



HAL
open science

Acceleration of the retrieval of past experiences in Case Based Reasoning: application for preliminary design in Chemical Engineering

Stéphane Negny, Jean-Marc Le Lann

► To cite this version:

Stéphane Negny, Jean-Marc Le Lann. Acceleration of the retrieval of past experiences in Case Based Reasoning: application for preliminary design in Chemical Engineering. 18th European Symposium on Computer-Aided Process Engineering ESCAPE 18, Jun 2008, Lyon, France. pp.1009-1014, 10.1016/S1570-7946(08)80174-5 . hal-04030236

HAL Id: hal-04030236

<https://hal.science/hal-04030236>

Submitted on 15 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Acceleration of the Retrieval of past experiences in Case Based Reasoning: application for preliminary design in Chemical Engineering.

Stephane Negny, Jean-Marc Le Lann

INPT- LGC- ENSIACET, UMR-CNRS 5503, PSE (Génie Industriel) 118, Route de Narbonne 31077 – Toulouse Cedex 04, France

Keywords: Knowledge Management, Case Based Reasoning, Design, Fuzzy Sets.

Abstract :

The way to manage knowledge accumulated is one of the firm's trends, in order to capitalize and to transmit this knowledge. Some Artificial Intelligence methods are devoted to preserve and to reuse past experiences. Case Based Reasoning (CBR) is one of these methods dedicated to problem solving, new knowledge acquisition and knowledge management. CBR is a cyclic method where the central notion is a case which represents an earlier experience. Several cases are collected and stored in a memory: the case base. The goal of this paper is to soften the way to describe problem and to increase the effectiveness of the system during the retrieval of relevant cases.

1. Introduction

The design of a process or a product includes several steps. It starts with the requirements formulation and ends with a product (process...) which satisfies most of the requirements. In the chemical engineering field, there are numerous studies dedicated to different design steps: detailed design, simulation, experimental tests or validation... Nevertheless, there are very few on the preliminary design, because this step is often based on the knowledge and past experiences of experts. But this step is essential for the remainder of the design because it gives a starting point for the future solution. In this context, there is a need for a method focused on capitalizing expert knowledge in order to propose quickly a preliminary solution with high quality. In an industrial context, seeking to reduce the time during the whole design process, an effective tool dedicated to preliminary design allows a saving of time thereafter.

Case Based Reasoning (CBR) is one method coming from Artificial Intelligence, very useful for capitalizing and reusing past and new experiences, and knowledge deployed in the resolution of problems. After several evolutions, nowadays it is commonly accepted that CBR is a cyclic method (figure 1, R^3 model) based on the general principle: *Similar problems have Similar solutions*. The problems and their solutions are objects of a CBR system, and a case is the representation of an episode problem solving. Most of the time, a case is composed of the descriptions of a problem and its associated solution (with eventually some comments). Many cases are gathered and stored in a memory called the case base. In practice, a new facing problem (target problem) is compared with other problems stored in the case base (source problem) and the most similar one and its solution are extracted, then adapted to propose a solution to the initial problem.

For the problem and its solution descriptions, we used a formalism with feature-value pairs: the features or attributes represent the main and the most relevant characteristics

of the problems and solutions. The first step is the filling of the problem attributes for the target problem (**represent**). The next step retrieves in the memory one or various similar cases with the help of a similarity measurement in order to rank them (**retrieve**). Because there are some differences between the source and target problems, the source solution must be adapted to correspond to the target problem (**reuse**). In the next step the previous adapted solution is tested and revised to eliminate the discrepancies between the desired and adapted solutions (**revise**). Finally, after its resolution, the target problem and its associated solution form a new case that is retained in the case base (**retain**). This is an advantage for this method because the storage of a new problem increases the effectiveness of the CBR system by enlarging the cover of the problem space. On the other hand it is also a drawback because by increasing the number of cases in the memory, the time to retrieve a similar case will be increased too. The goal of this paper is to help the user during the retrieval of past cases. This amelioration concerns the retrieved step and can be decomposed in two points: a method to soften case representation and similarity measurement in one hand, and to anticipate the adaptation step in the other hand.

