Possible Applications for Case-Based Reasoning in the Field of Wastewater Treatment

Content Areas: Case-Based Reasoning (CBR), Decision Support System (DSS), Real Time Control (RTC), Wastewater Treatment Plants (WWTP), Sequencing Batch Reactor (SBR)

Jürgen Wiese *#, Armin Stahl ** and Joachim Hansen +

# Corresponding Author

* Anlagen- und Sondermaschinen Automation GmbH (ASA GmbH), Zur Großen Heide 15, 32425 Minden, Germany, PHONE: +49-5704-164421, FAX: +49-5704-164422, email: wiese@asagmbh.de (formerly: +)

** German Research Center for Artificial Intelligence DFKI GmbH, Erwin-Schrödinger-Str., 67608 Kaiserslautern, Germany, email: stahl@informatik.uni-kl.de

+ tectraa – Center for Innovative Wastewater Technology, Technical University of Kaiserslautern, Paul-Ehrlich-Str. 14, 67663 Kaiserslautern, Germany, email: jhansen@rhrk.uni-kl.de

Abstract

For the last years, Artificial Intelligence (AI) approaches have become useful tools in environmental engineering. Here, one relevant application area is the optimization of wastewater treatment plants (WWTP). Besides the examination of technical aspects of the different environmental systems, their human managers' knowledge and experiences from past events gain more and more importance. In this paper, we will present several examples for approaches from Experience Management (EM) for control tasks and Decision Support Systems (DSS), specifically based on Case-Based Reasoning (CBR) in the field of wastewater treatment.

1 Introduction

During recent years, a rising complexity of the problems in the area of wastewater treatment can be observed. On the one hand, major reasons can be found in the increasing requirements for purification and the interweaving to a high degree by connections and dependencies of sewer system, WWTP, and receiving water. On the other hand, the technologies for measurements of the quality parameters as well as the process control systems have become more powerful and less expensive. Nevertheless, such systems are still a cost factor. Due to the fact of low public budgets, the use of latest technologies or even expensive enhancements in the WWTP infrastructure is often impossible.

Thus, approaches for optimization of existing plants attract more and more the attention, which make extensive use of the plant-inherent potentials. At this stage, methods and technologies from AI have been discovered to play an important role. Even though measuring and control technologies are improving, the problem of incomplete or missing data still exists because many parameters are difficult to be determined or cannot be determined at all. Furthermore, in specific cases, the measured data might not be representative for the overall system. Therefore, it often happens that the WWTP operator must control the plant rather with his experience from past events than with sophisticated machines. When it comes to capturing and especially drawing conclusions from experiences, AI offers with Case-Based Reasoning a powerful technology, which has already proved its potentials in different industrial applications (see, e.g., [Bergmann *et al.*, 1999]). In this paper, we will present several examples for possible applications for CBR in wastewater treatment.

The paper is structured as follows. In Section 2 we will describe an architecture for a predictive WWTP controller that bases its decisions for the plant control on past events and situations captured in cases. The system has been tailored to Sequencing Batch Reactors. We will also present results of three offline CBR models, which have been developed to predict the influent flow rate, the sludge settling curves, and the endogenous denitrification rate. Section 3 will focus on a DSS based on a CBR approach for Identification and Counteraction for Harmful Microorganisms in WWTPs. In Section 4 we will outline methods for the optimization of the prediction accuracy. In Section 5, we take a look at other CBR approaches in the field of wastewater treatment. Section 6 ends with the conclusions.

2 Example – Real Time Control (RTC)

The following examples will focus on possible applications for Case-Based Reasoning for control purposes of Sequencing Batch Reactor Plants.

2.1 Introduction

One of the several types of wastewater treatment technologies, which are commonly used in the world, is the SBR technology. In contrast to a continuous flow plant, in a SBR all treatment processes take place in one single reactor, step after step as illustrated in Figure 1. The time between the beginning of the fill and the end of the treatment process is called a cycle. The SBR technology has a high process flexibility and treatment efficiency, because with the help of modern computer-aided control devices (CACD) it is possible to adapt the duration of a cycle, the duration of the different phases (e.g., aerated react, settle) within each cycle and the volumetric exchange ratio (the fraction of the reactor volume, which is removed during draw, and replaced during fill) to the current requirements. This especially applies when sensors are used for control purposes. For instance, it is possible to vary the duration of the settle phase depending on the sludge settling characteristics. Unfortunately, most of the SBR plants are still using fixed time control strategies; until now, measuring devices are predominately only used for monitoring.

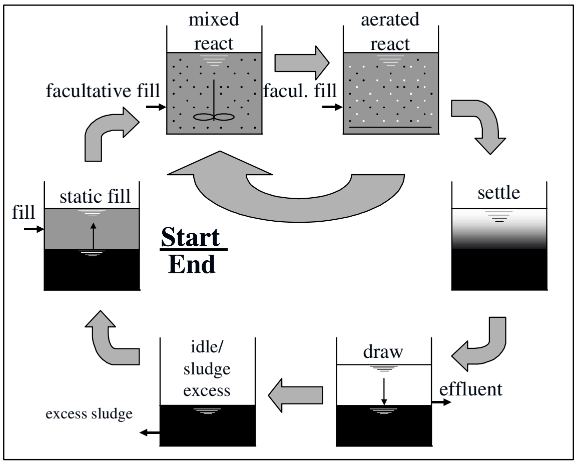


Figure 1: The concept of SBR.

2.2 Description of SBR-WWTP Messel

The WWTP Messel (schematically depicted in Fig. 2), which was put into operation in 2000, is a modern SBR plant with a primary treatment, one influent holding tank, two SBRs, one effluent buffer tank, and a final filter. Except for the filter, this configuration is often used in Germany. The plant was designed according to the German guidelines ATV A 131 [1991] and ATV M 210 [1997] for biological phosphorus removal, nitrification, denitrification, and a maximum flow rate of 230 m^3/h. The plant is equipped with a modern CACD and numerous online measurement equipment. According to the static dimensioning, the plant is operated with a cycle duration of 8 hours (h) during dry weather flow, but during combined sewage flow it is necessary to reduce the cycle duration to 6 h and thus, to increase the hydraulic capacity of the WWTP. The catchment area of WWTP Messel, which is typical for many other rural areas in Germany, covers 1.5 km^2, and a population of about 3,750. Most of the inhabitants are connected to a combined sewer. The wastewater can be characterised as domestic sewage, because there are only few commercial dischargers (500 p.e.). The effluent limits of WWTP Messel are very low (e.g.: 45 mg/l COD, 3 mg/l NH_4-N), because the receiving waters are very small and sensitive. Figure 3 shows a comparison between the old (trickling filter) and new (SBR) WWTP Messel. It becomes clear, even though the WWTP Messel is a small plant, it is a very complex technical system. Consequently, it is not quite easy to run such a system efficiently. This particularly applies, because the WWTP is not permanently manned and is operated by only one person.

