USING CITATION ANALYSIS TECHNIQUES FOR COMPUTER-ASSISTED LEGAL RESEARCH IN CONTINENTAL JURISDICTIONS

by

Anton Geist

Submitted for the degree of LLM by Research

The University of Edinburgh

2009

Table of Contents

Table of Figures

Abstract

The following research investigates the use of citation analysis techniques for relevance ranking in computer-assisted legal research systems.

Overviews on information retrieval, legal research, computer-assisted legal research (CALR), and the role of citations in legal research enable the formulation of a proposition: Relevance ranking in contemporary CALR systems could profit from the use of citation analysis techniques. After examining potential previous work in the areas of Web search, legal network analysis, and legal citation analysis, the proposition is further developed into a testable hypothesis: A basic citation-based algorithm, despite all its shortcomings, could be used to significantly improve relevance ranking in computer-assisted legal research. By computing and analysing the distribution of 242,078 headnote citations across 80,195 opinions written by the Austrian Supreme Court of Justice between 1985 and 2008, proof for this hypothesis is presented.

Declaration

This thesis has been composed by myself and has not been submitted for any other degree or professional qualification. The work reported was carried out by me, except where otherwise acknowledged.

Anton Geist

September 2009

Acknowledgements

Many people provided assistance while I have been working on this thesis. By offering to bear the vast majority of my expenses, my parents made my decision to pursue this research project much easier.

In terms of assistance with my actual research, three academics deserve to be pointed out specifically: Burkhard Schafer agreed to supervise this research topic, and provided continuous guidance throughout the year. Ulrich Bayer, a computer scientist, spent a vast number of hours coaching me on Python programming. Tamsin Maxwell offered valuable feedback from a computational linguist's perspective over many interesting lunches.

I have also been fortunate to get feedback and/or assistance from experts who have been in the field for decades: Jon Bing, Daniel Dabney, Graham Greenleaf, Erich Schweighofer, and Colin Tapper.

Throughout this year, I had the opportunity to present parts of this research at four conferences where I was welcomed by most sympathetic and open-minded people. I am grateful for all the suggestions that I could incorporate into my research:

- *The future of... Conference*, October 2008, Florence, Italy
- *IRIS 2009*, February 2009, Salzburg, Austria
- *Governance of New Technologies Conference*, March 2009, Edinburgh, Scotland
- *BILETA 2009*, April 2009, Winchester, England

I also want to express my deep gratitude to my partner, Alexandra Becker, for 8 years of understanding and support.

Finally I would like to thank you, my reader. I put this thesis online on my personal website (http://www.antongeist.com) as well as on the Social Science Research Network SSRN so that others can profit from my research. In return, I hope that you will be so kind to provide me with feedback at home@antongeist.com.

1. Introduction

"The coal miner with pick and shovel mined one ton per day a few years ago. Today he brings out over a hundred tons per day. The lawyer is still in the pick and shovel era."[1]

Louis O. Kelso, 1946

One year before Louis Kelso called for a "technological revolution" in law, Vannevar Bush had published a seminal paper entitled "As We May Think". In it, Bush shared his belief that due to an exponential increase of available information, no one was any longer able to make use of it by solely manual means.[2] By doing so, Bush was the first scientists to recognise "information overload" which describes a situation when we do not have too little, but rather too much information available to make sensible decisions. "As We May Think" fuelled research on electronic means to combat information overload[3], an area of interest which has been known as "information retrieval". Both academic and commercial information retrieval research continue to this day and have been constantly rising in their importance because "information defines one of the fastest growing markets on our planet. The issue is no longer lack of information, but an embarrassment of riches, and a lack of tools for organizing information, finding it, or selling it at the right price and the right time."[4] I give an overview about information retrieval in Chapter 2 of this thesis, starting the "Framework" section of my work.

Lawyering is a highly information-intensive profession[5], we might even argue that it is more information-intensive than any other industry or profession.[6] Therefore, in essence, legal research consists of the retrieval of relevant legal information.[7] Chapter 3 introduces and tries to compare and contrast legal research in common and civil/continental jurisdictions. For the latter, I focus on the situation in Austria, my home country. A legal adaptation of Bush's 1945 vision of an electronic "Memex" device that should support information retrieval, the "Lawdex", was conceived by Louis Kelso only months after Bush's

---

[1] Kelso, 1946, p. 392

[2] Bush, 1945, p. 101

[3] Herskovic, Iyengar, & Bernstam, 2007, p. 93

[4] Jackson & Moulinier, 2007, p. 3

[5] Halvorson & Basch, 2000, pp. ix–x

[6] Susskind, 1998, p. 79

[7] Smith, Gelbart, Maccrimmon, Atherton, Mcclean, Shinehoft et al., 1995, p. 57

aforementioned publication.[8]

This is where legal research met information retrieval, and computer-assisted legal research (CALR) was born. John Horty's 1959 project[9] to store health statutes on magnetic tapes at the University of Pittsburgh is commonly referred to as the first operational CALR system. The scope of CALR systems has been widening ever since, and I give a tour on computer-assisted legal research in Chapter 4. Again, differences between common and civil/continental jurisdictions are pointed out. Along the way, I present evidence for my conviction that current CALR systems have considerable room for improvement. Simon Chester provided an excellent summary of what I consider to be the basic problem in 1992. Tellingly, it is as valid today as it was back then:

"Whether you use Lexis or Westlaw, simply working out an effective search requires that you translate a legal problem into a line that looks like this:
offense crime violation /s crash accident /s 'parking lot'
Not how most lawyers talk or think."[10]
[Lexis(Nexis) and Westlaw are the two large, US-based CALR system providers]

In Chapter 5, I look at the use of citations in legal research. This last part of the "Framework" section also provides you with background information necessary so that you can follow and critically assess the proposition that I develop afterwards. More than twenty years ago, Daniel Dabney explicitly stated that in the legal domain "we cannot read all of the documents that might contain relevant information, so we rely on others to read the documents for us, and to note for us the texts that we will need to consult in the future."[11] I fully agree with Dabney. I think, however, that by now we have long gone past the point when it could still be other humans who pre-processed all potentially relevant documents for us. In my opinion, we have in reality long become dependent upon computer algorithms to help us fight legal information overload. Not making use of them involves very high risks: "Overload of information [...] has the potential to undermine law if something is not done"[12], Tamsin Maxwell and Burkhard Schafer recently observed.

This is the jumping-off place for my proposition (Chapter 6): Relevance ranking in contemporary computer-assisted legal research systems, I hold, could be improved by using

---

[8] Kelso, 1946, p.387

[9] Horty, 1959, p.31

[10] Chester, 1992, p.111

[11] Dabney, 1986a, p.6

[12] Maxwell & Schafer, 2008, pp. 63–64

citation analysis techniques.

In the following "Elements of a Theory" section I give an overview about three areas of research that I consider to be helpful for further developing the proposition. We look at Web search (Chapter 7), legal network analysis (Chapter 8) and legal citation analysis (Chapter 9). The experiences gained in each of the three areas further suggest that using citations should indeed be a promising method of improving relevance ranking and therefore search as a whole in an electronic legal research environment.

After the three chapters of the "Elements of a Theory" section, it is possible to further develop the proposition into a testable hypothesis (Chapter 10). I do this because one of the objectives with this thesis is to let real-world data speak for themselves. The third section "Testing My Hypothesis" is therefore mostly empirical in its nature.

In Chapter 11, the citation distribution of opinions written by the Austrian Supreme Court of Justice between 1985 and 2008 is computed in order to perform two experiments. In the first experiment, I align 80,195 opinions according to the number of citations that they receive from so-called headnote documents, using a total of 242,078 headnote citations. The second experiment sets out to test whether the computed citation distribution could successfully be used to prioritise "relevant" opinions in CALR relevance ranking.

I reach a final conclusion about my hypothesis in Chapter 12, and conclude the thesis by giving some ideas about possible avenues for future research (Chapter 13).

SECTION I - THE FRAMEWORK

The "Framework" section is organised to familiarise you with the key terminology and concepts of those areas that make up the background against which my proposition - and ultimately my hypothesis - of using citation analysis techniques for computer-assisted legal research will be set.

2. Information Retrieval

It was IBM computer scientist Hans Peter Luhn who, in the late 1950s, first[13] suggested that automatic text retrieval systems could be implemented based on a special comparison. A comparison, that is, of content identifiers attached both to electronically stored documents, as well as to search queries submitted by users. Generally, certain words extracted from the texts of documents and queries would be used as those content identifiers.

For the decades that followed, most people associated electronic information retrieval with librarians, or specialised business and legal analysts. Only those information specialists electronically worked with proprietary (online) information services in order to retrieve documents. Especially the development of the World Wide Web has changed this picture completely. Now, every one of us is his or her own document retriever, and we all work with search technology on a daily basis.[14] A Web search engine is nothing short of the embodiment of modern information retrieval.[15] This is why I use the terms "information retrieval" and "search", and also "retrieval system" and "search engine", interchangeably. Strictly speaking, this might be terminologically incorrect, but I think there is no need for introducing information retrieval subtleties in the context of this thesis.

2.1. Definitions of Information Retrieval

As Dietmar Wolfram observes, definitions of information retrieval abound.[16] Luhn's

---

[13] Salton & Buckley, 1988, p. 513

[14] Jackson & Moulinier, 2007, p. 23

[15] Weiss, Indurkhya, Zhang, & Damerau, 2005, p. 58

[16] Wolfram, 2003, p. 10

just mentioned suggestion provides the jumping-off place for a first approximation: "Information retrieval, referred to as "IR" by its practitioners, tries to retrieve relevant documents in response to a query."[17] Inspired by a similar decision of Peter Jackson and Isabelle Moulinier[18], when I talk about "information retrieval" or "search" in this thesis, I concentrate on document retrieval by full text search. This means that the user's query is matched against the actual texts of the stored documents, rather than against a set of keywords only. Full text searching, if you like, is the electronic equivalent to a huge (imaginary) back-of-the-book index in which you can look up every word of the book. A definition of information retrieval found in a leading textbook by Christopher Manning and others reads as follows:

"Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)."[19]

While Manning and others explicitly take into consideration that IR does not necessarily include computers, and that the stored elements do not necessarily have to be texts, Elizabeth Liddy takes a more relaxed approach:

"Document retrieval (more commonly referred to as "information retrieval" by researchers in the field) is the computerized process of producing a list of documents that are relevant to an inquirer's request by comparing the user's request to an automatically produced index of the textual content of documents in the system. These documents can then be accessed for use within the same system."[20]

2.2. Querying Models & Retrieval Models

Since Luhn's vision has been realised in the mid-1960s, various information retrieval models have been developed. Using them as categories, we can pinpoint an information retrieval project within the IR framework. A retrieval model indicates the document representations used and how they are matched - or compared - during the retrieval process.[21] Retrieval models are complemented by query(ing) models which deal with the

[17] Weiss et al., 2005, p. 85

[18] Jackson & Moulinier, 2007, p. 23

[19] Manning, Raghavan, & Schütze, 2008, p. 1

[20] Liddy, 2006, p. 748

[21] Pritchard-Schoch, 1993, p. 34

different ways in which search queries can be formed.

Retrieval Models in Information Retrieval

When talking about retrieval models, we often distinguish between three basic computer-aided techniques for searching information retrieval collections: Boolean models, vector space models, and probabilistic models.[22] I will not discuss the details of these three systems. One distinction among retrieval models is essential, however, when talking about search systems in the area of computer-assisted legal research. In his seminal book on "legal information retrieval" (what I decide to call "computer-assisted legal research"[23]), Jon Bing distinguishes between systems using "identity functions" (mentioning the Boolean system as one example) and systems using "nearness functions".[24] As much as I agree with the huge importance of the distinction itself, I believe that the category names "exact-match" (retrieval) models and "best-match" (or "partial-match") retrieval models as (for example) Keith van Rijsbergen uses them[25] should be preferred, as they are much more descriptive. Howard Turtle explains that by far the most common exact-match model is the Boolean model.[26]

Exact-match retrieval models use a way of matching the query against the document collection that partitions the stored texts into two sets, namely into those documents that match the query and into those that do not. A matching procedure like this is generally simple and efficient, and this is the reason why exact-match models have been forming the basis of many commercial retrieval solutions.[27] While commercial information retrieval systems have been relying on the Boolean retrieval model, researchers have at the same time been suggesting a number of alternative retrieval models[28], namely the just mentioned vector space and probabilistic retrieval models. Those two models are also collectively referred to as "best-match" (or "partial-match") models, which explains the exact-match / best-match distinction.

In best-match retrieval models, a document does not (necessarily) have to exactly match the query in order to be included in a result list. Documents are always returned in a ranked

---

[22] Langville & Meyer, 2006, p. 14

[23] See below, 4. Computer-Assisted Legal Research, p. 27

[24] Bing, 1984b, pp. 161; 164

[25] Rijsbergen, 1981, pp. 1–2

[26] Turtle, 1995, p. 24

[27] Turtle, 1995, p. 24

[28] Robert M. Losee, 1999, p. 882

order, according to their similarity with the query.[29] Despite the commercial introduction of those more sophisticated search systems especially on the Web, Boolean systems remain popular in many commercial and library applications.[30] In particular, this is also true for the legal domain.

Query(ing) Models in Information Retrieval

Retrieval models deal with methods and systems to relate a query to a document collection, query(ing) models account for the various query(ing) languages that exist.[31] When a user utilises an information retrieval system by entering a query that connects search terms with operators, such as AND, OR, and NOT, he or she is using a Boolean query(ing) language. This is sometimes also called a "terms and connectors" search, because a sharp distinction is made within the query between content-bearing terms and content-free operators based on Boolean logic.[32] When computers became more widely accepted, it was Boolean logic that was applied to information retrieval.[33]

As their names already suggest, the Boolean retrieval model and the Boolean query(ing) model work together very well. Boolean search statements can be applied to large sets of unstructured data easily, and the results exactly match the search terms and logical constraints put in place by the operators.[34] This is what has made exact-match, Boolean logic search engines so popular that they can still be considered the standard search model in many areas. Boolean querying coupled with Boolean retrieval, then often called "Boolean searching", is called exact-match because, for example, all concepts linked with an AND operator in the query must be present in a document for that document to be successfully retrieved. Documents that only contain, let's say, three out of four terms connected with AND, are just as lost as documents that contain only one, or even none of the query terms.[35]

This already hints at one of the major downsides of Boolean searching, its literalness. I will cover this and other Boolean IR issues in detail when talking about computer-assisted

---

[29] Paul & Baron, 2007, p. 22

[30] Jackson & Moulinier, 2007, p. 28

[31] Matthijssen, 1998, p. 82

[32] Jackson & Moulinier, 2007, p. 27

[33] The Sedona Conference Working Group on Best Practices for Document Retention and Production (WG1), 2007, p. 217

[34] The Sedona Conference Working Group on Best Practices for Document Retention and Production (WG1), 2007, p. 217

[35] Tenopir, 1993, p. 55

legal research systems.[36]

If users want to effectively use Boolean-based systems, they must be familiar with both Boolean retrieval operations and Boolean query construction.[37] The combination of Boolean retrieval with Boolean querying is not at all, however, the only one that we find today. Especially in Web search Boolean querying is often used, even though the underlying retrieval model is not an exact-match, but a best-match one. In commercial solutions, best-match search systems are often coupled with so-called "natural language" querying. "Natural language" in this case refers to the way in which we normally write or speak, and a best-match search system that lets its users form queries like this is often referred to as a "natural language search engine".[38]

2.3. Relevance Ranking and Information Retrieval

A basic characteristic of the Boolean retrieval model is that the returned documents in the result list are not ranked according to their potential importance or relevance. In other words, the search engine considers each document to be equally relevant to users.[39] What traditional commercial information retrieval systems therefore end up doing is that they sort and then present results ordered by date, author, journal name, or any other common database-specific data element.[40] Result lists sorted like this have one major advantage: They are transparent, and users quickly understand what has led to the ranking result that is being presented to them.[41]

Information Overload or The Need for Ranking

What we have to be aware of, however, is that as electronic document collections grow larger and larger, also the number of possibly relevant documents constantly increases. Irrespective of how big a document collection is, we neither have infinite time nor patience to sift through endless material in a search for relevant documents.[42] On the Web, where the discrepancy between retrievable documents and time available for manual review is most

---

[36] See below, 4.5. Selected Issues, p. 35

[37] Wolfram, 2003, p. 16

[38] Paul & Baron, 2007, p. 22

[39] The Sedona Conference Working Group on Best Practices for Document Retention and Production (WG1), 2007, p. 202

[40] Jacso, 2005, p. 676

[41] Jacso, 2005, p. 678

[42] Burson, 1987, p. 135

pressing, users rarely look beyond the first 10 or 20 documents in a result list.[43] Especially in Web search, but more and more in other IR environments as well, listing all potentially relevant results is therefore no longer sufficient. Users expect systems to present result lists in a meaningful order, even if that requires departing from the transparent and easy-to-understand traditional ways of ordering search results.

This is where "relevance ranking" comes into play. It describes various statistical methods for ordering documents that appear in a result list. Simply put, relevance ranking arranges the documents within a result list so that those most likely to be relevant to your request are shown to you first.[44] At this point, some further clarification is necessary to avoid potential confusion on your side: In best-match search engines, "relevance ranking" is not an additional technology that might or might not be used. It is in fact already an integral part of the retrieval model. The possibility of using or not using relevance ranking for ordering result lists exists only in exact-match information retrieval systems. Still, relevance ranking is essential in the context of this thesis: As we will see[45], Boolean searching still constitutes the standard technology used in computer-assisted legal research systems today. Any ideas of using citation analysis techniques for relevance ranking in CALR therefore essentially have to compete with Boolean retrieval systems and the relevance ranking techniques that they offer today. Peter Jacso observes that most Boolean information retrieval systems now also offer relevance-ranked result lists.[46]

Putting Returned Sets in Order

We will therefore try to find out how current relevance ranking actually works on those Boolean systems. We have to keep the following in mind, however: While common elements of relevance ranking are known, the exact nature of the "algorithms" used is not released to the public because it constitutes valuable intellectual property for its owners.[47] This above all applies to commercial Web search engines, where the relevance ranking algorithm is a huge factor in maintaining the search engine's competitive edge.[48] We do know, however, that most commercially available non-Web information retrieval systems that feature relevance ranking employ methodologies based on word frequencies. The search engine measures the

---

[43] Langville & Meyer, 2006, p. 10

[44] Bade, 2007, p. 831

[45] See below 4.5 Selected Issues, 35.

[46] Jacso, 2005, p. 676

[47] Bade, 2007, p. 831

[48] Wolfram, 2003, p. 23

total occurrences of all terms in each document, as well as the occurrences of all terms in the database as a whole.[49] After a result list has been built in response to a user's query, the search engine uses this frequency data of the query terms to score all retrieved documents. The basic idea here is to treat individual query terms as being more or less important according to how often they appear in individual documents, and in the document collection as a whole. Properties like the document's length, the terms' locations within the document (in the title or in subject headings, for example), and their proximity to one another within the document, are usually factored in as well.

2.4. Evaluation in Information Retrieval

The last issue that I need to address concerning information retrieval is the - quite problematic - issue of IR evaluation. Maybe the most important fact to keep in mind about measuring how an IR system is actually performing is that so far, there are no objective criteria for evaluating the performance of information retrieval systems that experts have agreed upon. When it comes to search engines, the notion of effectiveness is subjective. In any case, some kind of human judgment is ultimately always the criteria for the evaluation of whether or not an IR system returns the relevant information in a correct way.[50] This certainly does not mean, however, that we should stop evaluating information retrieval systems altogether. Exactly the contrary is true. We just have to be aware of the potential pitfalls, and the limitations of IR evaluation. The theoretical goal of any search system can, for example, perfectly be established. This ultimate goal, or "Holy Grail" of information retrieval is the perfect search that retrieves everything that the user is looking for, while retrieving nothing the user is not looking for. In more formal terms, this is a search that equals 100% "recall" and 100% "precision".[51]

Recall and Precision

Recall and precision are two of the earliest measures in information retrieval evaluation, and they are still the most widely used ones.[52] They are a pair of values, and calculated as follows: Recall is the proportion of relevant documents actually retrieved from a document

---

[49] Evans, 1994, p. 124

[50] The Sedona Conference Working Group on Best Practices for Document Retention and Production (WG1), 2007, p. 204

[51] Evans, 1994, p. 122

[52] Wolfram, 2003, p. 26

collection, while precision is the proportion of the retrieved documents that is found to be relevant to the user's needs.[53] Another way to think about recall and precision is: When determining recall, we ask: "Out of the total number of relevant documents in the whole document collection, how many were retrieved correctly?". When calculating precision, the question is "How much of the returned result set is on target?".[54] This leads us to the problematic nature of relevance, and other difficulties in information retrieval evaluation.

The Problematic Nature of Relevance and Other Difficulties

The calculations of recall and precision depend on complete relevance assessments of documents within a document collection. Therefore, relevance became a key notion, but also a "key headache" in information science.[55] A first difficulty lies in the fact that a document is considered to be either relevant or not, with no "grey area" in between.[56] Scott Burson correctly observes that relevance is in reality not an attribute of a document that is simply either present or not. He describes relevance as "a complex notion of how a particular document relates to a given line of inquiry", and brings forward our day-to-day observation that we think of documents as being more or less relevant to a particular issue. On top of that, we can profoundly disagree on the relevance of any particular document.[57] There is, however, another major difficulty concerning relevance judgments on top of the subjectivity of relevance. As Peter Jackson and Isabelle Moulinier point out, obtaining complete relevance judgments on all queries of interest is clearly impossible for modern document collections due to their size.[58] Lastly, different users may have completely differing needs when using an information retrieval system. When doing comprehensive research, a user may want to find all potentially relevant documents. If another user has less time at their disposal, he or she may - above all - want the system to correctly sort information by priority.[59]

---

[53] Salton, 1986, p. 648

[54] The Sedona Conference Working Group on Best Practices for Document Retention and Production (WG1), 2007, p. 207

[55] Saracevic, 1999, p. 1058

[56] The Sedona Conference Working Group on Best Practices for Document Retention and Production (WG1), 2007, p. 205

[57] Burson, 1987, p. 141

[58] Jackson & Moulinier, 2007, p. 48

[59] The Sedona Conference Working Group on Best Practices for Document Retention and Production (WG1), 2007, p. 204

3. Legal Research

Ian Gallacher observes that systemised legal research was developed because the legal universe in late nineteenth century America had become increasingly complex.[60] J.C. Smith and his colleagues, even though formally restricting themselves to so-called "common jurisdictions" (or "common law" systems)[61], in reality cover the whole legal landscape when they say that "The essence of legal research [...] is the retrieval of relevant legal information." They go on to point out a major problem that legal professionals have to deal with: The volumes of legal information are constantly increasing which can lead to information overload. But still, access to one relevant document can decide the fate of a whole research project.[62]

3.1. Primary, Secondary & Tertiary Legal Sources

No matter whether we deal with a "civil law" or "common law" system, we encounter primary, secondary and tertiary sources of law. Legal sources are different in terms of the relative weight that they are accorded. Some sources have binding authority, while others are only persuasive in varying degrees. A third kind of source is only useful as a tool for finding binding or persuasive sources. Legal professionals have to use each source with a sense of its place in this hierarchy of authority.[63]

When we talk about primary sources of law we refer to the law itself. This includes constitutions, statutes and regulations (called "legislation" as a group), and - at least in common jurisdictions - judicial opinions in case law.[64] Opinions are published in so-called "law reports". Secondary sources of law are legal commentaries and include practitioners' handbooks, looseleaf services[65], treatises, encyclopaedias, restatements, textbooks, monographs and journal articles. Finally, mere finding tools that include no substantive discussion of points of law are called tertiary sources.[66] Those tertiary sources have been

---

[60] Gallacher, 2006, pp. 160–161

[61] See right below, Differences Legal Research Austria - United Kingdom, p. 21

[62] Smith et al., 1995, p. 57

[63] Cohen & Olson, 2007, p. 7

[64] Hanson, 2002, p. 571

[65] Cohen & Olson, 2007, p. 414

[66] Hanson, 2002, p. 571

developed by legal publishers so that users can find information about the law. Examples include digests and indexes[67], namely the legal citation indexes, which we will cover in detail later[68].