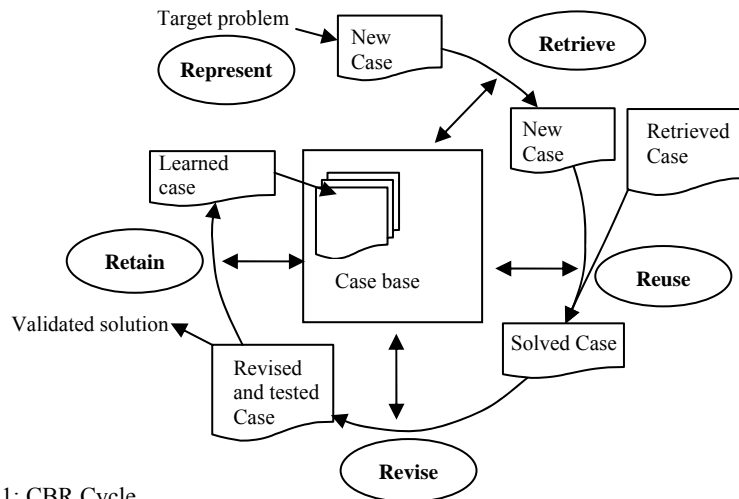


Figure 1: CBR Cycle

2. Retrieval

2.1. Case Base Organization

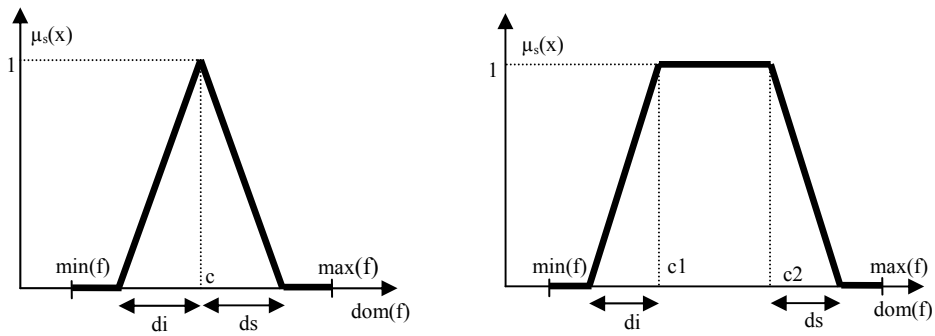
The number of cases in the case base is going to grow because of the Retain step or memorization of new cases. Without case base organization, the cost to estimate the global similarity between the target problem and all the source cases in the memory becomes prohibitive. In order to decrease the research time and to increase the effectiveness of the retrieval, the latter is decomposed in two steps. The first one consists in selecting a subset of relevant source cases. The second one is dedicated to the similarity measurement and the ranking of source cases included in the subset. To select the subset of the more relevant cases for aresearch, we index the case base to constrain the research space to the nearest source cases. The organization of the memory is based on the decision tree approach. In this approach, the case base is successively restricted thanks to decision sequences. All the cases of the base are gathered at a root node. Starting from this node, intermediate nodes are generated to restrict the number of case by an evaluation on a discriminate feature. And the end of the tree, at final nodes,

called the leaves, there are the source cases. Finally in this approach, leaves represent the classification and branches represent conjunction of features that lead to these classification. In the tool, the decision tree can be automatically built with an algorithm based on the ID3 algorithm. Nevertheless, the organization of the case base must reflect the point of view of the user, therefore he can generate its own decision tree corresponding to the aim of his retrieval.

2.2. Similarity Measurement

Generally in CBR systems, during the retrieval step, the most relevant case is the most similar one: the one which has the highest value for the similarity function. The global similarity between the target problem and some source ones is evaluated by a weighted sum of the local similarities on each attribute of the problem description.

$$SIM(X, Y) = \frac{\sum_i w_i \cdot sim(x_i, y_i)}{\sum_i w_i} \quad (1)$$



Figures 2: Fuzzy Sets representation

In chemical engineering, the attributes for the problem description can contain different type of values: semantic for the chemical compounds of mixture, and numerical values for operating condition. For the local similarity of chemical compounds, Avramenko et al, 2004 proposed an approach based on the chemical structure of compounds which is implemented in our system. Concerning the local similarity for attributes with numerical values, the most use formula is to measure the normalized distance (to avoid distortions of the results when features have different variation scales) between both source and target values on the same attribute. But during the preliminary design, the numerical values for the target problem description, are not often precisely known: an operating condition around a central value for example. Here we take into account this imprecision in the target problem description by the way of a percentage of imprecision around the central value specified by the user and a relation, for each attribute. Six different choices are available for the relation: *equ*, *sup*, *sup-equ*, *inf*, *inf-equ*, *between* the central value(s). For one attribute a_i , the local similarity measurement is achieved with the fuzzy set theory developed by Zadeh, 1965. We have considered two possible representations for the fuzzy sets: triangular (for the first five relations) or trapezoidal (only for the relation *between*) (figures 2). A fuzzy set S , on a domain D is defined by a characteristic function μ_s , which has values in $[0;1]$. $\mu_s(x)$ indicates the degree to which x is a possible value in S . With c (or $c1$ and $c2$); the central value for a_i for the target problem, the relation coupled with the imprecision allow to build the specific domain S_i