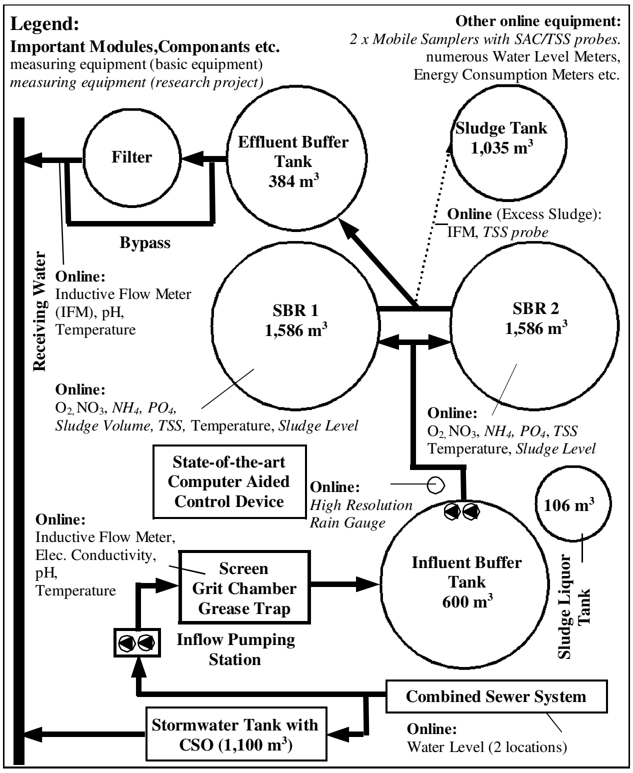

Legend:
Important Modules,Componants etc.
measuring equipment (basic equipment)
measuring equipment (research project)
Other online equipment:
2 x Mobile Samplers with SAC/TSS probes.
numerous Water Level Meters,
Energy Consumption Meters etc.
Filter
Effluent Buffer Tank 384 m³
Sludge Tank 1,035 m³
Online (Excess Sludge): IFM, TSS probe
Bypass
Receiving Water
Online: Inductive Flow Meter (IFM), pH, Temperature
SBR 1 1,586 m³
SBR 2 1,586 m³
Online: O2, NO3, NH4, PO4, Sludge Volume, TSS, Temperature, Sludge Level
Online: O2, NO3, NH4, PO4, TSS Temperature, Sludge Level
State-of-the-art Computer Aided Control Device
Online: High Resolution Rain Gauge
106 m³
Online: Inductive Flow Meter, Elec. Conductivity, pH, Temperature
Screen Grit Chamber Grease Trap
Influent Buffer Tank 600 m³
Sludge Liquor Tank
Inflow Pumping Station
Combined Sewer System
Stormwater Tank with CSO (1,100 m³)
Online: Water Level (2 locations)

Figure 2: Scheme of WWTP Messel.

	1975-2000	since 2000
Principle	Trickling Filter	SBR
Purification Efficiency	medium	very high
Control Device	none	state of the art
Actuators	1	> 15
Sensors, Status Indicators	3	> 50
Process Data	few	numerous
Complexity	low	very high
Control Possibilities	1	→ ∞
Optimization Potential	low	high
Requirements on Operators	low	high

Figure 3: Characteristics of the new WWTP Messel in comparison with the previous treatment facility.

Therefore, a research project has been initiated to develop RTC Strategies in simulation as well as in full-scale and to assess the economic and ecological benefits of such approaches [Wiese *et. al.*, 2004a]. The modeling procedure is described in detail in Wiese *et. al.*, [2004b].

In the first part of the research project, very detailed models of the combined sewer system and the WWTP have been developed. These models were calibrated and validated with monitoring data. With these models, several control strategies have been developed. These strategies are based on ammonia and nitrate sensors, as well as sludge blanket and suspended solids probes. Furthermore, a rain gauge has been integrated in the control strategies. The results of the WWTP simulation show that it seems to be possible to reduce the cycle duration during combined sewage flow in full-scale in almost every case to only 4 h without exceeding the low effluent limits. This leads to an increase of the hydraulic capacity of the plant up to 50 % by using the developed control strategies. In several cases, it should be even possible to reduce the cycle duration to less than 4 hours. I.e., it would be possible to increase the maximum flow rate to the WWTP from 230 up to more than 345 m^3/h.

In the second part of this project, the different control strategies were realized in full-scale. The results of the second phase are very good. With the help of the control strategies it was possible to further increase the treatment efficiency significantly. E.g., it was possible to reduce the average total nitrogen (TN) effluent concentration from 6,4 to only 2,9 mg/l TN (0,1 mg/l NH_4-N) and thus to reduce the nitrogen emissions into the receiving water by more than 50 %. But, despite these positive results, there are still several problems, e.g.:

- Due to the discontinuous principle and the limited capacity of the buffer tank, it is necessary in case of rainfall to reduce as early as possible the cycle duration.
- The optimization potential depends on several factors, e.g., influent load, wastewater temperature and sludge settling characteristics, but these parameters can vary strongly and sometimes rapidly.
- Furthermore, according to the German law, it is not allowed to exceed the official effluent limits.
- Depending on the actual operating conditions, it can be useful to use different optimization criterions (e.g., increase of treatment capacity vs. energy saving).

That means, the whole potential for optimization can only be used when a control strategy is used, which is able to act and not only to react. Consequently, we developed a method that is serviceable for a controller being able to predict as early as possible the duration of a cycle, which is necessary to achieve the treatment target. Furthermore, the controller also should be able to predict other important operating data (e.g., the maximum volumetric exchange ratio), and the influent flow rate.

2.3 A Case-Based Predictive Controller

From our point of view, it seemed to be promising to develop a predictive controller based on a CBR approach because of the following reasons:

- Beginning and end of the treatment process are exactly defined. With a few restrictions, this is also valid for the different treatment phases of the cycle, which helps to easily determine a case structure.
- It is important that the system works fast because the time delay between the beginning of a rainfall event and an increase of the inflow rate can be quite short.
- Numerous of online monitoring data are available. With cycle durations between 3 and 8 h the database will grow very fast, i.e. case and data acquisition is not a problem. In order to ensure efficient retrieval when dealing with huge case bases, one may apply different strategies. One possibility is to store only actually useful cases while throwing away redundant and less useful cases (e.g. see [Smyth, 1995]). Another possibility is to employ efficient retrieval approaches, for example, case retrieval nets [Lenz, 1999] or algorithms that build on top of relational databases [Schumacher, 2000].

