10th BILETA Conference Electronic: Communications

March 30th & March 31st, 1995 Business School, University of Strathclyde, Glasgow

The Use of Information Systems in Research for the Acquisition of Knowledge

C.J.M. Combrink-Kuiters and P.A.W. Piepers, Erasmus University Rotterdam

1. Introduction

The existence of a huge number of legal documents has been seen as a problem since modern, written law began(1).The amount of legal literature which is produced is increasing every day. Statutes, judicial decisions, regulations and other literary products such as text-books, academic articles and papers are being published and made available to lawyers, both practitioners and legal researchers. The problem which arises, is how to find the desired information in this overflow of legal data. The central problem in information retrieval is to find the required information at the right time, in a recognizable form and without it being accompanied by too much irrelevant extraneous information. To accomplish this task we need the right tools. Without the use of a computer it is almost impossible to select the right legal documents from the total supply of data. Even if a computer is used, selection will only be effective, however, if information is stored in a computer searchable form. A rapidly growing amount of legal data is stored in an electronic format, which makes it convenient to find the required information and unnecessary to search paper archives.

In a jurimetrical research project which is being carried out by members of the Centre for Computers and Law at the Erasmus University Rotterdam, an attempt is being made to obtain knowledge on judicial decision-making. This project is aimed at determining which factors are the most important and to what extent and in which direction these factors influence the outcome of court cases. This is done using the texts of previous verdicts as a source of information. With the help of this information, an assessment is made of which factors are really of importance to a judge in reaching a particular decision. This knowledge can then be used to predict court decisions.

The first step in this research project was to collect court cases. To achieve this purpose, it was decided to use an electronic databank as a source, because this was the most convenient way to obtain quick results. A selection of published Dutch case law is available on CD ROM in an electronic databank published by Kluwer Datalex.

Not only are there many advantages in using this kind of information system in research, there are also some disadvantages. To find out whether the shortcomings influence the results achieved in our research project, it was decided to carry out the same kind of research using more traditional sources of information, like the texts of unpublished verdicts of the lower courts, texts of court files and other kinds of written material. The results obtained were then compared. Both procedures are described below. After that we will present a comparison of the results using both kinds of sources and in the end we will draw some conclusions from the outcome of this comparison.

2. Using CD ROM in research

2.1 Current position regarding electronic information

At this moment several legal sources of information are available on CD ROM, for example legislation, case law (from 1965 onwards) and an electronic library containing articles and summaries of legal academic articles. The Kluwer Datalex case law databank consists of abstracts of verdicts which have already been published in legal magazines(2).

Those verdicts which have been published are usually of cases which have been referred to the Supreme Court. Verdicts of the lower courts are rarely published and, therefore, not yet obtainable in a computerized format. The editorial board of these legal magazines selects which cases will be published. The most important criterium is whether the verdicts are assumed to be of interest to a larger public. This may be because the factors of the case are remarkable, or because the legal construction of a particular case is markedly divergent or it is a sign that the administration of justice has been altered.

Only the most relevant parts - according to the editor - of the texts of these verdicts are published. In the near future, full texts of all the civil law verdicts of the Supreme Court will be available in a databank called CIV-DOC. Eventually the verdicts of lower courts will also be available in an electronic format, which, in our opinion, will have an enormous

impact on both legal research and legal practice.

Some people, however, state that there are also disadvantages to publishing all verdicts. Judges fear that they will be kept under constant surveillance. They are aware of the fact that their decisions are not always consistent. Others fear that this publication will make law less dynamic and new developments in law will be inhibited(3). The publication of all verdicts will nonetheless be important in contributing to assure equality and unity of justice and it will increase the predictability of a verdict; it will reveal and prevent judicial arbitrariness.

2.2 Some advantages and disadvantages

CD ROM has become an important format for the storage of legal data and for the availability of information. The number of legal documents - such as textbooks, academic articles, case law and legislation - provided in this way is growing quickly. This is no surprise when one bears in mind the enormous amount of data that can be stored with the help of this medium. One CD ROM can contain 600MB, which corresponds to 250,000 pages. Furthermore, not only does it have a considerable capacity, but it is also a reliable source of information because the (read-only) information cannot be erased or changed by any user.

Another advantage is that there are no restrictions with regard to the time of day information is available. When a user possesses his own CD ROM or when he has access to a CD ROM databank by means of a permanent network connection, he can obtain information 24 hours a day, 7 days a week.

