

## Combining Case Based Reasoning and Data Mining - A way of revealing and reusing RAMS experience

A. Aamodt

*NTNU/SINTEF, Dep. of Computer and Information Science, Trondheim, Norway*

H. A. Sandtorv

*SINTEF Industrial Management, Safety and Reliability, Trondheim, Norway*

O. M. Winnem

*SINTEF Telecom and Informatics, Trondheim, Norway*

**ABSTRACT:** The reuse of previous experience within safety, reliability and maintainability (RAMS) is an important input to complex decision making tasks. Statistical data are vital for many RAMS type of analyses, but frequently the investigation of a more limited number of past cases is preferable, enabling the reuse of relevant and concrete experience in dealing with a new task. In order to effectively utilise the increasing amount of information available in computerised databases, a combination of quantitative and qualitative methods are therefore called for. This paper describes an approach where statistical methods for extracting interesting relationships in data (data mining), are combined with knowledge-based methods for capturing and reusing problem solving knowledge in the form of specific experiences (case-based reasoning). The main idea is to combine the two technologies in order to reuse past user experience in dealing with a problem situation, and to extract relevant and focused information from large and scattered data sources. The work is done within the EU project NOEMIE<sup>1</sup>.

### 1 INTRODUCTION

The corporate information and knowledge of large industrial companies is scattered among many data sources at different geographical locations. Existing legacy databases were usually constructed for the purpose of storing and managing dedicated administrative or operational information, and targeted for specific uses and for specific needs. In today's rapidly changing information society, where the right decisions usually need to be based on a variety of information types, it has become increasingly important to make scattered information available in a coherent way according to the user's needs [Belkin and Croft 1992, Brachman et. al., 1996, Block-97].

Reliability and safety are key issues in most industrial and domestic areas. Although new plants or new activities are well designed and planned, inevitably everything from minor technical failures to fatal accidents may occur in operation. In addition to limiting the consequences of such unwanted events it is imperative to find the true causes in order to prevent an event from re-occurring. Hence, one way to improve safety and reliability is to utilize past experiences in a positive way in future designs and operations.

Traditionally experience on safety and reliability issues is implicitly contained in databases tailor made to its specific use. This may be company specific databases, but also generic databases are available within many areas. The use of such data is commonly limited to the analysis of data from

---

<sup>1</sup> NOEMIE is an Esprit project (P22312), where the partners are, from France: Schlumberger (prime contractor), Matra Cap Systèmes, Université Dauphine, Acknosoft; from Norway: SINTEF, Norsk Hydro; from Italy: JRC Ispra.

one databank at the time, and the analysis methods available are normally of statistical nature (e.g. failure rate, lifetime distribution, trend analysis, benchmarking). By statistical methods it is more difficult to utilize such databanks for more focused problem description and re-use of solution paths. Hence, we also need a way to extract information of more qualitative nature but limited to the useful cases that address our problem situation.

Further, in a huge amount of data that may be spread on several databases, it is normally impossible for the human brain to spot correlations between data that may aid in solving more complex explanatory causes. This may typically be events where the causes often are observed as technically related, but where the root cause may be related to some underlying human errors. One desirable objective will therefore be to provide knowledge discovery in databases that are created for different purposes and/or by different companies.

In an EU project termed NOEMIE we are developing methodology and tools, aimed at facilitating experience feedback and reuse within and between organisations, in order to improve the operational and managerial decision making process. The two main goals of the project are to enable an effective and problem-focused reuse of data and information stored in legacy databases, and - based on those results - to provide improved methods for the construction of future information and knowledge storage aimed at experience reuse.

A key feature of the methodology is a combination of data mining (DM) and case-based reasoning (CBR) technologies. The methodology has been demonstrated by the development of a demonstrator (described in section 3). Later the methodology will be applied in two pilot applications.

