9th BILETA Conference
Building Systems

1st & 2nd April 1993
John Moores University
Liverpool

The concept of concept in 'conceptual legal information retrieval'

R.V. De Mulder, M.J. van den Hoven and C. Wildemast

Keywords: legal infonnation retrieval - conceptual

Abstract: This paper provides an overview of a 'conceptual' legal information retrieval system developed at the Centre for Computers and Law at the Erasmus University Rotterdam. This, it is suggested, is a way of overcoming the short-comings of the basic keyword and logical connector based automated legal information retrieval systems.

Introduction

Many recent publications on legal information retrieval agree that traditional automated systems for this purpose do not satisfy the demands of lawyers. A multitude of suggestions have been made for improvement. Many of these suggestion have in common that legal information retrieval systems should be 'conceptual'. Roughly speaking this means that legal information retrieval systems should contain more knowledge about the law and be more 'intelligent'. Wildemast and De Mulder (1992) give an overview of attempts to build such systems. On the basis of the conclusions in that overview, in this paper we would like to give an alternative approach to legal information retrieval that has been developed by the Centre for Computers and Law at the Erasmus University in Rotterdam.

After reviewing the conclusions reached in the earlier article, which form our starting point, we continue our explanation with some basic remarks on the meaning of the term 'concept'. Conceptual models will be considered as an intermediate step between concrete systems (objects in the world of experience) and formal models. We then state and explain our preferred definition of the term 'concept' and show how this definition can be refined or 'operationalized' and therefore made more suitable for our specific purpose. We subsequently show that such concepts typically form a hierarchy in which the concepts are the categories of a classification. Finally, we propose a way to visualize our model. Visualization is particularly important if the model is used as the basis of a computerized legal information retrieval system.

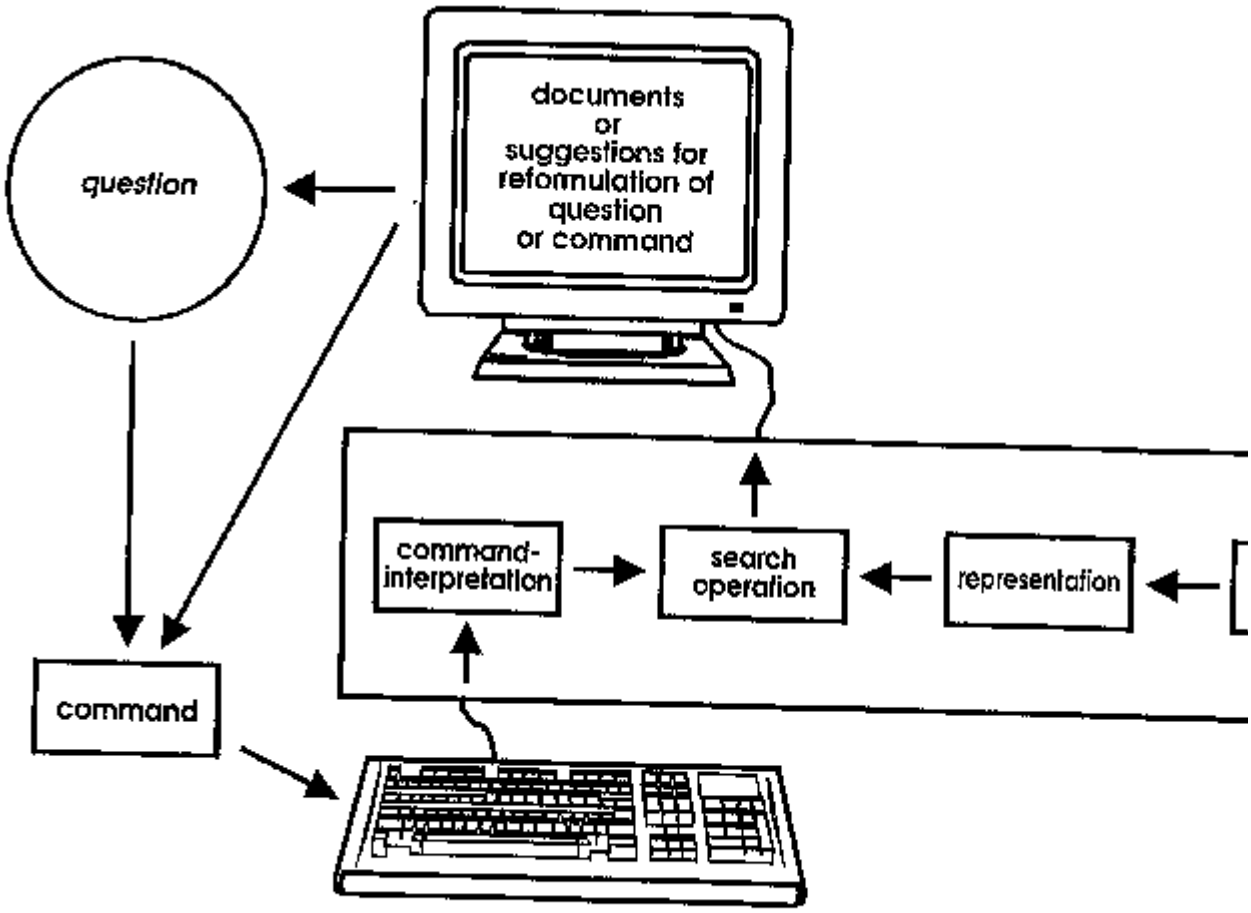
Starting point: conclusions from recent attempts to implement conceptual legal information retrieval systems.

