IJCAI-99 Workshop ML-5: Automating the Construction of Case-Based Reasoners, Stockholm 1999, S.S. Anand, A. Aamodt, D.W. Aha (eds.). pp 77-82

Learning Retrieval Knowledge from Data

Helge Langseth[1], Agnar Aamodt[2], Ole Martin Winnem[3]

[1]Norwegian University of Science and
Technology, Department of
Mathematical Sciences
N-7034 Trondheim, Norway
Helge.Langseth@stat.ntnu.no

[2]Norwegian University of Science and
Technology, Department of Computer
and Information Science
N-7034 Trondheim, Norway
Agnar.Aamodt@idi.ntnu.no

[3]Sintef Telecom and Informatics
N-7034 Trondheim, Norway
Ole.M.Winnem@informatics.sintef.no

Abstract

A challenge of future knowledge management and decision support systems is to combine the storage and effective reuse of *data,* systematically captured as process or system information, with *user experience* in dealing with problems and non-trivial situations. In CBR, situation-specific user experiences are typically captured in *cases*. In our approach, cases are linked within a semantic network of more general domain knowledge. In this paper we present a way to automate the construction and dynamical refinement of such a model of case-specific and general knowledge, on the basis of external process data continuously being generated. A data mining method based on a Bayesian Networks approach is used. We are also looking into how the notion of causality, being a central issue in both BNs and model-based AI, can be compared and better understood by relating it to such a combined model.

1. Background and motivation

Our research is conducted within the subarea of knowledge-intensive case-based reasoning, i.e. the Creek approach (Aamodt, 1995; Grimnes & Aamodt, 1996). Within this approach we are currently studying and experimenting with statistical data mining methods, primarily Bayesian Networks (Jensen, 1996; Aamodt & Langseth, 1998). This is a means to automate the construction of a case-base or its supporting background knowledge, on the basis of data dynamically generated from processes and activities that are part of the task domain. Example processes and activities are industrial production processes, problem solving operations, maintenance actions, planning activities, etc. We are in the process of studying and experimentally comparing various approaches to this integration, within the domain of petroleum engineering – more specifically oil well drilling - in cooperation with the Norwegian oil company Saga. Some initial results are described in this paper.

The motivation for the work reported here is two-fold, coming from the method side and the application side, respectively. At the method side there is a need for improved methods to dynamically modify and adapt the supporting general domain knowledge of knowledge-intensive CBR. So far, the Creek approach has been to learn by storing cases and linking them to the general domain knowledge, which in turn has been assumed static – or only subject to occasional manual updating. Since a major role of the general domain knowledge is to produce explanations to support and justify various CBR reasoning steps (two different approaches are described in (Sørum and Aamodt, 1999) and (Friese, 1999)), it is crucial that this knowledge is as updated as possible, always reflecting the current state of domain knowledge related to the task reality. In well-understood and static domains, this would introduce no problem, but since we are dealing with complex tasks within open-textured and changing domains; a static knowledge model will soon degrade and become less useful.

The other motivation comes from the primary type of application targeted by our methods, which is interactive intelligent systems for knowledge management, decision support, and learning support. Here we see a clear need to better combine the implicit ‘experience’ stored as data in databases with the more user-oriented experience that may be captured as cases. This is elaborated in the following section.

Our research is done within the scope of the Noemie EU project (Aamodt et. al., 1998). Here data mining and CBR are combined in order to improve the transfer and reuse of industrial experience. The aim of the project is to develop methods that utilize the two techniques in a combined way for decision support and for targeted information focusing over multiple databases. Application problems dealing with technical maintenance and tool design, and the prevention of unwanted events, are addressed. The domain of the research reported in this paper is diagnosis and repair related to the loss of drilling fluid into a geological formation during drilling (the so-called “lost circulation” problem).