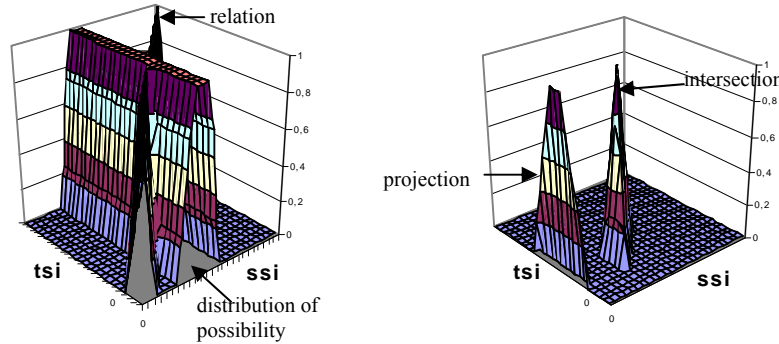
(to calculate d_i and d_s). The local similarity for a_i is calculated by $\mu_{s_i}(z_i)$ where z_i is the value of the source problem attribute corresponding to a_i .

2.3. Retrieval Guided by adaptation

The success of any CBR system is contingent on the retrieval of a case that can be successfully reused to solve the target problem. Consequently, the most similar case is unwarranted to be the most appropriate from the reuse point of view: it is not necessarily the easiest to adapt. Sometime, the most similar case may be difficult or impossible to adapt: technically, in terms of cost... Smyth and Keane, 1998, implement the idea of coupling the similarity measurement with a deeper adaptation knowledge traducing the easiness to modify a case to fit the target problem and to ensure case adaptation requirements. They called this technique: adaptation guided retrieval. With this technique, the research of source cases is based on two criteria: similarity and adaptability. Several methods exist to measure this adaptability, but we use the method proposed by Pralus and Gineste, 2006. For each attribute t_{s_i} of the target solution an adaptation domain is built from the definition domain of the same attribute in the source solution s_{s_i} , figure 3a. The definition domain of s_{s_i} is defined by a distribution possibility, which is specified when the case is retained in the base. The intersection of this distribution possibility with the relation (coming from the target problem description, previous part) is projected on the axis of the possible values for t_{s_i} , figure 3b. Finally we obtain a distribution of possible values for t_{s_i} . We state the assumption that the shape of this distribution determines the easiness to adapt this attribute. The more the range is large, the more this attribute is easiest to adapt (more choice to find an available value for t_{s_i}). The concept of specificity introduced by Yager, 1992, measures the degree to which a fuzzy set contains one and only one element (Specificity=1 for a very specific fuzzy set, with one value). Consequently the adaptability (ad_i) of each t_{s_i} is calculated with the measurement of the specificity of its distribution:

$$ad_i = 1 - S_p(F) \quad \text{with} \quad S_p(F) = \int_0^1 \frac{1}{\sup F_\alpha - \inf F_\alpha} d\alpha \quad (2)$$

$$\text{And the global adaptability of a case : } ad_s = \sum_{i=1}^n ad_i / n \quad (3)$$



Figures 3: Graphical representation of the adaptation domain

The user selects source cases with the two criteria and then the adaptation is made with the method described by Avramenko et al, 2004. This adaptation method is based on the main idea that the relative distances between the target problem and the selected source problems in the problem space are transferred in the solution space. To improve

adaptation, the Constraint Satisfaction Problem (CSP) method will be implemented, and it will also use the t_{s_i} distribution, this is why we choose this adaptability measurement.

3. Example

This example is presented in order to illustrate several parts of the method for the design of packing for separation. The mixture to separate is a three components distillation Methanol/Ethanol/Water. The target problem is a column which is operated at finite reflux, at atmospheric pressure, with feed flow rate between 0.1877 and 0.8123 mol/s. This distillation corresponds to the work of Mori et al, 2006. Moreover, in our problem description we impose that the distillation is at atmospheric pressure to exemplify the option EXACT, to impose a specific value to a feature. In their operating conditions, the authors do not give the range of temperature, consequently we suppose that it is not known. Of course, this range can be easily calculated with a thermodynamic analysis of the mixture at atmospheric pressure. But in order to show how our system treat the partial description of a problem, we do not fill this feature and we use the option IGNORE. The first five columns of table 1 sum up the problem description.