The methods proposed in the literature for conceptual retrieval are aimed at:

- the interface with the users

- the representation of documents
- the search operation (Wildemast and Mulder, 1992).

It is the interface which makes communication between the user and the computer possible. It assists in the translation of the user's question into an actual search instruction for the computer. When the search instruction has been carried out, it is responsible for the reproduction of the results. On the basis of these results it is then possible to assess the relevance 6f the documents which have been found and to reformulate the question (or the actual command) if necessary. An interface can also assist the user in the formulation of the question (Vries *et al.*, 1991).



Conceptual retrieval can be realized by assisting the user in (re)formulating a search request. This is done by assisting the user in finding the right words to describe the concept and by providing the legal context in which concepts are described.

One of the characteristics of intelligent interfaces is that the user must work with the the concepts which have been programmed into the interface. As P. Leith (1990) and others have convincingly argued, these concepts are not objective data but rather interpretations determined by social circumstances. De Mulder emphasizes the fact that, unlike the physical sciences, the concepts used are not derived from empirical observation nor, as in mathematics, are they based on explicit and unequivocal conventions (Mulder, 1984; Mulder et al., 1989). The methods of representation are based on the assumption that conceptual retrieval can be realized if the representations of the original texts are based on the legal importance or legal meaning of a text. There are two categories: those which (only) use manual methods and those which use automatic methods. Each of these categories can be further sub-divided between reduction and interpretation approaches. The reducing approach means that the original texts are not represented by all the words which appear in them but by words

which reflect most adequately the (legal) contents of the text. The interpretation approach uses methods which represent documents with the help of (legal) knowledge. The reduction methods for representation run the risk of loss of information which might be important for a conceptual search. The major drawback of the manual interpretation method (of knowledge representation) is that the

knowledge and concepts represented are fixed; they cannot be altered by the user. Only the opinion of one legal expert or a group of legal experts is represented.

The search operation is the function which ensures that the concrete search instruction (whether or not already re-worked in the interface) is carried out on the documents represented in the system. Most search operations (for instance the well known Boolean search) make use of the occurrence of a term rather than for example the term frequency in a document. The result of the Boolean search operation is the answering of a yes/no question for each document as to whether the document satisfies the search instruction. Other search operations look for a standard which indicates the extent to which the document satisfies the search instruction. This may possibly be expressed in the form of an estimate of probability (Salton, 1989; Bookstein and Klein, 1990).

A similar result is achieved by search techniques which make use of 'neural networks'. Conceptual retrieval with the help of neural networks was proposed by (Belew, 1987; Rose and Belew, 1989). The use of neural networks for searching as proposed by Belew and Rose has the advantage that no legally fixed representation is used. This method takes into account the open-structured character of (legal) concepts. The user is free to formulate (re formulating) his own concepts. Neural networks and their application to law however are still in the development stage. A topic of research could be how highly activated terms relate to legal concepts.

The analysis of the advantages and disadvantages of the techniques presented in current literature leads to the conclusion that it would be desirable if both the method of text representation and the interface would allow the user to define his/her own concepts. These concepts could then be more precisely re-defined on the basis of the results of search operations or interpretations by the interface. The system could store the user's concepts: thus becoming a 'learning' system.

The representation technique must, therefore, not only be objective but also be complete while ensuring that the search time does not become unpractically long. A full text storage supplemented with a combination of, on the one hand, a complete word list plus the corresponding word frequency in each document and, on the other hand, a document list including the frequency of each word of every document would seem to be a workable choice.

