The Open Revolution: Using Citation Analysis to Improve Legal Text Retrieval

Anton Geist[*]

I. Introduction

The legal discipline is an unusually information-rich one. Contrary to other disciplines, there is not only an enormous mass of texts about the law, but also the law itself is textual in nature.[1]

Consequently, being able to automatically retrieve texts from large document collections was one of the first applications of computer science in the field of law.[2]

Today, when legal professionals log onto online databases of legal service providers in order to retrieve the text of court cases, statutory material or other legal documents, the process of obtaining those documents is called legal text retrieval. Put simply, legal text retrieval systems like Westlaw (United States, United Kingdom), LexisNexis (United States, United Kingdom and other countries), Beck-Online (Germany) and RDB (Austria) provide electronic access to legal documents and enable users to search within their text collections.

Citation analysis looks at citations to and from documents. The theoretical foundation on which the indexing of citations is built upon, is as follows: if there is a citation between two documents, there is some kind of relationship between these texts.[3] This relationship can be further explored and used to learn more about the characteristics of the connected documents.

Citation analysis of World Wide Web documents enabled the Internet search engine Google to outperform all other Web search engines that did not yet make use of the citation structure of the Web. Google decides on the relevance of Web pages to users' queries not only because of the text of those Web pages, but also because of their link structure. So far, no similar technological change has occurred in the area of legal text retrieval.

II. Legal Text Retrieval Systems Today

[*] University of Edinburgh, United Kingdom

[1] **R. WIDDISON**, “New Perspectives in Legal Information Retrieval”, *Int J Law Info Tech*, *10*(1), 2002, pp. 41–70, at p. 4.

[2] **M.-F. MOENS**, “Innovative techniques for legal text retrieval”, *Artificial Intelligence and Law*, *9*(1), 2001, pp. 29–57, at p. 29.

[3] **J. FEATHER & R.P. STURGES**, *International encyclopedia of information and library science* (2. ed), London: Routledge, 2003,at p. 76.

When we take a closer look at the legal text retrieval systems just mentioned, it is worth trying to put their use in a wider perspective.

The global market for legal text retrieval is big, with revenues of about EUR 3 billion per year to a vast number of online services around the world.[4]

After some experimental work, both in the USA and in Europe, the first legal text retrieval systems were implemented in the 1970s. Today it would be difficult to imagine the practise of law without the use of online legal databases, be it to retrieve legislative texts, court cases or legal literature. Distributed computing made it possible for lawyers to perform so-called computer-assisted legal research tasks from their own desks, rather than from a shared terminal.[5]

When commercial legal search engines were introduced, they were custom-built, often proprietary, software. Professional legal users have been expected to undergo some kind of special training in order to effectively use the specific legal databases.[6]

Technologically speaking, legal database providers around the world generally still use the same search technologies that they did when the databases were built. Those technologies, that belong to the "extended Boolean model", require the users to formulate strict queries, using so called Boolean operators (AND, NOT, OR, and others) to tie together individual words or phrases that form the search query.[7]

A characteristic of Boolean retrieval models is their "exact match" nature. They do not take into account any partial matches of queries. The search engine will return results that fulfil all the criteria defined by the query, and only those. If a document fulfills all but one requirement, it will still not be returned as a result. Besides that, (extended) Boolean retrieval models do not employ any sophisticated technologies to rank the listed results.

Consequently, legal database providers still (have to) provide special training for their users outlining the peculiarities of Boolean syntax, proximity operators and field searching.

The main legal database providers have not ignored all technological advances in search technology in the past decades. It appears, however, as if both providers and customers have somehow agreed on settling for search technology that would be considered to be dated in most other disciplines.

[4] **L. KLASÉN**, “Legal IR services - from past to present”, In C. Magnusson Sjöberg (Ed.), *Legal Management of Information Systems - incorporating law in e-solutions,* Lund: Studentlitteratur, 2005, pp. 337–358, at p. 338.

[5] **C. TAPPER**, “Out of the box”, *International Review of Law, Computers & Technology*, *19*(1), 2005, pp. 5–11, at p. 8.