3.2. Differences Legal Research Austria - United Kingdom

Most countries of the world can be described as being either a common law or a civil law system. Within Europe, civil law systems are also being referred to as "continental jurisdictions". Each of the two systems has its own history, its own basic principles and procedures, and its publication practices for legal sources.[69] At least historically, the major difference between common law and civil law concerns the relative superiority that is given to court decisions on the one hand, and codified written rules on the other hand. Common law systems are largely based on the doctrines implicit in prior court decisions, while civil law countries rely heavily on codified written rules. Morris Cohen and Kent Olson observe that the differences between common law and civil law systems have become less apparent in recent years, because each system has been moving in the direction of the other. They mention that some American jurisdictions have enacted written codes, while some civil law countries have given greater weight to court decisions.[70] Similarly, J. Armstrong and Christopher Knott note that "American legal research today starts with statutes."[71] and Richard Susskind explained in 1998 that legislation had become central to all legal systems.[72] From a civil law point of view, Roland Wagner-Döbler from Germany stated as early as 1994 that "in the reality of continental law the importance of precedents can hardly be overestimated"[73].

In this thesis I propose making more extensive use of legal citations, in civil jurisdictions as well as in common ones. Therefore I now zoom in on two specific differences in legal research between Austria, a civil or continental jurisdiction, and the United Kingdom , a common jurisdiction. Before doing so, however, I want to briefly

---

[67] Arewa, 2006, p. 802

[68] See below, 5.2 Legal Citators, p. 43

[69] Cohen & Olson, 2007, pp. 397–398

[70] Cohen & Olson, 2007, pp. 398–399

[71] Armstrong & Knott, 2006, p. 13

[72] Susskind, 1998, p. 13

[73] Wagner-Döbler, 1994, p. 16

mention the basic issue that causes the differences in the two jurisdictions. It is in fact one identical problem that both systems have to deal with which expresses itself in the respective particularities in legal research.

A legal system that just consisted of clearly drafted rules would run into major practical difficulties. As a practical matter, it is simply impossible to think of every possible set of facts to be covered by a rule beforehand. Therefore legal drafters in both legal systems, in common as in continental jurisdictions, often turn to so-called "open-textured" concepts, like "reasonable care" or "malice".[74] Both legal systems then have to come up with a way to make the content of those abstract legal rules operational, and this is where the two systems still differ. Whereas in common law countries case law is often used to provide operational rules, civil or continental jurisdictions still to some degree prefer using more detailed legislation and interpretation by secondary sources for that purpose. This implies that basic differences remain in how legal research is conducted in civil and common law systems.[75]

Focusing on one specific difference, we now try to gain a better understanding about the perceptions of case law in civil and common jurisdictions. We will take a closer look at the importance of case law in both jurisdictions, followed by the related comparison of how law reporting works in both systems.

The Importance of Case Law

In my opinion, it already provides a strong implicit indication that the practical differences between civil and common law systems might be smaller than expected when Vincy Fon and Francesco Parisi observe: "There are substantial historical and conceptual differences between the doctrines of precedent in common law and civil law traditions."[76] The mere fact that they say that both systems do have "doctrines of precedent" already shows that precedent is not a concept exclusive to common jurisdictions. Common sense in fact also tends to tell us that it is nothing but human nature that judges in both jurisdictions are inclined to follow the decisions of earlier judges in similar cases.

What made common law develop was that over time, and in certain circumstances, judicial precedents started to become binding rather than merely being useful persuasive guidance for the future.[77] The common law doctrine of judicial precedent, or stare decisis, explains the special way in which decisions relate to each other in common jurisdictions. The

---

[74] Branting, 2003, pp. 61–62

[75] Cohen & Olson, 2007, p. 399

[76] Fon & Parisi, 2006, p. 521

[77] Finch & Fafinski, 2007, p. 90

latin phrase "stare decisis" means "let the decision stand". The general idea behind this doctrine is that like cases should be treated alike. Once a decision has been reached in a particular case, it becomes good law and should be relied upon in other future cases as an accurate statement of law.[78] As Richard Susskind observes even most non-lawyers are familiar with that notion of precedent.[79] American political scientists, legal scholars, and practicing lawyers all share the opinion that precedent is one of the central components of the American common law system.[80] The well-known American "Black's Law Dictionary" defines precedent as "[a] decided case that furnishes a basis for determining later cases involving similar facts or issues" and says that precedent may be divided between binding precedent that a court "must" follow and persuasive precedent that is "entitled to respect and careful consideration."[81] We should keep in mind, however, that also in common law countries, opinions written by top courts are considered to be more "important" than those written by lower courts.[82]

Morris Cohen and Kent Wilson give a general overview about civil jurisdictions: We find this type of legal system in continental Europe, Latin America, and parts of Africa and Asia. Cohen and Wilson, in my opinion, exaggerate a bit when naming a few distinctive characteristics of civil law systems, but the general trend is undoubtedly accurate: Civil jurisdictions usually have comprehensive and systematic codes governing large fields of law (civil law, criminal law, commercial law, civil procedure, criminal procedure). Concepts have strong influence; judicial decisions are given little weight as legal authority. Legal scholars who interpret, criticise and develop the law in their writings, particularly through commentaries on the codes are accorded great influence.[83] Also in civil law jurisdictions, as Roland Wagner-Döbler indicates, special attention is given to the decisions of the Supreme courts.[84] In Austria, for example, each of three Supreme courts is competent for a major area of law: the Constitutional Court ("Verfassungsgerichtshof") for constitutional law, the Supreme Court of Justice ("Oberster Gerichtshof") for civil and criminal law, and the Administrative Court ("Verwaltungsgerichtshof") for administrative law. That those courts are given special attention seems nothing but natural as they are the last instances for

---

[78] Finch & Fafinski, 2007, p. 145

[79] Susskind, 1998, p. 16

[80] Hansford & Spriggs, 2008, p. 5

[81] Black & Garner, 2007, pp. 1214–1215

[82] Finch & Fafinski, 2007, p. 145

[83] Cohen & Olson, 2007, p. 398

[84] Wagner-Döbler, 1994, p. 16

appeals. Wagner-Döbler holds that also in continental jurisdictions, facts cannot easily be subsumed under existing legislation, especially in complicated areas of law where matters change on a day-to-day basis. The necessary, more detailed legal norms and principles are, he ascertains, partly taken from precedents, just like in common jurisdictions.[85] Civil law systems gradually developed informal precedent law, which means that a sequence of analogous cases acquired persuasive force and became a source of law. Codifications had failed to bring certainty, consistency, and stability into the legal system, so judicial practice was supposed to compensate that. Vincy Fon and Francesco Parisi observe a "general tendency to accord persuasive force to a dominant trend of court decisions within civilian jurisdictions". Speaking of Germany - but this is true for Austria as well - they mention that legal professionals refer to a prevailing line of precedent which has been standing for some time as "permanent adjudication" ("ständige Rechtsprechung").[86]

Law Reporting

In 1987, before the World Wide Web took off, H. Patrick Glen pointed out a basic irony: Lawyers in common law systems were often surprised that their continental colleagues did not attach as much importance to case law as they did, while in reality, also common lawyers only cared about a small, specific selection of prior cases. Those cases that did not enter the published volumes of case law called "law reports" suffered an eternal neglect, just like in the purest of civilian traditions.[87] The general rule in common jurisdictions, still valid today, is as follows: A small portion of all the cases decided by the courts is published (or "reported") in so-called law reports. Only if a case raises a point of legal significance, it is selected for reporting.[88] As F. Allan Hanson recounts, this way of publishing legal cases was adopted in the United Kingdom with the explicit intent of restricting the growth of the legal literature. The publication of redundant cases was supposed to be prevented in order to keep the body of case law within manageable limits. Only those cases were meant to be published that "modify a principle of law, enunciate a new principle, settle a doubtful question, or that are in some other way particularly instructive".[89] Interestingly, however, there has been no official series of law reports in the United Kingdom. Law reporting has been in the hands of commercial publishers since the

[85] Wagner-Döbler, 1994, p. 16

[86] Fon & Parisi, 2006, pp. 522–523

[87] Glenn, 1987, p. 363

[88] Thomas & Knowles, 2006, p. 23

[89] Hanson, 2002, p. 565

last quarter of the thirteenth century.[90] In the UK, about one third of the Court of Appeal cases will subsequently be reported.[91] Similarly in the US "only the tip of the iceberg" of cases gets published.[92] Still, Philip Thomas and John Knowles observe that it is unusual that a case goes unreported if it raises a significant legal issue.[93] Computer-assisted legal research (CALR) systems and the development of the World Wide Web have, however, significantly changed the environment for law reporting. Due to the basically non-existing storage limitations of CALR systems and the World Wide Web, a vast number of decisions have become available that were not originally reported. In the United Kingdom, judgments from the High Court, the Court of Appeal and the House of Lords, are now available on the Internet on a variety of non-subscription websites, and have been available on commercial CALR systems even for some time before that[94]. Such judgments which are not reported but publicly available are referred to as "unreported judgments". They may be cited in court where it is believed that relevant legal issues are raised.[95] In the United States, the possibility of citing an unpublished case depends on the jurisdiction. But J. Armstrong and Christopher Knott point out a general fact that even if a lawyer is not permitted to cite a case, it may still serve as an indicator of how judges are likely to lean in similar circumstances in the future.[96] This new environment for law reporting does not imply, however, that law reports have ceased to be central to legal research in the United Kingdom and other common law countries.[97] It is the selection provided by the law reports, and the added explanatory material, that ensures that law reports remain a primary source of case law in common jurisdictions.[98]

As case law carries less weight in civil law jurisdictions, we generally find fewer decided law reports in them than in common law countries. Often legal periodicals publish court decisions or abstracts of case law in addition to legal articles.[99] When it comes to the selection task and the adding of explanatory material, however, the function of law reports or

---

[91] Clinch, 1990, p. 290

[91] Thomas & Knowles, 2006, p. 47

[92] Armstrong & Knott, 2006, p. 81

[93] Thomas & Knowles, 2006, p. 47

[94] See also below 4.4 The Situation in the United Kingdom Today, p. 32

[95] Thomas & Knowles, 2006, p. 23

[96] Armstrong & Knott, 2006, p. 81

[97] Cohen & Olson, 2007, p. 406

[98] Thomas & Knowles, 2006, p. 23

[99] Cohen & Olson, 2007, p. 419

legal journals doing law reporting in civil law countries exactly equals the function of their common law counterparts. In particular, those publications are responsible for creating special kinds of summaries called headnotes. Those headnotes depict legal principles arising from individual cases.

4. Computer-Assisted Legal Research

We have heard[100] that systemised paper-based legal research initially developed in late nineteenth century America out of a growing complexity of the legal system. By the early 1960s, American lawyers were again observing - even with the powerful secondary and tertiary print sources that they had available to them now - that reasonably finding relevant cases and secondary sources was becoming impossible. This time the solution lay in the further boost of CALR systems, and more generally in the introduction of artificial intelligence techniques to assist with legal information management.[101] At this point, I have to dispose of a terminological issue. Both in the world of companies and academia it seems to be popular to come up with a new name for - basically - the same concept every once in a while. This thesis, and this chapter in particular deal with the use of computers to access documents for legal research purposes. Whether we call that "Legal Information Retrieval"[102], "Computer-Assisted Legal Research (CALR)"[103], or even "Online Legal Resources"[104] does - at its core - not make a lot of difference in my opinion. I chose "Computer-Assisted Legal Research" because "Legal Information Retrieval" might today also comprise modern data mining tasks for "E-Discovery", and the term "Online Legal Resources" does not take into account that there are still offline legal research aids (for example on CD-ROM) available.

The basic keyword matching technique that Hans Peter Luhn suggested for general information retrieval has also been followed in computer-assisted legal research. [105,106,107] Throughout this chapter, I will often describe features of US computer-assisted legal research systems. This is due to, as Mario Ragona correctly observes, the exceptional position that US computer-assisted legal research systems have among all CALR systems. Both the long time span that those systems have been offering their services, and the sheer

---

[100] See above p. 20

[101] Hanson, 2002, p. 573

[102] Bing, 1984b

[103] Harrington, 1984

[104] Mason, 2006, p. 249

[105] Bing, 1984b, p. 152

[106] Moens, 2001, p. 34

[107] Turtle, 1995, p. 8

size of their searchable document collections covering all categories of legal information, make them exceptional.[108]

4.1. Initial Development

In the late 1950s and early 1960s it first appeared to be realisable to use computers to assist in legal research.[109] Usually people refer to John Horty who successfully demonstrated an operational system in 1959[110] as being the pioneer of computer-assisted legal research. Following his system, the first large-scale computer-assisted legal research systems were established.[111] Colin Tapper from the United Kingdom observes that one reason for the first development of computerised methods was that "the volume of legal material was increasing at such a startling rate that it could not be handled by conventional means"[112]. Describing the situation in the US in the 1960s, Bernard Hibbits explains that due to the continuing growth in the amount of published legal information "an eclectic variety of lawyers, legal academics, and law librarians [were looking] to emerging computer technology to facilitate the storage, accessing, and distribution of legal information."[113] As this thesis develops ideas across legal systems, I have to point out that there was an initial difference concerning the approaches of common and civil jurisdictions to the problems of computer-assisted legal research. The already mentioned[114] disparity especially concerning the relative weight of different legal sources in civil and common law systems was responsible for that inhomogeneous initial development. In civil jurisdictions, CALR systems were provided mainly by the governments because of their perceived role of having to provide the law for the people. In common law systems, on the other hand, entrepreneurs were simply responding to the demands of professional users.[115] Similarly, Jon Bing mentions that the initiative of creating CALR systems generally came from professional organisations or governments rather than from private companies in Europe.[116] One committee of the Council of Europe was central to the development of legal information systems in Europe: the "Committee of Experts on the

---

[108] Ragona, 2002, p. 14

[109] Tapper, 1976, p. 249

[110] Horty, 1959, p. 31

[111] Bing, 1984a, p. 30

[112] Tapper, 1976, p. 252

[113] Hibbitts, 1996, p. 655

[114] See above, 3.2 Differences Legal Research Austria - United Kingdom, p. 21

[115] Tapper, 1979-1980, pp. 37–38

[116] Bing, 1986, p. 3

Harmonisation of the means of Programming Legal Data into Computers". Also, the European Union (back then still called the "European Communities") has to be mentioned at this point: In 1971 the legal service of the European Commission began to work on its CELEX database (which forms part of EUR-Lex today).[117] In the 1970s, when in European countries there was still a slow movement from experimental to operational systems, large-scale commercial systems were already being set up in the US.[118] The LexisNexis system which is still active today started being developed in 1972 in the United States. It was also the first commercial system to allow full text searching rather than searching document surrogates only.[119] When commercial systems were finally being set up in Europe as well, the initial difference in the development of CALR systems between civil and common jurisdictions faded.[120]

4.2. Spotlights of Further Development

Due to space constraints, I am only able to provide some spotlights of the further development of computer-assisted legal research systems. I will first mention all developments which are essential for a basic understanding of the current situation of CALR in the United Kingdom and in Austria. Those specific issues that have a direct relation to the later parts of this thesis will be elaborated on individually after two short country overviews.

Up until the mid-1990s, the only way to electronically access CALR services was through proprietary software. You had to have a special program installed on your computer that would access a CALR provider's database. The Internet changed that as those software solutions gave way to Web browser based access. Now, the only technical requirements for using computer-assisted legal research systems are a personal computer with an internet connection and a Web browser. Steve Arnold and Lawrence Rosen already pointed at another development that still continues today in 1993: The luxury of having a legal information specialist as an intermediary to help with formulating queries, checking the first results, refining the search, and maybe even delivering the final results, is available to less and less legal professionals. Also, people with a day-to-day knowledge of using computers in general might also actually want to conduct their own searches without consulting anyone.[121]

---

[117] Nunn-Price, 1992, pp. 12–13

[118] Nunn-Price, 1992, p. 13

[119] Meadow, Boyce, & Kraft, 2007, p. 31

[120] Tapper, 1979-1980, pp. 37–38

[121] Arnold & Rosen, 1993

The technical term for that development, also in a legal setting, is "disintermediation". T.R. Halvorson and Reva Basch, describing the situation from a legal information specialist's point of view, recount that after a time when they had done searches exclusively themselves they started to train their customers to do their own searches. As of 2000, they observed that by that time the end users did most searches by themselves, with little or even no training at all, and they would just occasionally seek advice from an expert searcher.[122] In my opinion Arnold and Rosen draw the only correct conclusion from the disintermediation trend in computer-assisted legal research: they see a need to simplify the search and retrieval process, as a widening customer base means that software tools have to progressively become easier to use.[123]

This leads to a more in-depth look at the actual technology used by CALR systems. The changes over time have in fact been very modest here. Until the early 1990s, despite decades of academic research on the advantages of best-match systems, systems based on the Boolean retrieval model were the only search options offered by CALR providers. This conservative approach concerning technological innovations was still somewhat in line with the general trend observed in commercial information retrieval systems.[124] Then, however, while commercial IR systems in other sectors successfully started implementing best-match models[125] that were more technologically advanced, CALR systems decided to keep relying mainly on Boolean retrieval. The global CALR provider Westlaw, for example, added an alternative best-match, natural language querying search mode called "Natural Language" in 1992. Boolean searching (called "Terms and Connectors" by Westlaw), however, was - and still is - the default search mode and (therefore) used by a majority of users.[126] Pointing at the non-existent technological development of CALR systems in Europe, Erich Schweighofer observed in 1999 that "The retrieval algorithm [...] has not changed very much since the start of the first system of Horty"[127]. I will elaborate on those issues surrounding the dominance of Boolean retrieval, past experiences with natural language searching, and the reasons for the slow technological changes in computer-assisted legal research in more detail below[128]. Colin Tapper points out another important issue, brought about by the way the Internet and

---

[122] Halvorson & Basch, 2000, p. xii

[123] Arnold & Rosen, 1993

[124] Manning et al., 2008, p. 13

[125] See above 2.2 Querying Models & Retrieval Models, p. 13

[126] Manning et al., 2008, p. 14

[127] Schweighofer, 1999

[128] See 4.5 Selected Issues, p. 35

the World Wide Web have developed simultaneously with a dramatic decrease in the general cost of computing: As far as the distribution of non-copyrighted material, especially legislative texts and case law go, the development of many less expensive, and sometimes completely free, CALR services has become feasible. Many new CALR providers in those areas now supplement the traditional commercial CALR services.[129] Various government-run websites have been offering legal documents for free for a decade by now.[130] International examples of what Robin Widdison calls "second wave CALR systems"[131] include, following the prototypical American Cornell Legal Information Institute (LII; http://www.law.cornell.edu/), the Australasian Legal Information Institute (AUSTLII; http://www.austlii.edu.au/), the British and Irish Legal Information Institute (BAILII; http://www.bailii.org/), and the Canadian Legal Information Institute (CanLII; http://www.canlii.org/).[132] As the archives of those free services build up over time, big commercial providers of CALR services have to increasingly show that the material that they provide on top of publicly available documents is worth the additional investment.[133]

4.3. The Situation in Austria Today

CALR systems in Austria basically originated from two different sources.[134] During the 1980s, the Austrian Federal Chancellery increasingly served as a coordinator for electronic legal documentation projects in Austria. In the early 1990s, an initiative to build up a complete database of Austrian federal legislation was started, as well as the comprehensive electronic storage of case law from Austria's three High courts.[135] On the commercial front, the biggest Austrian legal publisher Manz created a limited company in 1982, the RDB Rechtsdatenbank ["RDB legal database"], in order to prepare the electronic distribution of legal print publications. All subsequent efforts to make the RDB CALR system operational came from private companies. Some of the big Austrian legal publishers started to load leftover data from their printed products onto a mainframe computer in 1986.[136] Up until 2006, the RDB system (http://www.rdb.at) continued to be the de-facto monopolist

[129] Tapper, 2005, p. 9

[130] Samborn, 1999, p. 75

[131] Widdison, 2002, p. 42

[132] Cohen & Olson, 2007, p. 407

[133] Norman, 2004, p. 97

[134] Niegl, 1996, p. 92

[135] Österreichischer Verfassungsdienst, 1994, p. 2

[136] Müller, 1991, pp. 257–258

concerning the electronic distribution of secondary legal sources in Austria. While the public RIS system was comprehensively providing legislation and case law for free on the Web, the RDB system was the only service to provide electronic access to a wide range of legal journals, and a growing number of legal commentaries.[137] The CALR landscape in Austria gradually changed around 2006 due to the acquisition of two legal publishers by LexisNexis. LexisNexis gradually withdrew the RDB the rights to its newly acquired legal texts, and now offers those documents on its own LexisNexis CALR service (http://www.lexisnexis.at).

Free Services in Austria

Initially, the objectives of the government-run Austrian Legal Information System RIS were defined as to provide "up-to-date, comprehensive, inexpensive legal information in an electronic format" to state organs and the public.[138] In 1997, the Federal Chancellery as the coordinator of this service decided to turn the RIS system, which had previously been accessible for free only to the public administration, into a free Web service (http://www.ris.bka.gv.at), available to everyone. As Elisabeth Staudegger correctly observes, the free service that the RIS system provides in Austria today is of an outstanding quality in an international context.[139] So far, none of the commercial CALR providers have even tried to compete with RIS in the area of primary legal sources by providing some kind of added value. In the area of case law from Austria's supreme courts, the RIS system even provides free access to all headnote documents that have been created by the High courts.