2. User and Data Views

Target systems for our methods are interactive systems aimed to support people in their daily job activities, by storing potentially relevant information and data, and capturing or deriving valuable knowledge, in order to make this easily available for later reuse and elaboration. People involved in this type of decision making and information/knowledge management today typically use computers, at least to some extent. In such companies large amounts of data are captured and stored on a routine basis, but often not in a form that make them useful for work support.

This growing store of data can be said to represent a certain view or slice of a real world description (sometimes referred to as the 'task reality'), determined by the type of data and the values registered. During oil well drilling, for example, a lot of data is continuously registered that describe state parameters such as bore hole pressure, fluid flow rate, lithology of the geological formation, operations being performed, drilling personnel involved, etc. The type and value of the data registered then represent a certain perspective or view to the reality being dealt with. Another view to this part of the real world is captured by the experiences that people gather as part of their daily information handling and problem solving effort. For example, whether a drilling process runs smoothly or has problems, what the actions available to deal with a critical situation are, and what competence people involved in an operation have or should have.

Essentially, then, in computer-assisted environments, the information about the task reality captured in databases and the understanding of the phenomena by the people in job situations represent two complementary 'views' to a task reality, as illustrated in Figure 1. A part of the two views, i.e. a part of the descriptors or submodels representing the two views, may be shared, other parts not. Note that the data bases pictured in the lower right of Figure 1 are standard company DBs, and different from, e.g. data bases storing experience cases or other knowledge bases. In the following section we will elaborate on this distinction between data and cases.

Looking at things in this way opens up for studying how the two views can form a basis for integrated decision support systems where user experience and information from data are synergistically combined.

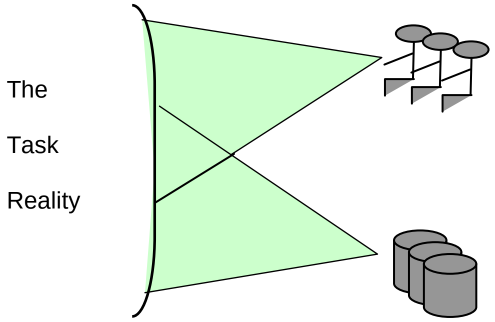

Figure 1: User and Data views of a part of the real world.

3. Data vs. Cases

We are studying how data mining methods may contribute to the construction of CBR systems on the basis of the two-view perspective outlined in the last section. As previously mentioned, the notion of data, as in the 'data view' reflects data of processes, state parameters, etc. as stored in standard company databases. Hence the notion of data in this sense does not include knowledge bases, containing cases or more general domain knowledge. This means that our view of a case is a user-oriented view, i.e. a case stores a past user experience. This is different from the view that a case is simply a data record. This latter view is adopted by some other CBR researchers, particularly those focusing on 'instance-based' methods, characterized by large case bases, simple case structures, and little if any background knowledge. The user-oriented case view, on the other hand, is characterized by fewer cases, larger and more complex case structures, and usually a significant portion of general domain knowledge to support the CBR processes. A clear distinction of the case vs. data issue is necessary in order not to confuse the mutual roles of DM and CBR methods in integrated systems.

4. Model representation

As stated, the topic of our research is to investigate how the construction of knowledge-intensive CBR systems may be automated by updating the general domain model on the basis of data from company data bases. Within Creek, general domain knowledge is represented in a frame-based system, where the frames constitute a densely coupled semantic network. Domain entities as well as relations are first class concepts, each represented in their own frame. Of the various candidate methods from the machine learning field that could be applicable for learning in this model, we have picked Bayesian networks as our initial method of investigation. There are several reasons for that. One is that the network structure of BNs has similarities with a semantic network structure, although there are significant differences (see next section). This is an important motivation, since the explanation-driven approach of Creek facilitates combined explanations coming from both type of networks, in an integrated way. Another is that statistical learning through data mining nicely complements the manually generated domain model. A third is that while we now are studying learning of general domain knowledge, we will in the future also investigate the automated re-construction of past cases (i.e. user experiences) from data. Here the BN model also provides possible solutions. However, once the BN method is implemented and tested, it will be interesting to study other DM/ML methods for this purpose.