2.3.1 Control System Architecture

Modern SBR plants often have a lot of online measurement equipment. However, as a consequence of higher treatment standards, reduced prices for sensors, etc., a further increase in online monitoring, especially for quality parameters (e.g., NH_4, NO_3) can be expected. Due to this fact, it will be possible to document the curves of important processes within each cycle. Later on, it would be possible to calculate the duration of each treatment phase, which would have been sufficient to reach predefined effluent standards. The opportunities for a Case-Based predictive SBR controller resulting from these circumstances are promising, especially in case of an integrated RTC strategy. E.g., at the beginning of a rainfall event, the controller could predict the required duration and composition of the next cycle, by comparing actual process information with historical data. In the next step, the maximum hydraulic capacity of the WWTP can be calculated. However, due to the enormous amount of measurement data, it would not make sense to use only one CBR model to predict the required cycle duration and composition, because the database would have to be extremely large. So, it is promising to work with multiple domain models. Figure 4 shows a part of our proposed system architecture. The specific process controlling units for the WWTP and the sewer system are connected via an interface (CACD) that mediates between our predictive control system and the controllers for the WWTP and the sewer system.

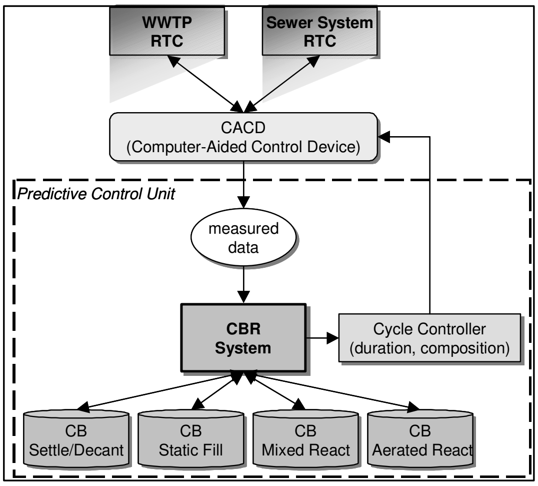


Figure 4: Principle of the predictive SBR controller.

The interface provides us with all measured data and forwards the control data resulting from the predictions depending on the current situation. Our predictive controller consists of a CBR system as the core part, which operates on multiple case bases and domain models, respectively, with respect to the WWTP subsystem to which the measured data (situation) belongs. Speaking more specifically, almost each process stage in the cycle depicted by Figure 1 is represented by its own case base. The exceptions are the "settle" and "draw" (also known as "decant") phases that are summarized in one case base and "idle/sludge excess" phase, which we will not support with respect to time optimization due to its very short duration.

New measured data is taken as an input to our CBR system, which generates the adequate problem descriptions for querying the different case bases. As we are dealing with an independent series of process phases in the regarded cycle, i.e. a phase can only be started after its predecessor having finished, we can optimize (predict) the processing time of each individual phase and add the predicted duration times of each single phase in order to obtain the overall cycle duration. This fact also allows us to query the single case bases simultaneously.

The cases are problem-solution pairs, where the current situation (measured data) represents the problem part and the solution is given by the respective control data for this situation. Due to the structure of the data, we are working with flat domain models.

The subsequent control data for each single phase is derived from the retrieval result of the *n* most similar cases from past situations. Adapting the solutions from the respective *n* cases generates the solution for the current situation. However, the adaptation method depends on the process phase. The new solutions are forwarded to the cycle controller unit, which processes them and gives the final solution back to the CACD. Depending on the results of the different case bases, the cycle controller will estimate the total duration of the cycle and create the composition of the cycle. Due to the fact, that the hydraulic capacity of the WWTP depends on the duration of each cycle and the current exchange volume, the maximum flow rate to the WWTP could be calculated in the next step. The system has been implemented with CBR-Works® (empolis – knowledge management, Inc.). Until now, we have only implemented a few test components of the described overall architecture. So far, our system only simulates the control process offline, i.e. the generated solutions are not to be returned to the CACD interface.

2.3.2 Example "Influent Flow Rate"

For specific tasks (e.g., energy saving) and questions (e.g., Which type of cycle (6 h, 8 h, 12 h) should be used next?), it is reasonable trying to predict the influent flow rate curve of the next few hours. Such an information can be useful to control the filling of a equalization basin etc. But due to several reasons (e.g., infiltration water), even during phase of dry weather flow, the influent curve can vary significantly (Figure 5).

Hence, it is not very helpful to base a control strategy on an average inflow rate curve. Consequently, a CBR model has been set up to predict the dry weather influent flow rate curve of WWTP Messel for the next 24 hours. The initial case base of this model were all influent flow rates, which have been measured in 2003 during dry weather flow conditions (124 curves). The following 5 attributes have been chosen for the model:

*Minimum of the daily influent flow rate of the past 21 days (*local similarity: polynomial function), because this attribute is suitable to estimate the influence of the infiltration water flow rate. *Weekday*, because the changing life rhythm of the people during the week has a significant impact on the influent rate curve of WWTP Messel as well as the different *school holidays* resp. *bank holidays*. Finally, the attribute *summertime/wintertime* was used. To describe the local similarities of the last four attributes, similarity matrices were used. The predicted influent flow rate curve is a weighted function of 3 historical curves, which have been measured under the most similar operation conditions. Even though the CBR model is simple, the results are very good (Fig. 6): In this figure the measured and the predicted influent flow rate curve for a 24 hour interval are almost identical.

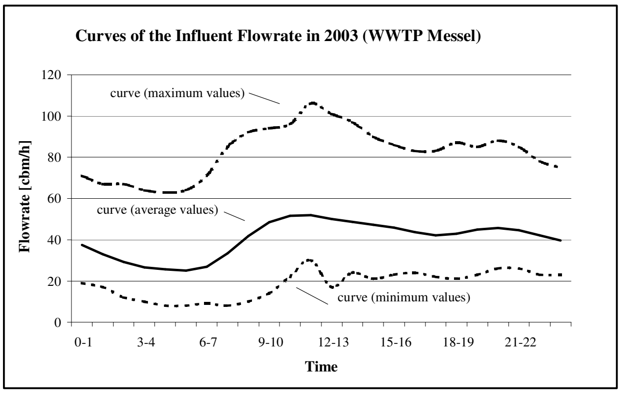


Figure 5: Bandwidth of influent flow rate curves during phases of dry weather flow in 2003 (WWTP Messel).