[6] **P. JACKSON & I. MOULINIER**, *Natural language processing for online applications: Text retrieval, extraction and categorization* (2nd rev. ed., Vol. 5), Amsterdam Philadelphia: John Benjamins Pub, 2007, at pp. 64-65.

Let us look at the example of Westlaw: The database provider is the largest commercial legal search service in terms of the number of paying subscribers. Over half a million subscribers perform millions of searches a day over tens of terabytes of text data.[8]

Today, Boolean search (called “Terms and Connectors” by Westlaw) is still the default search mode on Westlaw, and used by a large percentage of users. A technologically more advanced ranked retrieval search mode (called “Natural Language” by Westlaw) is available, but most users prefer the Boolean search.[9]

One is tempted to think that the legal material is simply not suited for more advanced search technologies. If the described Boolean legal text retrieval systems were performing well as legal research tools, there would really be nothing wrong with legal text retrieval using long established search methods. Ironically, however, the deficiencies of Boolean legal text retrieval have been known for decades.

The so-called STAIRS study conducted by Blair and Maron in the mid-1980s was the first legal retrieval experiment in a realistic operational environment. The searchers had a predefined goal to work on their queries until they were confident that the search had retrieved at least 75 percent of the total number of relevant documents. Further investigation, however, proved that while the result lists did not include a lot of off-topic results (“precision” was 79 percent), the average “recall” achieved by the Boolean system was no better than 20 percent. In other words: When the searchers thought that they had built a query that would give them 75 percent of all the relevant documents, the text retrieval system was only able to locate as little as 20 percent of those relevant documents.[10]

Although the Blair and Maron study has been triggering a passionate debate on the effectiveness of Boolean information retrieval systems in large full-text databases, legal text retrieval systems still build upon the same search technology that they did in the mid-1980s.

III. The Need for Better Legal Text Retrieval

A thorough discussion of why both providers and customers of legal text retrieval systems have (implicitly) accepted using dated search technologies is far too complex to be dealt with within this paper. A mix of several characteristics seems to be responsible for this. Let me just mention one of the major reasons for this fact. It not only explains why people do

[7] **C.D. MANNING, P. RAGHAVAN & H. SCHÜTZE**, *Introduction to information retrieval.* New York: Cambridge University Press, 2008, at pp. 14-15.

[8] **C.D. MANNING, P. RAGHAVAN & H. SCHÜTZE**, *Introduction to information retrieval,* at pp. 14-15.

[9] **P. JACKSON & I. MOULINIER**, *Natural language processing for online applications,* at p. 28.

[10] **D.C. BLAIR & M.E. MARON**, “An evaluation of retrieval effectiveness for a full-text document-retrieval system”, *Commun. ACM, 28*(3), 1985, pp. 289–299.

not use more up-to-date technology in the legal domain, but also shows very clearly why this situation must be changed.

Attorneys have the legal obligation to provide competent representation to their clients. In the U.S., but also in Europe, it has long been decided that the ability to perform adequate legal research is a component of this legal obligation. If an attorney fails to perform competent research, this constitutes a violation of their ethical duty and has led to malpractice suits against negligent lawyers.

As new computerised research tools have become available to lawyers, the standard of competence for attorneys concerning what "adequate legal research" means has arguably been increased.[11]

This puts the providers of legal text retrieval systems in a difficult situation: They want to provide their customers with electronic research tools, but at the same time they are eager to make sure that (sued) attorneys cannot argue that it was not them, but the search engine that failed to do "adequate legal research". The situation of the systems' users, on the other hand, is quite special too: They are certainly interested in tools that facilitate their legal research, but at the same time they want to avoid any increase in their standard of competence. A silent agreement that old search technology is "good enough" seems to be a plausible solution for both parties.

As soon as we remind ourselves that it is really not only the two mentioned parties who have to deal with the effects of more or less advanced search technology, it becomes obvious why it is in fact a necessity to use the most sophisticated search tools available. Legal documents build the foundation of every legal system. Better retrieval of legal texts effectively improves the work of all legal professionals, be it the judiciary, lawyers or legal researchers. This improvement, in turn, is useful for everyone who is affected by the legal system, which is society as a whole.