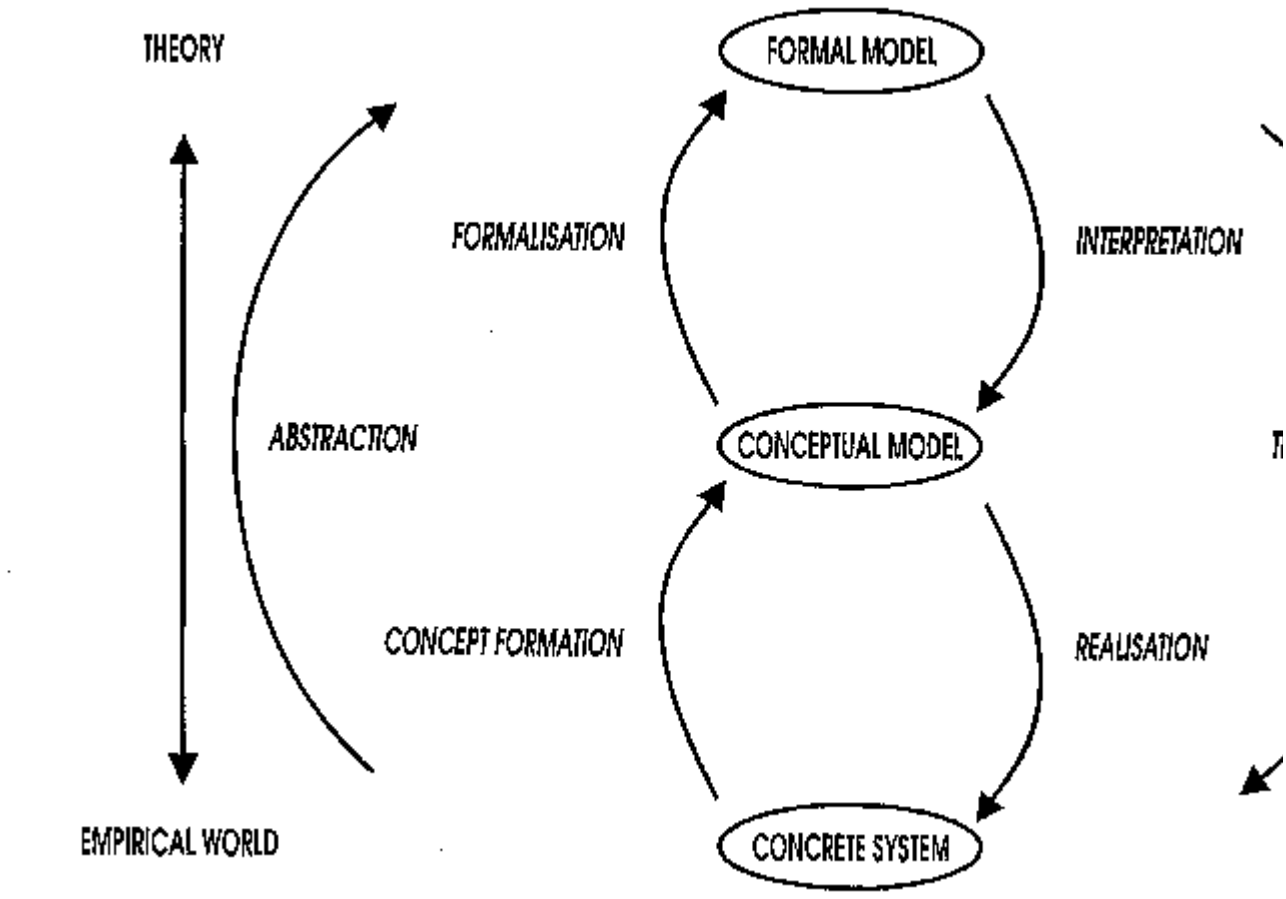
As regards the interface, it is especially important that the user can bring into the system and modify his own concepts. We would argue that the quality of the interface is therefore the constraining factor in conceptual legal information retrieval at present. Research efforts should concentrate in this area. A lot more can be done. For example, in the available literature there is little mention of an obvious method of allowing the user to make his own ideas explicit by providing examples of clearly relevant documents which are known to him (Bookstein and Klein, 1990; Gelbart and Smith, 1991: 229.) The choice of search technique is not a crucial design decision as, given the design choices for interface and document representation, various search techniques can be used as alternatives or supplements to each other.

A prototype containing a very large collection of legal cases and formal legislation, operating with techniques suggested above, is now nearing its completion at the Centre. We like to refer to it as a learning concept processor. The documents can be looked up via the interface and are given a relevance score using statistical techniques and indications by the user as to their relevance. Concepts are -roughly speaking - stored in terms of sets of relevant documents, with concept names, user name and date and time. Relationships between concepts can be traced and/or indicated by the users. In our opinion, such a concept processor is a necessary part of a legal conceptual retrieval

system because in law concepts do not have a fixed and objective content, but can vary from user to user, from problem to problem and from time to time.

Concrete, conceptual and formal systems

Knowledge acquisition in empirical science as well as learning in general are both characterized as processes of abstraction. Concepts are of a quite different kind from objects in the world of experience (the 'real world'.): they are abstract. This means that concepts can exist in human minds and possibly in a special 'world', the 'world of consciousness' (*cf.* Popper, 1973; Bruner *et al.*, 1986), but they cannot be observed or otherwise be a direct object of experience. We will call combinations of concepts a conceptual system (or conceptual model if it could represent something). As a consequence of their abstract nature, concepts as such cannot be entered into or processed by a computer. In order to use the result of knowledge acquisition in a computer it is necessary to construe a formal model which is a combination of mathematical or logical symbols. The two steps leading to a formal model (concept formation and formalisation) can be graphically represented in the following picture (fig. 2).