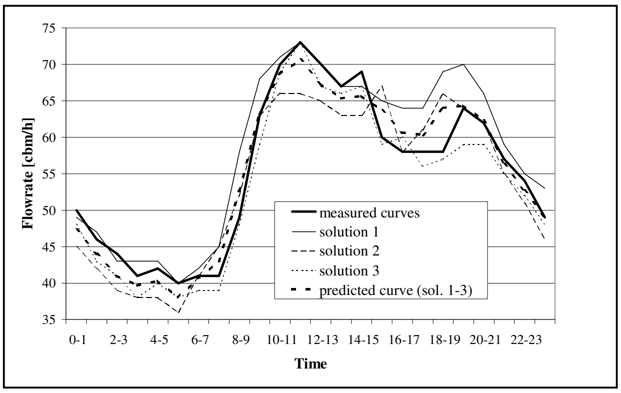


Figure 6: Example for a good curve prediction

Of course, not in every case it is possible to reach such good results. Nevertheless, even with this simple model, it is possible to predict the flow rate per hour in 80 % with a deviation of less than ± 5 m^3/h resp. in 95 % with less than ± 10 m^3/h; the maximum deviation was 33 m^3/h.

Since a few weeks, the organic load in the influent of WWTP Messel is measured online with the help of a dissolved organic load probe and a total suspended solids probe. So, in the next step, the authors will be trying to predict the influent organic load curve, too.

2.3.3 Example “Settle/Decant”

One of the results is that there is a huge potential for optimization of the settle and decant (draw) phase. During this phase, first the water/biomass separation takes place and then the treated wastewater will be decanted. Due to the fact that even a small sludge displacement from the reactor into the effluent of the plant can cause an exceeding of the required effluent standards, the settle and decant phase was dimensioned for unfavorable operational conditions. In order to point up the potential for optimization, an example is depicted in Figure 7. As a consequence of the static dimensioning, the duration of the settle and decant phase in case of WWTP Messel takes in total 140 min. In reality, however, the operational values are usually much better than the comparable design values. Therefore, sludge blanket and suspended solid probes were installed at the decant devices to investigate the potential for a reduction of the settle and draw phase. The results of this investigation show that in many cases it would be possible to reduce the settle and decant phase up to 70 min and thus to increase the hydraulic capacity up to almost 20 %. Furthermore, the monitoring shows that in most of the cases it would be possible to increase the volumetric exchange ratio from 40 % to approx. 50 % (+145 m^3; see Figure 7); this could further increase the hydraulic capacity. Due to the high optimization potential of the settle and draw phase, it was decided to develop the CBR subsystem “Settle/Decant” first.

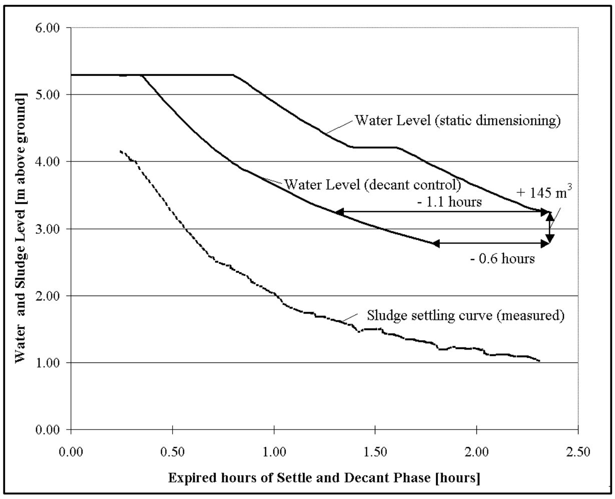


Figure 7: Potential for optimization of settle and decant phase.

In the first step, more than 120 sludge settling curves, which have been measured under different operational conditions, were analyzed and evaluated statistically. It could be observed that the settling velocity of the sludge blanket mainly depends on two factors. As already published by other authors (e.g., [Keudel and Dichtl, 2000]), the initial settling velocity mainly depends on the sludge volume at the beginning of the settle phase. For instance, the settling velocity in a full SBR is higher than in a barely filled tank, because the compression phase of the sludge starts later. Furthermore, it could be observed that the settling velocity depends on the last phase before the settle phase starts. For example, in case of a mixed react phase, it takes at least 10 min until the sedimentation begins. In case of an aerated react phase, the turbulence at the beginning of the sedimentation phase is smaller, thus the flocculation process is faster and the sedimentation process can start in less than 5 min. Consequently, the cycle type, the water level in the reactor, the sludge volume, and the water temperature were chosen as attributes in the respective CBR model (see Table 1). In order to create the case base, in the second step, 30 representative curves have been selected. Then, the calibration and validation process was started. The local similarity measures are mainly given by linear distance functions (Euclidean distances) between the query values and the respective case values. Only the cycle type with its two values ‘dry weather’ and ‘rain weather’ has been modeled as a simple similarity matrix. The global similarity function is a weighted sum of the local similarities. The solution part of the cases is given by the courses of the respective sludge heights, represented by curves (sludge settling curves). We simplified the representation of these curves approximating them by polynomials of degree six. The idea was to be able to easily compare the coefficients a_1 to a_6 of these polynomials with each other, in order to evaluate the quality of the generated solutions.

Table 1: Attributes and their value ranges

Attribute	Value Range
Cycle Type	dry weather, rain weather
Max. Water Level	3.32 m – 5.30 m
Sludge Volume	241 ml/l – 446 ml/l
Water Temperature	8.7 °C – 21.4 °C

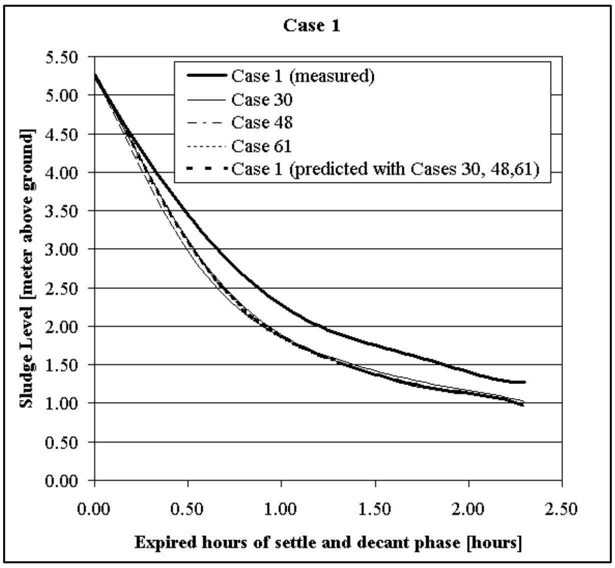


Figure 8: Good prediction of the sludge settling curve.

