College of Information Science and Technology

REPRESENTING CASES FROM TEXTS IN CASE-BASED REASONING

Rosina Weber-Lee*

Ricardo Miranda Barcia*

Roberto C. Pacheco*

Alejandro Martins*

Hugo C. Hoeschl*

Tania C. D'agostini Bueno*

Marcio C. da Costa*

Ilson W. Rodrigues Filho*

*Universidade Federal de Santa Catarina

R.Mestro Aldo Krieger,118/403 - Florianópolis, SC, BRAZIL 88.037-500, - rolee@eps.ufsc.br

Abstract. *Case representation is a Case-Based Reasoning (CBR) problem area that refers to selecting proper descriptors to describe and index cases. The complexity of case representation has been preventing CBR systems from solving problems when large case bases are required. We present the development and implementation of a methodology to automatically convert legal texts into cases based on indexing methods and domain expert knowledge. The methodology is tailored to the domain of law although it can be extended to be applied to other domains as well.*

keywords*: case-based reasoning, legal domain, automatic modeling of cases*

1. INTRODUCTION

Cases are units that describe an experience with dimension-value pairs (descriptors). Some of these descriptors guide retrieval and they are named indexes. Others describe lessons and solutions to solve the input problem. Identifying what dimensions better represent a case and which ones should be used for indexing comprehend the case representation problem in developing a CBR system.

The complexities of case representation prevent CBR systems from launching in several domains, particularly when large case bases are required and when the knowledge available is written in text format. Very usually this is how companies keep records and it represents a demand for applications that can provide novel solutions to make good use of the knowledge embedded in these records. To comprise an effective case representation, it should be guided by the task of the system, that is solving the input case. Usefulness requires that the indexes retrieve cases envisioning the reasoner task, making case indexing very complex. Based on the importance of considering the system's task, case representation depends essentially on the domain of the application; therefore, this can be

only performed using expert knowledge of the domain. Hence, an automatic tool to select and assign values for dimensions requires knowledge acquisition from domain experts. The later conclusion demonstrates a strong necessity of knowledge engineering requirements that are usually claimed to be reduced in CBR systems. This increased need takes place when the knowledge available to construct cases is textual and its volume prevents this task from being performed manually.

We claim that representing cases in CBR systems in which the knowledge is present in large corpus of texts is constrained by the subtask of text reading and interpretation to choose proper descriptors and the indexing vocabulary. Once we have figured out how to overcome these encumbrances, we can extend the application of CBR systems to several domains, comprising real world problems. Therefore, the breakthrough we require is a timely means of converting texts into a case-like representation, that is an automatic approach to read texts and extract descriptors to describe and index the experience of these texts as cases. To achieve it, we propose a methodology that uses an expert system that reads texts using domain expert knowledge, assigns values to previously defined dimensions and extracts other dimensions from texts. This methodology has two steps: the first is its development and the second is its implementation.

With the proposed methodology, every organization that has records kept in machine readable texts is eligible to use a CBR system to reuse the knowledge embedded in these records, no matter how large they are. Thus, it becomes feasible the use of CBR systems to solve real world problems. It represents a means of modeling case bases for leveraging existing expertise as suggested by Klahr (1996). Hence, the importance of CBR applications is enhanced.

In the next section we discuss in more detail the importance of the expert knowledge in the modeling of case representation. The indexing problem, its methodologies and the concern with usefulness are presented in section 3. Then, we introduce the proposed methodology . We conclude with remarks and future works.

