Legal Query Expansion using Ontologies and Relevance Feedback

Erich Schweighofer, Anton Geist
Centre for Computers and Law
Section for International Law and International Relations
University of Vienna, Austria

Abstract. The aim of our research is the improvement of Boolean search with query expansion using lexical ontologies and user feedback. User studies strongly suggest that standard search techniques have to be improved in order to meet legal particularities. Query expansion can exploit the potential of linguistic knowledge and successful user behaviour. First tentative results show the feasibility of our approach. A first search prototype has been built and tested in the area of European state aid law.

Keywords: Legal Information Retrieval, Ontologies, Query Expansion, Relevance Feedback

1. Introduction

Lawyers are knowledge workers and have to cope with a tremendous load of information of at least 1 GB of data (500 000 pages). In the legal domain, almost all available information is stored as text, most of the time in relatively unstructured forms (Stranieri and Zeleznikow, 2005). As work consists of solving legal problems, consultation of various texts is a prerequisite of legal work. This legal research can be outsourced to paralegals but in the very end good lawyers have to refine the quantity of relevant legal texts themselves.

Information and Communication Technology has dramatically altered legal research. Starting in the seventies with Boolean legal information systems, profiting from the internet revolution concerning on-line access, user interfaces and data handling, a very powerful and easygoing way of handling the mass of legal information was offered as the main ICT tool for legal knowledge management.

Boolean search has many advantages like rapidity, accuracy, and updating, but also one serious disadvantage. Users have to be very intelligent and highly trained in order to cope with the linguistic challenge of successful search. In order to get sufficiently good results users must know the appropriate terms and at least all synonyms, homonyms and polysems in a text corpus with more than 50 000 words. This more than Shakespearian endeavour (Shakespeare used about 20 000 words

in his works) usually ends in some failure, followed by iterative steps and frequent references to text books and commentaries.

Legal vocabularies contain open-textured terms, they are inherently dynamic. To a certain degree, this is necessary, because legal terms have to be flexible to be able to adapt to new life circumstances. Thus, legal concepts are ambiguous, their definitions vary depending on many factors like source and context. This allows for contradictions to arise from judicial problem solving. A "legal language", consisting of a complex structure of concepts, forms an abstraction from the text corpus as represented in legal databases. Such legal structural knowledge does not only contain interpretations of the meaning of legal terms, but also shows the (supposed) logical and conceptual structure. Bridging the gap between legal text archives and legal structural knowledge is a principal task of studying the law, and the key challenge in legal information retrieval.

Term frequencies do not help as much in law as in other domains. No redundancy exists in legal norms, but a lot of information is irrelevant in case law. Relevant texts parts may consist only of a short paragraph or even only of a single sentence in a very long legal document.

2. The Idea

The aim of our research is to improve the retrieval results of legal information systems.

On the one hand, we support the user with additional linguistic knowledge. In the last years, powerful legal ontologies have been developed that can be used for supporting querying as shown in the LOIS project (Dini et al., 2005). Legal text analysis has developed many methods that support the creation of ontologies.

One the other hand, we use search contexts to improve search queries. Legal information system providers have already stored information on search practices, and using query logs to improve search engine performance would be easy to implement.

Query expansion is a quite old technique for advanced search (Salton and McGill, 1986). But - unlike weighting citations ("Google's PageRank") - it never really took off, but remained in research labs.

The goal of our research project "Google the Law: Modern Text Retrieval in the Legal World" is the development of a methodology, a prototype and test applications for improved information retrieval using query expansion. This ambitious endeavour has reached the status of test applications although many improvements of our prototype are still

waiting for implementation. As a first test environment, we have chosen European State aid law.

The remainder of the paper is organized as follows: section 3 deals with related work, section 4 describes our methodology, section 5 the prototype and section 6 our test results. Last but not least section 7 draws conclusions and outlines future work.

3. Related work

Information retrieval (Salton and McGill, 1986; Frakes and Baeza-Yates, 1992) deals with the storage of documents in databases and their retrieval according to their relevancy to a query. This query is, at least in classical information retrieval systems, composed of key terms and subsequently matched with the index terms of all documents that are stored in the database. As a result to the query, the system returns those documents whose index terms match the query. It is important to note that only hints for relevant information are given.