The results produced by this subsystem are very promising. Despite the fact that the database is rather small, the model is able to predict the sludge settling curve well. Thereby, the predicted sludge settling curve is a weighted function, calculated with the help of 3 measured curves, which have been measured under the most similar operation conditions. Figure 8 shows an example for a good prediction of the sludge settling curve. The measured and the predicted curve are almost identical. Of course, not all predictions are as good as the example in Figure 8. Figure 9 shows an example for a worse prediction. However, even in this worse case the maximum difference between measured and predicted curve is only 0.5 m. It has to be taken into consideration that the measurement inaccuracy of the sludge blanket probe can be up to 0.2 m. Furthermore, in practice such worse predictions would not cause serious problems, because with the help of a sludge blanket probe-based and/or a suspended solids probe-based feedback decant controller, which survey the decant phase, it would be easily possible to close the decanter immediately, in case of a sludge displacement danger.

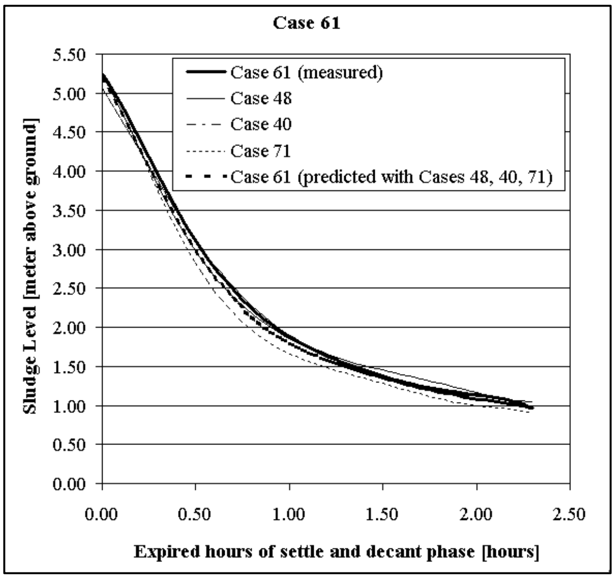


Figure 9: Example for a worse prediction.

2.3.4 Example "Endogenous Denitrification"

During the last years, several attempts were started to use online deterministic WWTP models (e.g., ASM 3 [IWA, 2000]) – which are able to simulate biological processes (e.g., nitrification, denitrification) – for control tasks. Unfortunately, these models are very complex. E.g., the ASM 3 model considered 12 processes, 14 model compounds and 36 kinetic and stoichiometric parameters; many of these parameters are difficult to be determined or cannot be measured at all until now. Consequently, the implementation and operation of deterministic online WWTP models are very costly. Therefore, the experiment was started to predict biological processes with the help of CBR: The endogenous denitrification process during the settle and decant phase of a batch cycle was selected as a test example. The initial case base consists of 137 settle and decant phases, in which all relevant measuring data for the calculation of the endogenous denitrification process could be determined. The result was, that the endogenous denitrification rate amounts between 0,1 and 5,0 kg NO_3-N per cycle (average value: 2,3 kg NO_3-N, standard deviation: 1,0 kg NO_3-N). The following 6 attributes have been chosen for the model:

NO_3-N load and NO_3-N concentration in the reactor at the beginning of the settle and decant phase, because these both attributes have a high influence on the endogenous denitrification rate. Other important attributes are the *water level*, the *total biomass* and the *wastewater temperature* in the reactor. Finally, the number of the SBR (where the data were measured) was chosen as an attribute, because the biological activities in the different reactors can be slightly different.

To describe the local similarities of the attributes, linear and polynomial similarity functions were used. The predicted endogenous denitrification rate is usually calculated as the mean value of 2 rates, which have been measured under the most similar operation conditions.

With the help of this model, it is possible to predict the endogenous denitrification rate quite good: The standard error of the estimation is 0,6 kg NO_3-N. For comparison: The standard error of the estimation with the help of a multiple regression model is 0,7 kg NO_3-N.

2.3.5 Future Work

As a consequence of the good results reached with the 3 different CBR models, other components of our architecture should be developed, i.e. we will create the domain models and the respective CBR subsystems. Thereby, the monitoring established within the research project serves as a data source for the other case bases. Furthermore, it is planned , to use the CBR software to explore the specific experiences of the operators and to use CBR as a training tool. In the next 2 years, our overall system should then be verified in full-scale by feeding the so generated control data into the modern CACD of WWTP Messel.

3 DSS Harmful Microorganisms

3.1 Introduction

Increasing quantities of wastewater made enlargements of treatment plants necessary. Then, trying to optimize the costs for running the plants by reducing the precipitation and minimizing the oxygen supply for the biological system in the plant sometimes leads to new problems; from the ecological and biological points of view, optimization can cause undesired side effects. Environmental conditions can appear that favor filamentous organisms, which can cause foam effects or later even lead to harmful bulking sludge or scum formation [Eikelboom, 2000]. We can observe this phenomenon in a growing number of WWTPs during recent years; especially during spring and autumn time. One crucial factor amongst others is the loss of biomass needed for the biological purification in the system. The responsible harmful microorganisms affect nearly all biological processes for WWT. Additionally, the bulking sludge problem does not only influence the WWT in a negative way but also the sludge treatment. If sludge dominated by filamentous bacteria reenters the anaerobic sludge treatment foaming of the digester contents can occur. As a consequence, the digester can over boil.

The managers of WWTPs with bulking sludge problem consider this one of the most important problems to be solved. Nowadays, various approaches for counteractions exist to eliminate the problem-generating microorganisms [Eikelboom, 2000], for instance, deployment of lime, polymers or pulverized lignite, installation of selectors, increasing or decreasing of the oxygen, etc. Usually, bulking sludge problems have their individual aspects depending on the WWTP where they occur. Therefore, the next problem has to be seen in finding the right solution. This task is even harder to solve, as different harmful types of microorganisms can exist in the sludge. The same counteraction that kills one of these types of bacteria can help the growth of others.

We conclude that the only efficient way for suppressing the excessive growth of the specifically responsible microorganisms is their identification and the closely related goal-directed selection of treatment means. Our starting points are the positive and negative experiences experts made in the treatment of bulking sludge problems. Their experiences serve as successful suggestions for solutions respectively the knowledge about unsuccessful treatments (failures). So, the aim was the development of an expert system that supports the decision process for the selection of adequate counteractions. The system is fed by a query that describes parameters of the WWTP. We will have a closer look at the technology behind the scenes of our expert system and the underlying domain model in Sections 3.2 and 3.3.

3.2 The Case Representation

It is typical for CBR applications that the case representation consists of two major parts: a problem description and a solution description, as mentioned before. In the following, we give an overview of the structure of these two parts that make up the domain model for our system.