Searching for information while sitting at a desk using a CD ROM instead of looking for it in a library saves much time and eventually, after the initial financial investment (to adapt hardware and to obtain software) much money.

Another advantage of the use of an electronic databank is that documents can be printed or that a private collection can be created by saving information on (hard) disk. A private data base of this nature is a very useful tool for research because the texts can be analyzed from different perspectives.

2.3 Searching for the right information using a legal CD ROM databank

The CD ROM databank on case law which has been used in our research project was provided with a user friendly full-text searching routine(4). By entering a search term - this can be anything: a word, a name, a number - all documents containing this particular word can be retrieved by means of Boolean processing. If the result of the first searching procedure is not yet satisfactory, a new search routine can be performed within the set which was already retrieved. The use of logic Boolean connectors makes it possible to search for documents containing two or more words at the same time, for example, both 'custody' and 'parent' or both 'access' and 'child'. It is also possible to search for documents containing either one word or the other, for example: 'contact' or 'access'. Sometimes the thesaurus which is added to the system can be of help in solving this problem.

In addition to traditional searching methods, it is also possible to use the special utilities available such as searching for a case number, date and a specific court of justice. With the help of a truncation, parts of a word can also be entered as a search term, for example: custo*. The addition of the truncation means that all documents in which the term appears as a part of a compound word are also indicated. Consequently, documents containing words like custodian, custodial, custody will also be retrieved. This Boolean retrieval system, however, has a few major shortcomings. It only indicates potentially relevant cases all containing the keyword, but does not give any indication of the measure of relevance. The occurrence of a word does not imply that a document containing this word is relevant. In order to obtain the best results when searching with the help of this system, a user should be well informed about the legal subject. The aim of using this kind of electronic information system is to find only those documents which are relevant within the data base and not to miss any of them.

The optimum result is obtained when both 'recall' (this is the number of documents which are retrieved and relevant divided by the total number of relevant documents) and 'precision' (this is the number of documents which are retrieved and relevant divided by the total number of retrieved documents) is a hundred percent(5). This result will never be achieved, which means that not all relevant documents are retrieved and that not all retrieved documents are relevant. It is, of course, impossible to determine the total number of relevant documents within a databank. The keyword will not correspond to the typing mismatch and, as a result, the document cannot be retrieved.

A disadvantage of the use of full-text data bases is that typing errors in the texts will interfere with the results or can even be fatal. The keyword will not correspond with the typing mismatch and as a result of this the document will not be retrieved.

2.4 The searching procedure in our research project

In order to carry out a practical examination of possible advantages and disadvantages of using different sources, research was undertaken in the field of family law. Custody and access(6) was chosen as sufficient case law was available on these topics. A first search showed that there would be enough cases available from 1965 onwards containing the word 'custody' (876) and 'access'. In order to establish the relevance of a verdict, the short summary at the beginning of the text must be read. If there is still some doubt, the complete text must be read. This reading is facilitated by the fact that the keywords are highlighted in the full-text, which makes it easy to conclude by reading the context whether the document is relevant.

As the cases which were presented as relevant appeared to be too dissimilar, the topic was narrowed down. Only custody and access conflicts between parents after divorce have been used in order to collect two homogeneous sets of approximately 50 comparable court cases which were appropriate for our research project.

The texts of these cases were analyzed in order to create a list of factors which might have been of importance to the judge in reaching his decision. By reading the verdicts, it was possible to determine whether or not each factor had appeared in the case. This was a time-consuming procedure, for the aim was to include in the list as many potentially important factors as possible. Methods are being developed for facilitating this coding in future using conceptual techniques (Mulder 1993).

If half-way through this procedure it was decided to add a new factor to the list, all previously coded cases had to be checked for the occurrence of this particular factor. Where the texts of the cases were stored in a private databank this procedure was facilitated, using computerized - for example WP51 - searching routines. In this way missing data were recovered.

This preparatory procedure resulted in a matrix with the cases in the rows and the factors (and the decision) in the columns. Factors which scored 'yes' or 'no' less than three times were removed from the list. This format made it possible to apply various kinds of statistical methods.

The results of these statistical procedures were threefold (Combrink, 1993) and (Piepers, 1994):

1. A ranking of the cases(7);
2. A prediction model: with the help of this model it was possible to predict a verdict in a certain case, using the factors of all the cases.
3. A ranking of the most significant factors.