Lawyers were eager to use information retrieval in working with the huge amounts of electronically available legal texts. It is no surprise that automated retrieval from large electronic legal document collections was one of the earliest applications of computer science to law (Moens, 2001). The limitations of information retrieval (only hints to information) and in particular of Boolean retrieval (need for exact terms and logical structure for queries) were never really liked. Single term searches seem to remain popular whereas theory considers them as quite unproductive, as they return many irrelevant hits and miss relevant ones.

Matthijssen developed a special interface for addressing four theoretical limitations in present legal information retrieval (Matthijssen, 1999): (1) the fact that the index of a database only partially describes its information contents, (2) the imperfect description of an information need by the query formulation, (3) the rough heuristics and tight closed world assumption of the matching function, and (4) the presence of the conceptual gap: the discrepancy between users' views of the subject matter of the stored documents in the context of their professional setting and the reduced formal view on these subjects as presented by information retrieval systems. Legal practitioners have to translate their information need - which they have in mind in the form of legal concepts - into a query, which must be put in technical database terms.

For the Norwegian jurisdiction (here two versions of the same language are used, Bokmål and Nynorsk), a special method called "conceptual text retrieval" was developed and is still successfully used. Queries

are described by a term class called "conceptor" consisting of a class of words representing the same idea (Bing, 1984). This idea derives from the NRCCL - Norwegian Research Center for Computers and Law - that has followed information retrieval research in law for more than 35 years and published famous books on that subject (Harvold and Bing, 1977; Bing, 1984) - and numerous articles (e.g. (Bing, 1987; Bing, 1995)).

The essential assumption of the so-called inference model is that the best retrieval quality is achieved with a ranking according to probability of relevance of the documents (Turtle, 1995). Bayesian inference nets are an elegant means of representing probabilistic dependencies and thus linguistic relations. The query representing information needs is extended via defined and computed dependencies.

Similar representations could also be achieved by a connectionist network containing nodes of terms, documents and authors. Synonym relations are represented in the nodes of terms (Rose, 1994). It may be also noted that a connectionist network seems to take most advantage of relevance feedback that may be used at a later stage of our project.

Legal publishers tried to cope with the linguistic problem by adding meta data (classification, thesauri, summaries etc.) to documents stored in legal information systems. European systems, in particular CELEX, are prominent for this approach that, however, did not get sufficient user support (Schweighofer, 2000). Users were simply not willing to learn all knowledge to use meta data. Hypertext (Bing, 1998) slightly improved the situation as browsing allowed easier use and learning in using meta data. The EUR-Lex (formerly CELEX) database still contains much meta data but it remains open if costs meet gains. Synonym lists are also partly added (e.g. in the Austrian LexisNexis system). Westlaw's WIN seems to have found the best and only solution: offer this support without interference by the user and at the highest quality available.

Ontologies (Gruber, 1993) constitute an explicit formal specification of a common conceptualization with term hierarchies, relations and attributes that makes it possible to reuse this knowledge for automated applications. The formalization must be on the one hand sufficiently powerful with regard to the knowledge representation, on the other hand it must offer functionalities for automation as well as tools to be produced automatically (see for lexically based ontologies (Hirst, 2003)).

Ontologies in law have some particularities. The motivations for the creation of legal ontologies are evident: common use of knowledge, examination of a knowledge base, knowledge acquisition, representation and reuse of knowledge up to the needs of software engineering (Bench-Capon and Visser, 1997).

