudential literature contains a rich set of citations, and thanks to the relatively uniform way of citing ECJ cases (through case numbers), extracting this data is not that difficult. Even general news or the larger web could act as a source of citations, but since these sources less often use case numbers. The gold standard set of judgments used for evaluating could also be better, e.g. by manual compilation of EU law experts.

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

- **Better variety**: Earlier studies on citation patterns in case law has come to similar conclusions regarding general citation patterns; primarily that citation networks tend to form scale-free networks, and that these networks in turn may be useful for determining relevance. But particularly considering the different performance between different configurations of the ranking algorithm, it might prove illuminating to see if the different configurations perform similarly when run on other case citation networks, such as US or continental case law.

- **Better algorithms**: The ranking algorithm used is fairly simple. In particular, the age compensation and graph selection options are implemented in a very simple way. Both these options can be implemented in various ways, some of which may increase the performance substantially. Legal cases are rich with metadata other than citations, which can be used both to augment the citation network (by for example giving extra weight to unusual citation patterns) and to influence ranking in general. One obvious thing to try is to combine the traditional keyword-based searching using e.g. probabilistic ranking with the authority scores that citation analysis yields.

- **Better theory**: We have seen that calculating relevance from citation yields better performance, and we have elements of a theory for why that should be the case. But we have not dug deep into the legal method as practiced by e.g. ECJ, and what that method might mean for the design of ranking systems. With better jurisprudential theory, we might be able to formulate a thesis on the causation between citations and relevance, not just the correlation.

- **Better interfaces**: In end-user IR systems, one key factor for producing end-user satisfaction is to give the user a controlled overview over vast information resources. Visualizing the citation network might be a novel but useful addition to the standard legal IR system interface.

- **Better topics**: In this study, we have for simplicity assumed that one article in the TFEU correspond to one topic of interest. In reality, a topic may be represented by only a part of an article, or several articles in unison. A case may touch on several different topics. By defining topics separately from articles, and using NLP techniques on case text to identify separate topical citation networks, we might create more realistic information needs for evaluations and thus better ranking for end users.