Over the last couple of hundred years demands have been put forward by scientific methodologists for the construction of conceptual and formal models of concrete systems (objects or combinations of objects of the world of experience). These demands have meant that statements which are made within the framework of a discussion aimed at the acquisition of empirical knowledge must be falsifiable except in so far as the content of the concepts is fixed by conventions. The concepts involved are empirical concepts (formed according to certain procedures based on observation), or concepts of which the content is fixed by conventions.

As the law is not an empirical science, we do not stipulate that concept formation and formalisation should, in all cases, be subjected to the demands mentioned above. It is, however, desirable to make use of the results of the methodology of science as even everyday learning such as the way children learn could be better understood and teaching could be made more effective by doing so (*cf.* Bruner *et al*.,1986). We have applied this idea in the process of building the 'concept processor' by making use of a notion of 'concept' that is in accordance with modern scientific methodology and a formal model that has shown its relevance in empirical science.

'Terms' and 'concepts'; initial definition of 'concept'

Concepts should be distinguished from the terms that refer to them.

Different terms could refer to one concept, and more than one concept could be referred to by one term. In our concept processor, the user is entirely free with respect to the terms that he wishes to employ to refer to a (legal) concept. These terms are not used as search terms (or keywords) in our system. In general, a user will have a more or less clear idea of the intension of the term that he uses for a certain concept, but he will have to define his concept precisely by entering the extension of the term into the computer.

The **extension** of a term is the totality of all things to which it applies (Rescher, 1969: 26). In logical terms, it is the class or set consisting of all the things, if any, to which it applies. The extension of a term cuts across boundaries of time: the extension of the general term 'lion', for example, includes not only living lions, but also the lions of the past, and those of the future. The **intension** of a term is the sum total of all the properties (or 'attributes') that must be possessed by every entity to which the term can be applied. Cats, for example, may differ in many ways, but anything to which the name is properly given will have to have certain specifiable properties: it must be a living thing, have a backbone, nourish the young by suckling etc. Let us, for example, use the extension of a concept denoting term to define 'concept'. Within the 'universe' of a data base of (legal) documents (cf. court decisions) the extension of the terms that the user employs to refer to his concepts consists entirely of documents. Therefore, we characterize 'concept' as follows:

A concept is a set of documents.

This means that in principle a user of our concept processor is required to define his (legal) concepts by entering a list of documents in the database that he considers to be relevant. (These documents, identified by the user as relevant to his concept, are called 'exemplars'). Consequently, the search facility of the system will search for documents that are similar to the exemplars. In order to fulfil this task, the program will compare the properties or attributes of potentially relevant documents with those of the exemplars. These attributes consist of the words used in the documents, their frequency, possibly the order in which the words appear etc.

It is outside the scope of this paper to discuss the standard that is used to decide whether and to what extent a document is similar to the exemplars. We will suppose that the method of representation of the documents combined with the search method applied are capable of ranking documents according to their relevance and, furthermore, compute some measure of the probability that the document is relevant. Those documents that are ranked at the top of the list are the ones that the user will be interested in most. If the system comes up with a document that the user identifies as relevant, he/she can decide to add it to the list of exemplars. The following search operation will then be based on more information than the initial one. There is, however, also a very important use for the documents that the system ranks highly, but that the user identifies as non-relevant.

Concept as ordered sets of exemplars and counter-exemplars

When teaching a child the meaning of the word 'cat', we would consider it a relative success if the child calls a dog a cat. The child may at least have learned that 'cat' refers to something living, furry, with four legs, an animal and a pet. For a more advanced level of knowledge, however, a fair demand would be that the child knows the difference between a cat and a dog. In order to teach this it is helpful if the child could be confronted with a dog and learn that this is not a cat. If we would just show a large dog to the child, it is not unlikely that it would still regard a small dog to be a cat. This means that in order to obtain a more precise notion of concept, it is important to have 'counter examples' that are very similar to examples of a cat. Translated to the problem of learning the meaning of a legal concept by the concept processor, it is important that is has at its disposal a set of counter exemplars of documents that are as similar as possible to the members of the set of relevant documents.

It is for this reason that although possibly for the initial stages of learning the system could compare the set of exemplars to all other documents in the total set, it is necessary that the user can provide the system with a set of counter exemplars that are as similar as possible to the exmplars of relevant documents. Typically, the user would inform the system that documents that are put forward as candidates for relevant documents are in fact counter exemplars. These non-relevant documents would 'teach' the system the finesses of the concept the user has in mind.

Let us consider a situation where a user is searching for documents concerning the concept 'eye-witnesses as evidence in criminal cases' in a case-law data base. The user provides the

system with some exemplars of the subject. The system compares the exemplars to all other documents and it is likely that, along with relevant documents, it will come up with documents that deal with criminal evidence in general as well as the use of eye-witnesses evidence in civil law cases. Both kinds of non-relevant documents will be taken into account as counter exemplars for the next stage of the search making a crucial difference to the final result. Therefore, a more precise definition of concept is desirable for use in the concept processor. The former definition 'a set of documents' should be refined to:

> A concept is an ordered pair of sets of documents consisting of exemplars (of relevant documents) and counter exemplars (a set of non-relevant documents that are as similar as possible to the relevant documents).