2. CASE REPRESENTATION AS A KNOWLEDGE-BASED TASK

According to Aamodt and Plaza (1994), “The challenge in CBR as elsewhere is to come up with methods that are suited for problem solving and learning in particular subject domains and for particular applications environments”. This sentence grounds the point that one can only develop a CBR system using expert knowledge of the application domain. Indeed, CBR systems avoid the knowledge engineering process the way it has to be dealt in the development of Expert Systems, for instance. This is why CBR researchers claim about the advantage of CBR systems over other symbolic AI techniques due to its reduced knowledge-engineering requirements. Hanney & Keane (1996) showed that such advantage does not hold in systems that require adaptation methods. Now, we shall demonstrate that the same holds for systems in which the case base requires a complex modeling. The fact that the knowledge is embedded in databases does not mean that the case base is modeled. The indexing of a case base is oriented by the usefulness of a case when performing the reasoner task, which differs essentially from the indexing in information retrieval and database systems.

Cases are represented through dimension-value pairs called *descriptors*. Selecting the proper descriptors characterizes case representation. These descriptors can describe the case, its solution, as well as other relevant information. Hence, the descriptors can have different functions, *i.e.*, describe the case, index the case to guide retrieval, describe solutions, and keep some important information about the outcome when the case is used

(Kolodner, 1993). Let us examine the influence of the expert's knowledge in modeling the descriptors with these three functions, describing the case, indexing the case and suggesting solutions to the case. All depends on the domain and task of the application and only the expert can point what and how to describe a case.

The indexing of a case resembles the indexing of books. Kolodner (1993.) points out that indexes should: (1) use the same vocabulary as the user; (2) anticitape the circumstances of the search; (3) use the concepts that are normally used in the domain; and (4) represent an interpretation of a situation. The vocabulary, the circumstances of a search, the concepts and the interpretation of the situations represent aspects of the domain knowledge; consequently, only an expert with deep understanding of this domain could perform the indexing task. The descriptors that present the solutions of the case have also to be designed by an expert because establishing what is a solution to a problem necessarily requires domain knowledge.

After describing the influence of the expertise over case representation, we conclude that its automatization is indeed an expert's task. Therefore, it is required to elicit from experts the proper knowledge and represent it in an intelligent (rule-based) system.

3. INDEXING AND USEFULNESS

The main characteristic of the retrieval in CBR systems is that it shall be guided by the usefulness of a given case in solving the input problem. This means that, when indexing, one has to envision the task of the system and the types of input problems that this given case might be able to solve. Determining descriptors that will be useful in solving a similar problem is not a trivial task. Kolodner (1993) defines the indexing vocabulary and the process of index assignment: identifying the indexing vocabulary consists of selecting dimensions that, when assigned, fulfill the desired functions of indexes. The latter process is the assignment of values to these dimensions.

Still according to Kolodner, indexes shall be predictive, abstract enough to be generic, and concrete enough to be recognizable and useful. A useful index carries out the purpose that may be either related to the solution of the input case, to some failure or to a result. For the selection of the indexing vocabulary two approaches are proposed, the *reminding* and the *functional* approach.

The *reminding* approach is very intuitive as it is the approach naturally used by human experts and knowledge engineers when asked to select an index: it searches for the issues that are brought up by experts of the domain when solving problems. The reminding approach has been applied in our system to designate dimensions before executing the automatic system.

The *functional* approach envisions the cases regarding three matters: (1) what dimensions cover the tasks intended to be carried out by the reasoner; (2) which cases provide different values for the dimensions; and (3) what levels of detail are necessary (now and foreseeing an expansion).

In extracting dimensions, we use the functional approach and the reminding approach, even though the dimensions we extract will not always be used as indexes. Some of them will be filled with expressions that we cannot expect to be automatically compared with other sentences in other legal texts. This is because we are trying to get the most out of the texts regarding the number of lessons that will be presented to the user. However, we cannot guarantee that all these lessons will be guiding the retrieval. Above all, many times lessons are the solutions, although we are using lessons to guide the retrieval too.

In the search for dimensions, Kolodner points out to the importance of identifying lessons the cases teach and the context they happen. This hint represented a major contribution to the development of our approach. Within a specific domain of knowledge, experts can indicate what situations are considered lessons. Besides, within a specific domain of speech (or writing) the experts can identify expressions that indicate lessons.