4.4. The Situation in the United Kingdom Today

Richard Susskind observes that until the late 1980s it was a rarity in the United Kingdom to see a terminal or a PC on a lawyer's desk. This is not to say that there were no IT applications for legal research purposes at all, but they literally stayed in the back office. Specifically built terminals to access CALR systems were usually located in legal libraries.[140] Today, the two main contenders for computer-assisted legal research services in the UK are LexisNexis and Westlaw UK.[141] In fact, those two companies do not only compete against each other on the UK market, but they are also "the heavyweights of legal

[137] Staudegger, 2006, p. 184

[138] Österreichischer Verfassungsdienst, 1994, p. 1

[139] Staudegger, 2006, p. 183

[140] Susskind, 1998, p. 73

[141] Mason, 2006, p. 246

research" in the United States.[142] For a few years, each of the two has been owned by a multinational corporation, LexisNexis by Reed-Elsevier (who in turn own UK publisher Butterworths) and Westlaw by Thomson (who in turn own UK publisher Sweet & Maxwell). This increasingly influences the range of materials offered by the two services, also resulting in the withdrawal of rights to each other's data.[143] Speaking of this related landscape in UK legal publishing, Cook has referred to the "general trend over the past few years towards the Butterworths / Sweet & Maxwell duopoly", resulting in a reduced range of resources.[144]

LexisNexis (http://www.lexisnexis.co.uk) was launched in the United Kingdom in 1980,[145] it was in fact the first general CALR service that was offered to legal professionals in the UK.[146] It originated out of the first international connection of the American LexisNexis service which was made to the UK legal publisher Butterworths.[147] Today, the LexisNexis service - in terms of secondary legal sources - holds the full texts of LexisNexis Butterworths titles, along with journals from some other publishers.[148] After having evaluated the latest LexisNexis platform, Janice Edwards summarises that LexisNexis offers a "good service". Access "isn't cheap, but you do get a lot of Butterworths commentary, as well as updated legislation and cases."[149]

The Westlaw UK CALR system (http://www.westlaw.co.uk) was launched in 1999.[150] After that, it rapidly became one of the most heavily used sources in online legal research. Legislation, case law, journal archives and general legal news are provided as well as access to UK, EU and US information.[151] What Westlaw UK does is that it uses the US-based Westlaw's retrieval system infrastracture in conjunction with data provided by UK publisher Sweet & Maxwell. As we see it with LexisNexis, the full texts of publicly available legislation and case law are being combined with a range of publications from a big UK legal publisher.[152] Just as Butterworths publications do not appear electronically in Westlaw

---

[142] Krause, 2004, p. 51

[143] Norman, 2004, p. 90

[144] Cook, 2002, p. 10

[145] Norman, 2004, p. 90

[146] Bing, 1984b, p. 452

[147] Bing, 1986, p. 4

[148] Thomas & Knowles, 2006, p. 82

[149] Edwards, 2005, p. 204

[150] Sayer, 2008, p. 299

[151] Monk, 2008, p. 147

[152] Thomas & Knowles, 2006, p. 11

UK, texts published by Sweet & Maxwell do not appear in the LexisNexis service.[153] One legal information professional who took part in a study done by Thomas Shaw mentioned that the entry of Westlaw into the UK market had brought "significant improvements" among large commercial CALR resources generally, especially concerning functionality and pricing policies.[154]

HeinOnline (http://www.heinonline.org) was introduced by William S. Hein & Co., an American legal publisher, in 2000. Since then, it has become popular in the United Kingdom as well. HeinOnline contains the full texts from a large number of legal journals.[155] Unlike other services, however, it provides access to digital reproductions of each page - that is, it offers full text searchable PDF images. Many law reviews are available from the beginning of their runs. The search facilities offered by HeinOnline are not as flexible as those offered by LexisNexis and Westlaw, but the extensive coverage is an extremely valuable addition to traditional CALR services.[156]

Free Services in the UK

Just like in other countries, the raw texts of primary sources of law, that is the legislation and the case law in the United Kingdom, are not the property of legal publishers or CALR providers.[157] The UK movement of publishing case law on the Internet was led by the House of Lords who since 1996 has made its own judgments available on the Parliament website (http://www.parliament.uk).[158] During the same year Her Majesty's Stationery Office (HMSO) started publishing new legislation online. Legislation dating back to 1988 was subsequently added to the website a few years later. Only having current and historical versions of legislative changes online was, however, only of limited practical value. Having an exhaustive archive of historical legislation, but also a database which shows the law in force, with all amendments taken into consideration, is what was really needed in practice. The Statute Law Database (SLD; http://www.statutelaw.gov.uk) partly solves this problem. The idea of such a system had first been put up for discussion as early as 1991, at the end of 2006 the system was finally released to the public.[159] Being a free online service, the UK

[153] Thomas & Knowles, 2006, p. 8

[154] Shaw, 2007, p. 28

[155] Finch & Fafinski, 2007, p. 188

[156] Armstrong & Knott, 2006, p. 93

[157] Thomas & Knowles, 2006, p. 16

[158] Thomas & Knowles, 2006, p. 45

[159] Sayer, 2008, p. 299

Statute Law Database is now the official revised edition of the primary legislation of the United Kingdom.[160] Commentators say that even though the SLD cannot yet compete with the commercial CALR services, it does still provide an excellent free alternative.[161] The already mentioned[162] British and Irish Legal Information Institute (BAILII; http://www.bailii.org/) website has to be elaborated on as well. BAILII recognised the public nature of primary sources of law, but also the proliferation of websites offering different document collections. Since its launch in 2000, BAILII has performed a particularly useful service by becoming a one-stop-shop for both UK legislation and case law.[163] Sharing his view on the future of BAILII, Trustee Philip Leith argues that the service will continue to be successful because "there is a considerable amount of goodwill amongst the judiciary and the profession, as well as amongst all those many users of legal information who cannot presently afford the costs of commercial systems"[164].

4.5. Selected Issues

CALR Evaluations

When the first full text retrieval systems were developed, many thought that there was nothing else to invent in the area of information retrieval, and that the ultimate solution to any retrieval problem had been found. In 1985, David Blair and M. E. Maron[165] reported on a large-scale study of full text retrieval for litigation support using the Boolean-based IBM STAIRS system. The study showed that full text retrieval did not live up to these irrational expectations.[166] The retrieval experiment was unique because it took place in a realistic operational environment. Searchers were told to keep searching until they located - in their own opinion, that was - at least 75% of all relevant documents.[167] The study involved a manual review of 350,000 pages (40,000 documents) of electronic text for the purpose of relevance assessments. It turned out that the legal professionals in the study greatly overestimated the effectiveness of the Boolean search system concerning finding relevant

[160] Finch & Fafinski, 2007, p. 45

[161] Sayer, 2008, p. 301

[162] See above, p. 31

[163] Thomas & Knowles, 2006, p. 17

[164] Leith, 2007, pp. 44–45

[165] Blair & Maron, 1985

[166] Thompson, 2008, p. 967

[167] Sormunen, 2001, p. 257

documents in response to their full text searches.[168] Attorneys and paralegals using the full text system, who believed that they were retrieving at least 75% of all relevant documents using STAIRS, were shown to be retrieving at best only 20%.[169] Irrespective of to what degree exactly the Blair and Maron study on a litigation support system was relevant to computer-assisted legal research systems as well, Jon Bing emphasised right away that the time had come to reassess full text document retrieval as a legal research tool.[170] The Blair and Maron study and especially its conclusions also did start off a passionate debate on the effectiveness of large-scale full text retrieval systems in general.[171]

In 1986, Daniel Dabney published "The curse of Thamus"[172] in the American "Law Library Journal". He essentially argued that the findings of Blair and Maron, even though the study had involved a litigation support and not a CALR system, should also be taken extremely seriously by the CALR community. Dabney observed that the existence of significant differences between litigation support systems and CALR applications was in fact doubtful. He conceded that only another experiment could really tell for sure, but argued that "for the time being, the similarities between the two kinds of systems appear to be much greater than their differences". His remark "The proponents of full-text searching now bear the burden of showing that the finding does not apply to CALR." was particularly powerful in my opinion.[173] "The Curse of Thamus" instantly touched off a spirited debate in the "Law Library Journal", and contributions from both LexisNexis[174] and Westlaw[175] disputed the applicability of the Blair and Maron findings to CALR systems. The representatives of the CALR providers argued that the low recall levels found in the STAIRS study were due to the wide variety in language that occurs in litigation support documents. As a result of the more "standardized vocabulary"[176] in the case law world, higher recall values were to be expected in CALR evaluations than those that had been found in the STAIRS study. In a response to those critics Dabney later cited a pilot study by himself that yielded even lower recall levels of 11.4% for LexisNexis (then just called "Lexis") and 19.7% for Westlaw, and precision

---

[168] Paul & Baron, 2007, p. 25

[169] Blair & Maron, 1985, p. 295

[170] Bing, 1987, p. 189

[171] Sormunen, 2001, p. 257

[172] Dabney, 1986a

[173] Dabney, 1986a, pp. 30–31

[174] McDermott, 1986, p. 343

[175] Runde & Lindberg, 1986, p. 345

[176] Runde & Lindberg, 1986, p. 345

levels of 26.1% for Lexis and 26.9% for Westlaw.[177] Looking back, it was the research of Blair and Maron and especially "The Curse of Thamus" that pointed out that recall with Boolean searching is much lower than expert searchers believe it to be, also in computer-assisted legal research. Even though the two big US CALR providers had initially tried to dispute Blair's and Maron's, as well as Dabney's findings, in the early 1990s both companies provided their users with a ranked retrieval search mode. The addition of that second search mode implicitly acknowledged the possibility of improving Boolean CALR retrieval techniques.[178] Furthermore, in a 1996 article, David Blair also mentioned personal communication by which he had been told that internal studies done by Westlaw had in fact corroborated the low recall levels initially found by Blair, Maron and Dabney.[179]

Summing up, we can establish that even though CALR providers do not usually tell us that, decades of information retrieval research have shown that the ability to retrieve all relevant documents (100% recall) from any given electronic document collection, also a legal one, is an unachievable goal.[180] That makes the assumption of many CALR users that present-day search methodology will find "all" or "nearly all" available documents nothing but an illusion.[181]

Dominance of Boolean Retrieval

We have already seen that Boolean search is still to some degree prevalent in general commercial settings, but it is only in computer-assisted legal research where it in fact still constitutes the main search technology used. We will now look at that situation in more detail. Amy Landville and Carl Meyer name three reasons for the prevalence of Boolean retrieval models: The implementation of a Boolean search engine is straightforward. The processing of queries works fast. Third, Boolean models scale well to very large document collections, which renders hosting a growing collection easy: In terms of programming things simply stay the same, only storage and parallel processing capabilities need to be increased.[182] Another reason might be that IR service providers did initially make respectable

---

[177] Dabney, 1986b, p. 349

[178] Thompson, 2008, p. 967

[179] Blair, 1996, p. 14

[180] The Sedona Conference Working Group on Best Practices for Document Retention and Production (WG1), 2007, p. 211

[181] Paul & Baron, 2007, p. 24

[182] Langville & Meyer, 2006, p. 6

investments in Boolean-based retrieval.[183]

When we turn more specifically to Boolean computer-assisted legal research systems, we can start off by observing that result lists in Boolean-based CALR systems are usually ordered in reverse chronological order, which means that the most recent documents are displayed first.[184] Even developers of best-match retrieval models for CALR systems have accorded that this kind of ordering is "surprisingly effective" for legal materials as more recent documents often modify or interpret earlier ones.[185] In my opinion, however, a conceivable conclusion laid out by Paul Thompson concerning this prevalence of Boolean systems in CALR would not be appropriate. One could argue, Thompson says, that the ranked retrieval of best-match systems is simply not applicable for carefully selected and manually augmented "premium content" in computer-assisted legal research systems, as opposed to extremely diverse document collections like the World Wide Web.[186]

In fact, the proposition that I will put forward shortly that novel relevance ranking techniques could improve traditional Boolean-based computer-assisted legal research systems directly opposes such an explanation. After having looked at some of the identified shortcomings of Boolean CALR systems, you should be able to better follow my reasoning.

Shortcomings of Boolean Retrieval in CALR

Carol Bast and Ransford Pyle presume that the most important fact for a researcher to understand when using a Boolean computer-assisted legal research system is that a relevant document must exactly match the search query in order to be retrieved.[187] In other words, the searcher has to come up with a query that selects as much as possible of the relevant material, but at the same time excludes as much as possible of the non-relevant documents. Most users grossly underestimate the difficulty of that task brought about by the literalness of Boolean CALR systems.[188] The search interfaces that are traditionally associated with Boolean queries often do not provide a lot of help in that respect. A level of understanding of Boolean logic is presupposed that the general users simply do not possess.[189] Constructing Boolean queries using AND operators to link individual query terms often brings back very

---

[183] Liddy, 2006, p. 749

[184] Thomas & Knowles, 2006, p. 16

[185] Turtle, 1994, p. 213

[186] Thompson, 2008, p. 968

[187] Bast & Pyle, 2001, p. 293

[188] Turtle, 1995, p. 24

[189] Smith et al., 1995, p. 82

few or no relevant documents, while queries with many OR operators tend to retrieve large volumes of documents, because of out-of-context occurrences of search terms in the documents' texts. Striking the right balance turns out to be an extremely difficult task.[190] To put it in more formal terms: A Boolean-based CALR system uses the existence or absence of individual search terms within a document to make a binary relevance judgement: If one query term - even out of many - happens to not be present in a document, that document is thought to be totally irrelevant to the query.[191]

Experiences with Natural Language Querying and Ranked Retrieval

Despite its shortcomings, Boolean retrieval formed the sole basis of commercial computer-assisted legal research systems up to the 1980s. In the early 1990s, the two major systems Westlaw and LexisNexis began to offer ranked retrieval options in addition to traditional Boolean retrieval.[192] Those systems actually pre-dated the widespread use of the Internet. Westlaw's system is called WIN (Westlaw Is Natural), LexisNexis' one Freestyle.[193] The way those systems are implemented is by coupling natural language input with best-match search techniques.[194] As regards the natural language querying part of those systems, however, some caution is in order. It is quite easy to set up a system that simply removes so-called "function words" (such as "the", or "a") from an initial query, and runs a very conventional search using the remaining, more content-bearing query terms.[195] Paul Thompson notes a somewhat ironic situation: Both Westlaw and LexisNexis, who have given their users the choice of using ranked retrieval instead of the default Boolean search, have found that the vast majority of users have preferred to stay with Boolean retrieval.[196] The irony comes from the fact that this does not mean that Boolean systems are more effective for professional searchers.[197] Experimenting on a Westlaw sub collection, Howard Turtle found that "on average a current generation natural language system provides better retrieval performance than expert searchers using a Boolean retrieval system when searching

---

[190] Gelbart & Smith, 1993, p. 20

[191] Wolfram, 2003, p. 16

[192] Turtle, 1995, p. 25

[193] Greenleaf, 2004, p. 65

[194] Tenopir, 1993, p. 54

[195] Weiss et al., 2005, p. 208

[196] Thompson, 2008, p. 968

[197] Manning et al., 2008, p. 15

full text legal materials."[198] Maybe it was because he was already aware of some of the issues that I will elaborate on shortly[199] that Turtle held that "despite the strong performance of natural language searching, Boolean query languages will not disappear anytime soon."[200] Morris Cohen and Kent Olsen give a general overview about the situation on the American versions of both LexisNexis and Westlaw today, even though they only explicitly mention Westlaw: Two basic methods of searching are being offered: natural language, and Boolean. When using the "natural language" option, not all query terms will necessarily appear in every document retrieved. It is possible, however, to define "required terms" that must appear in all documents.[201] As regards the UK versions of the two databases, the option of natural language searches continues to be offered by Westlaw. On the LexisNexis CALR system, however, it is no longer available.[202]

4.6. Reasons for the Slow Technological Change in CALR

At first sight, this ongoing reluctance to implement new, more sophisticated search technology in CALR systems seems odd. The economic environment of CALR systems should in fact be fostering technological development. Clients of legal researchers are used to being charged for high quality research. Therefore automated research tools that save time and / or produce better search results should be easy to justify.[203] Financial performance figures also indicate that the legal information business is quite profitable.[204] Big American law firms, after all, pay as much as $4 million a year for access to Westlaw and LexisNexis.[205] We might therefore be tempted to think that the CALR business should in fact be the very first place where big technological innovations take place in terms of search technology. I have come to the conclusion that a whole mix of reasons is responsible for the fact that quite the contrary has been the case. After observing that there is a lot of room for improvement in the Boolean retrieval systems of LexisNexis and Westlaw, Daniel Dabney

---

[198] Turtle, 1994, p. 212

[199] See right below, 4.6 Reasons for the Slow Technological Change in , p. 40

[200] Turtle, 1994, p. 219

[201] Cohen & Olson, 2007, pp. 20–21

[202] Norman, 2004, p. 94

[203] Turtle, 1995, p. 7

[204] Arewa, 2006, p. 827

[205] Fisher, 2008

first mentions that possible improvements might be expensive to implement. Besides that, it is the responsibility of users to "bring pressure to bear" in terms of technological innovations.[206] Robert Berring shares a candid revelation made to him by a CALR vendor representative. Making big investments in the search technology used in CALR systems could demonstrably improve the retrieval performance. The representative pointed out, however, that the system users were in fact not complaining about the retrieval performance, but only about the difficulty of using the system, and its cost. Therefore, from the provider's point of view, the introduction of a new layer of difficulty for the user that might also lead to additional costs "would be counterproductive. The money would be better spent in marketing."[207] Furthermore, we must not forget that any legal research support tool has to take the potential liability for bad expert advice into consideration. As an intellectually demanding process, legal research is an activity for which legal professionals can be held liable if they perform it inadequately. Document retrieval plays an essential part in that activity. If a CALR system looks "smart", or is marketed as being "intelligent", it might give the searcher a false sense of security, which is what the providers want to avoid.[208] Also, we can note that even critics of the Boolean-based CALR systems admit that for certain functions, those systems perform superbly. Searching for cases by using unique terms like the name of a particular judge, a particular date or a particular court brings back excellent results.[209] Finally, also in all situations when Boolean CALR systems do not function well at all, in principle Boolean operators and Boolean retrieval make it easy for users to understand why a particular document is retrieved, or is not retrieved. Best-match systems give better retrieval performance, but they are much more difficult to explain to the ones using them.[210]

Summing up, I want to say that the resulting situation has been a truly sad one in my opinion. The ongoing choice of further relying on Boolean search systems seems to be made entirely for the wrong reasons: Users are not aware of better search technology that is "out there", and providers are aware of it, but choose to ignore it because of the risks involved and the nonexistent pressure to adopt more sophisticated technology.

---

[206] Dabney, 1986b, p. 350

[207] Berring, 1997, p. 209

[208] Turtle, 1995, p. 48

[209] Berring, 1986, p. 42

[210] Turtle, 1995, p. 40

5. Citations and Legal Research

Fred Shapiro observes that legal communication effectively consists of nothing but two principal components: words and citations.[211] Citations are used to incorporate the language and power of one source in another one. Case law usually contains citation links to other case law as well as to legislation. Secondary sources usually cite other secondary sources, as well as case law and legislation.[212] What this means is that legal sources are in fact partly self-indexing, or chained. It is the interpretative process of legal professionals that is needed to understand this complex network of material, and to harmonise it. Jon Bing points out that legal sources therefore have, just like a textbook, a double nature: Besides their nature as a legal source, they also serve as an information system, and help with the retrieval of related relevant legal sources.[213] We have already looked at some basic differences between common and civil/continental jurisdictions[214]. Those differences, even though they might be smaller than expected in practice, still echo themselves in the availability and sophistication of citation-based legal research tools that are available to legal professionals in both legal systems.[215] Generally, the use of citation-based techniques is more developed in common law systems than in civil ones today. It is specifically the area of case law in common jurisdictions where the importance of legal citations is already extremely high in legal research. According to legal professionals, going to court citing an "outdated case" is arguably the most embarrassing experience one can make as a practising lawyer.[216] We will now therefore specifically look at the traditional and contemporary ways of retrieving case law in common jurisdictions. As I am going to propose the more extensive use of legal citations for relevance ranking purposes, I want to cover all existing uses of citations in legal research first. We will also look at possible other uses of citations within legal texts in the chapter on legal citation analysis[217]. There will be only little overlap, though, because in the current chapter we restrict ourselves to the role that legal citations already play in

---

[211] Shapiro, Mar., 1991, p. 1453

[212] Shapiro, 1996, p. 752

[213] Bing, 1988, p. 394

[214] See above, 3.2 Differences Legal Research Austria - United Kingdom, p. 21

[215] Tapper, 1976, p. 258

[216] Krause, 2004, p. 53

[217] See below, 9 Legal Citation Analysis, p. 66

contemporary legal research, not covering more extensive uses.

5.1. The Traditional Legal Retrieval Pattern

Stephen Marx describes what he calls the "traditional retrieval pattern" in case law research: After lawyers have collected the facts concerning the legal problem of their clients, they begin their search through the available tools to find relevant case law materials.[218] Having found one or more cases on point legal professionals start what has been described as "chaining"[219], or "footnote chasing" and "citation searching"[220]. Researchers of all scientific areas do that, but the activity is of particular significance in common law legal research: The notion of precedence implies that the relationships among cases - laid out by the citations between them - are of exceptional importance. Decisions of current cases most often need to be justified in terms of past decisions involving "similar" circumstances. Colin Tapper's description of common citation-based research procedures in law hints at the reality that this special nature of legal citations is in fact not restricted to common jurisdictions: If lawyers know that a current case addresses topics that they are interested in, they use citations in that case to look for recent material which may not have made it into legal commentaries like textbooks yet. Also, if a textbook, or an encyclopaedia, includes a "Table of Cases", lawyers use a known case to find the part of the publication relevant to their legal problem at hand.[221] Specific tools and services have been created in common jurisdictions with the aim of assisting attorneys in this kind of research. So-called citator services allow users to examine the whole list of citations that directly reference to a given case.[222]

5.2. Legal Citators

The practice of gaining legal authority through including citations dates back to at least the eleventh century, but the development of so-called legal citation indexes did not start until the early nineteenth century.[223] As the volume of cases grew exponentially, lawyers found it increasingly difficult to monitor the precedential value of cases just by themselves. Whenever they had found a relevant case, lawyers needed to determine whether it was still

---

[218] Marx, 1969-1970, p. 122

[219] Ellis, 1989, p. 183

[220] Bates, 1989, p. 412

[221] Tapper, 1984, pp. 94–95

[222] Zhang & Koppaka, 2007, p. 123

[223] Ogden, 1993, p. 7

"good law". The questions that they needed to find answers to were: "Had the case at hand been explicitly cited as precedent by subsequent cases?", "Were there no citations by subsequent cases to it at all?" or had it - and this was usually the worst case - been "overruled" by a later case?[224] Looking back[225], we might notice at this point that it was in fact simply the threat of "information overload" in case law that led to the development of legal citation indexes. Shepard's system of citators were the first widespread legal citation indexes and have become a familiar research aid for US lawyers. In 1873, Frank Shepard began to print citations to Illinois Supreme Court cases on gummed paper so that his subscribers could paste them into their law reports volumes. Eventually, Shepard's Citations Inc. expanded and published indexes that list subsequent citations to all US state and federal judicial decisions, statutes, and even other legal sources. By 1985, Fred Shapiro observed that diverse sources like "administrative regulations, court rules, law review articles, American Law Institute Restatements of the Law, and patents and trademarks" were all covered by Shepard's.[226] The original strips had eventually turned into printed red volumes of Shepard's citators, and those red books used to be found in every American law library before computer-assisted versions of Shepard's started to provide a more comfortable alternative.[227] Due to its de-facto monopoly, the word citator was long exclusively associated with the publications of Shepard's Citations in the minds of American lawyers.[228] Today, Shepard's information is available electronically on the US version of LexisNexis as well as in print, and the US CALR provider Westlaw has a competing electronic resource, KeyCite, that provides a similar citator service.[229] KeyCite, however, is only available electronically, therefore Shepard's Citations is the only choice for researchers who still want to use print citators.[230]

Overview about Legal Citators

Focusing on the purposes and uses of legal citators now, we can start off by observing that the legal citator is one example of a general class of tools known as "citation indexes". Fred Shapiro postulates that the citator is "among the crucial tools for legal research". It lists

