Daten verarbeitung im Recht

Band 5 Heft 3, 1976 ISSN 0301-2980

Bernt Bühnemann
Herbert Fiedler
Hermann Heussner
Adalbert Podlech
Spiros Simitis
Wilhelm Steinmüller
Sigmar Uhlig

Datenverarbeitung im Recht

Archiv für die gesamte Wissenschaft der Rechtsinformatik, der Rechtskybernetik und der Datenverarbeitung in Recht und Verwaltung.

Zitierweise: DVR

Geschäftsführender Herausgeber:

Dr. Sigmar Uhlig, An der Düne 13, D-5300 Bonn-Tannenbusch,
Telefon 0 22 21/66 13 78 (privat); 0 22 21/5 81 oder 58 48 27 (dienstlich)

Redaktioneller Mitarbeiter:

Dieter Hebebrand, Fliederweg 1, D-3501 Niestetal, Telefon 05 61/52 46 31 (privat); 05 61/30 73 62 (Bundessozialgericht)

Colin F. H. Tapper

Citation Patterns in Legal Information Retrieval

Übersicht

A. State of the Art[1]

1. The Established Method and Its Defects

In the late 1950's and early 1960's it first became possible to contemplate the use of computers to assist in the retrieval of legal information. Those years saw the formation of a special sub-committee of the American Bar Association, the launching of a number of journals specifically devoted to computer applications in law and predominantly to legal information retrieval,[2] and the establishment of a number of research programs.[3] Largely owing to the work of *John Horty* at the University of Pittsburgh the direction taken by most of these experiments, at least in the Anglo-American legal world, was towards the use of the full-text of legal documents for retrieval systems. There were a number of reasons for this. The main one was a distrust of any screen put between the lawyer confronted by the problem and the information made available to him. Indexing and abstracting are essentially methods of reducing the bulk of the information presented to the lawyer so as to make it of manageable magnitude.

Information regarded by the indexer or abstracter as less important is either discarded altogether or restated in a more general and more concise form. It is then possible for the lawyer to scan this reduced version and to select those parts which seem relevant to his problem. With the advent of the computer it seemed that things had changed. The machine could scan any amount of information in a very short space of time, and so long as its judgment of relevance was satisfactory could save the lawyer work by presenting him with all and with only relevant

1 See generally *Tapper* ‚Computers and the Law' chs. 5–7.

2 Jurimetrics Journal (formerly Modern Uses of Logic in Law), Law and Computer Technology, Rutgers Journal of Computers and Law, Datenverarbeitung im Recht.

3 Among the earliest experiments were those of Professor *Horty* at the University of Pittsburgh, *Colin Tapper* at Magdalen College, Oxford and *Aviezri Fraenkel* at the

material. It was argued that this method optimised the interaction of man and machine by restricting the machine to the mechanical job of searching for matches between words specified by the lawyer, and yet still allowed full scope for the creativity of the lawyer in selecting the words to be matched.

At first some felt that this approach was best suited to statutory materials because their volume was smaller and the use of words was relatively more precise than in case-law. Others took the view that case-law was more suitable than statute on the basis that the greater volatility of statutory materials more than compensated for their relatively smaller bulk, and that the relative poverty of statutory vocabulary could lead to failure to retrieve relevant information. Conversely it was felt that the trend towards ever cheaper storage of information reduced the force of the argument based on bulk. In fact the best argument in favour of concentrating on statutory materials within full-text systems was hardly ever deployed. This is that the sort of search which is commonly employed in the statutory area corresponds much more closely to the scanning technique of full-text than does the sort of search commonly required in the area of caselaw. Another argument is simply that there is less scope for reduction in statutory materials where every word is authoritative than in case-law where only the rule in a case can ever be authoritative. As usual however the decisive arguments were economic ones. No system offering access to statutes alone is economically viable or psychologically acceptable in an environment in which both statutes and caselaw have to be used. So working systems developed for widespread use catered for both statute and case-law as a data-base.

These systems were essentially full-text systems based on the principle of word matching. This requires the lawyer to state his problem in terms of the words and combinations of words which he would expect to find in any document relevant to his problem. All of these terms are pregnant with difficulty. What is to count as a word? When is one word different from another? What sorts of combination are allowed? What principle of individuation is to be applied to legal documents? These questions have received pragmatic solutions. In general a word is equated with a string of characters terminated either by a space or by some punctuation, though there are exceptions to this rough definition. Different strings are regarded as different words. Strings can be combined into lists and lists into final search formulations by Boolean operators and semantic distance measured in terms of document, sentence and words. In legislation the section is usually regarded as the basic unit, in case-law the case. In general also so as to reduce storage requirements in the concordances used by such systems ‚common words' are omitted. At first such systems were operated in batch mode, but increasingly they are offered on an interactive basis. This means that the user types in the words which are to characterise the answers to his problem, and is given the opportunity to review his characterisation in the light of interim results. The results are commonly expressed first as a numerical value representing the number of documents which satisfy the user's characterisation. The user then has the option of having the whole or part of the text of those documents displayed, or of

the responses obtained from the system at which stage he can get a hard copy, i. e., one printed on paper, of either the references to or text of those documents which satisfy his final characterisation.

It is plain that these methods harbour a number of defects. These may be classified into two broad groups. The first includes those which affect the level of performance in terms of the quality of the material produced, and the second those which otherwise affect the acceptability of the system to users. So far as the first is concerned it is easier to propound theoretical reasons for it than to produce empirical evidence, since there has been no published report of any systematic test of the more advanced systems now being offered commercially. Such empirical evidence as there is relates only to cruder and earlier experimental systems.[4] That evidence is somewhat equivocal in suggesting that while machine performance does tend to retrieve relevant information which cannot be recovered in any other way, it does so only at the expense of recovering a vast amount of irrelevant information also. The explanation lies in a combination of several factors. First it occurs because the nature of the procedure relies upon occurrence and co-occurrence of character strings as a unique indicator of meaning. This process has a number of drawbacks. In the first place, very similar meanings can be encapsulated in very different character strings, so all must be specified. Secondly, the same character string can encapsulate very different meanings in different contexts. The former raises problems of synonyms and, much more potently, of levels of abstraction. The latter that of homologues. Thus ‚auto', ‚automobile', and ‚car' although all different character strings have meanings which are substantially similar; so, too, ‚Chevrolet', ‚car' and ‚vehicle' can easily have the same meaning in the context of some legal problems though they may not do so in all; and then ‚jury' in the context of trial by jury has a different meaning from ‚jury' in the sense of a temporary maritime repair although the character strings are identical. This means that in order to characterise his meaning uniquely and accurately the user must specify all possible synonyms, particularisations and generalisations (including all their different grammatical forms), and exclude all identical character strings having different meanings. The latter task can only be accomplished by way of the context in which the character strings appear. So some lists of strings must be combined and some use made of combinatorial logic. It follows that the user must not only be able to specify in advance all the different ways of expressing the meaning he wishes to include, but also all the different ways of expressing the meaning of a least one other meaning which he wishes to find associated with the original meaning and the way in which the two are to be linked. It will be a great help to him in doing this to be able to think of the strings most likely to be associated with the other unwanted meanings of the string he wishes to use so as to be sure that they are excluded from the combined list. Thus if a lawyer wishes to find cases dealing with temporary maritime repairs, he must not only ask for occurrences of the string ‚jury' which if specified alone would deluge him with unwanted references to jury trial, but must also specify some association with, for

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.