5. Semantics of relations and links

Motivated by interesting results on network learning (Heckerman et. al. 1995), we are using a Bayesian method to generate a network structure from data, and use this either as a substitute or in cooperation with a user-generated semantic network. Several researchers have investigated different facets of this task. (Friedman,

1998) presents a method to learn BN structure when the data is prone to missing features. (Friedman and Goldszmidt 1997) offers a sequential method for structure refinement. (Koller & Pfeiffer, 1998) follow another path, as they extend the basic BN to a frame-based system. Hence, they are able to handle uncertain information in a structure that enlarges the expressive power of the graphical model. This construction raises hope that more complex structures than plain BNs can be extracted from data.

Given that search structures may be learned, we are especially concerned about the level of integration between this construction and the semantic network. To integrate the two types of domain models at any level, we must be assured that the semantics of the two models, as seen from that particular level of integration, can be inter-related.

Unfortunately, not all kinds of relations are simply learned from data. In fact, arcs in a BN are just carriers of statistic correlation, and it is – strictly speaking - the absence of an arc that can be given a semantic meaning. The BN semantics is defined by the joint statistical distribution function that it encodes, together with the conditional independencies that can be read directly from the graphical structure. However, it has been somewhat common to regard the arcs in a BN as a kind of "generalized causality". This definition is more loose than that traditionally used in AI, and is often defined as "A causes B if an atomic intervention of node A changes the probability distribution over node B". Important research has focused on whether such 'causality' can be learned from empirical data, (see, e.g., (Pearl, 1995)) for the foremost example. Pearl's conclusion was negative. For a two–node network of correlated nodes, for instance, it is not possible to infer which of the two nodes that is the cause and which is the effect by only using empirical data. The direction of the arc between them can be changed without altering the semantics of the Bayesian network. It seems counter–intuitive to call such arcs 'causal' in any way. Instead of labeling all arcs as 'causal', one can use algorithms like *Inferred Causation* (Pearl & Verma, 1991) to specifically test each arc in the network. This algorithm takes an estimated probability distribution as input, and returns an annotated graphical model in which a subset of the arcs is marked 'causal'. These arcs are exactly those, whose direction can not be changed without altering the BN semantics. (Neopolitan et. al., 1997) reports experiments which show that small children tend to investigate and learn causality in a way that supports the psychological plausibility of Pearl and Verma's algorithm.

From our work so far, we are reluctant to giving each arc in a BN a clear semantic meaning related to the semantic network relations. Therefore, it is not intuitively feasible to integrate the BN and the semantic network at the lowest level (i.e. the level of the meaning of single relations). However, when care is taken, i.e. a right suitable level of interpretation is found, we should be able to let the two domain models co-operate in a semantically meaningful way. For example, at the level of explanatory strength of a relation (semantic network notion) and, correspondingly, degree of belief (BN notion), the semantic mapping is easier. More research is needed to find an optimal level of integration.

6. Learning retrieval knowledge

At present, we regard the BN as a submodel of statistical relationships, which lives its own life in parallel with the semantic net. The BN generated submodel is dynamic in nature; i.e. we will continuously update the strengths of the dependencies as new data are seen. In this way, the system will be able to improve its ability to retrieve the best matching case given the input. The dynamic model suffers from its less complete structure (we will only include a term in the BN if it is linked via an influence-relation such as causes, indicates, etc.) but has an advantage through its sound statistic foundation and its dynamic nature. Hence, we view the domain model as an integration of two parts, a "static" and a "dynamic" one. The first consists of relations assumed not – or seldom - to change (like has-subclass, has-component, has-subprocess, has-function, always-causes, etc). The latter part is made up of dependencies of a stochastic nature. In changing environments, the strengths of these relations are expected to change over time.