---

[224] Foster & Kennedy, 2000, p. 277

[225] See above 1Introduction, p. 9

[226] Shapiro, Oct., 1985, pp. 1540–1541

[227] Foster & Kennedy, 2000, p. 277

[228] Shapiro, 2001, p. 177

[229] Cohen & Olson, 2007, p. 128

[230] Cohen & Olson, 2007, p. 133

subsequent sources that have cited a source, which allows researchers to not only verify the authority of a precedent but also to find additional sources relating to a given subject.[231] Morris Cohen and Kent Olson identify three major functions of legal citators, no matter whether they are used electronically or in print:

1. They provide parallel citations for the decision at hand and references to other proceedings in the same case, which allows a researcher to trace a case's judicial history.
2. They indicate if a subsequent case has overruled, limited, or otherwise diminished a case's precedent, which is the information that researchers need in order to find out whether a case is still "good law".
3. Their comprehensive listing network leads not only to later citing cases, but also to secondary legal sources, which enables researchers to find related cases and to trace the development of a legal doctrine forward from a known case to the present.[232]

We will look at those functions of legal citators in more detail now. Before doing so, however, I want to point out that we should not overestimate the importance of the limitation that the two big citator services Shepard's and KeyCite are only available on US CALR systems. Even though Shepard's and KeyCite do not have direct counterparts, there are tools available for finding later cases that have made a reference to earlier decisions, at least in other common law countries. In the United Kingdom, "Current Law" is a service useful for both finding and updating cases. The Current Law Case Citator enables users to check the judicial history of a case and to see where it has been reported. For those cases that have been judicially considered, the effects of later cases are indicated and put into categories like "overruled", "applied" or "considered".[233]

Citators for Verifying the Authority of a Precedent

Especially in common law jurisdictions, attorneys regard it as a "courtroom dilemma" to cite a decided case in court which has been overruled or reversed by a later authority. By using Shepard's or KeyCite to find references in a later authority to an earlier one, this can be avoided. Margaret Elliott and Rob Kling therefore rightly refer to the procedure of using a legal citator as being "imperative to avoid courtroom embarrassment or malpractice suits"[234]. In other words, before they rely on a case, attorneys must verify its current validity. Traditionally that was done by checking printed volumes of Shepard's Citations, as a result

---

[231] Shapiro, 2001, p. 177

[232] Cohen & Olson, 2007, p. 128

[233] Cohen & Olson, 2007, p. 410

[234] Elliott & Kling, 1997, p. 1026

the checking process is sometimes still referred to as "shepardizing".[235]

It is obvious that the way in which a case is treated when it is subsequently judicially considered has a direct effect on its importance and reliability.[236] J. Armstrong and Christopher Knott put this elegantly when they say: "The text of a case is not dynamic, but its significance is."[237] Shepard's therefore indicates how a particular court opinion is legally interpreted by the subsequently decided cases that cite it.[238] The treatment analysis that Shepard's provides ranges from strong negative ("overruled") to strong positive ("followed"). Between those two poles, we find a spectrum of potentially cautionary negative analysis (such as "criticized" and "distinguished") and more neutral analysis (including "explained" and "harmonized").[239] Attorneys are in fact hired and trained by the Shepard's Company to content-analyze court opinions for Shepard's Citations, a process that the company refers to as "letter editing".[240] LexisNexis' citator service KeyCite provides a similar service and also gives information as to how each citation to the current case was treated.[241]

Citators for Subject Searching

The main commercial purpose of Shepard's Citations might be the one I just described, to provide legal professionals with information about the legal authority of a case.[242] Expert legal researchers, however, use Shepard's and KeyCite not only to validate cases but also as a research device to find cases.[243] In fact, they even use them as research tools to collect various sources relating to a particular subject.[244] The principle that stands behind using a citator as a source finding tool is that, for example, a case that cites to an earlier case "must logically discuss the same legal issue as that for which it cites the earlier case". Consequently, by following the citations of a case for a particular point legal researchers can make out other cases on that point.[245] Cohen and Olson describe this use of citators as

---

[235] Cohen & Olson, 2007, p. 128

[236] Thomas & Knowles, 2006, p. 129

[237] Armstrong & Knott, 2006, p. 134

[238] Hansford & Spriggs, 2008, p. 44

[239] Morris, 2000, pp. 144–145

[240] Hansford & Spriggs, 2008, p. 46

[241] Zhang & Koppaka, 2007, p. 129

[242] Hansford & Spriggs, 2008, p. 49

[243] Halvorson & Basch, 2000, p. 13

[244] Shapiro, Oct., 1985, p. 1541

[245] Armstrong & Knott, 2006, p. 114

"providing one of the most effective ways to find sources for further research"[246].

Potential Problems and Evaluations

When thinking about potential issues when making use of legal citators, we have to keep in mind that the signals and editorial guidances that both KeyCite and Shepard's provide "are just tools for the researcher, not authoritative statements of the law". Cohen and Olson conclude, as a legal research rule, that reading a citing document and finding out for oneself its scope and effect must not be substituted by using a citator.[247] Similarly Armstrong and Knott maintain that even though all citators provide some kind of judgment regarding what the purpose of the citation for (some) citing cases is, researchers "must not rely on these". They explain that the rules that are applied by the citators so that they can provide editorial treatment analysis have caused individual choices with which few researchers would actually agree.[248] It is not necessarily a human choice that is responsible that treatment signals are attached to cases, they might also have been assigned by automated language analysis. Undoubtedly, however, citators can be helpful as a starting place in order to decide which cases to look at first.[249] Daniel Dabney tries to explain those limitations of citators from the perspective of the service provider. Individual legal researchers often do not agree among themselves concerning the exact way in which one case treats another one by citing it. As a consequence, there is no way that any CALR citator system could come up with generalisations that all researchers can agree with. Therefore, he concludes, citator services should not be relied upon exclusively, even though they provide great assistance to researchers.[250] For a strikingly long time, no one actually tested the reliability of Shepard's. In 2000, James Spriggs and Thomas Hansford finally observed that as "the reliability and validity of Shepard's is unknown, [...] we should therefore be appropriately sceptical of it." They went on to empirically test the reliability of Shepard's and to discuss the validity of its coding protocols. What they found was that Shepard's coding of legal treatment was "quite reliable".[251] After having tested the reliability of Shepard's Citations analysis of American Supreme Court opinions again, the two researchers found in 2008 that "Shepard's data on the

---

[246] Cohen & Olson, 2007, p. 128

[247] Cohen & Olson, 2007, p. 133

[248] Armstrong & Knott, 2006, p. 114

[249] Armstrong & Knott, 2006, p. 135

[250] Dabney, 2000, p. 385

[251] Spriggs, II & Hansford, 2000, p. 327

positive and negative interpretation of precedent are highly reliable".[252]

Citators and Computer-Assisted Legal Research

In some respects, the history of citator-like information and CALR actually begins before the computerisation of citator services. Colin Tapper recounts a much more radical approach to the retrieval of legal information than a citator-like service that was envisioned mainly at the beginning of the 1970s. At least in common law jurisdictions, it seemed feasible to actually derive meaning merely from the use of a document. In other words, the resemblance between documents by reference to common citation patterns was supposed to build the basis of the retrieval of legal documents. The reasoning behind that idea - which we still encounter today - was that two documents which both cite the same authorities and are themselves cited by the same authorities, are likely to have a lot in common with regards to their content.[253] As this radical idea of relying solely on citations for indexing purposes, however, did fade without further influencing the history of legal citators, we do not look at it more closely.

In 1975, LexisNexis offered the first online citation system, Auto-Cite, to subscribers in certain regions of the United States. Westlaw responded in 1980 by making Shepard's Citations available online, which can be regarded as the beginning of a race between the two companies concerning the electronic availability of legal citation information.[254] In the 1990s, LexisNexis acquired a remaining portion of Shepard's that it had not already owned before, and was thereby able to withdraw Westlaw the rights to provide access to Shepard's. Since then, Shepard's has been a service exclusive to LexisNexis, and Westlaw had to develop its own citator service.[255] The resulting KeyCite service was released in August 1997. Like Shepard's it can be used to trace case histories, retrieve secondary sources, find cases that cited the case at hand, and categorize citations by legal issue.[256] Since then, legal researchers have consequently had a choice between two online citators, the Shepard's system available exclusively on LexisNexis, and the KeyCite system unique to Westlaw.[257] The positive effects of merging computers and legal research become especially apparent when we look at the just mentioned electronic citator services. Already in 1993, Patty Ogden

---

[252] Hansford & Spriggs, 2008, pp. 46–47

[253] Tapper, 2005, p. 9

[254] Ogden, 1993, p. 37

[255] Dethman, 2002, p. 131

[256] Baker, 1998, p. 24

[257] Taylor, 2000, p. 127

observed that legal research had benefited greatly from online access to legal citation information.[258] This becomes extremely obvious when we contrast the manual shepardizing procedure with the electronic one. Manual shepardization is a time-consuming process that is prone to omissions.[259] The time, patience, and effort it requires in truth makes it rare that it is done thoroughly. CALR shepardization, on the other hand, makes looking up citations easy and fast to a degree that makes it much more likely that researchers persevere in it.[260] But this is in fact just one of the advantages that electronic citators have over the printed version of Shepard's Citations. Because space is basically an infinite resource in an electronic environment, names of publications and case treatments are spelled out rather than being abbreviated as in the print product. The searching of multiple volumes is also unnecessary, as the citing entries are compiled into one single listing. This includes the covered jurisdictions, as the coverage - unlike with the print version of Shepard's - is not divided into separate state and regional citators. The retrieval of specific treatments of headnote numbers is easily done, no scanning of lengthy lists of citations is necessary. Finally, hyperlinks make it possible to directly jump from the online citator to the text of citing cases.[261] As a consequence, Halvorson and Basch, after having interviewed various legal information professionals, concluded already in 2000 that legal researchers unanimously preferred online citators for validating case law.[262]

[258] Ogden, 1993, p. 38

[259] Elliott & Kling, 1997, p. 1026

[260] Dabney, 1986a, p. 37

[261] Cohen & Olson, 2007, pp. 135–136

[262] Halvorson & Basch, 2000, p. 11

6. Developing a Proposition

In 1993 Teresa Pritchard-Schoch writes

"As many legal researchers have noted, relevancy ranking is not as important in legal document retrieval [as in general document retrieval] - a perfect case, in terms of relevance, has little, if any, value if a subsequent statute has been enacted, or if the Supreme Court has reversed the case."[263]

I think that the conclusion that Pritchard-Schoch and those who share her opinion draw is wrong. In my opinion, because Boolean systems are still so prevalent in CALR, relevance ranking is even more important here than in other search domains. What Pritchard-Schoch correctly points out is that legal relevance ranking, rather than merely copying general relevance ranking techniques, has to specifically address the challenges that legal document collections entail. My view is in agreement with Graham Greenleaf's observation that one of the most "user friendly" developments in computer-assisted legal research has been the relevance ranking of result lists.[264] "Very large legal web sites like CanLII, BAILII or AustLII", Greenleaf holds, "would be far more difficult to use if they could not offer to the general public an "any of these words" search with relevance ranked results."[265]

Marie-Francine Moens states that current computer-assisted legal research systems provide users with the ability to search the full texts of stored documents. What current systems do not do, however, is to use "structured information", like citations, for best-match retrieval or relevance ranking purposes.[266] Based on this observation, I develop a proposition in the area of relevance ranking for computer-assisted legal research systems. Current ranking methods used by Boolean commercial CALR systems do not make use of human judgments inherent in legal citations.

I propose that the use of citation analysis concepts could improve the relevance ranking of those systems.

A whole number of overlapping scientific areas have been analysing link structures[267]in the sense of interconnections between documents. Citations constitute the prototypical way of connecting documents to each other. We will therefore take a closer look at three areas

---

[263] Pritchard-Schoch, 1993, pp. 35–36

[264] Greenleaf, 2004, p. 65

[265] Greenleaf, 2004, p. 66

[266] Moens, 2005, p. 225

that deal with interrelated document collections in the next section so that I will be able to develop my proposition into a testable hypothesis.

[267] Kleinberg, 1999, p. 617

SECTION II - ELEMENTS OF A THEORY

7. Developments in Web Search

It was largely due to the possibility to publish content with essentially no control of authorship that has led to the explosive growth of the World Wide Web. Ironically, this characteristic turned out to be the biggest challenge that Web search engines had to face, trying to make the Web's content searchable and retrievable.[268] The nature of the Web implies that we are in fact not really capable of saying how big it exactly is.[269] As early as 1998, Krishna Bharat and Andrei Broder observed that even though questions like "How many pages are out there and how many are indexed?" are of eminent scientific and public interest, few ways of objective and direct evaluation have been proposed.[270] Consequently, we have to content ourselves with extremely gross estimations. Those estimations, however, are still quite illustrative. In 2008, Google announced that its systems that process links on the Web to search for and find new content had hit a milestone: the so-called "index of unique Web addresses" had counted 1 trillion (= one million million or 1,000,000,000,000) unique URLs (Uniform Resource Locators).[271] But not only the sheer size of the World Wide Web causes challenges for Web search. We have to examine the typical Web search user as well. No matter how we feel about CALR and other specialised information retrieval areas, we can establish that the users of those systems are typically professionals in their fields. Often, they also have at least some basic training in the art of phrasing queries, and they usually understand what the main characteristics of the collections that they are searching are. Web search users, on the other hand, usually do not know, and do not care, about the characteristics of query languages, the art of phrasing queries, and the diversity of web content.[272] Unfortunately, publicly available large-scale studies using query log data are already several years old. There are, however, no indications that major changes have

[268] Manning et al., 2008, p. 387

[269] Manning et al., 2008, p. 388

[270] Bharat & Broder, 1998, p. 380

[271] Google, 2008

[272] Manning et al., 2008, p. 395

occurred in the meantime. Studying a query log made up of approximately 1 billion entries for search requests in 1999, Craig Silverstein and others observed that for 85% of the queries users only viewed the first result screen. Also, 77% of the sessions contained only 1 query, which means that the users did not modify their initial queries in those sessions.[273] Summarising the result of a study using a smaller, but different query log in 2001, Amanda Spink and others recognise that "a great majority of Web queries posed by the public are short, not much modified, and very simple in structure. Very few queries incorporate advanced search features, and when they do half of them are mistakes."[274]

At first, Web search engines made use of the classic model of information retrieval. Whether or not a document was relevant to a query, and how a returned document was ranked, depended exclusively on the document itself.[275] In other words, the first Web search engines used techniques for retrieval, including relevance ranking, that were based only on statistics of words in the document texts. One difference between traditional search engines and Web search engines that has originated quite early, however, is that popular search engines treat all queries without explicit connectors as "AND queries", as opposed to traditional information retrieval systems. In traditional IR environments, queries without operators are often still interpreted as "OR queries". As connecting query terms with AND operators is one way to reduce the length of result lists, this was in fact a first attempt to deal with sprawling result lists resulting from the sheer size of the Web.[276] Still, that search based solely on text techniques performed very poorly in the Web environment became obvious very soon.[277] This changed in 1998, when Google and with it so-called "link analysis" hit the Web search and information retrieval scene.[278]

7.1. Basic Technology: Link Analysis

Independently of one another, both Sergey Brin together with Larry Page[279], and Jon Kleinberg[280] recommended to exploit the hyperlink structure of the World Wide Web to improve the quality of Web search engines. Kleinberg described the limitations of a text-

---

[273] Silverstein, Marais, Henzinger, & Moricz, 1999, p. 12

[274] Spink, Wolfram, Jansen, & Saracevic, 2001, p. 233

[275] Witten, Gori, & Numerico, 2007, p. 135

[276] Witten et al., 2007, p. 118

[277] Henzinger, 2005, p. 1

[278] Langville & Meyer, 2006, p. 4

[279] Brin & Page, 1998, p. 108

[280] Kleinberg, 1999

only approach for IR in the Web environment using a striking example. At the time, the term "Harvard" was used by over one million pages on the World Wide Web. Unfortunately, the actual website of Harvard University, http://www.harvard.edu, was not a site that used the term most often, or most prominently, and also not in any other way that made the site succeed with text-based ranking functions.[281] Today, link analysis scores like those introduced by Google are being combined with more traditional information retrieval scores by almost all major search engines.[282]

7.2. Main Use: Ranking

The process of responding to a Web search query therefore goes well beyond any traditional information retrieval model.[283] Link analysis in Web search is thereby mainly used for relevance ranking, that is for the ordering of search results.[284] We have just covered the reasons for which order becomes exceedingly important in Web search: Free text searches tend to retrieve very large sets due to the size of the Web. Pretty much no user will examine the retrieved result set in detail.[285] It has become nothing but a necessity for Web search engines to use relevance ranking to order the results presented to its users.[286]

When we adopt a simplified view, a modern Web search engine in a first - invisible - step produces a set of relevant Web pages based on the occurrences of terms on those pages. Up to that point, the process is quite similar to the way in which current CALR systems with text-based relevance ranking capabilities operate.[287] Ricardo Baeza-Yates and Carlos Castillo fittingly describe this first stage as the "easy" step of Web searching. With most Web search queries being very broad, thousands of pages usually fulfil the criteria established by the entered queries. The "hard" part for the search engine comes next: Ranking the returned documents by their relevance and selecting the top hits to present to the user on the first results page is much more difficult, but it is also this what decides whether we use a particular Web search engine, or not.[288]

When we adopt a more sophisticated view, we have to look at the two stages of a Web

---

[281] Kleinberg, 1999, p. 606

[282] Langville & Meyer, 2006, pp. 4–5

[283] Jackson & Moulinier, 2007, p. 63

[284] Henzinger, 2005, p. 1

[285] Meadow et al., 2007, p. 161

[286] Jacso, 2005, p. 676

[287] Smith, 2007, p. 351

[288] Baeza-Yates & Castillo, 2006, p. 527

search in terms of two scores that are separately computed and then combined in order to arrive at the final relevance ranking scores. Those two scores are the content score and the popularity score of a Web page. The content score is, like just implied, comparable to the entire search process of traditional IR systems.[289] Even the computation of this content score can already be computationally intensive. To name only one example: Web pages that use a query term in their title are usually assigned a higher content score than ones that use a query term just in the main text. Overall, search engines already use a complex statistical analysis to determine the content score of a Web page, much like natural language search engines do in more traditional IR settings.[290] It goes without saying that the content score of a website is so-called query-dependent which means that it is computed by the search engine for each individual query after it has been submitted.[291] The popularity score, on the other hand, is often a query-independent value which means that Web search engines assign a score to each page independent of any specific query.[292] The popularity score is derived from an analysis of the Web's hyperlink structure. A combination of the content score and the popularity score finally leads to an overall score for each relevant Web page. The relevant pages resulting from a query are then presented to the user in decreasing order of their overall scores, in other words relevance-ranked.[293] We should keep in mind that due to the complexity of relevance ranking systems, even the engineers who developed an individual system would have a hard time to explain the precise reason why a particular result list at a particular time looks the way it does.[294] The basic ideas of relevance ranking algorithms, however, are often known even for commercial systems if they originated from publicised university or research lab work.[295] This is also the case for Google's popularity score technique PageRank, which we will now look at in more detail.

PageRank

In their 1998 paper[296] introducing PageRank, Google's founders Sergey Brin and Larry Page described PageRank as "an objective measure of [a Web page's] citation importance

---

[289] Langville & Meyer, 2006, p. 13

[290] Bast & Pyle, 2001, p. 296

[291] Langville & Meyer, 2006, p. 23

[292] Henzinger, 2005, p. 1

[293] Langville & Meyer, 2006, p. 13

[294] Witten et al., 2007, p. 9

[295] Weiss et al., 2005, p. 156

[296] Brin & Page, 1998

that corresponds well with people's subjective idea of importance. Because of this correspondence, PageRank is an excellent way to prioritize the results of Web keyword searches."[297] Still today, Google mentions its "breakthrough PageRank™ technology" prominently on its website.[298] The basic reasoning behind PageRank is strikingly simple. A Web page to which many hyperlinks point is thought to have higher chances of being authoritative on a topic than a page to which few, or no, hyperlinks point.[299] It is essential, however, to be aware of one more basic property of PageRank, namely its recursiveness. PageRank takes into consideration whether a page is being linked to by a Web page that has many incoming links itself, or by one with only few inlinks. This means that if a Web page is linked to by, let's say, http://www.cnn.com, it will have a higher relevance score than a Web page linked to by, for example, a personal Website.[300] Put differently, just counting the number of hyperlinks that point to a Web page ignores the fact that the pages that contain the hyperlinks can be of very different quality. PageRank therefore follows a so-called recursive approach: The PageRank value of a page A depends also on the PageRank values of those pages that point to A.[301]

The Random Surfer

When we look for metaphors to better understand PageRank values, we can view them as numbers that describe "how easy (or difficult) it is to find particular pages by a browsing-like activity".[302] Brin and Page themselves introduced[303] the notion of a random surfer. This idealised Web user randomly follows the hyperlink structure of the World Wide Web. Whenever he or she arrives at a page with several links to other pages (outlinks), he or she chooses any one of them at random, and continues this unplanned clicking process indefinitely. It is then the proportion of time that the random surfer spends on a given page in the long run that equals the PageRank value of that particular page. Even though the random surfer must never hit the "Back" button, he or she will still repeatedly find him- or herself returning to those pages that are well-connected to other ones, therefore spending more time

---

[297] Brin & Page, 1998, p. 109

[298] Google, 2009

[299] Henzinger, 2005, p. 1

[300] Smith, 2007, p. 351

[301] Henzinger, 2005, pp. 1–2

[302] Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata et al., 2000/6, p. 311

[303] Brin & Page, 1998, p. 110

on them than on others.[304]

7.3. The Underlying Assumption

In 2000, Brian Davison provided empirical evidence that Web pages sharing a link are more likely to be topically related than unconnected Web pages.[305] Connectivity-based ranking like that employed by Google's PageRank takes this idea further in order to arrive at the following key hypothesis. A hyperlink from a page X to a page Y means that the content of page Y is endorsed by the author of page X.[306] A considerable amount of latent human judgement is present in hyperlinks, and that judgement is utilised in order to capture a notion of authority, or importance, by connectivity-based ranking algorithms.[307] Web search engines are in a sense free-riding on the information that the people who create and manage web pages embed inside their hyperlinks.[308] Put yet another way, search engines like Google hold constant, huge elections, where each Web page votes for its favourite other Web pages, to find out which one of them is the most authoritative one.[309]

7.4. Some Challenges

Both Brin and Page, and Kleinberg already mentioned potential problems and challenges to using connectivity-based algorithms for ranking in their original papers: Our intuitive notion of authority that link analysis algorithms try to capture is made up of different criteria like relevance and popularity. Kleinberg observes that striking the right balance between those criteria is extremely difficult.[310] Similarly, Brin and Page state that "[c]ombining all of this information [that they store about Web pages] into a rank is difficult. We designed our ranking function so that no one factor can have too much influence."[311] It is obvious that a Web page that many links point to is popular, but what does that tell us? Richard Wiggins states that "popularity does not equate to authenticity, authoritativeness, accuracy, or currency. It doesn't even indicate that a source is believed or trusted -- people