The aim of the problem description is to characterize the current situation on a WWTP when a problem caused by uncontrolled reproduction of harmful microorganisms is observed. Unfortunately, even WWTP experts are not able to determine the relevant influences exactly. Therefore, all information that may have significant impact on the microorganism problem is considered in the problem description. Basically, the information of the problem description is divided into the following four parts, represented by particular concepts in an object-oriented domain model:

WWTP data: This part contains relevant information about the respective WWTP where the problem occurred. This kind of information includes attributes that describe the structure and operating parameters of the specific plant.

Already performed counteracts: Here, all available data about already performed counteracts against the sludge problem is stored. These pieces of information are also essential because it contains important hints about the responsible microorganism species. For example, if a counteraction that works usually very well against microorganism M has been applied, but the bulking sludge problem is still present, this is a clear advice that microorganism M is not the responsible species in the current situation.

Environmental data: Due to the fact, that the occurrence of microorganism problems crucially depends on the current environmental circumstances, this information is also a core component of the problem description.

Quality information: Additionally, some attributes describing the quality of the particular case data are introduced. Because the case base contains currently observed problems as well as problems described in specific WWT literature, it is useful to assign each case a respective confidence level.

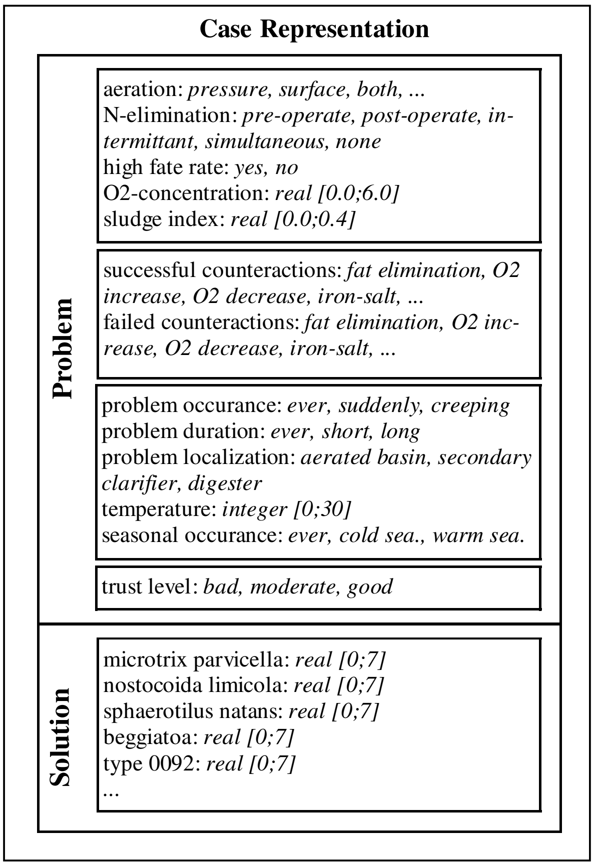
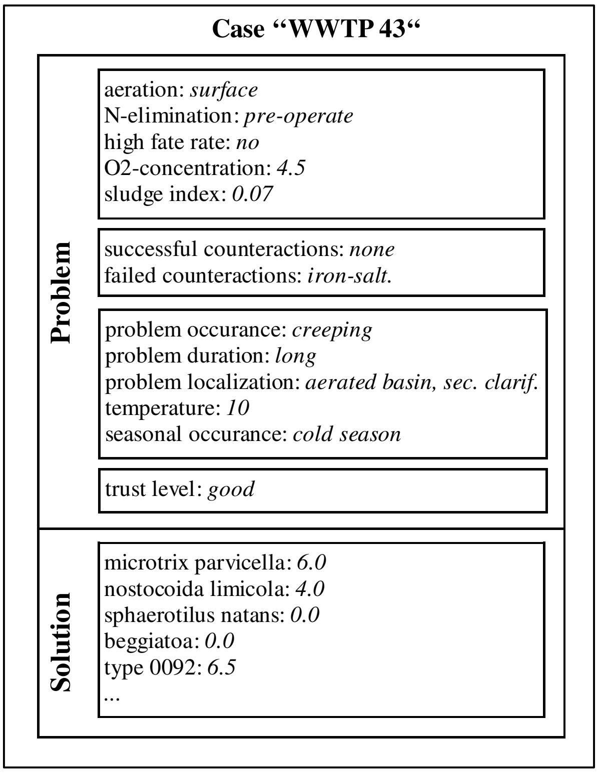

Case Representation
Problem
aeration: *pressure, surface, both, ...*
N-elimination: *pre-operate, post-operate, intermittant, simultaneous, none*
high fate rate: *yes, no*
O2-concentration: *real [0.0;6.0]*
sludge index: *real [0.0;0.4]*
successful counteractions: *fat elimination, O2 increase, O2 decrease, iron-salt, ...*
failed counteractions: *fat elimination, O2 increase, O2 decrease, iron-salt, ...*
problem occurance: *ever, suddenly, creeping*
problem duration: *ever, short, long*
problem localization: *aerated basin, secondary clarifier, digester*
temperature: *integer [0;30]*
seasonal occurance: *ever, cold sea., warm sea.*
trust level: *bad, moderate, good*
Solution
microtrix parvicella: *real [0;7]*
nostocoida limicola: *real [0;7]*
sphaerotilus natans: *real [0;7]*
beggiatoa: *real [0;7]*
type 0092: *real [0;7]*
...

Figure 10: Parts of the case representation.

The aim of the corresponding solution description is the qualitative and quantitative identification of the species of microorganisms measured in the described bulking sludge problem. Therefore, the solution description contains one attribute for each major microorganism species relevant with respect to the sludge difficulty. The value range of these attributes is the interval of real values. These values correspond to a particular measure used when carrying out a microscopic examination of sludge samples. Here, the value 0 states that the respective microorganism is absent, while high values correspond to a high concentration. Though the described application can be characterized as a classification task, the solution description is not a simple class identifier like in common similar applications. Instead, the solution itself is again a complex object in form of a 11-dimensional vector. The consequences of this complexity will be discussed in more detail in the next section. Figures 10 and 11 partially show the used case representation and an exemplary case. The complete representation consists of 40 attributes describing the problem part and 11 attributes describing the solution part. However, many cases contain some unknown attributes, especially the cases taken from scientific literature. The corresponding uncertainty about the quality of this case data is then explicitly remarked in the already mentioned additional attributes.



Case "WWTP 43"

Problem

aeration: *surface*
N-elimination: *pre-operate*
high fate rate: *no*
O2-concentration: *4.5*
sludge index: *0.07*

successful counteractions: *none*
failed counteractions: *iron-salt.*

problem occurance: *creeping*
problem duration: *long*
problem localization: *aerated basin, sec. clarif.*
temperature: *10*
seasonal occurance: *cold season*

trust level: *good*

Solution

microtrix parvicella: *6.0*
nostocoida limicola: *4.0*
sphaerotilus natans: *0.0*
beggiatoa: *0.0*
type 0092: *6.5*
...