## 2 THE NOEMIE APPROACH

Assume that an operator is faced with a problem related to failure of some technical system where the cause and possible solution to the problem is not obvious. He then wants to consult the NOEMIE computer system for assistance. The system contains in the experience case base a set of past experiences together with some general knowledge related to the problem domain (e.g. concept definition, taxonomies, decision rules). The cases

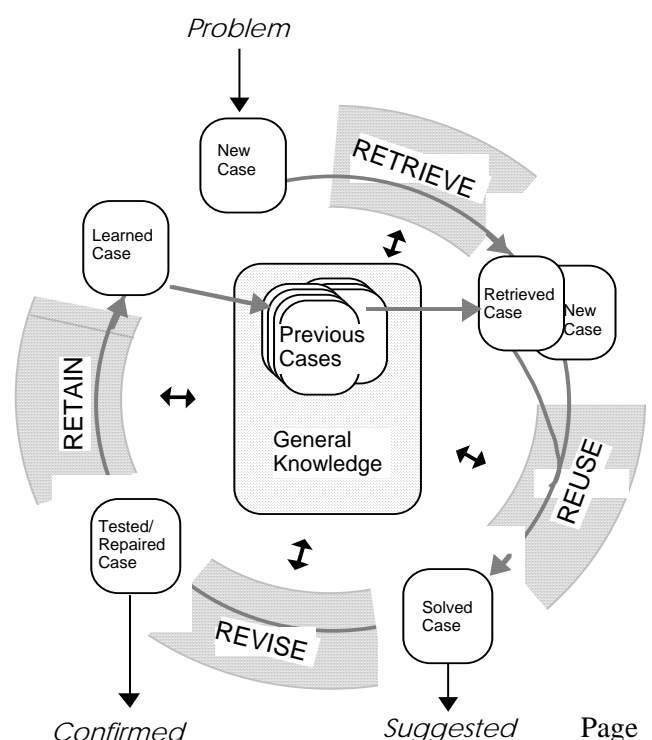
in the experience data base has been stored from previous cases of which some may have some similarity with the current problem situation. The NOEMIE user then set up a query to the system describing his current problem by some attributes (problem definition, measurements, key words etc. The more information he enters, the greater is the chance of retrieving useful cases.

By using some similarity measures the NOEMIE system first check whether a similar experience case has been solved before. If so, this case can be used:

1. To solve the problem directly by reusing and adapting the past solution
2. To focus the information search by using the past case as an automatic focusing 'lens' into the databases, i.e. reusing the information types that were found useful for solving the past case. If no relevant case is found in the experience case base, the source databases will be searched for single events (e.g. failures) which alone or in combinations may contain useful experience as to solving the current case.

Often the user may start with a course definition of his problem, then see what the system finds, and then add more information - or additional questions - depending on the result so far. For each new query the steps 1) and 2) above is repeated.

Below the CBR and DM processes are briefly outlined. This is followed by a description of how they are combined in order to realize the type of information retrieval and decision support just



described.

### 2.1 *The case-based reasoning process*

Case-based reasoning is the method of dealing with a new situation by retrieving and reusing a previous event. The process is usually described in terms of four steps [Aamodt and Plaza, 1994], see Figure 1.

Figure 1. The CBR-cycle

The user initiates the process by describing a problem situation, referred to as a *new case*. The *retrieve* step initiates a search in the experience database for one or more previous cases most similar to the new case. General domain knowledge may support this process, enabling retrieval based on semantic rather than syntactic similarity. A stored case typically contains identification of the event, description of the event, remedial actions initiated to solve the problem, and the success or failure of these actions.

The *reuse* step focuses on what part of the retrieved case that can be transferred to the new case, and the adaptation needed due to differences between the past and the current cases. The result is a suggested solution to the user's problem.

The *verify/revise* step is needed in order to give feedback to the system about the success of its suggested solution. This is necessary if the system is to learn from this event. It may involve the actual application of a suggested solution to the real problem, or other means of evaluation.

The *retain* step takes care of the final learning task, in which the system updates its case base from what it has just learned. A new case may be built, or the old case may be generalized.

The CBR method in NOEMIE is primarily based on the KATE tool [Acknosoft, 1996]. It has incorporated results from the INRECA EU-project [Auriol, 1994] in which case-based and inductive techniques were integrated. CBR can, in various ways, be combined with an inductive tree that acts partly as an indexing mechanism, partly as a problem solving mechanism in itself. In addition to indexing the cases in the case base, the inductive part also provides an additional, complementary method to data mining.

The Creek method [Aamodt, 1994] provides a complementary CBR method for knowledge-

intensive (explanation-based) similarity matching. Here a model of general domain knowledge supports the CBR processes of retrieval, reuse, and learning (retain). The model contains the relevant domain concepts and their dependencies and interrelations. An example of such a model is shown later (Figure 3). Using such a model the system is able to match two cases based on their contents (i.e. their meaning), rather than their appearance. Hence, the domain model is used to explain and justify why two cases are similar (even if they contain different parameters and values), how to modify a past solution if needed, and what to learn from a case just solved.