---

[304] Langville & Meyer, 2006, p. 36

[305] Brian D. Davison, 2000, p. 277

[306] Baeza-Yates & Castillo, 2006, p. 530

[307] Kleinberg, 1999, p. 606

[308] Smith, 2007, p. 342

[309] Weiss et al., 2005, p. 94

[310] Kleinberg, 1999, p. 606

[311] Brin & Page, 1998, p. 113

can link to a source that they distrust or even scorn."[312] Also, link analysis techniques have to take into consideration that huge amounts of hyperlinks can be created automatically by one individual.[313] Yet another observation is that popularity-based ranking algorithms might become self-fulfilling prophecies over time: As people mainly select results from the top of result lists, they might ultimately only be aware of those pages that appear on the first pages of result lists, and therefore link only to them. Like that, pages that are already popular remain popular automatically.[314]

7.5. General Utility of the Approach

In 1995 Howard Turtle observed that individual legal documents are embedded within a larger structure whose main characteristics users understand, but traditional retrieval models ignore. He therefore suggested the use of link structure based on citations for computer-assisted legal research.[315] As Marie-Francine Moens observes, however, his advice has never been followed.[316] Yet during the same time the importance of Web information retrieval has constantly been increasing since the mid-1990s. Researchers in both academia and industry have been putting a tremendous amount of research into finding strategies for effective search within hyperlinked environments.[317] I therefore fully agree with Moens. It is at least advisable to examine retrieval models based on link analysis carefully in terms of their potential use in computer-assisted legal research.[318]

---

[312] Wiggins, 2003

[313] Baeza-Yates & Castillo, 2006, p. 530

[314] Barmakian, 2000, p. 417

[315] Turtle, 1995, p. 17

[316] Moens, 2007, p. 1761

[317] Bernstam, Herskovic, JR, Aphinyanaphongs, Aliferis, Sriram, & Hersh, 2006, p. 96

[318] Moens, 2005, p. 227

8. Legal Network Analysis

We now want to look at the first one of two research areas that are law-specific in nature, legal network analysis. We can start off by observing that a huge number of constructions and events around us adopt a network organisation. To name but a few, airline routes, roadmaps, power grids, the Internet, and the World Wide Web are all characterised by their patterns of interconnection.[319] Graphs, which are structures studied by mathematicians, underlie any such concrete network.[320] Mathematical graph theory has therefore become the basis for a truly multidisciplinary approach of studying network structures with applications in sociology, the information sciences, the computer sciences, and many others. What network graphs stress is that entities are connected to other entities, instead of solely existing in isolation.[321]

Basics of Network Analysis

In order to develop an understanding of the relationships between various entities (individuals, groups, computers, information, and so on), network analysts focus on two concepts: They call the entities that they study "nodes", and refer to the connections between the nodes as "links".[322] Jon Kleinberg observes that a network structure can provide us with extensive information about the content of an interlinked environment. He further qualifies, however, that this only holds true as long as we have effective means of understanding the network structure at hand.[323] Without going into the - partly still controversial - details of network analysis, we can establish that while there are various differences between individual networks, some general patterns can be observed. Researchers in the field have therefore focused on identifying several such patterns.[324] This is extremely valuable because network systems obey certain laws irrespective of their particular domain. Therefore, we can apply knowledge that is standard in one field to other, less developed fields, if we observe

---

[319] Witten et al., 2007, p. 88

[320] Otte & Rousseau, 2002, p. 441

[321] Dibadj, 2008, p. 9

[322] Fowler, Johnson, Spriggs, Jeon, & Wahlbeck, 2007, p. 325

[323] Kleinberg, 1999, p. 604

[324] Chakrabarti & Faloutsos, 2006, no. 3

that both fields share the same network structure.[325]

8.1. Two Common Network Structures

A very important feature of any network is the number of links per node and the distribution of this number over all the nodes in the network. These network properties are called the node degree and the degree distribution, respectively.[326] According to the statistical properties of the degree distribution, two broad classes of networks have been identified. Random networks with a homogeneous connectivity pattern, and scale-free networks with a heterogeneous one.[327]

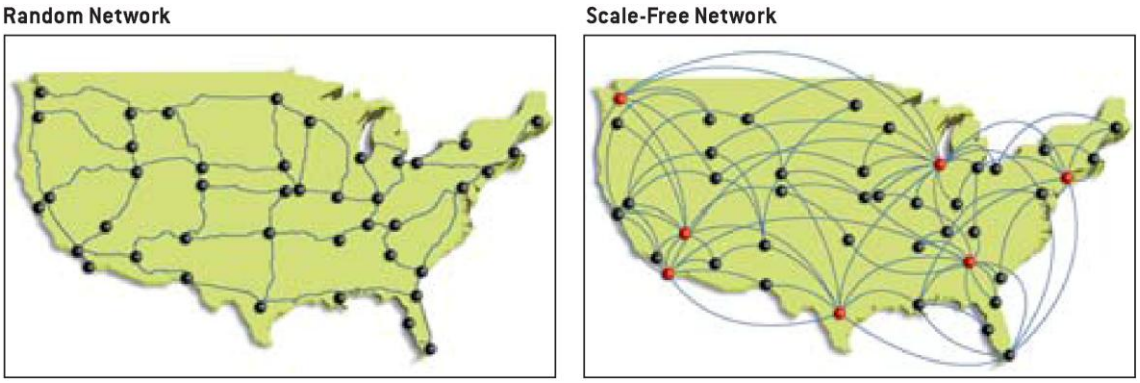


Figure 1: Random and Scale-Free Network Structures[328]

A simplified drawing (Figure 1, left map) of the U.S. highway system constitutes a random network. Nodes within the network are randomly placed, and are homogeneous in the sense that they all share approximately the same number of links. In contrast, the simplified American airline routing map (Figure 1, right map) resembles a scale-free network. It is characterised by a few important airports - hubs (red) - that are massively connected and therefore have a lot more links than the other nodes. This is a heterogeneous distribution.[329]

Random Networks - Bell Shape Curves

Looking at random networks (Figure 1, left map) more closely, we can establish that for

[325] Witten et al., 2007, p. 88

[326] Amaral & Ottino, 2004, p. 1659

[327] Katy Börner, 2007, p. 559

[328] Barabási & Bonabeau, 2003, p. 63

[329] Barabási & Bonabeau, 2003, p. 63

more than 40 years, scientists believed that all networks, however complex, followed that network structure. Trying to describe communication and biology networks, Paul Erdős and Alfréd Rényi had suggested that approach in 1959.[330] Their theoretical analysis of the properties of random graphs led to a number of important results, but most importantly, Erdős and Rényi observed that in random networks, it is extremely rare to find nodes that have significantly more or fewer links than the average. Using network terminology, each node will approximately have the same (link) degree.[331] When we remind ourselves that many of the entities that are measured by scientists have a typical size or "scale", the success of the random network theory is understandable. In many large networks, individual measurements are indeed centred around a typical value. Quantities are then distributed in a bell-shaped curve like the one shown in Figure 2 below. People's heights for example are distributed fairly narrowly, which leads to a bell-shaped distribution graph. Most adults are between 150 and 200 cm tall. There is some variation in our heights, but we actually never meet people who are only 10 centimetres, or 5 metres tall.[332]

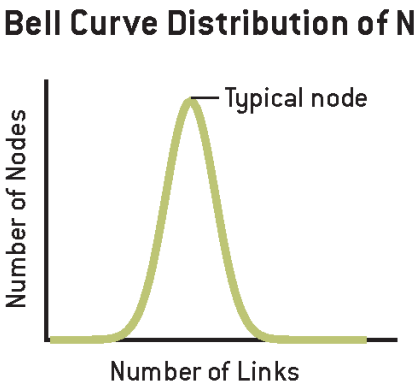


Figure 2: The Degree Distribution of Random Networks [333]

Traditionally, like I just mentioned above, also large and complex networks have been described with Erdős's and Rényi's random graph model. This was done, however, without possessing actual data on those large networks. When we finally had the technical means to obtain and process actual network data of large networks, the data indicated a degree distribution that had been unpredicted by all available random network models.[334]

---

[330] Barabási & Bonabeau, 2003, p. 62

[331] Smith, 2007, p. 318

[332] Witten et al., 2007, pp. 88–89

[333] Barabási & Bonabeau, 2003, p. 63

[334] Barabási & Albert, 1999, p. 510

Scale-Free Networks - Power Law Distributions

Albert-László Barabási and Réka Albert termed networks that share the observed particular degree distribution "scale-free"[335] ones. In contrast to random networks, they are characterised by an uneven distribution of links between the nodes.[336] Homogeneous random networks are common in nature, but there are also numerous cases of scale-free networks where - in terms of the degree distribution graph - the distribution of nodes far to the right of the mean is significantly higher than in random networks.[337] Carrying forward the height example from the last chapter, we can observe: If our heights followed the "power law" distribution of scale-free networks, being very short would be extremely common. The average height, however, would be a lot taller, also because some individuals would in fact be incredibly tall. As a result, it would be quite normal to run into someone five or ten times taller than the average from time to time.[338] Figure 3 right below shows the degree distribution found in scale-free networks.

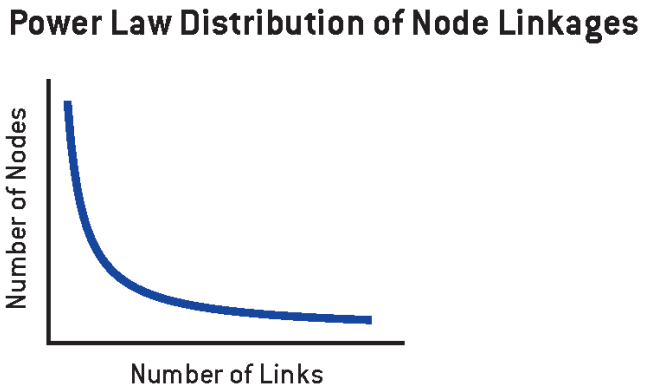


Figure 3: The Degree Distribution of Scale-Free Networks [339]

Following Barabási's and Albert's initial research, a quickly growing literature proved that many networks in the real world are scale-free ones. In cellular metabolism, in actors' collaborations in Hollywood, in the protein regulatory network, in scientists' publication collaborations, power-law distributions occur in an extremely diverse range of

---

[335] Barabási & Albert, 1999, p. 509

[336] Baeza-Yates & Castillo, 2006, p. 535

[337] Chakrabarti & Faloutsos, 2006, no. 3

[338] Witten et al., 2007, p. 90

[339] Barabási & Bonabeau, 2003, p. 63

phenomena.[340] Most important in terms of this thesis, however, is Barabási's and Albert's very first subject to testing: they investigated the structure of the World Wide Web. Their research revealed that the distribution of links pointing to Web pages (the in-degree of Web pages) did not fit a bell shaped curve as predicted by the random graph model, but resembled much more a power law distribution of a scale-free network.[341]

8.2. The Webgraph

Methodically, in order to examine the degree distribution of the World Wide Web, we have to view the static Web, that is the static HTML pages together with the hyperlinks between them, as a network graph. Each Web page is a node, and each hyperlink is a link.[342] The ones who first[343] observed the scale-free network structure of this so-called Webgraph were not only Barabási and Albert[344], but also Ravi Kumar and others[345]. As we have already seen the necessary development of modern Web search techniques has led to extensive research in the field of link analysis. Link analysis research, in turn, naturally examines the details of the network structure of the World Wide Web.[346] Even more so after the discovery of its unexpected nature, the newly emerged study of the Webgraph has therefore understandably attracted a large interest in the scientific community. This has been primarily due to the ongoing efforts to further develop Web search techniques. Looking at it from a network analysis point of view, we can in fact conclude that the heterogeneous link structure of the World Wide Web forms the very basis of Web search ranking algorithms such as PageRank.[347] As Dietmar Wolfram correctly observes, however, observations of inverse power laws applying to World Wide Web content might in fact not be as unexpected as some argue. Information scientists studying regularities in print-based and electronic literature have long been observing power law distributions.[348] This observation is one of the main reasons why we will look at legal citation analysis after this chapter[349].

---

[340] Barabási & Bonabeau, 2003, p. 64

[341] Smith, 2007, p. 324

[342] Manning et al., 2008, p. 389

[343] Donato, Laura, Leonardi, & Millozzi, 2004, p. 239

[344] Barabási & Albert, 1999, p. 509

[345] Kumar, Raghavan, Rajagopalan, & Tomkins, 1999, p. 1486

[346] Kleinberg, 2006, p. 210

[347] Donato et al., 2004, p. 239

[348] Wolfram, 2003, p. xii

[349] See below, 9 Legal Citation Analysis, p. 66

8.3. Network Analyses on Legal Document Collections

Summarising previous network analyses of legal document collections, Reza Dibadj observes that network analysis has so far been "vastly underutilized in the law". We can, however, find a few instances where legal scholars have made use of it.[350]

In 2000, David Post and Michael Eisen tried to further discover general principles about the structure of the US legal system by looking at the web of citations between cases. Their conviction that the web of citations in fact forms "a critical component of the network of rules that comprise "the law" in any area" provided the theoretical background for their citation study.[351] Post and Eisen performed the citation analysis using cases decided by the New York Court of Appeals and the United States Court of Appeals for the Seventh Circuit dated from 1930, 1950, 1970, and 1980.[352] In order to prove the "fractal structure" of common law legal systems, the two scientists developed the hypothesis that the specific fractal structure they were looking for would be reflected in a "power-law distribution of the output [...] of those systems."[353] We can establish at this point that what Post and Eisen were effectively hypothesising was that the citation network of the legal document collection that they studied would have a scale-free structure. It turned out that this hypothesis was in agreement with the empirical citation data. Post and Eisen did find that a very small number of the cases they examined received a large percentage of all citations, whereas the vast majority of cases were cited very little.[354]

In 2007, James Fowler and his colleagues constructed a complete network of 26,681 majority opinions written by the US Supreme Court and the cases that cite those majority opinions from 1791 to 2005. Observing a power-law link distribution of their data as well, the scientists - by that time aware of the abundance of scale-free networks in nature - surmised "that there is something systematic about the evolution of law that mimics the evolution of other network phenomena."[355]

Lastly, Thomas Smith reports in "The Web of Law" about his study applying network

---

[350] Dibadj, 2008, p. 9

[351] Post & Eisen, 2000, p. 545

[352] Post & Eisen, 2000, p. 571

[353] Post & Eisen, 2000, p. 570

[354] Post & Eisen, 2000, pp. 570-583; 574

[355] Fowler et al., 2007, p. 344

theory to a huge collection of more than four million US legal citations. Just like the other aforementioned scientists, Smith observes that "[the] American case law network [...] appears much like that of the Web and other citation networks, such as those of scientific papers."[356]

8.4. General Utility of the Approach

All the three studies that I just mentioned in the area of legal network analysis include evidence that network analysis is a useful approach also in the legal domain. Thomas Smith observes that the network of legal citations might very well be the "oldest, largest, and best documented citation network ever created". Lawyers have been using it to learn about the law on any given topic, but it has still been a lot less explored in terms of network analysis than other networks.[357] Post and Eisen correctly argue along the same lines. Legal citation data, being the raw material of which the network of law is made up of, is available in abundance. It should be used for network analysis purposes to "uncover general principles about the structure of the legal system".[358] Fowler and his colleagues take this idea even one step further. We could establish, they believe, that lawyers think of the law as an "interconnected set of rules" that evolves by the repeated use of some rules, and by ongoing interpretation over time. Focusing on case law, the scientists propose to look at how a particular opinion is incorporated into the network of law in order to find out how relevant an opinion is.[359] This proposition is fully in line with the main purpose of link analysis in Web search, and you have probably already realised that the three studies that I just categorised as legal network analysis studies could as well be mentioned as works done in legal citation analysis. We will therefore now take a more in-depth look at legal citation analysis. For a long time, citation analysis has been using citations to produce quantitative estimates of the "impact" of scientific papers and journals.[360]

---

[356] Smith, 2007, p. 313

[357] Smith, 2007, pp. 310; 354

[358] Post & Eisen, 2000, p. 545

[359] Fowler et al., 2007, p. 326

[360] Kleinberg, 1999, p. 618

9. Legal Citation Analysis

The chapter on the developments in Web search showed us that link analysis algorithms use inlinks as a measure of relevance assessment. The commercial success of Google to a certain extent proves the effectiveness of that approach.[361] When we look at it from a network analysis point of view, this number of inlinks is called a node's in-degree, and we can use network analysis concepts to learn more about the interlinked structure.[362] But in fact, we can trace the underlying idea of using inlinks or in-degree as a method of relevance assessment back further than both Web search and network analysis. Only when we go back to explore basic concepts of citation analysis, introduced by Eugene Garfield, we find the roots of the aforementioned technique.[363] In a sense, citation analysis methods have only been rediscovered and modified for both network analysis and Web search.[364]

9.1. Citation Indexing

In 1955, Garfield pointed out several shortcomings of the subject indexes and classified indexes that existed at the time. He claimed that neither one did properly address the facts that articles might deal with a variety of subjects, that terminology changes over time, and that specialised vocabularies exist within disciplines. Arguing that an "association-of-ideas" or a "thought index" was what was needed, Garfield proposed the introduction of a citation index. Interestingly, he modelled his proposal after the already well-established Shepard's citator.[365] Garfield himself described what happened after he had been advised of Shepard's like this:

> "I didn't know what Shepard's was so I went down to the Enoch Pratt Free Library and went into the reference room. I found Shepard's Citations and I literally screamed, "Eureka." [...] When I saw the Shepard's Citations I found the methodology that I needed for linking all these things."[366]

---

[361] Wolfram, 2003, p. 161

[362] Ma, Guan, & Zhao, 2008, p. 801

[363] Weiss et al., 2005, p. 101

[364] Wolfram, 2003, p. 53

[365] Garfield, 1955, p. 108

[366] Garfield, 1987, pp. 13–14

The concept of so-called "citation indexing" is in fact strikingly simple. When scientists document their own research, scientific tradition presupposes that they refer to earlier works that relate to the subject matter of research at hand.[367] As a result of that requirement, the publications found in scientific journals, be it papers, notes, reviews, or other documents, all contain citations. Those citations lead to documents that support, provide precedent for, illustrate, or elaborate on what the authors of the document at hand have to say. We can therefore think of citations as the formal, explicit linkages between papers that have particular ideas in common. Around these linkages a citation index is built. That index on the one hand lists publications that have been cited, and on the other hand identifies the sources of those citations. By just knowing one paper on a particular subject that has been cited, anyone can find more relevant documents. Those documents, in turn, provide a list of new citations that can be individually looked at. One of the main strengths of citation indexing is its simplicity.[368]

Possible Further Uses of Citation Indexing

Garfield proposed and introduced citation indexing with bibliographic purposes in mind. Still, he himself speculated that "the most important application of citation indexing may prove to be nonbibliographic." As the activities of science manifest themselves in the science literature, a comprehensive and multidisciplinary citation index might offer valuable insights into science as a whole. Both the structure and the development of science could be examined. When thinking along those lines, the possibilities of using a comprehensive citation index seem to be almost endless: We might be able to evaluate the research role of individual journals, individual scientists, scientific organisations and communities. It might become possible to explore the relationships among journals and between journals and fields of study. The impact of current research could be determined. It might even become possible to establish an alert system for important, new interdisciplinary relationships, as well as for fields of study whose development accelerates. Historically, the sequences of developments that have led to major scientific advances could be further explored.[369] Pranas Zunde, thinking along the same lines, identified three broad application areas where citation indexes could be successfully used as early as 1971:

1. Scientists, publications, and scientific institutions could be quantitatively and qualitatively

[367] Nicolaisen, 2007, p. 610

[368] Garfield, 1979, p. 1

[369] Garfield, 1979, p. 62

evaluated.

2. The historical development of science and technology could be modelled.

3. Information search and retrieval could be supported.[370]

9.2. Citation Analysis

Every one of those ideas of using citation indexing for more than just bibliographic purposes has been realised in some way by now. For decades, citation indexing has been providing the raw material that information scientists have analysed in various ways.[371] Those developed sophisticated methods of using citation data fall within the area of "citation analysis", which in turn is a cornerstone of so-called bibliometrics.[372] In addition to studying the citation structures of documents, bibliometrics also deals with their actual texts.[373] As far as the merely bibliogaphic use of citation data is concerned, the merits of citations are clear, and their use is not controversial. The use of citation analysis to provide quantitative measures on top of that, however, has always been very controversial.[374]

9.3. Citation Analyses on Legal Document Collections

When we again focus on citation analyses on legal document collections, we can begin by looking at the citator services Shepard's and KeyCite that are used in legal research once again[375]. From what we have established by now, both citation indexes also analyse legal citations, which puts them close to being legal citation analysis tools.[376] Forms of citation analysis had therefore commonly been practised in the legal world long before the technique was re-discovered for science. As legal publications and their interconnections constitute the law itself, rather than being mere by-products of the research enterprise like in other scientific disciplines, this should not come as a big surprise.[377] When talking about citations and the law, I already mentioned[378] that comparatively early, truly radical approaches were

---

[370] Zunde, 1971, p. 11

[371] Shapiro, Mar., 1991, p. 1457

[372] Shapiro, 1992, p. 338

[373] Langville & Meyer, 2006, p. 123

[374] Pinski & Narin, 1976, p. 297

[375] See also above 5.2 Legal Citators, p. 43

[376] Chiorazzi, 2002, p. 5

[377] Shapiro, 1992, p. 337

[378] See above p. 48

proposed in the area of legal citation analysis. In 1970 Stephen Marx recommended[379] that CALR users should not rely on "the superimposed key word systems", but should rather make use of "the cross-citation structure because it provides a more rational linkage between relevant cases". Similarly, Colin Tapper argued[380] that common citation patterns, at least in common law jurisdictions, could provide the basis for the retrieval of legal documents. In the mid-1990s, Stuart Sutton looked back at those approaches and concluded that previous research into the area of legal citation analysis had been "encouraging [...] but not without problems".[381] By the same time, however, information scientists had been working on more sophisticated theoretical foundations for citation analysis and its potential applications for some time[382], which prompted Fred Shapiro to observe that "in law, the birthplace of citation study, even richer results may be possible than in the other fields to which that study has subsequently been applied."[383] Similarly, Howard Turle stated in 1995 that citations "are important in the legal domain but they remain under exploited in retrieval." He pointed out ongoing work by Daniel Dabney[384], but was still of the opinion that much work remained to be done.[385] Somewhat ironically, Dabney himself, even though he had done seminal research in the area by then, subsequently mentioned in 2001 that "relatively little has been reported on the science of legal citation analysis."[386] As I said before, the studies that I summarised in the chapter on legal network analysis[387] do certainly at the same time constitute work done in legal citation analysis. Even so, however, much work remains in legal citation analysis. The scarcity of projects in this area might have to do with an observation that I also frequently made myself during the course of my studies. Basic citation analysis concepts do appear simple, but only at first sight. The following collection of controversies and problems surrounding citation analysis, and legal citation analysis in particular, is intended to give an impression of that.