A concept can be referred to by a term that indicates the membership of the first set of the pair, possibly by expressing its intension. For example: '(documents that contain) civil law (cases)'.

The hierarchical structure of a conceptual system: categories

Our reasoning leads to a further conclusion. In order for a child to learn the difference between a cat and a dog, it is helpful if the child knows about furry animals, or possibly about pets. This means that concepts are part of a hierarchy or classification (*cf.* Rescher, 1969: 48) and are learned and clarified by their use. Which of the several possible hierarchies (e.g. 'furry animals' or 'pets' as the broader concept) would lead to more effective learning is, of course, dependent upon the aim of the learning process and the child in question, but probably more important is the clarity of the concepts (or 'categories') that form the hierarchy.

An example would be a set of documents dealing with court cases using eye-witnesses in evidence in criminal law trials, and the set of counter exemplars consisting of documents about evidence in criminal law, but not about eye-witnesses. The hierarchically higher category would be 'evidence in criminal law'. The set of exemplars would consist of both exemplars and counter exemplars of the lower concept, and the set of counter exemplars would consist of documents dealing with cases

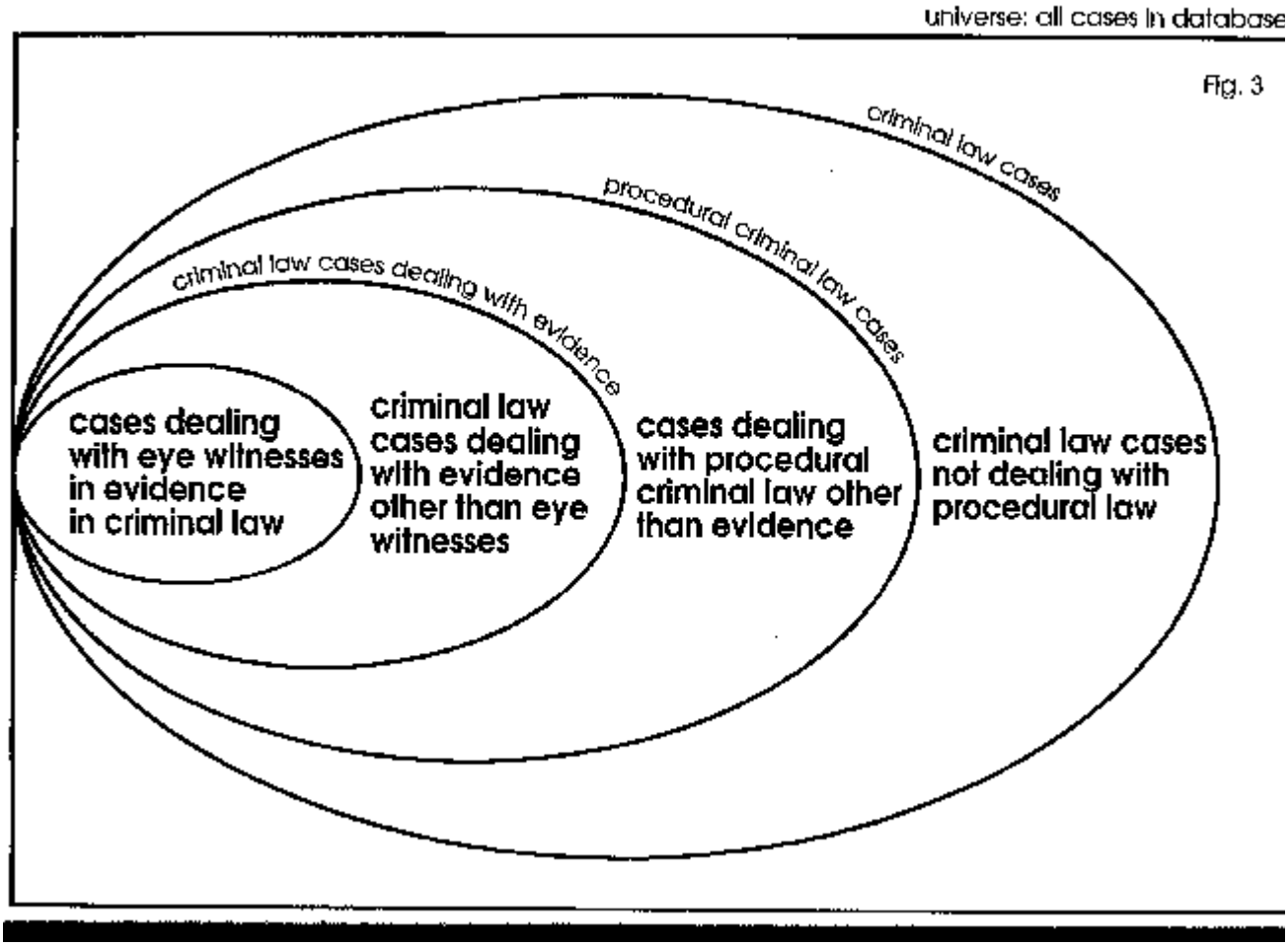
about procedural criminal law, but not about evidence (See Figure 3).