The results of the second and third group were suitable for use as a criterium to compare both procedures described in this paper.

3. Using paper court files

3.1 Some advantages and disadvantages

The same preparatory method was followed using traditional paper court files as a source. Instead of sitting at a computer and looking at a screen, the researcher had to visit the archives of the Dutch courts to find and examine the files. First the approval of the court had to be asked in order to have access to this extremely confidential material which still contains the names of the people involved. For this reason a declaration of secrecy had to be signed. Nothing was to be copied or scanned unless the names had been removed so that it could not be recognized.

Compared to CD ROM verdicts, the paper files kept by the Dutch courts contain much more detailed information, such as a report of the Council of Child Care, a protocol of each court session, the petition and the defence, letters written by the children involved and, finally, the verdict reached in the case. Unlike electronic CD ROM verdicts, the complete text of the verdict is available. Nothing is removed from it and it has not been summarized by any editor. Using the court files, the researcher has at his disposal the same written information the judge had at his disposal. The only difference is that the judge has seen and heard the parties so he also has an impression of their behaviour in court when reaching his final decision.

3.2 Searching for the right information using paper court files

We started with verdicts on access conflicts reached by one court only. There were four different juvenile judges involved in the decision-making process in this court. As there were more than enough juvenile court cases available on access conflicts between parents, we only used cases in which there was a request to establish a right of access and not cases in which there had already been an arrangement but one of the parties wanted this to be changed. In order to obtain as much information as possible we only used files containing a report of the Council of Child Care. Until now a sample

of 45 cases had been taken at random from all verdicts pronounced during a period of three years (1991, 1992 and 1993) 8. In total 50 cases will be collected.

3.3 The analysis

Basically the same list was used as described above to analyze and code these cases. However, some factors were removed because they would never occur with regard to lower courts' procedures (for example the outcome of the previous court, the residence of the court) and some specific lower court factors were added. Due to the fact that the material was far more detailed, some new factors were also added to the list. Reading and analyzing the complete court files was even more time-consuming than using electronic information. Reading too many cases after each other is not to be recommended because this increases the chance of making mistakes during the procedure: cases and factors might become mixed up.

By the application of statistical linear regression techniques, the factors with a high correlation value to the verdict could be retrieved. This made it possible to compare the ranking lists of the most important factors in both methods: using traditional paper files and electronic databank verdicts.

4. Comparison of the results

4.1 The ranking of the facts

In the cases which were collected from the CD ROM set and analyzed, several factors seemed to have been of significant importance to the judge. In order to compare the results, the paper files list was checked in order to ascertain whether these particular factors also occurred on its list and, when this was the case, whether they were also of significance with regard to lower juvenile court decision.

This was done in the following way. The correlation coefficients (p.m.c.) which proved to be statistically significant (Xn) were compared to the correlation coefficient of the same factor based on lower court files only (Ym).

Table 1: The highest factors on the CD ROM list compared to the paper files list

The results of this comparison were reasonably satisfactory, although there were some differences.

The most highly correlating factors X8 and X117 and also the factors X138, X82, X105 and X76 were specifically related to the Supreme Court procedure and, therefore, did not appear on the paper files ranking list. The factors X27/Y66 and X26/Y65, dealing with the relation between the decision and the advice, respectively negative and positive, of the Council of Child Care appeared on both lists. When comparing these values, one has to bear in mind that the lower court paper files all contain such advice from the Council. With regard to the CD ROM cases there was mention of advice in only 50% of the cases. When the p.m.c. is calculated taking only these cases into account the values of X27 and X26 appear to be substantially higher.

Table 2: The highest factors on the paper court file list compared to the CD ROM list

A remarkable similarity between the results obtained using both systems was seen with regard to the age of the children. Non-custodial parents with children under the age of 7 were more likely to be denied access. Non-custodial parents with children aged between 7 and 12 years old, on the contrary, are more likely to obtain the right of access to their child after divorce. With regard to children above the age of twelve, there was no correlation between age and the judge's decision on access. After that the p.m.c. of the factors which proved to be statistically significant on the paper files list (Ym) were compared to their value on the CD ROM list (Xn). This yielded the results shown in Table 2.

The result of this comparison proved to be less satisfactory than the comparison described above. Several highly correlating factors did not occur at all on the CD ROM list. This could be because the fact scored 'yes' or 'no' less than three times and was removed from the list or because it was added while analyzing and coding lower court cases because this material provided much more detailed information.

