College of Information Science and Technology

Reusing Cases to the Automatic Index Assignment from Textual Documents

Rosina Weber **Alejandro Martins** **Eduardo Mattos**
Tania Bueno **Hugo Hoeschl** **Christiane**
Roberto Pacheco **Ricardo M. Barcia**

Federal University of Santa Catarina - Graduate Program in Production Engineering
(PPGEP/UFSC), Florianópolis, SC - BRAZIL, Phone/Fax: +55 48-331-7010
rosina@eps.ufsc.br

Abstract. This paper describes one solution developed to convert textual documents into formlike representations of cases. The experiences described by cases are textual descriptions of legal decisions. Indexing vocabulary and assignment theory contributed in gathering expert knowledge to define attributes and values as well as the required elements to employ template mining. Most index values are automatically extracted by the use of template mining. The multi-purpose index *Theme* is automatically assigned by reusing cases through an elaboration process. Seed cases are used to indicate values if the new case is a partial match to one in the case base.

1. Introduction

Textual descriptions of legal decisions are the knowledge source of jurisprudence research. Judicial cases are described with natural language text what represents a hard-to-use form, therefore demanding case engineering efforts. The review of the research brings forward intelligent applications in the legal domain that use complex indexes that demand extensive case engineering such as HYPO (Ashley, 1990), GREBE (Branting, 1991), CABARET (Rissland & Skalak, 1991) and BankXX (Rissland et al, 1993 & 1996). The extensive case engineering requirements limit the implementation of these systems as effective commercial tools to solve real world problems comprising the universe of such domains. The significance of this research is the reduction of case engineering requirements, facilitating the development of case-based systems to the legal domain what eventually improves the judicial system's quality and celerity.

This paper describes one method of the system PRUDENTIA (Weber et. al., 1997a; Weber 1997, and Weber et. al., 1997b). PRUDENTIA is a Case-Based Reasoner for jurisprudence research. The feasibility of this system depends upon the development of a methodology that converts the cases in textual form into a formlike representation that is amenable to be handled by the case-based reasoner. Following this target, we have developed methods (Weber et. al., 1997) based on expert knowledge to automatically assign values to indexes using template mining (Lawson et. al., 1996). Template Mining is a Natural Language Programming technique that extracts data from texts when the text forms recognizable patterns

from the target to be extracted or its surroundings. A template carries information on what to search in the text and it is triggered to extract the parts indicated.
The formlike cases in the reasoner represent an interpretation of the experiences described in the textual decisions. Multi-purpose indexes refer to sets of indexes in which one value indicates the category of the index what raises difficulties to their assignment by template mining. Elaboration[1] consists of inferring values that are not on the surface of cases. Reusing cases that have been assigned semi-automatically as seed cases to an elaboration process supports the automatic assignment of multi-purpose indexes. These seed cases provide the required knowledge to find the proper values for the indexes in the new cases.

2. The system: PRUDENTIA

Humans employ analogical reasoning to perform jurisprudence research as the research is motivated by a new legal situation that is compared to the past legal decisions. Thus, the choice of a CBR system to model such task is not only appropriate but obvious since each legal decision describes an experience. PRUDENTIA is an interpretive case-based reasoner designed to represent analogical reasoning to provide efficient jurisprudence research. These days the main source of jurisprudence research consists in text database systems. The Information Retrieval literature (Blair & Maron, 1985) points out that these systems are limited to a recall of around 25%, that is, only this percentage of relevant cases are actually retrieved. Besides, the low precision (the second parameter used for evaluation of database systems) forces the user to search among several cases to find the relevant ones. These two parameters indicate that the use of these systems makes the task time-consuming and imprecise.
Cases in PRUDENTIA are interpretations of legal decisions that are originally written in natural language text. The interpretation of each legal decision is performed by a legal professional whose expertise is needful to interpret experiences within the legal domain. Cases in this reasoner are modeled with a formlike representation, i.e., a set of fields (attributes) properly valued. Experts have suggested attributes to represent the interpretation of the legal decisions and this is the point when we have been confronted with the strongest encumbrance, that is about the impossibility of hand-coding the cases. The automatic indexing of multi-purpose legal indexes was still one difficulty left to be solved; we describe its solution to this problem in next section.