IV. Importance of Citations for Legal Research

For a long time lawyers have consequently developed very sophisticated forms of citation handling. One just has to think of the various manuals of legal citations that are published, contrary to other disciplines. Part of the reason for the great importance of citations

[11] **M. WHITEMAN**, “The Impact of the Internet and Other Electronic Sources on an Attorney's Duty of Competence Under the Rules of Professional Conduct”, *Albany Law Journal of Science and Technology, 11*, 2000-2001, pp. 89–103,at p. 90.

in legal texts lies in the nature of legal texts themselves. There is no physical "legal object" that is described in texts, the texts themselves constitute the law.[12]

When judges write opinions, they perpetually cite cases and other authorities. Legal scholars write articles and treatises that cite cases and other authorities, and which in turn are, at least sometimes, cited by cases and other authorities. These citations hold a great amount of information in them. Judges cite those cases that they think are the most relevant ones to the case they are deciding. Therefore, when two judges who are deciding different cases, cite some of the same authorities, this does mean that those cases are, at least somehow, relevant to each other.[13]

In common law countries, because of the eminent importance that citations play in those legal systems, there are specific tools and services that aim to assist attorneys in citation research. So-called Citator services allow users to see all citations that directly refer to a given case. Globally, the two biggest Citator services are LexisNexis's Shepard's, and WestLaw's KeyCite.[14]

But citators not only allow researchers to verify the authority of a precedent by listing subsequent sources that have cited a source. They are also used by legal professionals to find additional sources relating to a given subject, by using the compiled citations as references to (somehow) related material.

Online legal databases certainly do provide access to citator services online, but given the enormous practical use of citators, it seems strange that they do not make use of legal citations to improve their (keyword) search results. So far, the result lists of legal text retrieval systems like Westlaw or LexisNexis are based solely on word occurrences, not on citations between legal documents.

V. Using Citations for Legal Text Retrieval

Taking into account the importance of citations in legal research and the deficiencies of current legal text retrieval systems, it appears to be just a question of time that legal databases will have to make use of citation analysis of their document collection to improve their search results. A few observations further underline this claim.

[12] **C. TAPPER**, "Citation Patterns in Legal Information Retrieval", at p. 258.

[13] **T.A. SMITH**, "The Web of Law", *The San Diego Law Review, 44*, 2007, pp. 309–354,at p. 341.

[14] **P. ZHANG & L. KOPPAKA**, "Semantics-based legal citation network", In *Proceedings of the 11th international conference on Artificial intelligence and law,* Stanford, California: ACM, 2007, pp. 123–130, at p. 123.

Citation analysis is part of the academic science of bibliometrics, which is a set of methods used to study or measure texts and patterns of publication.[15]

In his 1955 paper Eugene Garfield, one of the founders of bibliometrics, actually used the already existent legal citation research tool, Shepard's Citations, to explain his then futuristic ideas of citation indexes for science.[16]

Today, those citation indexes have long become reality in many research areas and their use has been expanded to improving text retrieval systems in various areas. In the legal domain, however, the use of Shepard's citations as a mere reference collection has stayed the same, except for the possibility to use Shepard's online. The use of Shepard's has therefore remained limited, while the tools that it inspired have been, among other areas, been used to improve the performance of text retrieval in their respective areas.

Thus, it seems strange that legal citations are not used for indexing. Every single citation that an author includes in their texts can also be seen as an act of indexing. When authors put references in their work, they include terminological interpretations in their texts.[17]

Also, returning to the comparison of modern Web search and legal text retrieval, I want to highlight one more aspect of legal citations. Google has shown us that link analysis in Web search has the potential to greatly improve search results.

As already shown, however, connections between documents are more central to the discipline of law than they are to any other field. Lawyers, legal scholars, and judges all pepper their writings with links to earlier sources, the only difference being that these links are called "citations" or "quotations."[18]

In fact, one might even argue that tightly linked court cases have to be even more closely related than equally linked Web pages: Citations within court decisions are nothing but arguments themselves. Citations have to persuade higher courts that the decision the judges made was correct. Therefore, judges have strong motivations to include citations thoughtfully.[19]

VI. Case Study: Austrian Supreme Court Decisions

[15] **. FEATHER & R.P. STURGES**, *International encyclopedia of information and library science* (2. ed), at p. 38.