After important preliminary work (e.g. (McCarty, 1989), (Hafner, 1981), (Stamper, 1991)), the frame-based ontology FBO of (Van Kralingen, 1995) and (Visser, 1995) as well as the functional ontology FOLaw of (Valente, 1995) achieved some prominence. Both were formalized with the description language ONTOLINGUA (Gruber, 1992; Gruber, 1993) and represent a rather epistemic approach. FOLaw has been used in the follow-up projects like ON-LINE, an architecture for artificial case solving, and CLIME/MILE with the test applications of classification of ships and maritime law (Winkels et al., 2002). The central difficulty of the FOLaw proved to be the modelling of the "world knowledge". The knowledge gained from FOLaw was used in the project E-Court and in the development of a core legal ontology, LRI-Core. Within the framework of this project, a flexible, multilingual information retrieval system using heterogeneous sources (audio, video, text) has been developed in the field of criminal procedure. The LRI-Core also finds experimental use in the projects E-Power (Van Engers et al., 2001) and DIRECT (Breuker and Hoekstra, 2004).

The main task of the EU-funded e-Content project LOIS (Lexical Ontologies for legal Information Sharing) was building a multi-lingual legal WordNet with concepts in six European languages for the purpose of facilitating legal information retrieval. Thus, the LOIS project focus was limited to one piece of the "cake of problems", the thesaurus problem. Up-to-date thesaurus and lexical ontologies research was used to develop a cross-lingual ontology with 5000 thesaurus entries in 6 languages in order to improve legal information retrieval (Dini et al., 2005).

In the very end, legal information systems should develop into dynamic electronic commentaries (Schweighofer, 2006) summarizing, structuring and indexing relevant legal information as required by users. Standard text books comply with this aim but are not sufficiently dynamic. Quite often, they are only updated every few years. The same methodology as described in the next section may be used for developing such electronic commentaries but for the time being ontologies and text analysis methods are not sufficiently developed for an implementation in practice.

4. Supplementing Boolean search: Query Expansion and Relevance Feedback

Our model should not replace but supplement current legal information retrieval systems. As the quality of the query is the main problem query improvement is the first logical step for improving retrieval performance.

Two methods have been developed and tested so far: query expansion using ontologies, and using relevance feedback.

4.1. Query Expansion using Ontologies

Improving the user's query with additional terms is called query expansion. For quite some time, query expansion has been seen as an effective way to improve retrieval performance (Salton and McGill, 1986). New words and phrases are added to the existing search term(s) to generate an expanded query.

In the LOIS project, some sort of query expansion was used for searching with appropriate terms in other jurisdictions. Our approach is similar but more focused on the terminology of the same legal jurisdiction. A lexical ontology was built for providing the knowledge base containing about 5500 terms, definitions and relations between concepts. Most of the terms were reused from the LOIS database; the extensions concern mostly competition law, European law and international law. It has to be noted that 3 types of relevant lexical information are stored in the database: terms, definitions and relations that could be weighted differently. The ILI concept of LOIS was also reused.

The one or two (or more) words provided in a query are searched in the knowledge base and weighted: The easy case concerns the search for a synonym. If the term exists and a synonym relation is established, a weight of 1 is given. More difficult is the case if several subterms exist. These terms are given a weight of 0.5. All meaningful terms in a definition are selected and given a weight of 0.25. All these assigned weights for terms are added. It would be fine if these weights could be reused but Boolean retrieval does not allow that. So weights greater than 1 are reduced to 1, weights greater than 0.5 are enlarged to 1 and the rest is simply not taken into account. No linguistic pre-processing besides automatic use of truncation exists at the moment.

Example: Knowledge base entry for term "animal welfare"
Animal welfare:
ILI: Tierschutz=Tierwohlfahrt (DE), le bien-être des animaux (FR), el bienestar de los animales (ES) etc.
Sub-terms: Artenschutz (DE), Tierhaltung (DE), Tiertransporte (DE), Schlachtung (DE), Tierversuche (DE)
Definition: payments for additional costs and income foregone for treatment of animals beyond the relevant mandatory standards established pursuant to Art. 4 of and Annex III of Regulation 1782/2003 (Directives

91/629, 91/630 and 98/58) and other mandatory requirements

4.2. Relevance Feedback: Using Search Context Information

In classic relevance feedback, relevance information is collected from the documents retrieved using an initial query, in order to form a second query. We think, however, that relevance feedback potential lies within the search context of the different users.