---

[379] Marx, 1969-1970, p. 124

[380] Tapper, 1974, p. 30

[381] Sutton, 1994, p. 199

[382] Ogden, 1993, p. 44

[383] Shapiro, 1992, p. 339

[384] Dabney, 1993, p. 1

[385] Turtle, 1995, p. 48

[386] Conrad & Dabney, 2001, p. 288

[387] See above 8.3 Network Analyses on Legal Document Collections, p. 64

9.4. Controversies and Problems Surrounding Legal Citation Analysis

The following concept lies at the heart of citation analysis techniques: A citation, so the argument goes, usually indicates that the author has read the corresponding document and made the decision that it merited an explicit reference.[388] Critics of citation-based rating methods often attack this fundamental assumption that underlies them, namely that a person who cites an article has actually read it, and thought that it was significant enough to warrant a reference.[389] Having done an extensive review of studies on citing behaviour, Lutz Bornmann and Hans-Dieter Daniel summarise the general trend of their findings: The acknowledgement of intellectual and cognitive debts to colleagues is not the sole reason for citing. The inclusion of a citation can also be caused by a number of other factors.[390] In fact, it was already Eugene Garfield himself who published the earliest paper that lists a whole range of possible motivations of citers.[391] He observes that in general, citations are used to provide "documentation" or support for specific statements at hand. Citations in scientific papers are, however, also provided for a number of other reasons including

1. Paying homage to pioneers
2. Giving credit for related work (homage to peers)
3. Identifying methodology, equipment, etc.
4. Providing background reading
5. Correcting one's own work
6. Correcting the work of others
7. Criticizing previous work
8. Substantiating claims
9. Alerting to forthcoming work
10. Providing leads to poorly disseminated, poorly indexed, or uncited work
11. Authenticating data and classes of fact - physical constants, etc.
12. Identifying original publications in which an idea or concept was discussed.
13. Identifying original publication or other work describing an eponymic concept or term
[...]
14. Disclaiming work or ideas of others (negative claims)
15. Disputing priority claims of others (negative homage)"[392]

---

[388] Perry, 2006, p. 21

[389] Perry, 2006, p. 24

[390] Bornmann & Daniel, 2008, p. 66

[391] Bornmann & Daniel, 2008, p. 51

[392] Garfield, 1977, p. 85

A principal criticism of citation studies is that the mere number of citations is a poor proxy for what citation analysts really intend to measure. In many citation studies, all citations are treated as uniformly positive recognitions, which causes problems similar to the ones that I mentioned when talking about relevance in information retrieval[393]. Neither the specific significance of the citation at hand, nor the citer's motivation for including a citation, are taken into consideration when citations are simply counted.[394] William Landes and Richard Posner correctly note that there are various seemingly extraneous considerations that influence the number of citations a document obtains. The number of journals in a particular field of study, the development of that number over time, and also the citation conventions of different areas all have to be taken into consideration.[395] This causes researchers who want to measure the scientific impact of documents or individuals to be subdivided into two groups.

One camp believes that citation analysis is suitable as a means of assessment for scientific impact. Working in their favour is that a considerable amount of literature has proven that the number of citations that a scientist receives correlates nicely with other assessments of the scientist's impact or influence. Bornmann and Daniel list studies that have established a correlation between the number of citations and awards, Nobel laureateships, departmental prestige, research grants, academic rank, and peer judgments.[396] An overview about even more studies on the association between citations to scientists' publications and other assessments of the scientists' scientific impact or influence can be for example found in a recent book[397] by Blaise Cronin.

The opposing camp interested in obtaining impact measures stresses that citation counts depend on many factors that have nothing to do with scientific impact. Studies have identified various factors that influence the probability of a work being cited or not, irrespective of its merit: time-dependent factors, field-dependent factors, journal-dependent factors, article-dependent factors, author/reader-dependent factors, and even the availability of publications and technical problems have been shown to have an influence on citation counts.[398]

It might not come as a surprise to you that I consider myself to be a strong supporter of

---

[393] See above 2.4 Evaluation in Information Retrieval, p. 18

[394] Wolfram, 2003, p. 50

[395] Landes & Posner, 2000, p. 320

[396] Bornmann & Daniel, 2008, p. 46

[397] Cronin, 2005, pp. 125–129

[398] Bornmann & Daniel, 2008, pp. 46–47

citation analysis. Still, I am convinced that we must not ignore any of the issues raised by opponents of citation analysis techniques, even though there is a general correlation between citation counts and impact. David Adam phrased this perfectly, I think.

"Important papers, the argument goes, will be cited more frequently. As a general rule, that is a reasonable assumption. But apply it blindly, without regard to the quality and limitations of the raw data, and the conclusions you draw may be far from reasonable."[399]

I will now try to categorise and address some of the limitations and difficulties that citation analyses in the legal domain necessarily have to deal with if they want to draw reasonable conclusions beyond general observations, that means especially on a document- or author-specific level.

Obliteration

Firstly, the phenomenon of "obliteration by incorporation" poses a possible problem to citation counting and citation analysis. American sociologist Robert King Merton first identified this process[400] where previous works do have an influence on documents at hand, but that influence is not reflected in explicit citations. When the work of previous authors has become so influential that it is generally viewed as being part of the common body of knowledge, scholars tend to believe that they no longer need to cite it explicitly.[401] Already in 1973, Michael J. Moravcsik tellingly observed: "Anybody today who cited Einstein's original paper when he writes down $E = mc^2$ would be laughed at."[402] Thomas Smith observes that obliteration is also a common phenomenon in legal contexts. Today, citing Marbury versus Madison, a landmark case in US law, in a routine case when reviewing a statute for constitutionality would appear "somewhat pedantic or cranky". Smith, however, at the same time points out that this problem is in fact not a serious one. Taken as a percentage of all cases, very few fall into this rarified category.[403]

Self-Citations

Fred Shapiro, whose "Collected Papers on Legal Citation Analysis"[404] still constitute

---

[399] Adam, 2002, p. 727

[400] Merton, 1968, p. 35

[401] Shapiro, Mar., 1991, p. 1453

[402] Moravcsik, 1973, p. 269

[403] Smith, 2007, p. 348

[404] Shapiro, 2001, p. i

the prime research collection in the area of legal citation analysis, identifies self-citations and negative citations as the most problematic motivations for citing in terms of citation analyses. Self-citations "may inflate an author's citation total".[405] The general concern here is that authors might over-cite their already published works, regardless of their quality. It is true that the personal development of scientific ideas necessarily leads to entirely legitimate self-citations, but it is evident that ubiquitous citation rankings effectively encourage self-citation. There are, however, two main reasons why self-citations should not pose an unsolvable problem to citation analyses. First, abundant unjustified reference to his or her own work should harm any author's academic prestige, and should therefore not be common in the first place. Also, superfluous citations will usually be omitted during any editorial publication process.[406] Second, as also Shaprio observes, self-citations are unlikely to have much effect because of their small percentage compared to usually large citation totals.[407] In a study done by James Leonard, for example, self-citations constituted only 5.6% of all citations that were examined.[408]

Negative Citations

Moving on to the problems surrounding negative citations, Shapiro strikingly describes one potential threat that they might pose for legal citation analyses: Within legal periodicals, a "high [citation] total for a shoddy piece of scholarship" could be the misleading result of citation counting.[409] At first sight, it seems plausible that critics might indeed cite a bad article quite often, which would really mean that the particular citation frequency could not be indicative of the document's quality. Another observation, however, seems to move things back into perspective. Most of the times, a low-quality document will simply not be cited at all in subsequent writings. Legal scholars will simply ignore insignificant papers.[410] An article that manages to receive hundreds of critical citations probably triggers advances in the professional discourse and therefore deserves to be considered as a high-impact document, albeit being a controversial work.[411] Thinking along the same lines, William Landes and his colleagues apply this general argument to the case law environment. In their citation study,

---

[405] Shapiro, Oct., 1985, p. 1543

[406] Perry, 2006, pp. 24–25

[407] Shapiro, Jan., 2000, pp. 412–413

[408] Leonard, 1989-1990, p. 191

[409] Shapiro, Oct., 1985, p. 1543

[410] Posner, 2000, p. 387

[411] Perry, 2006, p. 24

they decided not to distinguish between different kinds of citations. Their reasoning for treating critical citations the same way as positive citations is simple. They believe that "it is easier to ignore an unimportant decision than to spell out reasons for not following it."[412] Looking at Leonard's citation study again, we can observe that negative citations only accounted for 10.6% of the citation total.[413]

Age of Document and Size of Literature

Ian Ayres and Fredrick Vars point out another major difficulty for citation studies. It's obvious that the number of citations a document receives also depends on how many chances it has had to obtain citations. Firstly, this means that recent documents are penalised in citation rankings as long as those rankings do not take time factors into consideration.[414] In the area of case law, for example, this means that simple counts of citations to an individual judge's opinions as a measure of his or her judicial impact are misleading as long as they do not take each judge's length of stay on the bench into account. Those judges with longer tenure automatically have more opportunities to collect citations, independent of their judicial impact. It therefore seems that a more suitable measure for similar citation analyses is often the average number of citations per year.[415] Looking at possible problems regarding "opportunities to get cited", there is, however, another bias tilting the playing field against certain documents. Taking secondary sources as an example, some topics have a far larger scholarly literature than others, resulting in countless opportunities to pick up citations. Other areas are a lot less frequented by legal journals, which makes it a lot harder for documents in that area to get cited.[416]

9.5. General Utility of the Approach

Like in the previous two chapters, I want to end this one as well by summing up the general utility of legal citation analysis in the context of my thesis. First of all, we can observe that in the legal domain, several databases store citation data, which makes obtaining large data sets for citation analyses comparatively easy.[417] Full-text online CALR systems

---

[412] Landes, Lessig, & Solimine, 1998, p. 273

[413] Leonard, 1989-1990, p. 191

[414] Ayres & Vars, 2000, p. 430

[415] Landes et al., 1998, pp. 279–280

[416] Shapiro, Mar., 1991, pp. 1459–1460

[417] Post & Eisen, 2000, p. 570

have generally made retrieving references much more convenient, even the retrieval of exotic combinations of references in conjunction with keywords has become possible.[418] In terms of the availability of citation data, we can therefore observe that the legal domain is well-suited for citation analyses.

As regards the basic, perpetual discussion about whether or not citation analysis should be used as an evaluative tool in law at all, I think that those legal academics who have performed large-scale citation studies have developed an educated approach to the assessment and use of citation analysis in the legal domain. Fred Shapiro concedes that citation counting "falls somewhere between historiography and parlor game"[419], but also points out that it has been shown that citation counts in the area of legal journals do "correlate highly with peer judgments of scholarly influence. Lists of most-cited works therefore serve to draw attention to authors and publications that, by a rough measure, have had the most extensive impact on scholarship."[420] He furthermore mentions that almost all citation analysts state that counting citation measures a "quality" that is socially defined, reflecting the utility of documents at hand to other people, rather than measuring their intrinsic merit.[421] Likewise, William Landes and his colleagues note that we only get a crude and rough proxy for measuring influence when we look at citations.[422] Limitations like that do not, however, prevent the aforementioned legal citation analysts from strongly believing in the legitimacy of methodically sound legal citation analyses. The core assumptions that underlay two recent legal citation studies provide perfect examples of what I would call a "knowledgeable application" of legal citation analysis. Thomas Smith observes

> "While it may be hazardous to conclude that one actively cited case is more important or authoritative than another, it certainly seems plausible to distinguish between cases that are very actively cited, modestly cited, rarely cited, and never cited."[423]

---

[418] Shapiro, 1992, p. 339

[419] Shapiro, Oct., 1985, p. 1540

[420] Shapiro, Mar., 1991, pp. 1450–1451

[421] Shapiro, Mar., 1991, p. 1454

[422] Landes et al., 1998, p. 271

[423] Smith, 2007, p. 347

and David Post and Michael Eisen add a telling practical example to that approach:

"The difference between a case that cites to 100 previously decided cases and one that cites to 10 is surely due to many factors [...] but we think it reasonable to assert that the former case in some sense raises (and resolves) more questions than the latter."[424]

[424] Post & Eisen, 2000, pp. 570–571

10. Developing a Hypothesis

In the "Elements of a Theory" section, we have seen that when the goal is to measure the "impact" or "relevance" of individual entities within an interlinked structure, common ideas underlie the efforts of various scientific fields. This constitutes further theoretical evidence that those common concepts should also be applicable to computer-assisted legal research.

My exact proposition has been that the use of citation analysis techniques could improve current CALR systems in terms of the relevance ranking of retrieval results. At this point, I further develop this initial proposition into a hypothesis that I can then test on an actual document collection. The "Elements of a Theory" section showed that when we use citation analysis techniques to obtain accurate results on a document-, or author-specific level, we must be aware of and deal with many potential pitfalls that can lead to wrong results. For my hypothesis, however, I go back to one of the most basic citation analysis concepts which reads very simple: Counting citations is a way to measure importance. In the particular context of this thesis, I believe that such a simplistic approach is permissible. I am only trying to provide proof for a general potential of citation-based techniques to improve relevance ranking in the legal domain, without claiming to obtain accurate results on an individual document level. I take inspiration from James Fowler and his colleagues who also observe that "at the most basic level", the number of citations that a case receives from other cases can be used to measure how important the case at hand is.[425]

I therefore hypothesise that a basic citation-based algorithm, despite all its shortcomings, could already be used to improve relevance ranking in computer-assisted legal research.

[425] Fowler et al., 2007, p. 329

SECTION III - TESTING THE HYPOTHESIS

11. Supreme Court of Justice Experiment

For my experiment, I have computed the degree distribution of 80,195 opinions written by the Austrian Supreme Court of Justice between 1985 and 2008. Each opinion represents a node, 242,078 references from headnotes to those opinions constitute the links. I describe the resulting distribution graph and explore whether it could successfully be used to prioritise legally relevant Supreme Court cases when relevance ranking retrieval results. An opinion is considered to be "legally relevant" if it has been included in an official law report published by the Austrian Supreme Court of Justice itself. To my knowledge, there have not been any analyses of citation frequency distributions in the realm of Austrian law so far. The main inspiration for my experiment came from Thomas Smith's article "The Web of Law"[426] in which he describes a study applying network theory to over four million US legal citations[427]. Smith in turn modelled[428] his work on the research of statistical physicists that I already mentioned[429] in the chapter on legal network analysis.

11.1. The "RIS Justice" Database

Just like the other case law databases of the Austrian legal information system RIS, the so-called Justice ("Justiz") database (http://www.ris.bka.gv.at/Jus/) consists of two sub-databases: A headnote database ("Rechtssätze (RS)") stores headnote documents created by the court. Secondly, an opinions database ("Entscheidungstexte (TE)") contains the actual full texts of most of the court's decisions.[430] The Austrian Supreme Court, like the country's other two High Courts, has traditionally been carrying out extensive legal documentation tasks, coordinated by the court's records office. A massive amount of efforts goes especially

---

[426] Smith, 2007

[427] See above, 8.3 Network Analyses on Legal Document Collections, p. 64

[428] Smith, 2007, p. 312

[429] See above 8.1Two Common Network Structures, p. 60

[430] Österreichischer Verfassungsdienst, 1994, p. 10

into the creation and modification of headnote documents at the court, before they become available in the RIS Justice headnote database. Each document is intellectually processed by several legal specialists. Comparable to the way it is handled in the United States, individual headnotes are created to summarise significant legal points made by the court.[431] More specifically this means that if the legal specialists at the Austrian Supreme Court consider points decided in an opinion to be of legal relevance, each single point leads either to the creation of a new individual headnote, or a reference to the opinion at hand is added to a pre-existing headnote. The number of headnotes written or modified in response to an opinion therefore depends on the number of legally relevant issues the opinion addresses. On average, a case in my document collection is cited by three headnotes, although a lot of cases do not have headnotes referring to them at all, and some opinions have an unusually high "headnote citation count".[432]

11.2. The Text Collection

We have just seen that a lot of intellectual effort goes into the creation of headnotes in Austria, just like in other countries. Legal professionals value headnotes everywhere where they are created because they make it possible to quickly assess the content of a case.[433] What makes the Austrian situation unique, however, is that all the headnote documents created by the three High Courts are available to everyone free of charge on the legal information system RIS. Similar, for example, to the manually created hyperlinks which connectivity-based Web search ranking algorithms make use of,[434] a monumental amount of latent human expert judgement is present in those headnotes. Not only the texts of headnotes contain priceless information, but also the manually created connections between headnotes and opinions. I therefore decided to try to utilise those connections for my experiment. Even though it is certainly not the intended use of the RIS Justice ("Justiz") database (http://www.ris.bka.gv.at/Jus/), minor tweaking made it possible to download Supreme Court of Justice documents from it in bulk. The compilation of my document collection still required continuous searches on 1 and 2 January 2009.

My document collection contains all the opinion full texts written by the Austrian Supreme Court between 1985 and 2008 that were available as of 1 January 2009. The

---

[431] Armstrong & Knott, 2006, p. 10

[432] See right below 11.3 The First Experiment: Power-Law Distribution, p. 80

[433] Moens, 2007, p. 1749

[434] Kleinberg, 1999, p. 606

starting point had to be 1985 because according to the Court[435], this is the current start date for the comprehensive online publication of all opinion full texts in the RIS Justice database. As regards the headnote documents, I had to download all available headnotes, irrespective of their creation dates. A lot of headnote documents created before 1985 include references to "younger" opinions, therefore all available headnote documents had to be examined for references to all opinions found in my collection.

In total, my corpus contains 80,195 opinion full texts (dated 1985-2008), and 121,699 headnote documents (dated 1914-2008). I downloaded the documents in HTML format, in which they span over two gigabytes of information.

11.3. The First Experiment: Power-Law Distribution

For my first experiment I formalise the view of the opinion full texts and headnote documents as a graph. I ignore the text in all the documents, and focus instead entirely on the citations between headnotes and opinions. The experiment asks whether the distribution of opinions according to the number of headnote citations that they receive resembles a power law. If the citation distribution graph of my case law document collection looked like that of scale-free networks in general, it would share this basic property with the World Wide Web. Like Thomas Smith observes, this would be a strong indication that searching for relevant cases in the "Web of Law" could be improved by using techniques initially developed for Web search.[436]

Distribution Graph Construction

The documents that I downloaded are partitioned into fields, which made it easier to extract only the citation data of interest in order to compute the citation distribution graph. The details of how I managed to do so, especially the source code used, are somewhat beyond the scope of this text. I will, however, explain the main steps of my work. I realised the task of computing the frequency distribution mainly by programming three so-called "scripts" (comparable, but in detail different from programs) using the programming language Python. The main script was based on so-called regular expressions. Using regular expressions is a common approach in different programming languages for parsing, that is analysing, free text.[437] Having had no previous experience in computer programming, I chose

[435] Supreme Court of Justice [Oberster Gerichtshof]

[436] Smith, 2007, p. 350

[437] Jackson & Moulinier, 2007, p. 74

to use Python as the programming language because it is a "rather simple language at heart"[438], but still offers the "power and general applicability of traditional compiled [programming] languages".[439] Moreover, Python is open source software and freely available at http://www.python.org. All my experimentation was done on a single laptop with an Intel Core 2 Duo 2 Ghz processor, with 2 GB of memory, running Windows Vista.

The first, most important Python script extracted all references to opinions from the 121,699 headnote documents. It then returned the docket numbers of all Supreme Court opinions (not only those written between 1985 and 2008) that had ever been referenced in headnotes. Along with those docket numbers, the Python script returned the citation counts for each opinion. For obvious reasons this was a computationally complex task and took 3 hours and 22 minutes on the laptop that I used.

My second Python script filtered out those docket numbers that identified opinions which were part of my 1985 - 2008 opinion full text selection. The reason for that filtering step is as follows: As mentioned above[440], the RIS Justice database only stores selected opinion full texts for the years before 1985. I therefore had no way of telling how many opinions had been written in total in the years prior to 1985. I only had citation information as to how many, and which opinions had been cited by headnotes. This meant that I had to discard headnote citations to pre-1985 cases, because a major goal of my distribution graph is to give an accurate picture of the ratio between cited and uncited Supreme Court opinions.

At last, a third Python script took the 80,195 individual citation counts (opinions that had not been cited at all had automatically been assigned a zero "headnote citation score") and computed the citation distribution.

---

[438] Mertz, 2003, p. xi

[439] Chun, 2008, p. 5

[440] See just above, p. 80

The First Experiment's Distribution Graph

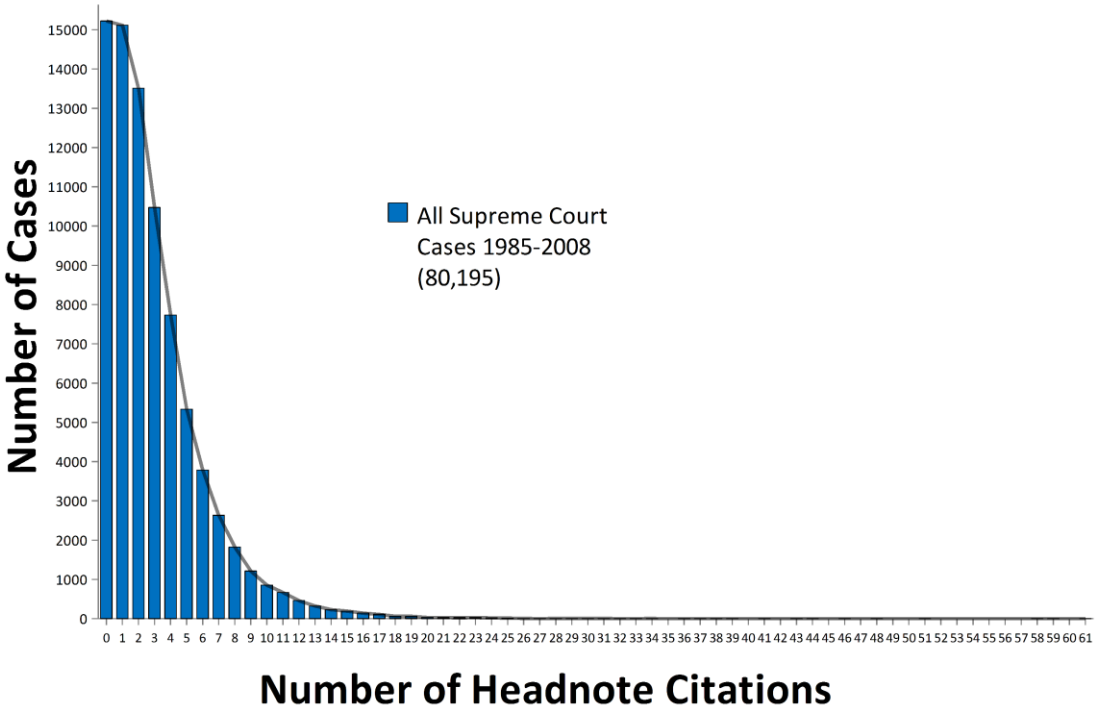


Figure 4: The Distribution of All Supreme Court Opinions According to the Number of Headnote Citations they Receive[441]

Figure 4 shows the distribution graph that I computed at the end of my first experiment. The Supreme Court of Justice opinions are aligned along the x-axis according to the number of headnote citations that each of them receives. Comparing this distribution graph to the one we saw in Figure 3[442], simple inspection reveals that the distribution bears a marked resemblance to the typical power law distribution that we find in scale-free network structures. We notice that the vast majority of opinions are either not cited at all or only cited by a few headnotes. On the other hand, there are a few opinions that are cited by a high number of headnotes. I would like to point out some empirical properties of this Supreme Court citation network: 15,222 opinions, or 19% of all cases, are not being referred to by headnotes at all. The majority of opinions receive very few headnote citations, 68% of all cases are being referred to only 3 times or less. The average number of citations per opinion, however, nevertheless equals 3.02 due to the high headnote citation counts of a chosen few cases.

---

[441] See also below, Appendix 1: Larger Version of Figure 4, p. 102

[442] See above p. 62

My first experiment confirms that the headnote citation distribution of my case law document collection follows a power law, which makes the network structure a scale-free one. As we find a similar scale-free nature in the World Wide Web, this does already suggest the efficiency of citation-analysis techniques for computer-assisted legal research systems. It does not yet, however, prove it.