3. Multi-purpose Legal Indexes

The expert's interpretation of the legal decisions resulted in several attributes that must be defined in the index vocabulary to ensure efficient retrieval. Defining the index vocabulary is one of the indexing task steps. CBR literature (Kolodner, 1993) suggests that the knowledge engineer envisions the whole collection of

cases seeking for possible values that might comprise the requirements for a good index. To define an index, one should ensure that this index has a value in each case. Since the task in our system requires a large data base that comprises several different types of cases, accomplishing such guideline becomes unfeasible. There are some relevant indexes that must be defined in order to ensure proper retrieval. On the other hand, there are a great amount of values that occur in the cases that must be represented as indexes in order to ensure an efficient similarity assessment that results in the retrieval of the most useful cases. We believe that the expert interpretation can indeed represent all cases if we efficiently overcome the indexing constraint. The indexing constraint refers to the impossibility to value one same index in all cases since they represent experiences that vary in nature. Evaluating these values, we have noticed that they can be classified into groups according to their nature. For example, there are some specific purposes that motivate different petition types. An *habeas corpus* petition is motivated (in Brazilian law) by four different applications, i.e., annulment, canceling, mistrial, and abatement. In a *criminal appeal*, there are other motivations such as the ones related to categorizing offenses according to their severity, which might be of great relevance within the representation of those types of experiences. However, one may find a case describing another type of petition in which the motivation (application class) is not so important. Rules assigning different weights could be a choice if there weren't so many important index values to be used to index some cases.

The solution we have found is to define multi-purpose indexes. These indexes comprise multi-purpose values, that can be classified in different classes. Following the previous example, if the index value is the *habeas corpus* motivation, the index value will be for instance canceling and the case representation will store the class motivation in another attribute which is not an index. As a result, by giving to this index a generic name theme, we assign as many values as necessary to represent one case, overcoming the necessity of acknowledging all the values to every index in all cases.

The index *theme* refers to some secondary aspects or circumstances that interpret cases. The complexity of the these indexes stems from the fact that they were defined to complete the universe of the attributes in describing the content and context of the experiences on the legal decisions. The values for the attribute theme that may be present in a legal case are grouped into classes of the same nature, namely: class of tests required (mental health evaluation required, evaluation of drug dependency required); class of application (application for annulment, application for canceling, abatement or mistrial); class of external context (traffic accident, strikes, penalty reduction); severity category (cruelty, break into), etc.

In next section we describe how the reuse of cases can support the automatic index assignment of values for the index *theme*.

4. Reusing Cases to the Automatic Index Assignment

One of the advantages of CBR concerns to knowledge representation. It is claimed that one does not need to model all the interrelations and details of the cases for the system to work. Testing and working with this case base of 138 cases supported such statement. Some new cases described with an argument of insanity, for instance, will cause the retrieval of cases that have issues associated to mental health such as a court order for mental health evaluation. It does not happen because we have explicitly represented this correspondence. Our system does not evaluate the similarity of every argument evaluating the semantics of each value. It happens because there are concepts that are inherent to cases where the defendant was sent to a health institution that are also present when insanity is pleaded. This is a result of dealing with actual experiences.

At this point, we attempt to make use of this advantage in another fashion. Since judicial decisions describe experiences that comprise inherently correlated issues, we may comprehend these cases as consistent in nature. Therefore, it is valid to attempt to search for values to secondary issues and circumstances in cases that are somehow similar to a given new one we are attempting to ascribe. This is why we can reuse seed cases, to point what types of values to search in the new case.

The reuse of cases is motivated by the elaboration method [1]. The idea of reusing cases to suggest possible new values can be extended to suggest values to new cases to be incorporated in the case base.

The last prototype of the system PRUDENTIA has a case base in which its cases were converted semi-automatically. One type of index - *theme*- was hand-coded. This has limited the case base size to only 138 cases. The next step is to increase this case base by indexing automatically 3,700 cases, which have all the attributes valued except for the index *theme*. These 138 seed cases have been valued

---

[1] *From K. Branting, (1991), case elaboration consists of inferring facts that are not explicitly stated in cases in order to improve their match. Case elaboration was termed knowledge-based pattern matching in (Porter et al., 1990) In the context of analogical reasoning, the process of determining "implicit shared properties" of cases from differing explicit representations has been termed reformulation (Russel, 1986). A related notion in machine learning is constructive induction which has been defined as "any form of induction that generates new descriptors not present in the input data"(Dietterich and Michalski, 83). Kolodner (1993) describes the elaboration process to support situation assessment; when the value of an index has to be derived, one can perform a partial retrieval and search throughout the similar cases for possible values to such features.*

automatically except for the index theme. From a case base of 138 seed cases we aim at generating automatically a case base with 3,700 cases within the same sub domain (criminal appeals).