The BN indexes its cases in a way quite different from how it is done in Creek. Cases are leaf nodes (i.e. they have no children), and they are sparsely connected to the case features. In Creek, a case frame is connected to the frames of all its features. In the BN on the other hand, effort is taken to minimize the number of arcs pointing to a case node. The BN inference mechanism works just as easily over long paths of influence as it does on a one-step path, hence the direct remindings are not necessary. This difference is illustrated in Figure 2.

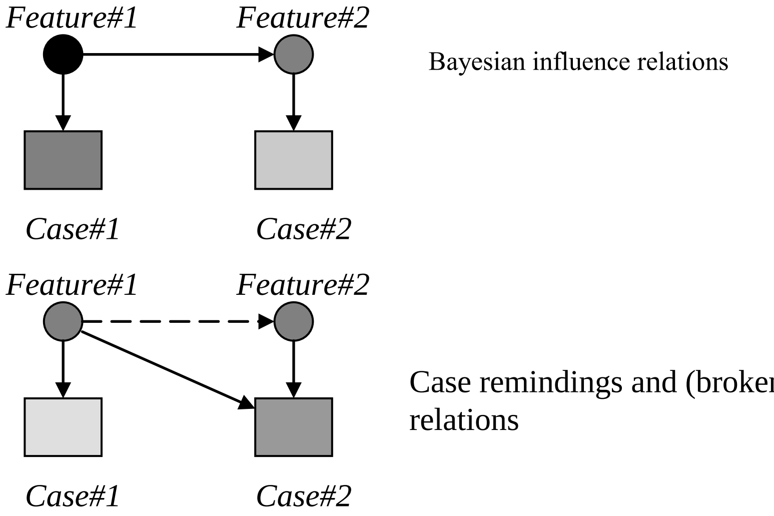

Figure 2: Case indexing in Bayesian and semantic networks.

Each case is indexed by a binary feature link (ON or OFF, with probability). The standard Creek process of choosing index features is adopted, taking both the predictive strength and necessity of a feature into account.

As seen in the top of Figure 2, the BN does not index Case#2 directly from Feature#1, since the information flow from Feature#1 through Feature#2 already indicates Feature#1's influence over Case#2. In the semantic net, however, both features are remindings to Case #2. If Feature#1 is observed, both Case#1 and Case#2 are affected in the BN according to the strength of the path from Feature#1 to the respective case. If Feature#2 is then observed, Feature#1 is no longer influencing the relevance of Case#2, since Feature#1 is independent of Case#2 conditioned on Feature#2. In the semantic network, however, conditional independence does not come to play. When both features are observed, both the cases are affected. Case#2, having 2 remindings, is likely to be more strongly reminded, but this depends on the strength of the individual remindings. The case with the strongest combined reminding will be selected as first choice.

Calculations within a BN are performed using a compiled structure referred to as a *junction tree*. This is basically a tree structured graphoid where the nodes are the cliques in the BN, i.e. the maximally connected subgraphs of an undirected version of the BN, see (Jensen, 1996) for details. Both the size and complexity of the compiled structure is depending on how densely connected the BN is. If the BN is very densely connected, the cliques grow larger, which will increase the computational costs of the BN inference. To avoid escalating memory requirements, arcs that are not necessary to link a case to its features are removed from the BN, resulting in a simpler structure as illustrated in Figure 2. We also employ a particular spreading activation algorithm (van de Stadt, 1995) to compile only those parts of the BN which are required for a given inference task, reducing the size of the memory required for the BN structures.

7. Experimental evaluation

In this section we describe some initial results of the experimental evaluation of our method. In the experiment, we started off with a reasonably well elaborated semantic network describing the "lost circulation problem" of oil well drilling. The semantic network consisted of 2434 relationships between a total of 1254 entities. The case-base consisted of 45 cases, which captured the whole recorded history of lost circulation incidents in the oil company.