At the moment we are working on extending the current CBR methods to provide this type of knowledge-intensive functionality.

### 2.2 *The data mining process*

The overall process of discovering and extracting valid, implicit and previously unknown knowledge from large databases is often referred to as Data Mining (DM).

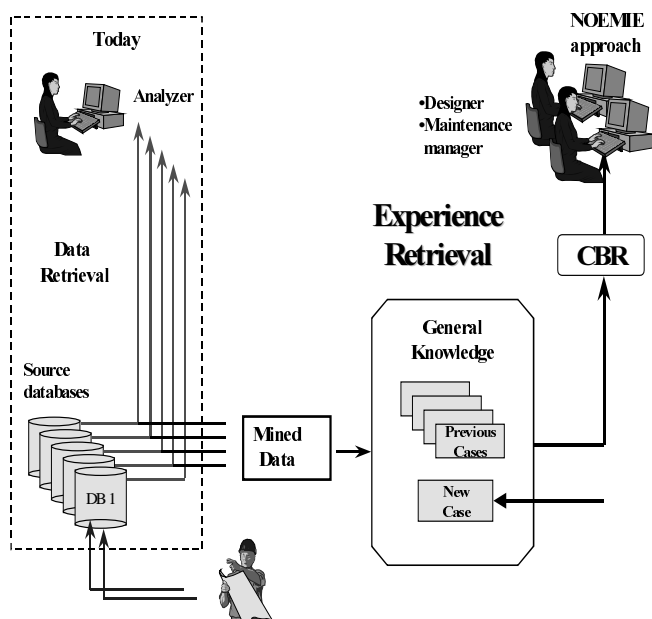
The term "knowledge discovery in databases" essentially refers to the same process, but is also used for a larger process including also the preparation and selection of data for data mining as well as the post-processing and presentation of the information from the mining process (e.g. [Fayyad, 1996]).

The actual data mining (extraction, generalisation) step is basically handled by a statistical clustering method. An extension of the classical clustering methods now under way is provided by what is called symbolic objects. Symbolic objects [Diday, 1995] extend standard objects of data analysis, in at least, two ways: first, in the case of individuals of varying complexity, by giving the possibility of introducing structured information into their definition; second, in the case of concepts or classes, by being "intentionally" defined.

Bayesian network has been identified as an addition to these methods. A Bayesian network (BN, also referred to as belief networks) is a network of statistical dependencies between parameters. They may model a network of causal relationships, of other types of influences between

concepts. BNs are constructed partly by manually defining the important concepts/parameters and their dependencies, partly by automatic methods that update the strengths of the dependencies. Our goal is to generate a BN containing the observable attributes of the domain model both to discover new relationships between them, and to verify the dependencies already entered by the domain modeller. The result is a submodel of statistical relationships, which will live their own life in parallel within the more general domain knowledge model. Hence, the Bayesian type of domain model can also contribute to knowledge-intensive CBR as described in the previous section. The BN generated submodel is dynamic in nature i.e. we will continuously update the strengths of the dependencies as new data is seen. This is opposed to the static relationships of a classic domain model.

In NOEMIE, data from different legacy databases are captured, abstracted and integrated into what is called a data warehouse. Two functions are required for the data warehouse: to organise the indexes, and to define how to access the information from the different sources. A common terminology – an object model – has been defined, containing indexes to corresponding terms used in the various data base schemes. We here make use of results derived from the POSC<sup>2</sup> standardisation effort. It defines a set of data types and relationships for equipment, processes, activities, etc., related to offshore oil and gas production.



reutechnical Open Software Corporation, a membership corporation founded in 1990 by BP Exploration, Chevron, Elf Aquitaine. Mobil and Texaco in order to improve the quality and

Figure 2. Current approach and the NOEMIE approach

### 2.3 Combining CBR and DM methods

The following functionality is incorporated in the NOEMIE concept:

- Modeling the problem domain
- Establish parameters and rules for indexing cases for efficient retrieval of relevant cases
- Selection of methods for assessing the similarity between a new problem and the past cases
- Testing and improving the measure of similarity through a learning process
- Retaining and generalising a new case as they are solved

The combination of CBR and DM methods may be envisaged as part of the following steps, spanning the gap from single databases (DB) to the end user:

**DB -> MDB<sup>3</sup>-> DM -> CBR -> RAMS -> User**

Human factors (HF) as part of the RAMS domain has two roles in the methodology. One is related to the interface between the user and the NOEMIE system i.e. optimise the system's functionality and user friendliness. The other role is HF as a type of information in the databases. (This is dealt with in the next section).