This example provides a possible alternative hierarchy as well. The set of exemplars could again consist of documents dealing with court cases about the use of eye witnesses in criminal law trials, but the set of counter exemplars would contain documents about the use of eye-witnesses in cases other than criminal law cases. This shows that for the success of a search actvity it would be important to use the most appropriate hierarchy of concepts. For in the second example the hierarchically higher category would be 'criminal law and non criminal law cases dealing with eye witnesses', for which counter exemplars would be easy to find as they would form a large group, but they would probably not be very similar to the relevant documents. This means that the concepts 'evidence' and 'procedural law' would be necessary for a succesful search. Note that the next higher concept ('criminal law and non criminal law cases' could already cover the whole 'universe' of the data base (unless the database dealt with documents other than legal cases).

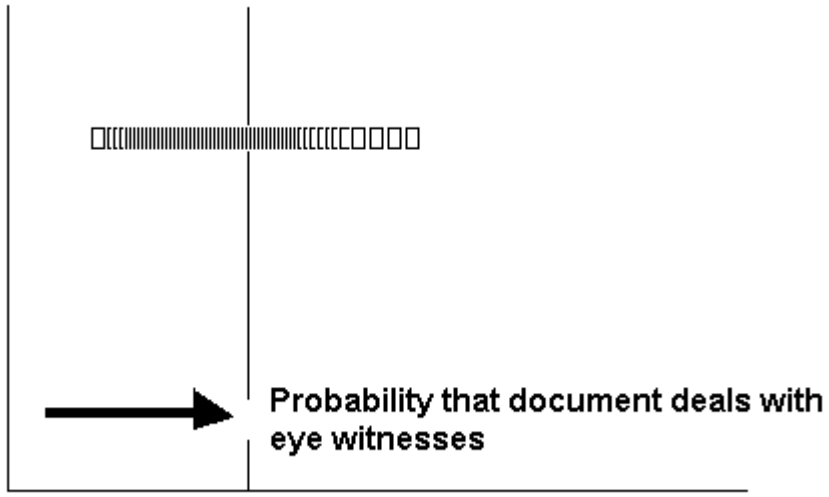
Finally, our initial definition of concept itself is an example of the use of an hierarchy. Objects within the extension set of any concept introduced by a user will be documents. Therefore, we characterised 'concept' as a set of documents rather than things or objects.

Visualisation of the model

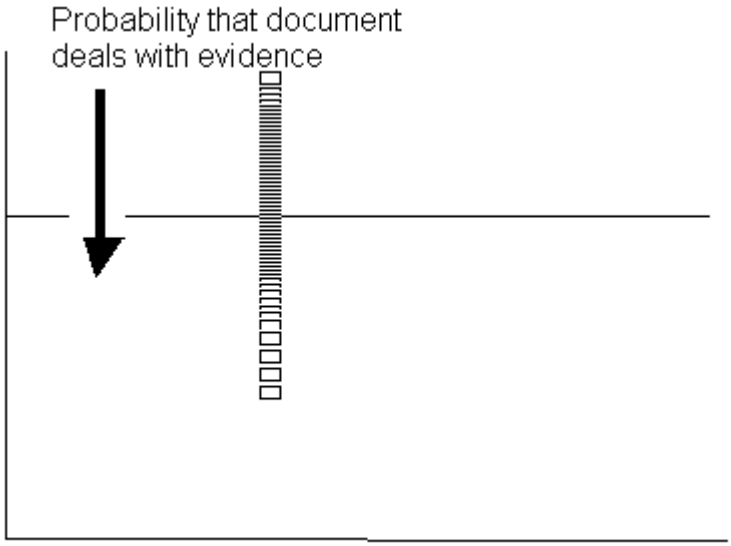
As the 'concept processor' is meant to be part of the user interface of a computer program and, furthermore, the amount of information to be processed is large it is of the utmost importance that understandable visualisations can be created. The use of a concept of 'concept' taken from set theory, enables such a visualisation. The graph in fig. 3 shows a so called Venn-diagram which provides a clear picture of the hierarchy of concepts. The broader concepts, i.e. the ones 'higher' in the hierarchy, are shown as larger ellipses. Larger ellipses correspond to larger extensions and simpler intensions.



