1 A Hybrid CBR-IR Approach to Legal Information Retrieval, Edwina Rissland and Jody Daniels (Adam Wyner)

Retrieving relevant legal opinions from the large corpora has long been a key problem in common law jurisdictions, where the opinions of judges are distributed amongst the courts. Legal professionals are faced with the difficult and significant task of identifying the cases from the corpora which are relevant to a problem case. Law libraries have long been the center of law firms, containing extensive indexed collections of cases or collocations according to precedential relationships, e.g. shepardized cases. In the current period, large legal informatics companies, LexisNexis and Thomson Reuters, provide services at cost to the legal industry to support legal professionals in finding relevant opinions. The need and value of successful textual search tools has always been important to the legal industry.

[3] is set in the context of research from the late 1980s to mid 1990s, where the problems and limitations of information retrieval (IR) approaches using Boolean keyword search in corpora of legal information were starting becoming clearer and more pressing, and where the tools and techniques of Artificial Intelligence applied to legal information were becoming more well developed, e.g. case-based reasoning (CBR). Boolean keyword search was widely used for IR, yet had a range of problems - it depended on the user knowing what terms to use, how to formulate and refine the query, and there was no assurance that relevant documents were returned (recall) or that all those returned were relevant (precision). In general, such an approach had no knowledge representation. Another approach enriched the texts in the corpus with some conceptual information [1], where cases were associated with frame information that was used for search. CBR focussed on the analysis of factual aspects of cases [5, 4]. IR and CBR were at opposite ends of the knowledge representation spectrum: IR can apply to any case in a large case base, but is shallow (only textual), does not support reasoning, and does not have a strong sense of the relevance of the documents that are returned; CBR can only apply to the cases that have been annotated, but supports reasoning, and indicates high relevance. [3] represents an effort to bring these strands together by using the knowledge representation of CBR to refine IR queries. The objective is to combine the best of both - the refined queries derived from CBR can be used to quickly search large corpora using standard IR techniques.

[3] propose and develop a technique in which a problem case is represented as a generic case frame filled in with the specific facts of the case. It outputs a set of documents considered relevant to the problem case. It does this by comparing the problem case to cases in a claim lattice, which are cases sorted according to similarity and difference of case factors along the lines of HYPO [2]. The most on-point cases are selected, from among which a 'best exemplar' is selected. The full-text of this best exemplar fed into a processor which selects the top unigrams or bigrams and generates queries, which are then used to query a larger corpus of cases. A relevance feedback method is used to improve results: a user judges the relevance of the documents returned given initial queries; by tagging documents as relevant, this causes the query processors weights to be altered, which returns a more refined query.

For corpora, [3] use cases that they have previously analysed in, e.g. the 25 cases of [4]. These are used to derive the query. An "answer key" is created, being the cases that are being searched for; for one domain, there are 128 cases bearing on the home office deduction and as identified by a keyword search. The keyword search also sets a baseline return average precision (the average of precision values from amongst the different levels of recall).

They report that the approach improves search results significantly over the baseline, meaning that by using the CBR case base to support refinement of the query can improve results. Using this approach, a legal professional must still evaluation the relevance of the documents returned, but it has relieved her of formulating queries and gives her access to large document collections in a problem-specific manner.

Case-based reasoning and textual information extraction and retrieval have continued to be important topics in AI and Law [6, 7]. The fundamental motivations have, if anything, increased, for there is now greater access to a wider spectrum of legal information on the Internet. Case facts, one of the basic ingredients of common law opinions, are essential to identify, organise, and compared. However, while a variety of approaches have been applied, the core issues remain, for facts are represented in a variety of linguistic forms and represent complex knowledge. However, there has been a signficant change from the 1990s that bodes well for future progress in the area - the movement to open source data (legal corpora available unencumbered and online) and software development (where tools are made openly available for research development). In this way, researchers are able to reuse, develop, evaluate, collaborate, and integrate research as never before. Thereby, the research community can decompose the large, knowledge intensive, complex problems of legal informatics into smaller problems and engineered solutions.

References

[1] C. Hafner. Conceptual organization of case law knowledge bases. In *ICAIL '87: Proceedings of the 1st International Conference on Artificial Intelligence and Law*, pages 35–42, New York, NY, USA, 1987. ACM.

[2] E. L. Rissland and K. D. Ashley. A case-based system for trade secrets law. In *ICAIL '87: Proceedings of the 1st International Conference on Artificial Intelligence and Law*, pages 60–66, New York, NY, USA, 1987. ACM Press.

[3] E. L. Rissland and J. J. Daniels. A hybrid CBR-IR approach to legal information retrieval. In *ICAIL*, pages 52–61, 1995.

[4] E. L. Rissland and D. B. Skalak. CABARET: rule interpretation in a hybrid architecture. *International Journal of Man-Machine Studies*, 34(6):839–887, 1991.

[5] E. L. Rissland, D. B. Skalak, and M. T. Friedman. BankXX: Supporting legal arguments through heuristic retrieval. *Artificial Intelligence and Law*, 4(1):1–71, 1996.

[6] R. O. Weber, K. D. Ashley, and S. Brüninghaus. Textual case-based reasoning. *Knowledge Engineering Review*, 20(3):255–260, 2005.

[7] A. Wyner and W. Peters. Lexical semantics and expert legal knowledge towards the identification of legal case factors. In R. Winkels, editor, *Proceedings of Legal Knowledge and Information Systems (JURIX 2010)*, pages 127–136. IOS Press, 2010.