4. FROM LEGAL TEXTS TO LEGAL CASES

In this section we present our methodology and illustrate it within the domain we are testing it. The domain of our system is the State Court of Justice (SCJ) of Santa Catarina, Brazil. All legal cases are described by an official reporter, henceforth referred to as legal texts. The legal texts are either Civil or Criminal cases. This primary classification heads the "tree of categories" of cases of the court. This tree has branches that lead cases such as adoption, murder or larceny. The legal cases of this Court since 1990 are being typed and they amount to about 90,000 legal texts.

The proposed methodology converts these texts into cases deploying indexing methods. From the knowledge acquisition with domain experts, results the indexing vocabulary and the index assignment is performed automatically through rules. The knowledge elicitation is not the same for texts from the Civil and Criminal areas although several rules are reused.

The legal cases of the Court are the description of petitions that comprehend parts of lawsuits. All these texts are written by judges who work for the Court and they have very similar backgrounds. Besides, there are some rules they are supposed to follow when writing these texts such as mentioning about the result of the petition in a paragraph right after describing what the parts are and which part applies for the petition. According to the self-explaining documents framework proposed by Branting and Lester (1996), the knowledge of the illocutionary and rhetorical structures of complex documents can be used for indexing. The knowledge acquisition step allowed us to come up with a rhetorical structure of these texts as well as the identification of the moments in the texts where illocutionary expressions orient the identification of relevant dimensions.

The result of this approach is the case base of a retrieval-only CBR system. The input of this system is a legal situation. The output is a set of legal cases that share similarities with the input case and bring useful lessons to the input case. The main reasoning of the system is the search for the most similar legal cases. The reasoner's task is to retrieve the most useful cases to suggest solutions to the input problem. Within the domain of law, queries are made by judicial professionals who are able to understand and use the suggested lessons. The system we are discussing does not create arguments, but retrieve the most useful cases to help solving an input problem. It plays the role of a decision support system in the legal domain.

The methodology comprehends text analysis, definition and assignment of fixed surface features and dimensions, and extraction of dimensions. These are all knowledge-based steps and we have been working at two levels; firstly, the development level refers to knowledge acquisition to design the methodology; and, secondly, the implementation of the designed system to perform the steps developed.

4.1 Text Analysis

The analysis of the legal texts aims at identifying the rhetorical structure of the documents. In the development of text analysis, experts indicate what structures to recognize in the texts and how to identify them. The final goal is to identify recognizable parts in the text, that is where to find relevant information. Experts who write these legal cases follow some rules making this task easier. The development would require revision only if there is any change in the structure of the texts. Otherwise it would be only necessary to be repeated when a new domain is chosen.

We have performed sample tests to ensure that each substructure is actually present in every legal text. The rhetorical structure of the legal texts in our example is presented in Table 1.

Identification: surface features such as date, city, reporter and petition type.
:
Abstract: varies in its length, starts after the end of the *identification* and ends with two paragraphs, the first indicates the applicant and the second presents the result.
:
:
Body: This is where the search for illocutionary expressions takes place. Upper paragraphs describe details of the situation, indicating the laws that categorize the subject, and points to foundations.
: in its conclusion it is usually the court decision and its foundations.
Closing: starts with one paragraph about votes followed by date, place and names of participating attorneys.

Table 1: Rhetorical Structure of Legal Texts.

The implementation of texts analysis is employed through rules in a logic programming module using Natural Language Processing (NLP) techniques. The module receives the text as input and outputs the content of every structure and substructure that will be the input in the following steps of implementation.

4.2 Definition and Assignment of Fixed Surface Features and Dimensions: Some Results

The development of this second step started with the knowledge elicitation from domain experts who have defined a small set of fixed attributes to describe all legal texts. Experts expect them to be valued in all cases. It is important to point out that the definition of the attributes do not require the experts to examine a significant amount of texts. Their capability of pointing out these attributes relies on their expert knowledge of the domain. Next, experts were asked to point the substructure where the value of each attribute is informed.