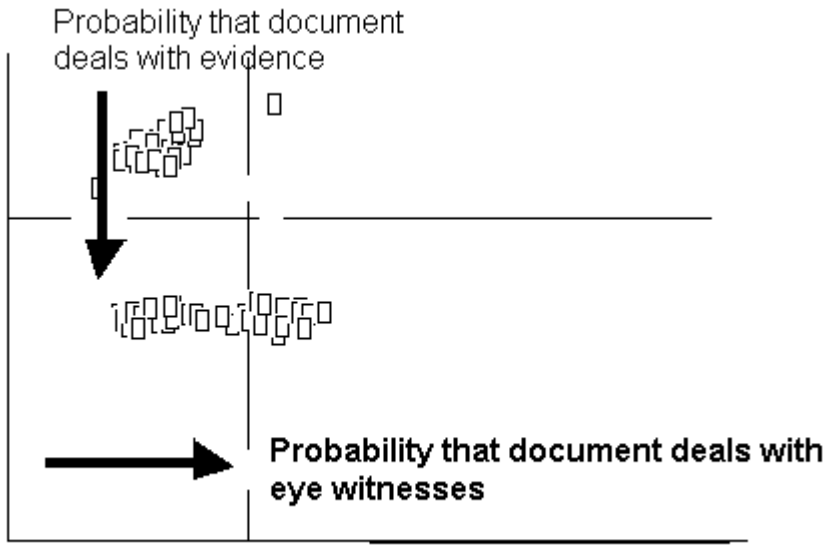
On the basis of our hypothesis, that the system is capable of ranking documents according to their relevance, it is possible to draw even more interesting graphs of the relationships between the concepts as well as the position of individual documents. Suppose that a user is looking for documents about 'eye witnesses in criminal law'. Within a previously selected set of documents the user has indicated a number of exemplars and counter exemplars. On that basis, the computer has calculated a measure of the probability that a document is relevant. In fig. 4 all the documents under investigation have been drawn parallel to the x - axis, each rectangle representing a document. Unfortunately, on the basis of this picture it would be hard to decide where to 'draw the line' between relevant and irrelevant documents.



The picture of fig. 5 is much clearer. In this case for the same set of documents a measure for the probability of relevance with respect to another concept, namely 'evidence' has been computed. The documents are ranked along the y - axis in this case.



In fig 6, finally, both dimensions are combined. Apparently, if the cases (probably) dealing with evidence are singled out, there are two rather clear clusters of documents. Similar curves as in fig. 3 could be drawn around these clusters in order to obtain a Venn diagram.



In order to obtain a clearer picture, the user has to introduce to the system more knowledge about the subject. He can either add more exemplars and counter exemplars, or introduce a new category. If the knowledge has some empirical relevance to the content of the documents (as 'perceived' by the system on the basis of the user's information), the documents will show up as clusters rather than as a shapeless group or line.

Conclusion

The paper was based on a conclusion reached in a former article: that for conceptual legal information retrieval a system should have a user interface which enables the user to define his own concepts. We called this facility a 'concept processor'. It should be able to extract a maximum of information from the user, and be able to store this information for later use. It should also be flexible enough to be able to deal with a variety of concepts and with changes to existing concepts over time. Furthermore, it would be desirable if the system could help the user to define clear concepts, i.e. it should make use of a notion of 'concept' that is consistent with the insights of contemporary scientific methodology and those of cognitive psychology.

Set theory provides such a notion of 'concept'. 'Concept' was characterized as a 'set of objects' and in our subject of conceptual legal information retrieval, consequently operationalized as a 'set of documents'. Further investigation led to a more refined definition of concept, namely 'an ordered pair of sets of documents'. The first set of the pair consists of (exemplars of) relevant documents and the second set consists of counter exemplars (exemplars of non relevant documents) that are as similar as possible to the relevant documents.

For reasons of effectiveness and efficiency of the searching it would be desirable to be able to limit the search to the set of all potential exemplars and counter-exemplars. Exemplars and counter exemplars together form another concept. For this common concept, which would typically be 'higher' in an hierarchical order of concepts, counter exemplars could be found in turn. Thus a hierarchy of concepts (or 'categories', or 'classes') could be built. Alternative hierarchies could be built to obtain the same result in searching for the desired documents. We would like to put forward as a hypothesis that a clear and detailed hierarchy produces faster and more unequivocal results. Thus by using the system the user would not only enable the system to 'learn', but he would also clarify his own concepts and, therefore, possibly increase his knowledge, in addition to what he could learn from the end result of his search activities.

The use in this model of clear and formalized concepts in general and set theory in particular, enables a clear means of visualisation. This is an important part of the user interface and the system as a whole as it can help the user define concepts in the most effective way; a way in which the hierarchy of concepts has the clearest image.

References

Bakel J. van (1991). 'Meaning, Prototypes and the future of Cognitive science'. *Minds and*

Machines 1: 233-257.

Belew R.K. (1987). 'A connectionist approach to conceptual information retrieval', in:

Proceedings of the First International Conference on AI and Law, Boston, 116-26.

Bing, J. (1987). 'Designing Text Retrieval Systems for 'Conceptual Searching", in: *Proceedings of the First International Conference on AI and Law*, Boston, 43-51.