Legal information systems store - for billing purposes - accumulated information on user interactions consisting of query, results and downloaded documents. As a start, we - in our system - only consider the quantitative most important queries and documents. Even quite irrelevant terms are taken into account in order to support those with some "erroneous imagination" (e.g. the term subvention takes into account also Community support that is technically not State aid).

In the near future, this approach of relevance feedback will be tested in a sub-domain of Austrian law, tax law.

5. Prototype

The prototype consists of a database of about 1770 Commission decision on State aid in the agriculture sector covering the period of 2000 to 2006 but also the relevant guidelines and case law. 22 Community languages should be covered, however, still with strong focus on English, French, German and Spanish. This text corpus simulates an index covering all relevant sources on State aid (websites EUR-Lex, Directorates-General Competition, Agriculture and Secretariat-General). It may be noted that users get easily frustrated by the complex structures of publication (e.g. in EUR-Lex, the term "animal welfare" produces 1497 hits but relevant information can only be found if the user knows that a restriction to "Other Documents" leading to 299 documents; only if the user is aware that the Guidelines for State aid in the agriculture sector have been recently published and Commission decisions are summarised under "Summary information communicated by Member States Ĕ" then a more detailed analysis of results can be done).

This text corpus is stored in an information retrieval system (we are using askSam and the Open Source free text standalone enterprise search server Solr). The core of value-added constitutes the knowledge base containing a lexical ontology (similar to that developed in the LOIS project, stored in askSam and XML) and some statistical tools.

Quite valuable support for improving the lexical ontology provided also the GATE tools for linguistic analysis (www.gate.ac.uk). In addition to that, programs developed within the LOIS and KONTERM projects are reused if possible (e.g. term clustering using context, document classification, clustering and labelling of documents etc. (Schweighofer, 1999).

Solr is based on the Lucene Java search library providing also indexing XML documents. The Lucene Query Language is sufficiently powerful and flexible to offer standard legal search options but also query expansion ranking functions.

The overall objective of our prototype is to show that the search result quality of legal information systems can be significantly improved by using artificial intelligence and natural language processing techniques, in a first step in particular by query expansion.

6. Experimental test results

First tests have been done in the domain of State aid law using a highly sophisticated lexical ontology. Evaluation results are still tentative and mostly based on the so-called Delphi method (Linstone and Turoff, 1975). The first tests concerned the improvement of retrieval results using query expansion with synonyms in the other Community languages. The results were - not really surprising - quite good. If the knowledge base has sufficient coverage and quality it remains the best way of finding and summarising documents in other languages than the query language. Using sub-terms, umbrella terms and definition terms delivers a much higher number of relevant results, thus more information hints - but results have to be properly presented.

A typical example of our test series: A Czech farmer is displeased by high subsidies given to German farmers doing animal welfare. In particular, he does not understand why 150 euros are paid every year for each cow that has a bigger stable, can get out in free air as it wants and is offered free access to drinking water. He is considering a State aid complaint. The Czech query is extended using the ILI synsets of the other 22 Community languages (e.g. animal welfare, Tierschutz, le bien-être des animaux, el bienestar animal) but also synonyms, umbrella terms, sub-terms and definition terms and weighted accordingly (see example above). This quite complex search is done on the test information retrieval system (in practice it would be accomplished using the indexes of the various databases and websites). Relevant documents are grouped according to main term and Member State and then presented according to document type and chronological order.