The knowledge acquisition process elicits from experts how the values appear in the texts. Rules were developed to be applied on each substructure to extract values for the

attributes. The resulting rules are programmed in with NLP techniques in a module reads the proper substructure and assigns values to the attributes.

One of the fixed dimensions is category. The assignment of this value requires the use of NLP techniques because the category is not written clearly in a specified part of the text as the feature reporter is. However all possible values to be assigned to the dimension category are available in a "tree of categories", making the assignment easier. One type of assignment is the one when the values are limited to a list according to some other value already assigned. In the assignment of the dimension outcome there are different limited lists that are chosen depending on the value assigned to the feature appeal. The expressions used to present the result may be refute, impugn, sanction or accept for one type of appeal whereas for another, the expressions used may be traverse, concede or disclaim.

Still under the development of the assignment phase, there is the rule validation. For instance, to test the rule set oriented to extract the result of habeas corpus cases, we have gathered 684 texts – referring to all cases of this type from 1990 to 1996. The first rule set stemmed from a sample of 17 texts. Applying this rule set on the 684 texts, generated a 63% rate (434) of proper assignments. Two new rules were added and the execution of this new rule set resulted in 678 assignments. Out of the 6 cases left without values, 5 of those referred to cases where no result has been decided – the court relegated the decision to another court; only one (1) case provided no information about the decision in the proper substructure. We consider this 99% (678 out of 684) good enough.

The implementation of the assignment phase can be performed since all rules are tested. In this phase, the rule-based system receives the texts and assigns values for all surface features and dimensions.

4.3 Extraction of Dimensions

During the early stages of knowledge acquisition, experts pointed some expressions that indicate that there is either an illocutionary statement or a lesson. This has motivated us to add another process to the methodology. This step refers to the modeling of rules that enable the system to automatically search for new dimensions using indicative expressions. *Indicative expressions* were pointed by the experts after an analysis of the samples of the text. Two examples of *indicative expressions* that the rule-based system searches for are the noun "impossibility" and the verb "certify". The experts provided knowledge with which the knowledge engineers could design heuristics to be deployed by the system in defining dimensions from each *indicative expression* found. An heuristic for the noun "impossibility" is based upon the idea that "impossibility indicates the condition of doing the action represented by the main verb in the sentence where impossibility is used."

When experts were asked about how to use these expressions, they suggested heuristics. One example is the noun *impossibility*. According to experts, nouns derived from adjectives indicate the presence of a lesson. Suppose the legal case reads, "…the impossibility of penalizing the defendant stems from the fact the defendant is under legal age and therefore his imprisonment constitutes a misfeasance…". This sentence clearly teaches the lesson that the defendant who is under age cannot be kept imprisoned. The sentence following the expression impossibility will usually inform about an illegal fact, whereas the sentence following therefore can either inform an illocutionary expression or expose reasons for the assertions, *i.e.*, reveal the grounds for such impossibility. From this fact we can determine another dimension concerned to the grounds of the condition. Hence, the dimensions extracted from this first example would be: *penalizing condition* and *penalizing condition grounds*; and the values would be respectively *impossible* and

defendant is under legal age (for more details see *A Large Case-Based Reasoner for Legal Cases*, Weber-Lee et al, 1997).

The implementation of this phase is employed by the module that searches for these expressions applying the proper heuristics whenever an expression is found. As explained above, the result of the rule set is the new dimension and its value.

5. CONCLUDING REMARKS

Figuring out that representing cases is an expert's task made possible the development of a knowledge-based approach to build cases from texts. We have illustrated our approach in a CBR system applied to the complex domain of Law. As a result of this methodology, we expect to overcome the difficulty in modeling cases that has been avoiding the launching of CBR systems in many real world applications, particularly where the knowledge available is in large corpus of texts.