Figure 11: Parts of an example case.

3.3 Project Summary

The approach presented in this example is implemented in the research project ZERBERUS. In a preliminary stage of the project, the WWTP managers' experiences had been learned using a mail questionnaire. All relevant data was extracted from the questionnaires and transformed into cases. So, we gathered approximately 70 cases until now. Starting from this point, we divided the project into two major stages. On the first stage, we concentrated on the identification of the harmful microorganisms that caused the bulking sludge problem. A WWTP manager can specify a current problem and query the system's experiences to find out what might be the responsible bacteria. The second stage can generate an individual treatment solution for the queried problem situation. The solution will be based on the specific WWTP conditions and the retrieved solutions from the most similar experiences in the case base. The WWTP manager's feedback on the quality of the generated solution will be used to improve our system by a certain learning effect. If the generated suggestion – which counteraction to take – was successful or unsuccessful this new experience will be integrated in the case base. In 2003, the implementation of the DSS was completed (see *www.zerberus-online.de*).

4 Optimizing Prediction Accuracy

The success of any CBR application crucially depends on the quality of the employed similarity measure used to retrieve the most *useful* cases with respect to the current problem situation. Unfortunately, the actual utility of a case or its solution part, respectively, is first known, once it has been applied to the current problem situation. Hence, a similarity measure only represents a heuristics to approximate the a priori unknown utility function during retrieval. In CBR this heuristic is based on the assumption that similar problems have similar solutions, where the "similarity" between problems is often interpreted as similar appearance measured by simple distance metrics. However, as typical for any heuristics, its quality usually can be increased significantly if it is possible to incorporate meaningful domain knowledge. This can be realized in two different ways:

- One asks a domain expert to provide the required knowledge and then encodes it manually into the similarity measure, for example, by defining accurate local similarity measures and feature weights.
- One applies machine learning approaches to extract knowledge from particular training data, and to generate accurate similarity measures automatically.

In the example applications described previously, up to now we have applied the first approach. However, for several reasons we plan to optimize the employed similarity measures, and therewith also the prediction accuracy of our systems, by applying machine learning techniques:

- Depending on the particular application, we have to deal with very complex problem descriptions. Here, it is very hard to define an optimal similarity measure manually.
- Often the relationships and influences of the different parameters are unknown and hence also domain experts are unable to define accurate similarity measures.
- Even if the impact of the parameters is known in principle, the determination of quantitative aspects of similarity measures, such as exact feature weights or numerical parameters of local similarity measures, is a very difficult job that often can only be done intuitively.

A lot of approaches to learn one important part of the similarity measure, namely the feature weights, have been developed up to now [Wettschereck and Aha, 1995]. Núñez et al. [2002] have presented some statistical-based weighting techniques and they have evaluated them also using two environmental databases. However, all these approaches address classification tasks only. In general, they try to find a measure that assigns a higher similarity to cases containing a “correct” classification than to cases containing an “incorrect” classification. However, this approach is only applicable when the occurring classes are quite simple (e.g., only described by a simple class identifier represented by a string) and disjunctive. Nevertheless, as described in the precedent sections our “classes” are really complex objects (e.g. influent flow rate curves or 11-dimensional vectors). Therefore, a hard distinction between “correct” and “incorrect” classes is insufficient. In our scenario, cases or solutions, resp. can rather be judged as “better” or “worse” predictions of the actual solution while an exact match is very unlikely due to the complexity of the solution descriptions.

Another problem is that existing learning approaches are not suited to learn local similarity measures, which are usually represented as similarity functions or similarity tables. However, in particular local similarity measures can be used to encode a lot of domain knowledge in order to obtain a good approximation of the cases utility.

4.1 Learning from Utility-Feedback

To avoid this problem, we plan to apply a novel learning approach for optimizing the prediction accuracy of the described CBR applications in the field of wastewater treatment. The advantage of this alternative learning approach (see [Stahl, 2003] for a detailed description) is that it allows flexible learning of both, feature weights and local similarity measures and that it is not restricted to traditional classification tasks.

The basic assumption of this approach is the existence of some *similarity teacher* who is able to estimate the relative utility of retrieved cases with respect to a given set of training queries. This means the teacher has not to decide absolutely whether a given case is useful or not, but must only be able to compare given cases with respect to their utility resulting in statements like “case x is more useful than case y”. Such a kind of utility feedback leads to partially ordered lists of cases representing the desired outcome of a similarity-based retrieval for given training queries. The task of the learning algorithm is then to find a similarity measure leading to these optimal retrieval results as close as possible. Here, genetic algorithms have been applied successfully [Stahl and Gabel, 2003].

4.2 Exploiting Solution Similarity

To apply the described learning approach in the previously described application scenarios we need some similarity teacher who is able to provide the required utility feedback. Basically, such a similarity teacher has not necessarily to be represented by a human expert but can also be realized by some evaluation procedure. For our applications we plan to apply an approach based on a leave-one-out test and a novel concept, that we call *solution similarity* [Stahl and Schmitt, 2002] represented by an additional similarity measure that compares solution parts of cases instead of problem parts (see Figure 12). This concept allows us to exploit utility knowledge implicitly contained in the huge amount of available case data by measuring the utility of retrieved cases during a leave-one-out test. This approach assumes that it is much easier to define a reasonable solution similarity measure than a problem similarity measure. In fact, if we recall the structure of the solution parts occurring in our applications, we see that it is easy to define meaningful solution similarity measures. For comparing influent flow rate curves one could use, for example, the integral of the difference between two curves. Since we know the correct solution (here a curve) of a given problem during a leave-one-out test, such a measure allows us to estimate the prediction quality of retrieved curves and hence the utility of the corresponding cases. This allows us to generate utility feedback automatically to be used as input for the learning algorithm.

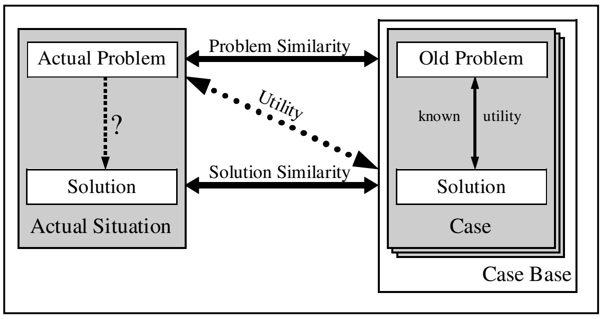


Figure 12: The concept of solution similarity.