As a starting point for the BN construction, we used a subset of the semantic network. We extracted all relationships which could be regarded as describing generalized causality, i.e. the relations causes, has-consequence, enables, involves, occurs-with and indicates, together with the nodes on either side of these relations. This resulted in a BN consisting of 128 nodes and 146 links. Simple statistical formulas were used to generate the local probability tables of the BNs from the strength of the relations. Afterwards, the complete case-base was indexed by the BN. The mean number of links to a case (average number of remindings) was 4.0 in the BN compared to 44.9 in the semantic network. The semantic network uses 55 different relations, in the BN we only have one. These numbers indicate that the BN is only reflecting a small part of this task reality, compared to the broader scope of the semantic network.

Because of very strict confidentiality of the data for this domain, we could only access a small part of the total set of databases that are intended to be used in the final application for the company. The reduced data material made learning of the BNs network structure unfeasible, so we where not able to update the structure of the domain model through data mining. We were, however, able to fine-tune the parameters in the model, using an algorithm by (Binder et. al., 1997).

Below, the two screen excerpts of Figure 3and Figure 4 illustrate how an example case (Case-16) is indexed in the general domain model. Figure 3 indicates the sparsely connected structure of the BN, while Figure 4 shows that a case is more densely linked within a semantic network – corresponding to a more complex case structure than what is employed by the BN method. In the semantic network we find that both Induced-Fracture-Lc and Tripping-In are remindings to Case#16. From the general domain model (not shown) we know that Tripping-In causes Large-ECD causes Very-Small-Leak-Off/Mw-Margin-<0.02kg/L causes Induced-Fracture-Lc. Interpreting Bayesian inference as a kind of causal inference, it is not necessary to link Induced-Fracture-Lc directly to Case-16 in the BN model.

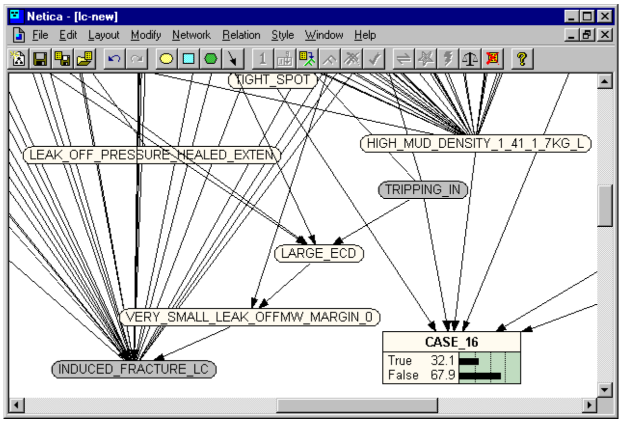
Netica - [lc-new]
File Edit Layout Modify Network Relation Style Window Help
TIGHT_SPOT
HIGH_MUD_DENSITY_1_41_1_7KG_L
LEAK_OFF_PRESSURE_HEALED_EXTEN
TRIPPING_IN
LARGE_ECD
VERY_SMALL_LEAK_OFFMW_MARGIN_0
INDUCED_FRACTURE_LC
CASE_16
True 32.1
False 67.9

Figure 3: Bayesian Model. Grey nodes are activated; white nodes are not. Current belief in Case#16 is 32.1%.

To look further into the behavior of the two domain models we have designed an experimental setup, where each of the two domain models retrieves cases separately, and the results are compared. As a measure for the success of a retrieval method, we use the difference in calculated similarities; i.e. we assess both the systems ability to give high score to the similar cases as well as to give the poor matches a low score.

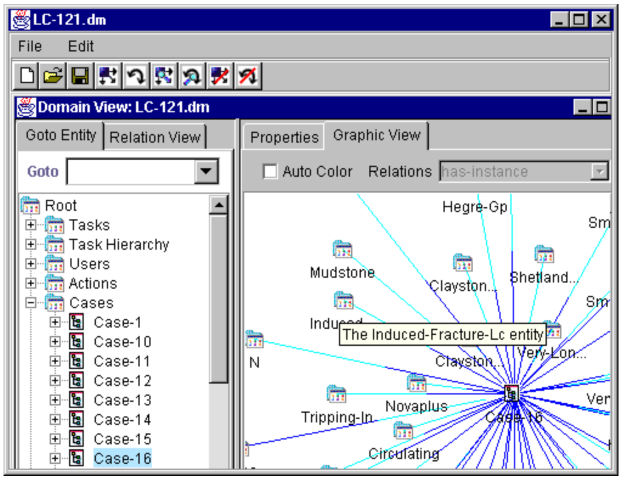

Figure 4: Semantic Network Model. Shows the features pointing to Case-16. Relation names and feature values are not shown.