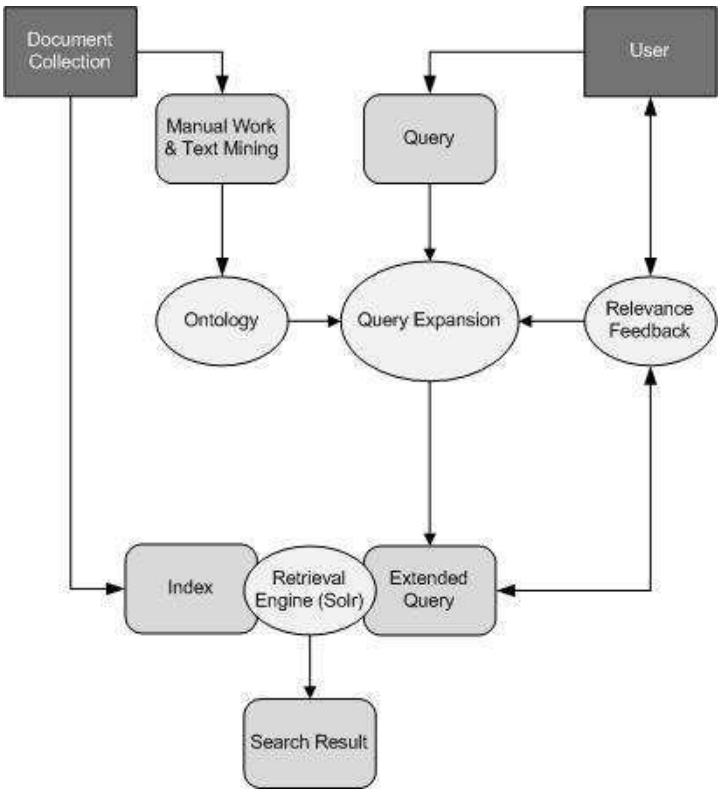

Figure 1. Sketch of Prototype IR system with query extension and relevance feedback

For improvement, some models of better presentation and visualisation are currently under examination in order to address the problem of the lower precision of the search results. The clustering of documents allows an easy browsing through Commission decisions and Member States concerning animal welfare leaving beside relevant legislation (Community guidelines for State aid in agriculture and Regulation 1698/2005). Thus, it is quite easy to find the document that is really relevant: the State aid approval of the German notification of State aid for "Gemeinschaftsaufgabe für Agrarstrukturen und Küstenschutz (Common Task for Agrarian Structures and Costal Protection)".

The main improvement consists in the much broader coverage of documents found and thus a broader scope of hints of useful information (e.g. aspects of standards, additional costs and income foregone in all relevant topics, e.g. also in agri-environment). For a practical implementation, the presentation and filtering of results seems to be decisive that has to be a strong focus in future research. Some experimental checks revealed that users may not be able to find a proper search term for the knowledge base as common language may use a different term. Here, integration of terms from relevance feedback research may help but no results of experiments can be reported so far.

7. Conclusions and future work

Our research is still quite at the beginning. A sound methodology and a first prototype are now available that are presented in the paper. At the moment, database and knowledge base are focused on the domain of State aid in agriculture. In the future, this application should be enlarged, covering the whole of EU competition law and also the general part of EU law. A text corpus exists also for international law but has to be enlarged substantially. The relevance feedback methodology will be tested in Austrian tax law. At the moment it is still too early to address questions of scaling-up as further test results are still pending. However, it does not seem insurmountable to achieve the required number of entries of a lexical database (also including ILI entries). The success of this approach depends on the quality of the knowledge base and the ability of the knowledge team to build and constantly update the lexical ontology. A (semi)automatic approach seems to be required and, therefore, tests on that will also be part of our future research.

References

Bench-Capon, T. J. M. and Visser, P. R. S. Ontologies in legal information systems; the need for explicit specifications of domain conceptualisations. In *ICAIL '97: Proceedings of the 6th international conference on Artificial intelligence and law*, ACM Press, 1997.

Bing, J., editor. *Handbook of Legal Information Retrieval*. Elsevier Science Inc., 1984.

Bing, J. Designing text retrieval systems for conceptual searching. In *ICAIL '87: Proceedings of the 1st international conference on Artificial intelligence and law.*, ACM Press, 1987.

Bing, J. Legal Text Retrieval and Information Services. In *25 Years Anniversary Anthology In Computers and Law*, TANO, 1995.

Bing, J. Text Retrieval and Hypertext: The deep Structure (opening and invited talk). In *DEXA '98 Ũ Data Bases and Expert System Applications*, Wien, 1998.

Breuker, J. and Hoekstra, R. Direct: Ontology based discovery of responsibility and causality in legal case descriptions. In *Legal Knowledge and Information Systems. Jurix 2004: The Seventeenth Annual Conference*, IOS Press, 2004.