11.4. The Second Experiment: Comparing the First Distribution to a Relevant Subset

With my first experiment, I established that the distribution of Austrian Supreme Court opinions according to their headnote citation counts follows power laws. The question that I want to ask in my second experiment is "How do those headnote citation counts compare to expert relevance assessments?" This question is of vital importance because, like I just mentioned, the scale-free nature of the computed degree distribution does not yet provide actual proof that citation-analysis techniques could in fact improve relevance ranking in computer-assisted legal research systems. What it takes to provide actual proof is to show that the citation counts that I computed for the first graph could in some way help to prioritise "relevant" court cases. The decision of whether or not a case is "relevant" would have to conform to expert opinions (a so-called "Gold Standard").

In my second experiment I therefore try to show that cases with certain authority scores are (much) more likely than others to be considered as "relevant" by legal experts. In more formal words, I hypothesise that there is a correlation between the number of headnote citations that an opinion receives, and its legal relevance.

Defining a Gold Standard

We have already heard about the difficulties surrounding the information retrieval concept of relevance[443]. Following Peter Jackson and Isabelle Moulinier[444] my goal for this second experiment is simply to obtain workable expert relevance judgments (a Gold Standard) in order to evaluate my citation-based authority scores. On a practical level, this meant that I needed to obtain relevance judgements which the average user of computer-assisted legal research systems is likely to agree with. Obtaining and using such relevance judgements was complicated by the precondition that they had to be freely available, just like all the other parts of the document collection. There were no resources available for

[443] See above, 2.4 Evaluation in Information Retrieval, p. 18

[444] Jackson & Moulinier, 2007, p. 24

obtaining relevance judgements, and any potential verification of my experiments should stay as easy as possible for others.

When I talked about law reports[445], I mentioned that one of their primary purposes consists in the selection of "relevant" opinions. As domain experts decide about which cases to publish and which not, I argue that those Supreme Court opinions that are published in a major law report are considered to be "relevant" by experts. Luckily, the Austrian Supreme Court of Justice in fact publishes its own official law report, called "Entscheidungen des Österreichischen Obersten Gerichtshofes in Zivilsachen - amtlich veröffentlicht [Decisions of the Austrian Supreme Court of Justice in Civil Matters - officially published]"[446] In practice, legal professionals usually refer to the widely used report as "Sammlung Zivilrecht (SZ) [Collection Civil Law (SZ)]". Even though in a somewhat hidden way, my existing document collection also contains information about which opinions written by the Supreme Court of Justice had subsequently been published in the law report. In addition to containing the references to opinions that I used for my first experiment, headnote documents also mention which decisions have been selected for publication in the law report.

Gold Standard Distribution Graph Construction

Using techniques similar to the ones employed for the construction of the citation distribution in the first experiment, I was therefore able to identify 4,842 opinions out of all 80,195 Supreme Court cases which had been published in the "Sammlung Zivilrecht (SZ)" between 1985 and 2006. For the years 2007 and 2008, the law report has not been published yet.

[445] See above p. 24

[446] Oberster Gerichtshof [Supreme Court of Justice], 1922-

The Second Experiment's Distribution Graphs

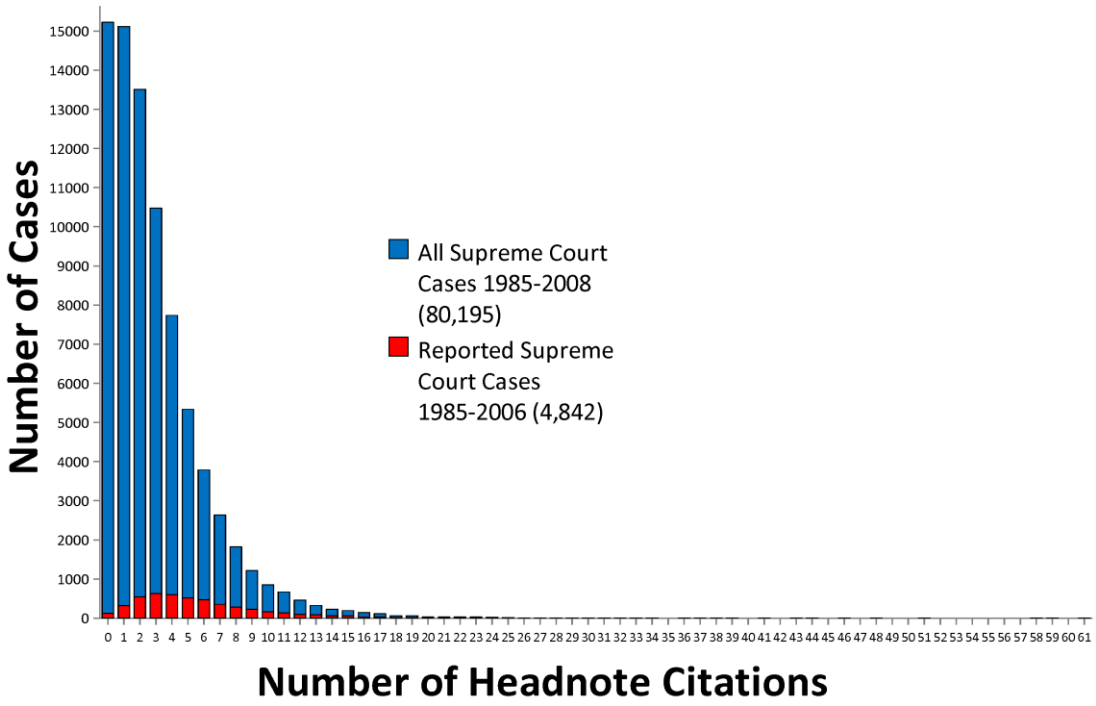


Figure 5: The Headnote-Citation Distributions of All (Blue Bars) and Only the Published (Red Bars) Supreme Court Opinions[447]

Figure 5 repeats the distribution graph which was computed in the course of the first experiment as blue bars in the background. All the 80,195 Supreme Court of Justice opinions written between 1985 and 2008 are again aligned according to the number of headnote citations which each of them receives. The red bars represent the subset of 4,842 opinions that have been officially published in the "Sammlung Zivilrecht (SZ)" law report. The red bars therefore highlight legally relevant opinions.

The objective for the second experiment is to show that cases with certain headnote citation counts are (much) more likely than others to be considered as "relevant" by legal experts. If there is such a correlation between the number of headnote citations that an opinion receives, and its probability of being relevant, my initial (blue) distribution graph could be successfully used for relevance-ranking purposes in computer-assisted legal research.

The two distribution graphs shown in Figure 5 clearly provide proof for that correlation: The correlation is found by comparing the ratios between blue and the red bars at different

[447] See also below, Appendix 2: Larger Version of Figure 5, p. 103

points along the x-axis. Simple inspection reveals those ratios are highly uneven. At 0 headnote citations, the ratio between all non-cited opinions and the published non-cited opinions is extremely high. As we move along the x-axis, it is obvious that "relevant opinions" increase in ratio to the total number of opinions with a certain headnote citation count.

This variation in the ratio between all, and only legally relevant opinions at different points along the x-axis, can perfectly provide the basis for effective relevance ranking. I will give one example: CALR system providers could automatically move non-cited opinions to the bottom of result lists. This would, on average, free the users from manually examining a big portion of their result lists, as non-cited opinions account for 19%[448] of all opinions. The drawbacks, on the other hand, would be very slim: With the bulk of non-cited (blue) opinions, only extremely few legally relevant (red) cases would be put further down in the result lists. Clearly, a relevance ranking rule like this would yield far more advantages than disadvantages.

[448] See above p. 82

12. Reaching a Conclusion

I hypothesised that a simple citation-based algorithm, despite all its shortcomings, could already significantly improve relevance ranking in contemporary computer-assisted legal research systems.

With my first experiment, I show that the headnote citation distribution of the Supreme Court opinions follows a power law. We have seen[449] that the Webgraph also follows a power law. This further suggests that algorithms that have successfully been used on the World Wide Web could also be applied to computer-assisted legal research.

By showing that the distribution graph from the first experiment could be used to prioritise groups of cases that contain significantly more "relevant" opinions than others, I then prove the effectiveness of basic citation-analysis techniques for relevance ranking in CALR systems.

[449] See above, 8.2 The Webgraph, p. 63

13. Further Research

By far the most important result of this thesis lies in its demonstration of the effectiveness of using citation-analysis techniques for computer-assisted legal research systems. In doing so, this thesis constitutes an initial feasibility study. It is "only" a feasibility study, because in my experiments, I opted for quantity rather than for quality. I only tried to prove a most general hypothesis so that all the potential pitfalls inherent in any more fine-grained legal citation analysis did not have to be dealt with. Also, I did not clean up my citation data. The "Elements of a Theory" section did, however, discuss many of those potential pitfalls for citation analyses, both general and law-specific[450]. I tried to hint at possible solutions to them, but every area (self-citations, negative citations, and so on) constitutes a whole avenue for future research by itself. I will now point out those additional research topics that in my opinion appear especially fruitful.

Taking the Recursive Nature of Citations into Account

Starting off from my experiment, a logical next step would be to try to develop a more sophisticated, recursive citation analysis approach. When we looked at[451] relevance ranking in Web search, we observed that Google's PageRank technology employs such a recursive approach. The authority score of a Web page A depends also on the authority scores of those pages that point to A. Just like we have seen throughout the "Elements of a Theory" section, this approach in fact also goes back further than Web search. As early as 1976, Gabriel Pinski and Francis Narin[452] proposed a more fine-grained citation-based measure of authority, stemming from their observation that not all citations are equally important. They argued that a journal is "influential" if, recursively, it is heavily cited by other influential journals. Common sense in fact suggests that as well, just think of other recommendation systems such as letters of reference. Not only the number of recommendations, but also the status of the recommender is important.[453] You might very well ask at this point "How can we calculate authority scores of documents without knowing the scores of all the other pages?" It turns out that even though this often seems miraculous to non-specialists,

---

[450] See above 9.4 Controversies and Problems Surrounding Legal Citation Analysis, p. 70

[451] See also above 7.2 Main Use: Ranking, p. 54

[452] Pinsky & Narin, 1976

[453] Langville & Meyer, 2006, p. 27

"iterative" methods are employed by mathematicians on a daily basis. Starting with crude approximations, calculations are repeated over and over again until all individual results level off and build a coherent whole. Other legal citation analysts have voiced similar plans for future research: James Fowler and his colleagues observe that "Ideally, we should be able to use information about the importance of citing cases to improve our estimate of the importance of the cases that they, in turn, cite."[454]

Jurisprudential Significance of Power-Law Distributions

The graph of the citation distribution of Austrian Supreme Court cases shows how relatively very few cases are the basis for the vast majority of all references found in headnotes. The vast majority of legal influence is concentrated in a relatively small number of cases. As Thomas Smith points out, this is jurisprudentially significant.[455] It shows that, even when we only look at case law from a country's High court, not all cases are created equal. A chosen few cases decide the direction of law, just like a chosen few Web pages decide what is important on the Web. In the context of this thesis, that observation also has a strong significance for comparative law.

Significance for Comparative Law

The studies that applied network analysis to US case law document collections[456] showed that power law distributions are ubiquitous phenomena in American jurisdictions. As long as those distributions are only observed in common jurisdictions, we might be tempted to think that the common law doctrine of precedent is responsible for the scale-free nature of case law citation networks in common jurisdictions: This special feature of the common law system, we might argue, accounts for the fact that precedential authority, measured by citation frequency, is highly concentrated in a relatively very small core of cases. My experiments, however, confirm that we find the same power-law distribution in Austrian case law, which means in a continental jurisdiction. This result certainly poses a challenge to the view that the network structure that case law adopts is considerably different between common and continental jurisdictions. Thinking along the lines of Wagner-Döbler[457], we might draw the conclusion that this shows that the informal doctrines of precedent in continental jurisdictions lead to the exact same practical results as the explicit doctrines of

---

[454] Fowler et al., 2007, p. 329

[455] Smith, 2007, p. 325

[456] See above 8.3 Network Analyses on Legal Document Collections, p. 64

[457] See above, p. 21

stare decisis in common jurisdictions. Much further research is needed, however, to be able to say for sure. A much more radical explanation is perfectly conceivable: It might simply be insignificant for case law citation patterns what kind of doctrine of precedent a jurisdiction follows. The World Wide Web adopted a scale-free network structure autonomously, so maybe citation distributions in case law just automatically follow a power law distribution as well.

Concluding Remarks

This leads me back to some concluding remarks about this research project. Despite its limitations, my feasibility study proves the effectiveness of citation-analysis techniques for relevance ranking in computer-assisted legal research. When discussing my research with CALR system providers especially in Austria, they continually raised concerns about using citation analysis techniques in their non-Web IR environments. This research indicates that those concerns are unfounded. We can successfully transfer citation-analysis concepts developed in other areas to CALR systems. The analogy is a sound one. Paraphrasing Daniel Dabney's previously mentioned[458] 1986 quote, I do believe that the ball is now in the CALR providers' court. They should start providing strong evidence why citation analysis concepts cannot be used in computer-assisted legal research.

[458] See above, p. 36

Reference List

Adam, D. (2002). Citation analysis: The counting house. *Nature, 415*(6873), 726–729, from http://dx.doi.org/10.1038/415726a.

Amaral, L. A. N., & Ottino, J. M. (2004). Complex systems and networks: challenges and opportunities for chemical and biological engineers: Complex Systems and Multi-scale Methodology. *Chemical Engineering Science, 59*(8-9), 1653–1666, from http://dx.doi.org/10.1016/j.ces.2004.01.043.

Arewa, O. B. (2006). Open Access in a Closed Universe: Lexis, Westlaw, Law Schools and the Legal Information Market. *Lewis & Clark Law Review, 10*(4), 797–839. Retrieved April 28, 2009, from http://www.lclark.edu/org/lclr/issue_10_4.html.

Armstrong, J. D. S., & Knott, C. A. (2006). *Where the law is: An introduction to advanced legal research* (2nd ed.). *American casebook series.* St. Paul MN: Thomson/West.

Arnold, S., & Rosen, L. (1993). Bye bye, Boolean: natural language and electronic information retrieval. *Searcher, 1*(5).

Ayres, I., & Vars, F. E. (2000). Determinants of Citations to Articles in Elite Law Reviews. *The Journal of Legal Studies, 29*(1), 427–450, from http://dx.doi.org/10.1086/468081.

Bade, D. (2007). Relevance ranking is not relevance ranking or, when the user is not the user, the search results are not search results. *Online Information Review, 31*, 831–844, from http://dx.doi.org/10.1108/14684520710841793.

Baeza-Yates, R., & Castillo, C. (2006). Web Searching. In Keith Brown (Ed.), *Encyclopedia of Language & Linguistics* (pp. 527–538). Oxford: Elsevier.

Baker, D. (1998). The enemy is in their cites. *ABA Journal, 84*, 24.

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science, 286*(5439), 509–512, from http://dx.doi.org/10.1126/science.286.5439.509.

Barabási, A.-L., & Bonabeau, E. (2003). Scale-free networks. *Scientific American, 288*(5), 60–69.

Barmakian, D. (2000). Better Search Engines for Law. *Law Library Journal, 92*(4), 399–438.

Bast, C. M., & Pyle, R. C. (2001). Legal research in the computer age: A paradigm shift? *Law Library Journal, 93*(2), 285–302.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*(5), 407–424.

Bernstam, E. V., Herskovic, JR, Aphinyanaphongs, Y., Aliferis, C. F., Sriram, M. G., & Hersh, W. R. (2006). Using citation data to improve retrieval from MEDLINE. *Journal of the American Medical Informatics Association, 13*(1), 96–105, from http://dx.doi.org/10.1197/jamia.M1909.

Berring, R. C. (1986). Full-Text Databases and Legal Research: Backing into the Future.

High Technology Law Journal, *1*, 27–60.

Berring, R. C. (1997). Chaos, Cyberspace and Tradition: Legal Information Transmogrified. *Berkeley Technology Law Journal*, *12*(1), 189–212.

Bharat, K., & Broder, A. (1998). A technique for measuring the relative size and overlap of public Web search engines: Proceedings of the Seventh International World Wide Web Conference. *Computer Networks (and ISDN Systems)*, *30*(1-7), 379–388, from http://dx.doi.org/10.1016/S0169-7552(98)00127-5.

Bing, J. (1984a). Legal information services: some trends and characteristics. In C. Campbell (Ed.), *Data processing and the law* (pp. 29–45). London: Sweet & Maxwell.

Bing, J. (1984b). *Handbook of legal information retrieval.* Amsterdam: North-Holland.

Bing, J. (1986). Legal Text Retrieval Systems the Unsatisfactory State of the Art. *Journal of Law and Information Science*, *2*(1), 1–17.

Bing, J. (1987). Performance of Legal Text Retrieval Systems: The Curse of Boole. *Law Library Journal*, *79*(2), 187–202.

Bing, J. (1988). Computerized Legal Information Services: An Introduction. *Nordic Journal of International law*, *57*(4), 393–404.

Black, H. C., & Garner, B. A. (2007). *Black's law dictionary* (8th ed.). St. Paul, Minn.: Thomson-West.

Blair, D. C. (1996). STAIRS redux: Thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science*, *47*(1), 4–22, from http://dx.doi.org/10.1002/(SICI)1097-4571(199601)47:1<4::AID-ASI2>3.0.CO;2-3.

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, *28*(3), 289–299, from http://dx.doi.org/10.1145/3166.3197.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*, 45–80, from http://dx.doi.org/10.1108/00220410810844150.

Branting, L. K. (2003). A reduction-graph model of precedent in legal analysis: AI and Law. *Artificial Intelligence*, *150*(1-2), 59–95, from http://dx.doi.org/10.1016/S0004-3702(03)00102-4.

Brian D. Davison (2000). Topical locality in the Web. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 272–279). Athens, Greece: ACM.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks (and ISDN Systems)*, *30*(1-7), 107–117, from http://dx.doi.org/10.1016/S0169-7552(98)00110-X.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000/6). Graph structure in the Web. *Computer Networks (and ISDN Systems)*, *33*(1-6), 309–320, from http://dx.doi.org/10.1016/S1389-1286(00)00083-9.

Burson, S. F. (1987). A Reconstruction of Thamus - Comments on the Evaluation of Legal Information Retrieval Systems. *Law Library Journal*, *79*(1), 133–143.

Bush, V. (1945). As We May Think. *The Atlantic*, (176), 101–108.

Chakrabarti, D., & Faloutsos, C. (2006). Graph mining: Laws, generators, and algorithms. *ACM Computing Surveys*, *38*(1), 2, from http://doi.acm.org/http://doi.acm.org/10.1145/1132952.1132954.

Chester, S. (1992). The Natural Language Way. *ABA Journal*, *78*, 111.

Chiorazzi, M. G. (2002). Books, Bytes, Bricks and Bodies: Thinking About Collection Use in Academic Law Libraries. In M. G. Chiorazzi & G. Russell (Eds.), *Law library collection development in the digital age* (pp. 1–28). Binghamton N.Y.: Haworth Information Press.

Chun, W. J. (2008). *Core Python programming* (2. ed., reprint. with corr.). Upper Saddle River, NJ: Prentice Hall.

Clinch, P. (1990). The Use of Authority - Citation Patterns in the English Courts. *Journal of Documentation*, *46*(4), 287–317.

Cohen, M. L., & Olson, K. C. (2007). *Legal research in a nutshell* (9th ed.). *West nutshell series.* St. Paul, MN: Thomson/West.

Conrad, J. G., & Dabney, D. P. (2001). Automatic recognition of distinguishing negative indirect history language in judicial opinions. In *Proceedings of the tenth international conference on Information and knowledge management* (pp. 287–294). Atlanta, Georgia, USA: ACM.

Cook, R. (2002). Another one bites the dust? *Legal Information Management*, *2*(03), 10–11.

Cronin, B. (2005). *The hand of science: Academic writing and its rewards.* Lanham, Md.: Scarecrow Press.

Dabney, D. (2000). Another response to Taylor's comparison of KeyCite and Shepard's. *Law Library Journal*, *92*(3), 381–385.

Dabney, D. (1993). *Statistical Modeling of Relevance Judgments for Probabilistic Retrieval of American Case Law*. PhD thesis, University of California, Berkeley.

Dabney, D. P. (1986a). The Curse of Thamus: An Analysis of Full-Text Legal Document Retrieval. *Law Library Journal*, *78*(1), 5–40.

Dabney, D. P. (1986b). A Reply to West Publishing Company and Mead Data Central on The Curse of Thamus. *Law Library Journal*, *78*(2), 349–350.

Dethman, J. (2002). Trust v. Antitrust: Consolidation in the Legal Publishing Industry. In M. G. Chiorazzi & G. Russell (Eds.), *Law library collection development in the digital age* (pp. 123–151). Binghamton N.Y.: Haworth Information Press.

Dibadj, R. (2008). Networks of Fairness Review in Corporate Law. *San Diego Law Review*, *45*(1), 1–32.

Donato, D., Laura, L., Leonardi, S., & Millozzi, S. (2004). Large scale properties of the Webgraph. *The European Physical Journal B - Condensed Matter and Complex Systems*, *38*(2), 239–243, from http://dx.doi.org/10.1140/epjb/e2004-00056-6.

Edwards, J. (2005). Review of LexisNexis Butterworths new online platform. *Legal Information Management*, *5*(03), 202–204, from http://dx.doi.org/10.1017/S1472669605880880.

Elliott, M., & Kling, R. (1997). Organizational Usability of Digital Libraries: Case Study of Legal Research in Civil and Criminal Courts. *Journal of the American Society for Information Science, 48*(11), p1023 - 1035, from http://dx.doi.org/10.1002/(SICI)1097-4571(199711)48:11<1023::AID-ASI5>3.0.CO;2-Y.

Ellis, D. (1989). A behavioral approach to information retrieval system design. *Journal of Documentation, 45*(3), 171–212, from http://dx.doi.org/10.1108/eb026843.

Evans, R. (1994). Beyond Boolean: Relevance Ranking, Natural Language and the New Search Paradigm. In M. E. Williams & M. E. Williams (Eds.), *15th National Online Meeting, Proceedings 1994 /// Proceedings 1994* (pp. 121–128). Medford, NJ /// Medford, N.J.: INFORMATION TODAY INC.

Finch, E., & Fafinski, S. (2007). *Legal skills.* Oxford: Oxford University Press.

Fisher, D. (2008). Open-Sourcing The Law. *Forbes, 181*(13), 70–73.

Fon, V., & Parisi, F. (2006). Judicial precedents in civil law systems: A dynamic analysis. *International Review of Law and Economics, 26*(4), 519–535.

Foster, L., & Kennedy, B. (2000). Technological Developments in Legal Research. *The Journal of Appellate Practice and Process, 2*(2), 275–303.

Fowler, J. H., & Jeon, S. (2008). The authority of Supreme Court precedent. *Social Networks, 30*(1), 16–30.

Fowler, J. H., Johnson, T. R., Spriggs, I. J. F., Jeon, S., & Wahlbeck, P. J. (2007). Network analysis and the law: Measuring the legal importance of precedents at the US Supreme Court. *Political Analysis, 15*(3), 324–346, from http://dx.doi.org/10.1093/pan/mpm011.