In the Appendix the main content of Case-16 is shown. In the initial experiment a subset of this case was entered as the "new case", in order to compare how the to methods behaved on a simple, controlled retrieval task. As expected, both systems retrieved Case-16 as their best choice. On the second best choice there was a difference, however. The BN tends to give higher values of belief to cases than the semantic network-based retrieval does. The most prominent reason for this is that the domain expert has given stronger reminding strengths than what is justified by the data. Nevertheless, the BN-based system is capable of recognizing both a poor match as well as a good one.

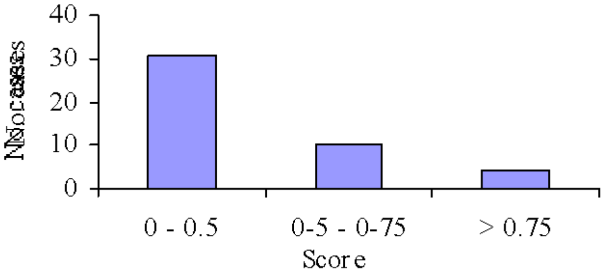

Figure 5: Histogram of the belief that the BN gives the cases during retrieve

A histogram showing the distribution over the cases of the degree of belief in the retrieved case the over the cases is shown in Figure 5.

8. Conclusions and future research

Initial research on the use of BNs to learn retrieval knowledge from data has been described. The retrieval knowledge is learned by updating a general domain model used to generate explanations in knowledge-intensive CBR. We are currently in a phase where we compare the abilities of the two different network models, both regarding retrieval and retain. It should be clear that both models have strong and weaker sides, and continued experimentation is needed in order to understand how they best should be combined into an integrated model. Future research should also include comparative studies of other machine learning methods for the purpose of updating the general domain knowledge as well as (re-)constructing experience cases from company data bases. The two views introduced early in the paper, the data and user views, has already shown to be a fruitful model for discussing possible ways of automating the construction of knowledge-intensive CBR systems.

Acknowledgements

Several members of the Noemie project team contributed to the results of this paper. Pål Skalle, Jostein Sveen, and Trond Gravem built the domain model, collected the cases, and Truls Paulsen contributed to the modeling editor and user interface.

References

Aamodt, A. (1995): Knowledge Acquisition and Learning from Experience - The Role of Case-Specific Knowledge, In Gheorge Tecuci and Yves Kodratoff (Eds.): Machine learning and knowledge acquisition; Integrated approaches, (Chapter 8), Academic Press, pp. 197-245.

Aamodt A., H. A. Sandtorv, and O. M. Winnem (1998): Combining Case Based Reasoning and Data Mining - A way of revealing and reusing RAMS experience. In Lydersen, Hansen, Sandtorv (Eds.), Safety and Reliability; Proceedings of ESREL '98, Trondheim, June 16-19, 1998. Balkema, Rotterdam. ISBN 90-5410-966-1. pp. 1345-1351.

Aamodt, A and Langseth, H (1998): Integrating Bayesian Networks into Knowledge-Intensive CBR. Proceedings from AAAI-98 Workshop on CBR Integration, Madison, pp. 1-6.

Binder, J, D. Koller, S. Russell and K. Kanazawa (1997): "Adaptive Probabilistic Networks with Hidden Variables" *Machine Learning* 29: 213-244.

Friedman, N. (1998): The Bayesian Structural EM Algorithm. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI), pp. 129-138.