Bookstein A. and S.T. Klein (1990). 'Information Retrieval Tools for Literary Analysis', in: Tjoa, A.M., and R. Wagner (eds.), *Database and Expert Systems Appilcations (DEXA), Proceedings of the International Conference in Vienna,* Austria, 1-7.

Bruner J.S., J.J. Goodnow and G.A. Austin (1986). *A study of thinking*, New Brunswick and Oxford.

Davis D. (1986). 'Semantic Analysis in Legal Text Information Retrieval', in: *Automatic Analysis of Legal Text; Logic, Informatics and Law*, Elsevier Science Publishers B.V. (North-Holland), 473-81.

Dick J.P. (1987). 'Conceptual Retrieval and Case Law', in: *Proceedings of the First International Conference on AI and Law*, Boston, 106-14.

Dick J.P. (1991). Representation of Legal Text for Conceptual Retrieval, in: *Proceedings of the Third International Conference on AI and Law*, Oxford, 244-53.

Gelbart D. and J.C. Smith (1990). 'Toward A Comprehensive Legal Information Retrieval System', in: *Database and Expert Systems Applications (DEXA)*, ed. by Tjoa and Wagner, 121-25.

Gelbart D. and, J.C. Smith (1991). FLEXICON, 'A Legal Text-Based Intelligent System', in: *Proceedings of the Third International Conference on Artificial Intelligence and Law,* Oxford, 225-33.

Guldotti P., L. Lucchesi, P. Marlani, M. Ragona and D. Ticornia, (1990). 'A Simple Intelligent Interface to Data Bases on Environmental Law', in: Tjoa, A.M., and R. Wagner (eds.), *Database and Expert Systems Applications (DEXA)*, 285-289.

Hafner C.D. (1978). *An Information Retrieval System, Based on a Computer Model of Legal Knowledge*, UMI research press, Michigan.

Hafner C.D. (1987). 'Conceptual Organization of Case Law Knowledge Bases', in: *Proceedings of the First International Conference on AI and Law*, Boston, 35-42.

Hempel C.G. (1967). 'Fundamentals of Concept Formation in Empirical Science'. In: *International Encyclopedia of Unified Science*, vol II, no.71-62.

Kripke S. (1972). 'Naming and Necessity'. In: G. Harman and D. Davidson *Semantics of Natural Language*. Dordrecht: Reidel Publ. Comp.

Leith P. (1990). *Formalism in AI and Computer Science*, Ellis Horwood, Simon and Schuster.

Martin R. (1987). *The Meaning of Language*. Cambridge, Mass.: MIT Press.

McCauly R.N. (1988). 'Epistemology in an Age of Cognitive Science'. *Philosophical Psychology* 1, no.2 143-52.

Merkl D., A.M. Tjoa and S. Vieweg (1992). 'BRAND - An Approach for Knowledge Based Document Classification in the Information Retrieval Domain', in: *DEXA* 145-259.

Merkl D., S. Vieweg and A.Karapetjan (1990). 'KELP: A Hypertext oriented User-Interface for an Intelligent Legal Fulltext Information Retrieval System', in: *Database and Expert System Applications (DEXA)*, ed. Tjoa and Wagner, 399-404.

Mulder De R.V. (1984). *Een model voorjuridische informatica*, Koninklijke Vermande bv.

Mulder De R.V., c. van Noordwijk and H.O. Kerkmeester (1989). 'Knowledge Systems and Law-The JURICAS Project', in: *pre-proceedings of the Third International Conference on Logica, Informatica, Diritto, Legal Expert Systems*, vol.1, ed. by A.A. Martino, Florence, 205-16.

Popper K. (1973). *Objective knowledge*, Oxford: Oxford University Press.

Putnam H. (1975). *Mind, Language and Reality*. Cambridge: Cambridge University Press.

Rescher N. (1969). *Introduction to logic*, New York.

Rosch E. (1975). 'Cognitive Reference Points'. *Cognitlve Psychology* 7 532-47.

Rose, D.E. and R.K. Selew (1989). 'Legal Information Retrieval: a Hybrid approach', in: *Proceedings of the Second International Conference on AI and Law*, Vancouver, 138-146.

Salton G. (1989). *Automatic Text Processing; The transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley Publishing Company, US.

Vries de WS., H.J. van den Herik and A.H.J. Schmidt (1991). 'Separate Modelling of User-System Cooperation', in: *Legal Knowledge Based Systems. Model-Based Legal Reasoning*, Jurix, eds.

Breuker J.A., Mulder De, R.V., Hage, J.C., Koninklilke Vermande BV, Lelystad, 28-39. Wildemast, C.A.M. and R.V. De Mulder (1992). 'Some design Considerations for a Conceptual Legal Information Retrieval System', in: Gr&uutters, C.A.F.M., J.A.P.J. Breuker, H.J. Van den Herik, A.H.J. Schmidt, C.N.J. de Vey Mestdagh (eds), *Legal Knowledge Based Systems: Information Technology and Law*, JURIX'92, Koninklijke Vermande, Lelystad, NI, 81-92.