The type of data we need to deal with includes both structured and unstructured data. In parallel to DM methods, we also investigate new methods for free text search in large databases, and how to integrate such methods with DM and CBR.

NOEMIE is based on the integrated utilisation of two technologies: Case Based Reasoning (CBR) and Data Mining (DM) viz.:

- CBR enables retrieval of relevant data by comparing the user's current problem situation with previous situations (cases). Cases are situation-specific knowledge, stored in an *experience base* together with the necessary general knowledge (concept hierarchies, relationships, decision rules, associations, etc.).

<sup>3</sup> (MBD = merged database)

- DM is the knowledge 'creator' through its input to the experience database by mining information from several databases.

Together CBR and DM fill the gap between the scattered, hidden and hardly understandable data in existing databases, and the users' needs. The data selections, aggregations, and abstractions resulting from the data mining process is put in a form suitable for case-based reasoning.

#### 2.4 Technical and human reliability data.

Our methodology is targeted at two types of data, purely technical data and data related to human operations. For some databases the two types are intermixed and difficult to distinguish. But since one of the foci of the project is to increase the awareness of human factors related to operation of technical equipment, it makes sense to make this distinction. As an example of such a situation, imagine a leakage in a joint between two pipes. What is the cause of the failure, and how can it be prevented in the future? Well, the cause may be purely technical, i.e. mechanical wear or faulty components, but it may also be due to the human operator who failed to tighten the bolts optimally. The actual cause will have consequences for how to design new parts for pipeline joints, e.g. should there be some torque or tool specification to ensure that a certain bolt will be tightened within the acceptable range? Being able to combine the necessary data - technical as well as human factors, and irrespective of their location - in order to make the right decision, is at the core of the NOEMIE technology.

Traditionally Human Factors data bases have been concerned with human errors, with the purpose of collecting errors rates to be used in human reliability analysis. While this is important in system design, it gives little information when you want to improve human performance in an existing system. In addition to focusing on human factors data in the existing databases, we will develop a general HF database that, together with a specific HF company data base can produce useful information for the development and implementation of solutions [Bodsberg et al., 1993, Steen and Ulleberg, 1992].

### 3 THE DEMONSTRATOR

A demonstrator has been developed in NOEMIE to show the feasibility of CBR and DM. The demonstrator includes two applications from the oil industry viz. from the two participating companies Schlumberger and Norsk Hydro. Below is given an example from the Norsk Hydro scenario.

This scenario relates to unwanted events that may affect oil and gas production safety and availability. More specifically unwanted gas leaks have been used as base case. Such event is characterised by great impact on safety and often a complicated cause/effect relationship. In order to index the database for cases related to such events we had to define a domain model which contain the main attributes that characterise such event:

- Where the leakage occur
- Detection method of the leakage
- Descriptive parameters characterizing the leakage (hole size, gas pressure)
- Cause factors (technical, human related)
- Consequence (shutdown level, evacuation, production loss, fire/explosion risk)
- Remedial actions (repair, procedure revisions, training)
- Effect of remedial action (short/long term)

This information is included in a so-called domain model. This model contains each type of information describing the event plus the relations between this information. A principal example is shown in figure 3.



Figure 3. Domain model example.

The domain model may serve three purposes:

1. As an interface to the user for identification of relevant problem descriptors
2. As a formal model of terms and their interrelations
3. As a navigational tool into the problem area

In the Demonstrator we have addressed the two following situations:

- a) A high number of gas leaks is experienced on one offshore installation.

We want by the NOEMIE approach to:

- Check if the frequency of leaks is above some acceptance criteria
- Access the experience case base to see if this problem has been addressed in a broad scale before and if there is contained some general knowledge that can assist in solving this problem
- Use DM/CBR in combination to reveal any dominating cause factors and related solutions
- Devise solution for the most dominating causes (if any)
- Store the solution(s) in the experience case base including the steps taken to arrive at a solution

The basic steps are shown in figure 4. The main step is to set up a query which describe the gas leak problem we want to solve.