The factors which appeared high on the paper files ranking list are of a different character to those highly ranked on the CD ROM list. The first are primarily related to the substance of the case, while the latter are mainly related to the procedural aspects.

Five factors appeared in both tables: X56/Y139, X27/Y66, X25/Y61, X130/Y53 and X26/Y65, which indicates that they are all of great importance.

4.2 Result of the cross validation method

The next step in the research is to compare the predictive capabilities of the models based on the factors of the cases. The cross validation method which has been used is described in (Combrink, 1993) and (Piepers, 1994) with regard to CD ROM custody cases.

Results achieved using the CD ROM cases

Results achieved using the Paper Files cases

In spite of the rather high a priori percentages of 55.56% and 68.89%, there was still a substantial gain in correct predictions using these models. Although there are those who contend that a collection of published cases is not representative and should only be used for practice and not for this kind of legal research, the results obtained using these cases are promising and even better than those obtained using the paper files.

5. Conclusion

The availability of legal texts in electronic form provides us with enormous opportunities for legal research. The number of electronically stored Dutch verdicts is still rather limited, but this will change in the near future. All the verdicts of the Supreme Court will be stored in electronic format and verdicts of the lower courts will also be available. It will be necessary to de-personalize the material with respect to family law cases heard by the lower courts.

The results of the comparison described in this paper prove that, although published CD ROM verdicts can be used as a source for this kind of jurimetrical research, unpublished lower courts' paper files provide us with more substantial information.

This makes the availability of these paper files in computerized format for research in the near future very desirable in order to build a model to predict the verdicts of courts on the basis the information obtained from material of lower courts. This is of importance because the number of people applying to the lower courts is higher than those applying to a Court of Appeal or even the Supreme Court.

The publication of all Supreme Court verdicts and also those of the lower courts will give a new dimension to research. The fact that the advice of the Council of Child Care is of decisive importance makes its report in which the advice is contained an interesting subject for further research.

Notes

1 See also (Morrison 1992), p.89.
2 Verdicts published in a magazine called Nederlandse Jurisprudentie were used in our research project.
3 See (Goodhart 1934, p.61) on precedents: 'It will make the judge a slave to the past and a despot for the future'.
4 See also (Leith 1991, p.88) for a more detailed outline of the use of full-text databases.
5 See (Mulder 1984, p.120) for a further outline on 'recall' and 'precision'.
6 Custody orders are now referred to as 'care, residence and supervision orders' and access as 'contact orders' in English family law. However our research only involved Dutch family law and it was not felt necessary to update the translation terms as these changes in terminology did not effect our research.
7 Due to the fact that cases are available in machine searchable form, other conceptual techniques can be applied, which makes it possible, even without analyzing the texts thoroughly, to make a ranking of the cases according to their relevance. This is done by comparing word frequency and word patterns in the verdicts (Mulder 1993).
8 In total 50 verdicts will be collected, but the last verdicts from 1993 are not available yet.

References

Combrink-Kuiters, C.J.M. and P.A.W. Piepers (1993). The Implementation of Predictive Capabilities into Legal Computer Advice Systems. Paper for the 8th BILETA Conference, Building Systems, 1-2 April, John Moores University, Liverpool. Published in: Pre-Proceedings of the Conference, pp.63-72.
Goodhart, A.L. (1934). Precedent in English and Continental Law, Law Quarterly Review.
Morison, John and Philip Leith (1992). The barrister's world and the nature of law. Open University Press, Milton Keynes.
Leith, Philip (1991). The Computerised Lawyer: A Guide to the Use of Computers in the Legal Profession, Springer-Verlag, London.
Mulder De, R.V. (1984). Een model voor juridische informatica (A model for the application of computer science to law), Vermande, Lelystad.

Mulder De, R.V., M.J. van den Hoven and C. Wildemast (1993). The concept of concept in 'conceptual legal information retrieval'. Paper for the 8th BILETA Conference, Building Systems, 1-2 April, John Moores University, Liverpool. Published in: Pre-Proceedings of the Conference, p.79-91.
Piepers, P.A.W. and C.J.M. Combrink-Kuiters (1994). Statistically Analysing Court Decisions on Custody Disputes. Paper for the 9th BILETA Conference, The Changing Legal Information Environment, 11-12 April, University of Warwick, Coventry. Published in: Pre-Proceedings of the Conference, pp. 65-72.