Gallacher, I. (2006). Forty-Two: The Hitchhiker's Guide to Teaching Legal Research to the Google Generation. *Akron Law Review, 39*(1), 151–205.

Garfield, E. (1955). Citation Indexes for Science - A New Dimension in Documentation through Association of Ideas. *Science, 122*(3159), 108–111, from http://dx.doi.org/10.1126/science.122.3159.108.

Garfield, E. (1977). Can Citation Indexing Be Automated?: *Vol. Vol. 1 (1962-1973), Essays of an information scientist*, pp. 84–90. Philadelphia Pa.: ISI Press.

Garfield, E. (1979). *Citation indexing, its theory and application in science, technology and humanities. Information sciences series.* New York: Wiley.

Garfield, E. (1987). *Oral history: Transcript of an interview conducted by A. Thackray and J. Sturchio at the Institute for Scientific Information, Philadelphia, PA, on 16 November 1987.* Retrieved April 28, 2009, from Beckman Center for the History of Chemistry: www.garfield.library.upenn.edu/oralhistory/interview.html.

Gelbart, D., & Smith, J. C. (1993). Towards Combining Automated Text Retrieval and Case-Based Expert Legal Advice. *Law Technology Journal, 1*(2), 19–24.

Glenn, H. P. (1987). The Use of Computers: Quantitative Case Law Analysis in the Civil and Common Law. *The International and Comparative Law Quarterly, 36*(2), 362–368.

Google (2008). *We knew the web was big…* Retrieved April 28, 2009, from http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.

Google (2009). *About Google - Corporate Info - Technology Overview.* Retrieved April 28,

2009, from http://www.google.com/intl/en/corporate/tech.html.

Greenleaf, G. (2004). Jon Bing and the History of Computerised Legal Research – Some Missing Links. In O. Torvund, L. A. Bygrave, & J. Bing (Eds.), *Et tilbakeblikk på fremtiden: artikler samlet i anledning Jon Bings 60-årsdag 30.april 2004* (pp. 61–75). Oslo: Institutt for rettsinformatikk.

Halvorson, T. R., & Basch, R. (2000). *Law of the super searchers: The online secrets of top legal researchers.* Medford NJ: CyberAge Books.

Hansford, T. G., & Spriggs, J. F. (2008). *The politics of precedent on the U.S. Supreme Court* (2nd print.). Princeton: Princeton University Pr.

Hanson, F. A. (2002). From key numbers to keywords: How automation has transformed the law. *Law Library Journal, 94*(4), 563–600.

Harrington, W. G. (1984). A Brief History of Computer-Assisted Legal Research. *Law Library Journal, 77*(3), 543–556.

Henzinger, M. (2005). Hyperlink analysis on the world wide web. In *HYPERTEXT '05: Proceedings of the sixteenth ACM conference on Hypertext and hypermedia* (pp. 1–3). New York, NY, USA: ACM.

Herskovic, J. R., Iyengar, M. S., & Bernstam, E. V. (2007). Using hit curves to compare search algorithm performance. *Journal of Biomedical Informatics, 40*(2), 93–99, from http://dx.doi.org/10.1016/j.jbi.2005.12.007.

Hibbitts, B. J. (1996). Last Writes? Reassessing the Law Review in the Age of Cyberspace. *New York University Law Review, 71*, 615–688.

Horty, J. F. (1959). Research Report: University of Pittsburgh Health Law Center. *Modern Uses of Logic in Law (M.U.L.L.), 1*(1), 31–32.

Jackson, P., & Moulinier, I. (2007). *Natural language processing for online applications: Text retrieval, extraction and categorization* (2. rev. ed.). *Natural language processing: Vol. 5.* Amsterdam: Benjamins.

Jacso, P. (2005). Relevance in the eye of the search software. *Online Information Review, 29*(6), 676–682.

Katy Börner, S. S. A. V. (2007). Network science. *Annual Review of Information Science and Technology, 41*(1), 537–607, from http://dx.doi.org/10.1002/aris.2007.1440410119.

Kelso, L. O. (1946). Does the Law Need a Technological Revolution? *Rocky Mountain Law Review, 18*(4), 378–392.

Kleinberg, J. (2006). Social networks, incentives, and search. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 210–211). Seattle, Washington, USA: ACM.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46*(5), 604–632, from http://dx.doi.org/10.1145/324133.324140.

Krause, J. (2004). Towering Titans. *ABA Journal, 90*(5), 50–54.

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the Web for emerging cyber-communities. *Computer Networks (and ISDN Systems), 31*(11-16), 1481–1493, from http://dx.doi.org/10.1016/S1389-1286(99)00040-7.

Landes, W. M., Lessig, L., & Solimine, M. E. (1998). Judicial Influence: A Citation Analysis of Federal Courts of Appeals Judges. *Journal of Legal Studies, 27*(2), 271–332, from http://dx.doi.org/10.1086/468022.

Landes, W. M., & Posner, R. A. (2000). Citations, Age, Fame, and the Web. *Journal of Legal Studies, 29*(1), 319, from http://dx.doi.org/10.1086/468075.

Langville, A. N., & Meyer, C. D. (2006). *Google's PageRank and beyond: The science of search engine rankings.* Princeton, NJ: Princeton Univ. Press.

Leith, P. (2007). BAILII - Towards a National Law Library? *Legal Information Management, 7*(01), 42–45, from http://dx.doi.org/10.1017/S1472669606001113.

Leonard, J. (1989-1990). Seein' the Cites: A Guided Tour of Citation Patterns in Recent American Law Review Articles. *Saint Louis University Law Journal, 34*, 181–239.

Liddy, E. D. (2006). Document Retrieval, Automatic. In Keith Brown (Ed.), *Encyclopedia of Language & Linguistics* (pp. 748–755). Oxford: Elsevier.

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development, 1*(4), 309–317.

Ma, N., Guan, J., & Zhao, Y. (2008). Bringing PageRank to the citation analysis: Evaluating Exploratory Search Systems; Digital Libraries in the Context of Users' Broader Activities. *Information Processing & Management, 44*(2), 800–810, from http://dx.doi.org/10.1016/j.ipm.2007.06.006.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval.* Cambridge: Cambridge Univ. Press.

Marx, S. M. (1969-1970). Citation Networks in the Law. *Jurimetrics Journal, 10*(4), 121–137.

Mason, D. (2006). Legal Information Retrieval Study - Lexis Professional and Westlaw UK. *Legal Information Management, 6*(04), 246–250, from http://dx.doi.org/10.1017/S1472669606000831.

Matthijssen, L. (1998). A Task-Based Interface to Legal Databases. *Artificial Intelligence and Law, 6*(1), 81–103.

Maxwell, K. T., & Schafer, B. (2008). Concept and Context in Legal Information Retrieval. In E. Francesconi, G. Sartor, & D. Tiscornia (Eds.), *Frontiers in artificial intelligence and applications: Vol. 189. Legal Knowledge and Information Systems - JURIX 2008: The Twenty-First Annual Conference on Legal Knowledge and Information Systems, Florence, Italy, 10-13 December 2008* (pp. 63–72). IOS Press.

McDermott, J. (1986). Another Analysis of Full-Text Legal Document Retrieval. *Law Library Journal, 78*(2), 337–344.

Meadow, C. T., Boyce, B. R., & Kraft, D. H. (2007). *Text information retrieval systems* (3. ed. /). *Library and information science.* Amsterdam: Elsevier.

Merton, R. K. (1968). *Social theory and social structure* (Enlarged Edition). New York: Free Press.

Mertz, D. (2003). *Text processing in Python.* Boston: Addison-Wesley.

Moens, M.-F. (2001). Innovative techniques for legal text retrieval. *Artificial Intelligence*

and Law, 9(1), 29–57, from http://dx.doi.org/10.1023/A:1011297104922.

Moens, M.-F. (2005). Retrieval of Legal Documents: Combining Structured and Unstructured Information. In M. Dobreva (Ed.), *From author to reader. Challenges fro the digital content chain ; proceedings of the 9th ICCC International Conference on Electronic Publishing, Leuven, Arenberg Castle, June 8-10,2005* (pp. 223–228). Leuven: Peeters.

Moens, M.-F. (2007). Summarizing court decisions: Text Summarization. *Information Processing & Management, 43*(6), 1748–1764.

Monk, C. (2008). Westlaw UK ? a Review of the New Platform. *Legal Information Management, 8*(02), 147–150, from http://dx.doi.org/10.1017/S1472669608000376.

Moravcsik, M. J. (1973). Measures of scientific growth. *Research Policy, 2*(3), 266–275.

Morris, J. W. (2000). A response to Taylor's comparison of Shepard's and KeyCite. *Law Library Journal, 92*(2), 143–145.

Müller, H. (1991). Legal Information Systems and Other Law-Related Databases in Germany, Austria and Switzerland. *Law Library Journal, 83*(2), 253–267.

Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology, 41*(1), 609–641, from http://dx.doi.org/10.1002/aris.2007.1440410120.

Niegl, R. (1996). Legal Databases in Austria. *International Journal of Legal Information, 24*(1), 92–96.

Norman, P. (2004). The Big Match - Lexis v Westlaw. *Legal Information Management, 4*(02), 90–97, from http://dx.doi.org/10.1017/S1472669604001392.

Nunn-Price, N. (1992). Computers and Law 1960 - 1990. *Law Technology Journal, 1*(2), 11–15.

Oberster Gerichtshof [Supreme Court of Justice] (1922-). *Entscheidungen des Österreichischen Obersten Gerichtshofes in Zivilsachen [Decisions of the Austrian Supreme Court of Justice in Civil Matters]: amtlich veröffentlicht [officially published]*. Wien: Verlag Österreich.

Ogden, P. (1993). "Mastering the Lawless Science of Our Law": A Story of Legal Citation Indexes. *Law Library Journal, 85*(1), 1–48.

Österreichischer Verfassungsdienst (1994). *Rechtsinformationssystem des Bundes: RIS ; eine kurze Einführung.* Wien: Bund, Bundeskanzleramt, Verfassungsdienst.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science, 28*(6), 441–453, from http://dx.doi.org/10.1177/016555150202800601.

Paul, G. L., & Baron, J. R. (2007). Information Inflation: Can the Legal System Adapt? *Richmond Journal of Law & Technology, 13*(10), 1–41, from http://law.richmond.edu/jolt/v13i3/article10.pdf.

Perry, R. (2006). The Relative Value of American Law Reviews: A Critical Appraisal of Ranking Methods. *Virginia Journal of Law and Technology, 11*(1).

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications - Theory, with application to the literature of physics. *Information*

Processing & Management, 12(5), 297–312.

Posner, R. A. (2000). An economic analysis of the use of citations in the law. *Am Law Econ Rev, 2*(2), 381–406.

Post, D. G., & Eisen, M. B. (2000). How long is the coastline of the law? Thoughts on the fractal nature of legal systems. *Journal of Legal Studies, 29*(1, Part 2), 545–584.

Pritchard-Schoch, T. (1993). Natural Language Comes Of Age. *Online, 17*(3), 33–43.

Ragona, M. (2002). Managing legal information - from databases to the Internet. *European Legal Forum*, (1), 10–19.

Rijsbergen, C. J. v. (1981). *Information retrieval* (2. ed., repr.). London: Butterworth.

Robert M. Losee, L. A. H. P. (1999). Measuring search-engine quality and query difficulty: Ranking with target and freestyle. *Journal of the American Society for Information Science, 50*(10), 882–889, from http://dx.doi.org/10.1002/(SICI)1097-4571(1999)50:10<882::AID-ASI5>3.0.CO;2-6.

Runde, C. E., & Lindberg, W. H. (1986). The Curse of Thamus: A Response. *Law Library Journal, 78*(2), 345–347.

Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM, 29*(7), 648–656, from http://dx.doi.org/10.1145/6138.6149.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523, from http://dx.doi.org/10.1016/0306-4573(88)90021-0.

Samborn, H. V. (1999). Rev up your research engines. *ABA Journal, 85*, 74–79.

Saracevic, T. (1999). Information science. *Journal of the American Society for Information Science, 50*(12), 1051–1063, from http://dx.doi.org/10.1002/(SICI)1097-4571(1999)50:12<1051::AID-ASI2>3.0.CO;2-Z.

Sayer, J. (2008). Review Article on Statute Law Database. *Legal Information Management, 8*(04), 299–301, from http://dx.doi.org/10.1017/S1472669608000741.

Schweighofer, E. (1999). The Revolution in Legal Information Retrieval or: The Empire Strikes Back. *Journal of Information, Law and Technology, 1999*(1). Retrieved April 28, 2009, from http://www2.warwick.ac.uk/fac/soc/law/elj/jilt/1999_1/schweighofer/.

Shapiro, F. R. (Oct., 1985). The Most-Cited Law Review Articles. *California Law Review, 73*(5), 1540–1554.

Shapiro, F. R. (Mar., 1991). The Most-Cited Articles from The Yale Law Journal. *The Yale Law Journal, 100*(5), 1449–1514.

Shapiro, F. R. (1992). Origins of Bibliometrics, Citation Indexing, and Citation Analysis: The Neglected Legal Literature. *Journal of the American Society for Information Science, 43*(5), 337–339, from http://dx.doi.org/10.1002/(SICI)1097-4571(199206)43:5<337::AID-ASI2>3.0.CO;2-T.

Shapiro, F. R. (1996). The Most-Cited Law Review Articles Revisited. *Chicago Kent Law Review, 71*(3), 751–780.

Shapiro, F. R. (Jan., 2000). The Most-Cited Legal Scholars. *The Journal of Legal Studies,*

29(1), 409–426.

Shapiro, F. R. (2001). *Collected papers on legal citation analysis.* Littleton Colo.: F.B. Rothman.

Shaw, T. (2007). Free v Fee: Drivers and Barriers to the Use of Free and Paid-For Legal Information Resources. *Legal Information Management, 7*(1), 23–29, from http://dx.doi.org/10.1017/S1472669606001083.

Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. *SIGIR Forum, 33*(1), 6–12, from http://dx.doi.org/10.1145/331403.331405.

Smith, J. C., Gelbart, D., Maccrimmon, K., Atherton, B., Mcclean, J., Shinehoft, M., et al. (1995). Artificial intelligence and legal discourse: The Flexlaw legal text management system. *Artificial Intelligence and Law, 3*(1), 55–95, from http://dx.doi.org/10.1007/BF00877695.

Smith, T. A. (2007). The Web of Law. *San Diego Law Review, 44*(2), 309–354.

Solla Price, D., de (1976). A General Theory of Bibliometric and Other Cumulative Advantage Processes. *American Society for Information Science Journal, 27*(5), 292–306, from http://dx.doi.org/10.1002/asi.4630270505.

Sormunen, E. (2001). Extensions to the STAIRS Study – Empirical Evidence for the Hypothesised Ineffectiveness of Boolean Queries in Large Full-Text Databases. *Information Retrieval, 4*(3-4), 257–273, from http://dx.doi.org/10.1023/A:1011950323099.

Spink, A., Wolfram, D., Jansen, M. B. J., & Saracevic, T. (2001). Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology, 52*(3), 226–234, from http://dx.doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.0.CO;2-R.

Spriggs, J. F., II, & Hansford, T. G. (2000). Measuring Legal Change: The Reliability and Validity of Shepard's Citations. *Political Research Quarterly, 53*(2), 327–341, from http://dx.doi.org/10.1177/106591290005300206.

Staudegger, E. (2006). Rechtsdatenbanken in Österreich. *Medien und Recht*, (4), 183–185.

Supreme Court of Justice [Oberster Gerichtshof]. *Accessibility of Decisions [Zugänglichkeit von Entscheidungen]*. Retrieved April 28, 2009, from http://www.ogh.gv.at/entscheidungen/index.php?nav=47.

Susskind, R. (1998). *The future of law: Facing the challenges of information technology* (Rev. paperback ed.). Oxford: Oxford Univ. Press.

Sutton, S. A. (1994). The role of attorney mental models of law in case relevance determinations: An exploratory analysis. *Journal of the American Society for Information Science, 45*(3), 186–200, from http://dx.doi.org/10.1002/(SICI)1097-4571(199404)45:3<186::AID-ASI8>3.0.CO;2-F.

Tapper, C. (1984). An Experiment with Citation Vectors. In C. Campbell (Ed.), *Data processing and the law* (pp. 90–109). London: Sweet & Maxwell.

Tapper, C. (2005). Out of the box. *International Review of Law, Computers & Technology, 19*(1), 5–11. Retrieved July 30, 2008, from

http://www.informaworld.com/10.1080/13600860500049162.

Tapper, C. F. H. (1974). Legal Information Retrieval by Computer - Applications and Implications. *McGill Law Journal, 20*(1), 26–43.

Tapper, C. F. H. (1976). Citation Patterns in Legal Information Retrieval. *Datenverarbeitung im Recht, 5*(3), 249–275.

Tapper, C. F. H. (1979-1980). Computerised Legal Information Retrieval: its Past, Present and Future. *Poly Law Review, 5*(2), 37–40.

Taylor, W. L. (2000). Comparing KeyCite and Shepard's for completeness, currency, and accuracy. *Law Library Journal, 92*(2), 127–141.

Tenopir, C. (1993). Natural language searching with WIN. *Library Journal, 118*(18), 54–55.

The Sedona Conference Working Group on Best Practices for Document Retention and Production (WG1), S. a. R. S. S. P. T. (2007). The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery. *The Sedona Conference Journal, 8*, 189–223.

Thomas, P. A., & Knowles, J. (2006). *Effective legal research* (1. ed.). *Legal skills series.* London: Thomson Sweet & Maxwell.

Thompson, P. (2008). Looking back: On relevance, probabilistic indexing and information retrieval. *Information Processing & Management, 44*(2), 963–970, from http://dx.doi.org/10.1016/j.ipm.2007.10.002.

Turtle, H. (1994). Natural language vs. Boolean query evaluation: a comparison of retrieval performance. In W. B. Croft (Ed.), *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '94* (pp. 212–220). Dublin, Ireland: Springer-Verlag New York, Inc.; Springer.

Turtle, H. (1995). Text retrieval in the legal world. *Artificial Intelligence and Law, 3*(1), 5–54, from http://dx.doi.org/10.1007/BF00877694.

Wagner-Döbler, R. (1994). The frequency distribution of legal decision citations in the German jurisdiction. *Scientometrics, 29*(1), 15–26, from http://dx.doi.org/10.1007/BF02018381.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. J. (2005). *Text mining: Predictive methods for analyzing unstructured information.* New York: Springer.

Widdison, R. (2002). New Perspectives in Legal Information Retrieval. *International Journal of Law and Information Technology, 10*, 41-70(30), from http://dx.doi.org/10.1093/ijlit/10.1.41.

Wiggins, R. (2003). The Privilege of Ranking: Google Plays Ball. *Searcher, 11*(7), 23.

Witten, I. H., Gori, M., & Numerico, T. (2007). *Web dragons: Inside the myths of search engine technology. Morgan Kaufmann a series in multimedia and information systems.* Amsterdam: Morgan Kaufmann/Elsevier.

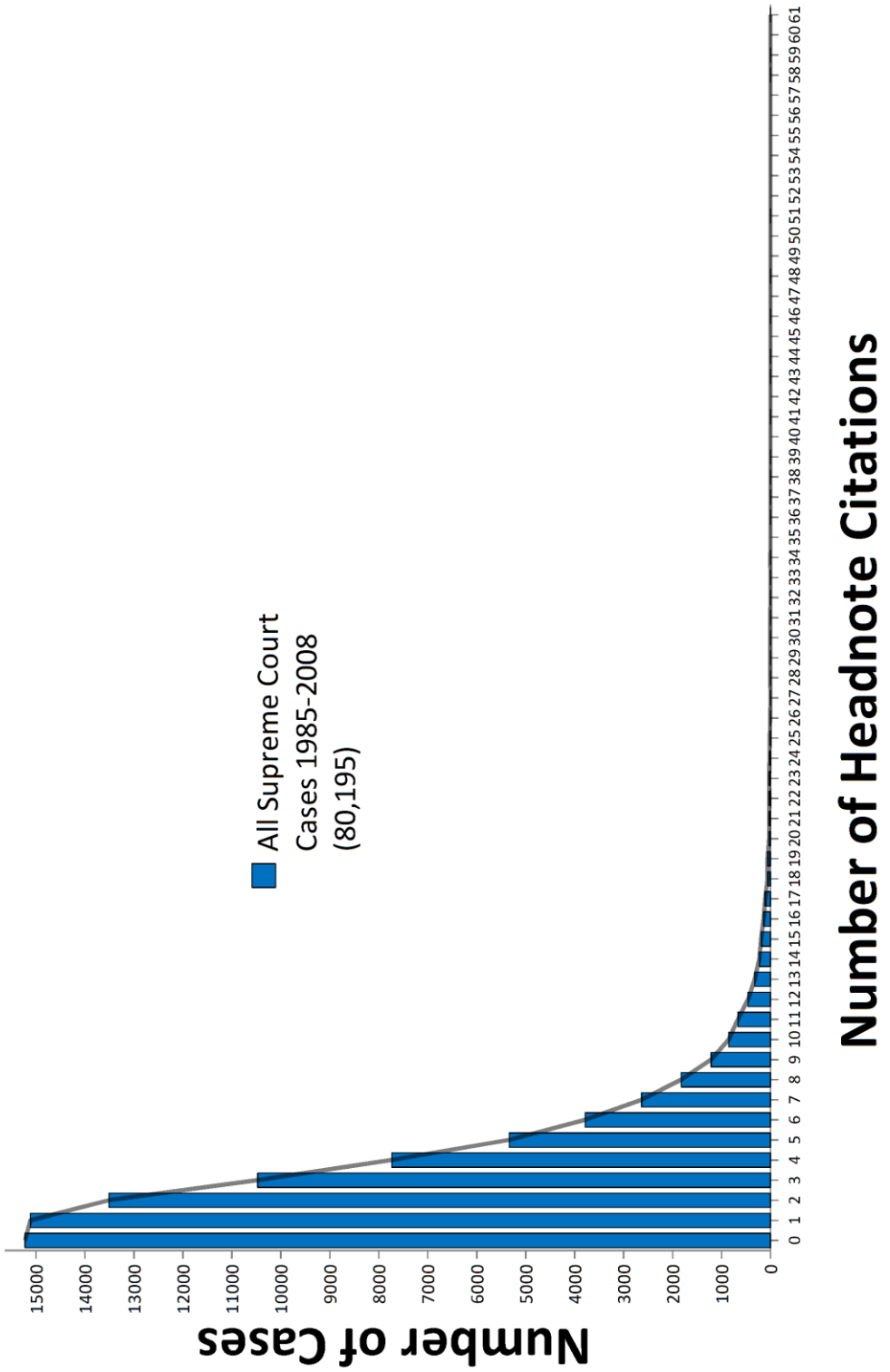
Wolfram, D. (2003). *Applied informetrics for information retrieval research.* Westport, Conn.: Libraries Unlimited.

Zhang, P., & Koppaka, L. (2007). Semantics-based legal citation network. In *Proceedings of the 11th international conference on Artificial intelligence and law* (pp. 123–130).

Stanford, California: ACM, from http://doi.acm.org/10.1145/1276318.1276342.

Zunde, P. (1971). Structural models of complex information sources. *Information Storage and Retrieval, 7*(1), 1–18, from http://dx.doi.org/10.1016/0020-0271(71)90023-4.

Appendix 1: Larger Version of Figure 4



Appendix 2: Larger Version of Figure 5