Friedman, N. and M. Goldszmidt (1997): Sequential update of Bayesian network structure. In Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI), pp.165-174.

Friese, T. (1999): Utilization of Bayesian Belief Networks for Explanation-Driven CBR. NTNU, Trondheim. Unpublished paper (submitted to this workshop).

Heckerman, D., D. Geiger and M. Chickering (1995): Learning Bayesian networks, the combination of knowledge and statistical data, *Machine Learning* 20: 197-243.

Grimnes M. and A. Aamodt (1996): "A two layer case-based reasoning architecture for medical image understanding" in "Advances in Case-Based Reasoning, Third European Workshop, EWCBR-96" by Smith and Faltings (eds.), pp. 164-178, Springer, ISBN 3-540-61955-0, 1996

Jensen, F.V. (1996): An introduction to Bayesian Networks UCL Press, London.

Koller, D. and A. Pfeffer (1998): Probabilistic Frame-based Systems, Proceedings of AAAI-98, pp. 580-586.

Neopolitan, R. E., S. Morris and D. Cork (1997): Cognitive Processing of Causal Knowledge. In Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI), pp. 384-391

Pearl, J. and T.S. Verma (1991): A Theory of Inferred Causation. In J.A Allen, R. Fikes, and E. Sandewall (Eds.), Principles of Knowledge Representation and Reasoning: Proceeding of the Second International Conference, San Mateo, CA: Morgan Kaufmann, pp. 441-452.

Pearl, J. (1995): "Causal Diagrams for Empirical Research" *Biometrika*, Vol. 82, No. 4, pp. 669-709.

van de Stadt, E. C. (1995): Problem-directed decomposition of Bayesian belief networks. Ph.D. Thesis, Technische Universiteit Delft, 1995. ISBN: 90-900852-9-7

Sørmo, F. and A. Aamodt (1999): Improving CBR through Knowledge Elaboration. IJCAI-99 Workshop on Automating the Construction of CBR Systems, Stockholm, August 1999.

APPENDIX

Below the main contents of Case-16 is shown. Platform identification data has been removed for neutralisation reasons.

```
case-16
  instance-of                 value   case
  has-activity                value   tripping-in circulating
  has-geological-formation    value   shetland-gp cromer-knoll-gp hegre-gp claystone-with-dolomitestringe
                                      claystone-with-limestone-stringers sandstone mudstone
  has-depth-of-occurrence     value   5318
  has-country-location        value   n
  has-task                    value   solve-lc-problem
  has-observable-parameter    value   high-pump-pressure high-mud-density-1.41-1.7kg/l
                                      high-viscosity-30-40cp normal-yield-point-10-30-lb/100ft2
                                      large-final-pit-volume-loss->100m3 long-lc-repair-time->15h
                                      low-pump-rate low-running-in-speed-<2m/s complete-initial-loss
                                      decreasing-loss-when-pump-off very-depleted-reservoir->0.3kg/l
                                      tight-spot high-mud-solids-content->20%
                                      small-annular-hydraulic-diameter-2-4in
                                      small-leak-off/mw-margin-0.021-0.050kg/l
                                      very-long-stands-still-time->2h
  has-well-section-position   value   in-reservoir-section
  has-drilling-fluid          value   novaplus
  has-failure                 value   induced-fracture-lc
  has-outcome                 value   squeeze-job-acceptable
  has-well-section            value   8.5-inch-hole
  has-repair-activity         value   pooh-to-casing-shoe waited-<1h increased-pump-rate-stepwise
                                      lost-circulation-again pumped-numerous-lcm-pills
                                      no-return-obtained set-and-squeezed-balanced-cement-plug
  has-operators-explanation   value   "we tripped in and lost circulation.the mud was unstable and barite
                                       settled probly out and tended to pack around bha. we also know that
                                       depletion lowers fracture resistance and this combined is sufficient
                                       to explain the losses. we also probably crossed faults"
```