The CBR literature provided us with two very important clues. One was from Kolodner's approach to search for lessons that cases teach and the context they happen. The other was the document representation proposed by Branting and Lester (1996) that called our attention to the structure of the texts and to look deeper at the significance of goals and expressions in the text. These ideas and the understanding that only expert domain knowledge orients to the proper selection of descriptors made the automatization possible. The result is an intelligent system that converts texts into cases within a specific domain of knowledge.

An important issue to be addressed regards to the reusability of the knowledge engineering effort expended in such an application. The reuse of the knowledge acquisition is not possible in different domains although the architecture of implementation is. One may benefit in other domain from the knowledge about what to elicit from the experts; that is what the indications of lessons in these texts are.

Comments are provided to partial results due to the early stage of the present work. Further results will be available in the next months, along with the presentation of the subsequent developments.

6. REFERENCES

Aamodt, A. & Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *Artificial Intelligence Communications,* 7 (1), 39-59.

Branting, L. Karl & Lester, James C. (1996) Justification Structures for Document Reuse *Advances in Case-Based Reasoning: third European Workshop*; proceedings/ EWCBR-96, Lausanne,Switzerland, November 14-16, 1996. Ian Smith; Boi Faltings (ed.)-Berlin; Springer,1996.

Hanney, Kathleen & Keane, Mark (1996). Learning Adaptation Rules From a Case-Base. *Advances in Case-Based Reasoning: third European Workshop*; proceedings/ EWCBR-96, Lausanne,Switzerland, November 14-16, 1996. Ian Smith; Boi Faltings (ed.)-Berlin; Springer,1996.

Klahr, Philip (1996). Global Case-Base Development and Deployment. *Advances in Case-Based Reasoning: third European Workshop*; proceedings/ EWCBR-96, Lausanne,Switzerland, November 14-16, 1996. Ian Smith; Boi Faltings (ed.)-Berlin; Springer,1996.

Daniels, J. J. and Rissland, E. L. (1995). A Case-Based Approach to Intelligent Information Retrieval. Proceedings of the *SIGIR '95 Conference* SIGIR '95 Seattle WA USA 1995 ACM.

Kolodner, J. (1993). *Case-Based Reasoning*. Morgan Kaufmann, Los Altos, CA.

Lehnert, Wendy (1991). A Performance Evaluation of Text Analysis Technologies. *AI Magazine*, Fall, 81-94.

Ram, Ashwin. (1991). Interest-Based Information Filtering And Extraction In Natural Language Understanding Systems. *Bellcore Workshop on High-Performance Information Filtering*, Morristown, NJ, November, 1991.

Riloff, Ellen (1993). Automatically Constructing a Dictionary for Information Extraction Tasks. *Proceedings of the Eleventh National Conference on Artificial Intelligence*, 1993, AAAI Press/MIT Press, 811-816.

Smeaton, Alan F. (1995). Low Level Language Processing for Large Scale Information Retrieval: What Techniques Actually Work. *Proceedings of a workshop "Terminology, Information Retrieval and Linguistics", CNR*, Rome, October, 1995.

Uyttendaele, Caroline, Moens, Marie-Francine & Dumortier, Jos. (1996). SALOMON: Automatic Abstracting of Legal Cases for Effective Access to Court Decisions. (*JURIX 1996)*

Masand, B.; Linoff, G. & Waltz, David L. (1992). Classifying News Stories using Memory Based Reasoning. *Proceedings of the SIGIR*, Copenhagen, 59-65.

Waltz, David L. & Pollack, Jordan B. (1985). Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation. *Cognitive Science*, 9, 51-74.

Weber-Lee, Rosina; Barcia, Ricardo M.; Costa, Marcio C.; Hoeschl, Hugo C.; Bueno, Tania D.; Rodrigues Filho, Ilson W.; Martins, Alejandro; Pacheco, Roberto. (1997). A Large Case-Based Reasoner for Legal Cases. *Submitted to the ICCBR97*.