First the system search for similar cases in the past which has been solved and/or search for useful knowledge in the general knowledge base. The cases that may be relevant are listed in descending order according to a similarity measure. These cases may then be used as guide for a new and possible more focused search. If no relevant cases are found in the experience case base, a search will be made on all the data in the different source databases that are hooked up to this system.

- b) Single gas leakage case

The second main scenario addressed is a situation where one single gas leakage has occurred and the operator want to see if he can find cases of similarity that he may re-use to solve the current problem. In this case he address the NOEMIE tool by inputting problem descriptors that describe as

adequate as possible this specific problem. Otherwise the steps are the same as above.

The Demonstrator has been set up on a computer in France and is run in Norway via Internet. In France the company Schlumberger has applied the Demonstrator for assessing the reliability of a large variety of their products (tools) used in oil well logging. Four different databases are connected to the NOEMIE software. (The result from this Demonstrator is not described in this Paper).

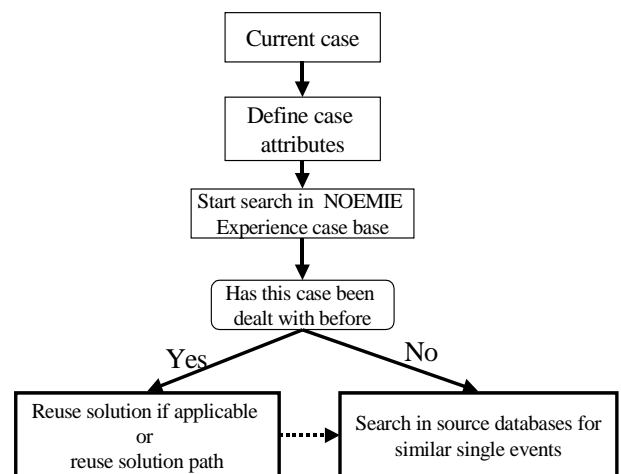


Figure 4. Basic approach Norsk Hydro Demonstrator

#### References:

- Aamodt A. and Plaza E. (1994). Case-Based Reasoning: Foundational issues, current state, and future trends, *AI Communications*, Vol 7, no.1, pp. 39-59.
- Aamodt A. (1994). Explanation-driven case-based reasoning: In S.Wess, K.Althoff, M.Richter (eds.), *Topics in case-based reasoning*. Springer. pp 274-288.
- Acknosoft (1996). *Introduction to KATE 5.02*. AcknoSoft, Paris, 1996.
- Auriol, E., Manago, M., Wess, S., Althoff, K-D., Dittrich, S. (1994). Integrating induction and case-based reasoning: Methodological approach and first evaluations. In M. Keane, J-P. Haton, M. Manago (eds.), *Proc. EWCBR '94*, Acknosoft Press. pp.145-157.

- Belkin, J. and Croft, B. (1992). Information filtering and information retrieval: Two sides of the same coin?, *Communication of the ACM*, Vol.35, nO.12, December 1992. pp 29-38.
- Block, J. (1997). Data warehouses: Clarifying the hype and confusion. *Inside Gartner Group*, Vol.13, no.3, January 22, 1997. pp 14-17.
- Bodsberg, L., Rosness, R., Øien, K. (1993). *Guideline of reduction of human errors during maintenance of safety systems, maintainability and maintenance support*. SINTEF report STF75 A93065.
- Brachman, R.J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., Simoudis, E. (1996). Mining business databases. *Communications of the ACM*, Vol.39, no.11, November 1996. pp 42-48.
- Diday, E. (1995). From data to knowledge: Probabilistic objects for a symbolic data analysis. *DIMACS, Series in Discrete Mathematics and Theoretical Computer Science*, 19, 1995.
- Fayad. U. (1996). From data mining to knowledge discovery in data bases, *AI Magazine*, Vol.17, no.3, Fall 1996. pp 37-54.
- POSC (1996). POSC/CAESAR Project, Oil and gas facilities data model: Snapshot B. POSC/CAESAR Document PCO-15, June 1996.
- Sten, T., Ulleberg T. (1992). Top-down approach to human factors. In *Proceedings from International Conference on Hazard Identification and Risk Analysis, Human Factors and Human Reliability in Process Safety*. American Institute of Chemical Engineers, 1992.