[16] **E. GARFIELD**, “Citation indexes for science; a new dimension in documentation through association of ideas”, *Science, 122*(3159), 1955, pp. 108–111, at p. 108.

[17] **E. GARFIELD**, “Citation indexes for science; at p. 110.

[18] **F.R.. SHAPIRO**, *Collected papers on legal citation analysis*, Littleton Colo.: F.B. Rothman., 2001, at p. 161.

[19] **T.A. SMITH**, “The Web of Law”, at p. 345.

In order to show that there is a direct correlation between the number of citations to legal documents and their legal impact, I have started to conduct a study with Austrian Supreme Court cases.

At www.ris.bka.gv.at, the Austrian Federal Chancellery operates the Legal Information System of the Republic of Austria ("Rechtsinformationssystem", abbreviated "RIS").

The objectives of RIS have been to provide up-to-date and exhaustive legal information in an electronic format to both state organs and the general public.[20]

Another objective of providing "cost effective legal information"[21] has become obsolete because since 1997 the respective content of RIS has been accessible for everyone free of charge via the Internet.

The RIS application "Case Law" ("Judikatur", http://www.ris2.bka.gv.at/Judikatur/) contains - among other documents - the full texts of the Austrian Supreme Court decisions. The RIS system is, however, able to provide not only the full texts of the decisions, but also headnote documents created by the Publication Office of the Supreme Court. Each decision document is intellectually processed by a legal specialist at the court, and if a decision has introduced a new interpretation of Austrian legal statutes, special headnotes called "Rechtssatzdokumente" are created and cite to the respective court decisions.

Using a document collection that consists of more than 100,000 full-text decision and headnote documents, I have constructed a network of all Austrian Supreme Court cases since 1985.

Although various practical problems did arise, my methodology has been quite simple so far: The computer code I developed counts the number of citations from within headnote documents to Supreme Court decisions, and ranks the court decisions according to their citation totals. By doing that, I was already able to show - for the first time - that this network is a scale-free one, which means that a few court decisions have many headnotes attached to them, while most decisions have only a few headnotes, or none at all. This observation by itself already suggests further looking into the possibility of using Web search ranking algorithms in legal information retrieval systems, because the network structure of the World Wide Web is a scale-free one as well.

I am now turning to traditional indications of impact concerning Supreme Court cases, including the publication in an official legal reporter or high citation counts in annotated

[20] **BUNDESKANZLERAMT**, *Rechtsinformationssystem des Bundes: RIS ; eine kurze Einführung,* Wien: Bundeskanzkeramt-Verfassungsdienst, 1994, at p. 1.

[21] **BUNDESKANZLERAMT**, *Rechtsinformationssystem des Bundes: RIS ; eine kurze Einführung,* at p. 1.

codes. By examining the cases returned by those means of traditional legal research with respect to their position in the "Web of Supreme Court Cases", I will be able to further explore the exact nature of the correlation between the number of headnote citations to Supreme Court decisions and their legal impact.

VII. Conclusions

Court cases, statutes and other legal authorities are linked together by citations. This legal citation network is in fact not only extremely large, but also one of the best-documented existing human-created networks.[22]

Legal text retrieval systems have not made any major changes to their search technologies within decades, although the shortcomings of current legal text retrieval have been well-documented.

This paper has set out that possible improvements for legal text retrieval constitute a necessary area for research, and that legal citation analysis can serve as a means to improve current systems.

Using only freely available data from the Legal Information System of the Republic of Austria, I have been able to show that the network structure of Austrian Supreme Court decisions and their headnotes is similar to the one of the World Wide Web. This already suggests the general feasibility of using Web ranking algorithms in legal information retrieval systems. The next step will be the computation of a more sophisticated automated ranking of Austrian Supreme Court cases that is fully in line with methods of traditional legal research.

[22] **T.A. SMITH**, "The Web of Law", at pp. 310-311.