5 Related Work

Recently, an increasing number of publications can be found that deal with WWTP control and optimization respectively, using knowledge-based techniques, sometimes also Case-Based Reasoning, e.g.:

Sànchez-Marrè [1996] presents the DAI-DEPUR system. The system is based on an integrated multi-level architecture for WWTP supervision in real-time. Like the SBR controller approach to use multiple case bases for the different control tasks, DAI-DEPUR maintains several knowledge bases that are connected for solving the global control task. In contrast to the SBR controller, DAI-DEPUR is kept more general with respect to the supported WWTPs. Furthermore, different knowledge-based approaches besides CBR are deployed.

Cortés et al. [2000] presents an approach to put forward a Knowledge Management Methodology for EDSS.

Fenner and Saward [2002] describe a methodology to produce a performance assessment model. They identify changes in the internal conditions of sewer pipes. Amongst other data, they build up a case base of performance histories. The past performances are used to predict suitable management strategies in the current situation.

6 Conclusions

Despite the fact, that CBR is a powerful technology, which has already proved its potentials in different industrial applications, CBR is not widely used in the field of wastewater treatment until now. Although approaches for optimization of existing plants attract more and more the attention, they are still based in almost all cases on Fuzzy Logic, Neuro Fuzzy, Genetic Algorithms, and Neural Networks. Nevertheless, there are some examples that the use of CBR in the field of wastewater treatment could be very promising, especially in case of Decision Support Systems and Real Time Control. Consequently, there is a good chance, that CBR will be far more common in environmental engineering during coming years.

Acknowledgments

The Project "ZERBERUS" was funded by the Ministerium für Umwelt und Forsten, Rheinland-Pfalz, Germany. The Project "Messel" is being undertaken with financial support from the Deutsche Bundesstiftung Umwelt, Germany. The authors are also greatly indebted to all persons, partners, and institutions, which made both projects possible.

References

[ATV, 1991, 2000] German Association of Water, Wastewater, and Waste (ATV). *ATV-Arbeitsblatt A 131 "Bemessung von einstufigen Belebungsanlagen" (Guideline ATV-A 131 for the Design of Aeration Plants)*, Hennef, Germany, 1991 and 2000 (new version).

[ATV, 1997] German Association of Water, Wastewater and Waste (ATV). *ATV-Merkblatt M 210 Belebungsanlagen mit Aufstaubetrieb (Guideline ATV-M 210 for the Design of SBR Plants)*, Germany, 1997.

[Bergmann *et al.*, 1999] R. Bergmann, S. Breen, M. Göker, M. Manago, and S. Wess. *Developing Industrial Case-Based Reasoning Applications. The INRECA-Methodology.* LNAI 1612, Springer, 1999.

[Cortés *et al.*, 2000] U. Cortés, M. Sànchez-Marrè, J. Comas, I. R-Roda, and M. Poch. *Knowledge Management in Environmental Decision Support Systems. Workshop Papers,* ECAI workshop on Binding Environmental Sciences and Artificial Intelligence (BESAI2000), 2000.

[Eikelboom, 2000] D.H. Eikelboom, *Process Control of Activated Sludge Plants by Microscopic Investigation.* IWA Publishing, 2000.

[Fenner and Saward, 2002] R.A. Fenner and G. Saward. *Towards Assessing Sewer Performance and Serviceability using Knowledge Based Systems.* Proc., 9TH Intern. Conference on Urban Drainage, Portland, 2002.

[IWA, 2000] N.N., *Activated Sludge Models ASM 1, ASM 2, ASM 2d and ASM 3.* IWA Publishing, 2000.

[Keudel and Dichtl, 2000] L.O. Keudel and N. Dichtl. *Settling Characteristics of Activated Sludge in Sequencing Batch Reactors obtained from Full-scale Experiences,* Proc. 2nd International Symposium on Sequencing Batch Reactor Technology, Narbonne, France, Vol. 1, pp. 75-83

[Lenz, 1999] M. Lenz. *Case Retrieval Nets as a Model for Building Flexible Information Systems.* Ph.D. Thesis, Humbolt University Berlin, 1999.

[Núñez et al., 2002] H. Núñez, M. Sànchez-Marrè, U. Cortés, J. Comas, I. Rodríguez-Roda, M.Poch. *Feature Weighting Techniques for Prediction Tasks in Environmental Processes.* ECAI Workshop on Binding Environmental Sciences and Artificial Intelligence (BESAI'2002). Lyon, France, July 2002.

[Sànchez-Marrè, 1996] M. Sànchez-Marrè. *DAI-DEPUR – An integrated Supervisory Multi-level Architecture. PhD thesis.* Universitat Politècnica de Catalunya, 1996.

[Schumacher, 2000] J. Schumacher and R. Bergmann. *An Efficient Approach to Similarity-Based Retrieval on Top of Relational Databases.* In : Proceedings of the 5th European Workshop on Case-Based Reasoning EWCBR 2000. Springer Verlag, 2000.

[Smyth, 1995] B. Smyth and M.T. Keane. *Remembering to Forget : A Competence Preserving Case Deletion Policy for CBR Systems.* In : Proceedings of the 14th International Joint Conference on Artificial Intelligence IJCAI-95. Morgan Kaufmann Publishers, 1995.

[Stahl, 2003] A. Stahl. *Learning Knowledge-Intensive Similarity Measures in Case-Based Reasoning. Ph.D. Thesis*, Technical University of Kaiserslautern, 2003.

[Stahl and Gabel, 2003] A. Stahl and T. Gabel, *Using Evolution Programs to Learn Local Similarity Measures.* In: Proceedings of the 5th International Conference on Case-Based Reasoning ICCBR-03. Springer Verlag, 2003

[Stahl and Schmitt, 2002] A. Stahl and S. Schmitt. *Optimizing Retrieval in CBR by Introducing Solution Similarity.* In*:* Proc. of the International Conference on Artificial Intelligence IC-AI'02. CSREA Press.

[Wiese *et al.*, 2004a] J. Wiese, J. Simon, and T.G. Schmitt. *Integrated Real-Time Control for a Sequencing Batch Reactor Plant and a Combined Sewer System.* In: Proc. of the 6th International Conference on Urban Drainage Modeling, Dresden, FRG, 2004.

[Wiese et al., 2004b] J. Wiese and J. Simon. *Dynamic simulation of a SBR plant and a Combined Sewer System – Description of the modeling procedure.* A practical application for the HSG simulation guidelines (publication planned)

[Wettschereck and Aha, 1995] D. Wettschereck and D.W. Aha. *Weighting features.* In: Proceedings of the 1st International Conference on Case-Based Reasoning, ICCBR-95. Springer Verlag, 1995.