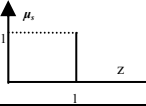
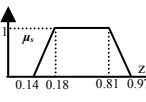
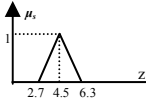
| | Relation | Central Value(s) | Imprecision | Ignore | μ_s Function |
|-----------------|----------------|------------------------|-------------|--------|---|
| Mixture | <i>equ</i> | Methanol/Ethanol/Water | | Off | |
| Pressure | <i>equ</i> | 1 | EXACT | Off |  |
| Temperature | -- | -- | -- | On | |
| Inlet Flow Rate | <i>between</i> | 0.1877 and 0.8123 | 20% | Off |  |
| Reflux | <i>equ</i> | 4.5 | 40% | Off |  |

Table 1: Problem description

For the retrieved step, the first work is to build automatically the function μ_s for each numerical feature, except for the temperature because the option IGNORE is activated. Therefore, this feature is not included in the global similarity calculation. These functions are represented in the last column of table 1. Before to calculate the global similarity, the case base is restricted to the subset of the most relevant cases thanks to a decision tree with the following succession of feature evaluation: at the root node the evaluation is on the Reflux, then the Pressure, then the Inlet Flow rate. Here again the temperature is ignored. For each cases in the selected subset, the global similarity measurement is calculated on four features; compounds, pressure, inlet flow, reflux, with the same weight for each one.

After the retrieved step, the ranking gives three structured packing (and two random packing, which are eliminated). The two different Montz-pak B1 are retained for adaptation. Finally, after adaptation, the proposed target solution is the Montz-pak B1 30, table 2. The second column of table 2 gives the characteristics of the structured packing used by Mori et al, 2006. In this example, the tools gives a good starting point for the resolution of the initial problem. It is to notice that the material of the two retrieved cases selected are: stainless steel (in case 1) and carbon steel (in case 2).

Consequently, for adaptation we search in the subset of metal. Then, the choice is oriented to the stainless steel because, under operating conditions in the same magnitude, the mixture of case 1 is most similar to the mixture of the target problem than the one of case 2. Therefore the choice is made with the following assumption: under operating conditions in the same magnitude, the most the mixtures are similar, the most the risk of degradation is reduced. This way to proceed is just a first approximation, and of course it needs to be improved because this assumption is not completely satisfactory. The CSP method will be useful for that.

| | Proposed Solution | (Mori et al, 2006) Solution |
|--|-------------------------------------|-------------------------------------|
| Type of Packing | Structured Packing Montz pak B1 300 | Structured Packing Montz pak B1 250 |
| Material | Stainless Steel | Metal (not specified) |
| Specific Area (m²/m³) | 350 | 247 |
| Geometrical Characteristics | | |
| angle | 45° | 45° |
| element height (m) | 0.201 | 0.197 |
| corrugation height(m) | 0.008 | 0.012 |
| corrugation base (m) | 0.0167 | 0.0219 |
| corrugation side length (m) | 0.0116 | 0.016 |

Table 6: Solution description

4. Conclusion

This paper focuses on the retrieval step in CBR and gives two ways to improve this step. In one way, it proposes a method to soften case representation by taking into account some imprecisions during the problem description. It also defines another criteria to determine if a retrieved case is relevant or not. Most of the time the similarity is the only criteria to choose a source case. Unfortunately, the most similar case is not often the most adaptable one. Here the similarity is coupled with an adaptability criteria. An improvement of our system concerns the next step of the CBR cycle, i.e. adaptation. Currently there is a general method which gives good results if the retrieved cases are very near the target problem (like in the presented example). With the adaptability criteria calculation, we generate the definition range for all the attribute of the target solution, this is a first step. The second one, is to fix the values of these attributes (numerical or not) in these intervals, with CSP method. Moreover with this method some constraints would be added like: user preferences, technical constraints...

References

- Avramenko Y., Nyström L., Kraslawski A., 2004, Selection of internals for reactive distillation column – case based reasoning approach, *Comp. and Chem.Eng.*, 28, 37-44.
- Pralus M., Gineste L., 2006, Recherche et adaptation d'expériences structurées, imprécises et incomplètes, *Raisonnement à Partir de Cas 1*, Hermes, 65-93.
- Mori H., Ibuki R., Taguchi K., Futuma K., Olujić Z., 2006, Three-component distillation using structured packing: performance evaluation and model validation, *Chem. Eng. Sci.*, 61, 1760-1766
- Smyth B., Kean M. T., 1998, Adaptation-guided retrieval: questioning the similarity assumption in reasoning, *Artificial Intelligence*, 102, 249-293.
- Yager R.R., 1992, On the specificity of a possibility distribution, *Fuzzy Sets and Systems*, 50, 179-292.
- Zadeh L.A., 1965, Fuzzy sets, *Information and Control*, 8, 338-353.