Dini, L. and Peters, W. and Liebwald, D. and Schweighofer, E. and Mommers, L. and Voermans, W. Cross-lingual legal information retrieval using a WordNet architecture. In *ICAIL '05: Proceedings of the 10th international conference on Artificial intelligence and law*, ACM Press, 2005.

Frakes, W. B. and Baeza-Yates, R., editors *Information Retrieval: Data Structures and Algorithms*. Prentice Hall PTR, 1992.

Gruber, T. R. Ontolingua: A mechanism to support portable ontologies. Stanford University, Knowledge Systems Laboratory, Technical Report KSL-91-66, 1992.

Gruber, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

Hafner, C. D. *An Information Retrieval System Based on a Computer Model of Legal Knowledge*. UMI Research Press, 1981.

Hirst, G. Ontology and the Lexicon. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies in Information Systems*, Springer, 2003.

Harvold, T. and Bing, J. *Legal decisions and information systems (Publications of the Norwegian Research Center for Computers and Law)*. Universitetsforlaget, 1977.

Linstone, H. and Turoff, M., editors *The Delphi Method: Techniques and Applications)*. Addison Wesley, 1975.

McCarty, L. T. A language for legal Discourse I. basic features. In *ICAIL '89: Proceedings of the 2nd international conference on Artificial intelligence and law*, ACM Press, 1989.

Matthijssen, L. *Interfacing Between Lawyers and Computers:An Architecture for Knowledge-Based Interfaces to Legal Databases (Law and Electronic Commerce)*. Kluwer Law International, 1999.

Moens, M. F. Innovative techniques for legal text retrieval. *Artificial Intelligence and Law*, 9(1):29–57, 2001.

Rose, D. E. *A Symbolic and Connectionist Approach To Legal Information Retrieval*. LEA Inc., 1994.

Salton, G. and McGill, M. J. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1986.

Schweighofer, E. *Legal Knowledge Representation:Automatic Text Analysis in Public International and European Law (Law and Electronic Commerce)*. Kluwer Law International, 1999.

Schweighofer, E. *Wissensrepräsentation in Information Retrieval-Systemen am Beispiel des EU-Rechts (Dissertationen der Universität Wien)*. WUV, 2000.

Schweighofer, E. Computing Law: From Legal Information Systems to Dynamic Legal Electronic Commentaries. In C. M. Sjöberg and P. Wahlgren, editors *Festskrift till Peter Seipel*, Norsteds Juridik AB, 2006.

Stranieri, A. and Zeleznikow, J. *Knowledge Discovery from Legal Databases (Law and Philosophy Library)*. Springer, 2005.

Stamper, R. K. The Role of Semantics in Legal Expert Systems and Legal Reasoning. *Ratio Juris*, 4(2):219–244, 1991.

Turtle, H. Text retrieval in the legal world. *Artificial Intelligence and Law*, 3(1):5–54, 1995.

Valente, A. *Legal Knowledge Engineering, A Modelling Approach*. IOS Press, 1995.

Van Engers, T. and Gerrits, R. and Boekenoogen, M. and Glassee, E. and Kordelaar, P. POWER: using UML/OCL for modeling legislation - an application report. In *ICAIL '01: Proceedings of the 8th international conference on Artificial intelligence and law*, ACM Press, 2001.

Van Kralingen, R. W. *Frame-Based Conceptual Models of Statute Law.* Kluwer Law Intl, 1995.

Visser, P. R. S. *Knowledge Specification for Multiple Legal Tasks: A Case Study of the Interaction Problem in the Legal Domain.* Kluwer Law Intl, 1995.

Winkels, R. and Boer, A. and Hoekstra, R. CLIME: Lessons Learned in Legal Information Serving. In *ECAI 2002: Proceedings of the 15th Eureopean Conference on Artificial Intelligence*, IOS Press, 2002.

Zeleznikow, John and Hunter, Dan *Building Intelligent Legal Information Systems (Computer Law, No 13).* Springer, 1994.