Our problem refers to finding possible values of the attribute *theme* in 3,700 cases that are represented through other attributes. Once we assess the similarity of each new case in comparison to the 138 in the complete case base, we reuse the knowledge embedded in this case base to derive hints on what types of values to search in the new cases. Therefore, we develop a reasoner to perform incremental learning to enlarge the original case base.

The case base of 138 cases embodies an amount of 300 values for attribute *theme* that are being used as seeds in the learning process. These 300 values are represented in terms of concepts, objects and context and of their relationships with the case which they are derived.

The input cases are composed by new decisions that have been automatically modeled and indexed except for the index theme. When a new case is input, the reasoner searches the case base for similar decisions, relying only on the indexes that have been valued in the new case. The outcome of the retrieval process is a set of similar cases that is presented to the user sorted by a similarity measure, indicating the given values for attribute theme. The user (expert) chooses the cases and values the system is supposed to attempt to find in the new case. This is an iterative process that ends up with a number of suggested values to assign in the new case. The user is asked to review these new values before adding this new case to the base.

If a similar value for theme is found, the user (the expert) is asked whether this new value can be added to the new case to be learned. As a result, we can reuse the knowledge embedded in the cases of the case base and besides, use the knowledge present in the new cases. This is possible because the content of the legal decisions are somehow interconnected. Hence, instead of trying to represent symbolically these interconnections, we infer on it making use of the context of the new cases.

5. Concluding Remarks

The motivation for the implementation introduced here is another way of making use of the well known advantage of case-based reasoners that bypasses explicit knowledge representation. The analogical reasoning behind the use of cases comprise indeed an effective approach when we can certify the inherent correspondence of concepts within case descriptions.

The definition of multi-purpose indexes is an alternative when one cannot comprise a value for an index in every case of the collection. Large case-based reasoners can make use of such alternative.

6. References

Ashley, Kevin D. (1990). Modeling Legal Argument: reasoning with cases and hypotheticals. A Bradford book. The MIT Press, Cambridge, Massachussetts.

Blair, D. C. and Maron, M.E. (1985). "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System". Communications of the ACM, 28, 3, 280-299.

Branting, L. Karl (1991). Reasoning with Portions of Precedents, Proceedings of the Third International Conference on Artificial Intelligence and Law, Oxford, England, June 25--28, 1991.

Branting, L. Karl (1991b). Building explanations from rules and structured cases. International Journal of Man-Machine Studies, 34, 797-837.

Leake, David. (1996). Case-Based Reasoning: Experiences, Lessons, & Future Directions. AAAI Press/The MIT Press, Menlo Park, California, 1996.

Rissland, E. L., & Skalak, D. B. (1991). "CABARET: Statutory Interpretation in a Hybrid Architecture." International Journal of Man-Machine Studies, June, 1991. (34):839-887.

Rissland, E.L., Skalak, D.B., & Friedman. (1993). M.T. "BANKXX: A Program to Generate Argument through Case-Based Search." Proceedings of the Fourth International Conference on AI and Law (ICAIL-93) , Amsterdam, The Netherlands, June 1993. pp. 117-124.

Rissland, E., Skalak, D. & Friedman, M. (1996) Using Heuristic Search to Retrieve Cases that Support Arguments. In Case-Based Reasoning: Experiences, Lessons, & Future Directions. David Leake, ed. AAAI Press/The MIT Press.

Weber-Lee, R.; Barcia, R.; Costa, M.; Rodrigues Filho, I.; Hoeschl, H. C.; Bueno, T.; Martins, A; & Pacheco, R. (1997a). A Large Case-Based Reasoner for Legal Cases. Lecture Notes in Artificial Intelligence: 2nd Int. Conference on CBR, ICCBR97. David Leake, Enric Plaza (ed.)-Berlin: Springer.

Weber, (1997) PRUDENTIA: Enabling a real world application of Case-Based Reasoning to jurisprudence research. Qualifying Examination approved by the Graduate Program of Production Engineering at the Federal University of Santa Catarina. In Activities. Available online http://www.eps.ufsc.br/~rosina/html/activities.html, [October, 1997].

Weber, Rosina; Barcia, Ricardo; Pacheco, Roberto; Martins, Alejandro; Hoeschl, Hugo; Bueno, Tania; Costa, Marcio; Rodrigues Filho, Ilson. (1997b). Representing Cases From Texts In Case-Based Reasoning. *III Congresso Internacional de Engenharia Industrial e XVII ENEGEP*, 6 a 9 de Outubro de 1997, Canela, RS, Brasil.