Applying and Optimizing Case-Based Reasoning for Wastewater Treatment Systems

Jürgen Wiese [a], Armin Stahl [b] and Joachim Hansen [c]

[a] *Anlagen- und Sondermaschinen Automation GmbH (ASA GmbH) Robert-Bosch-Street 7, 32547 Bad Oeynhausen, Germany, E-mail: wiese@asagmbh.de*

[b] *German Research Center for Artificial Intelligence (DFKI) GmbH, Image Understanding and Pattern Recognition Group, Erwin-Schrödinger-Str., 67608 Kaiserslautern, Germany, E-mail: Armin.Stahl@dfki.de*

[c] *tectraa - Center for Innovative Wastewater Technology, University of Kaiserslautern, Paul-Ehrlich-Str. 14, 67663 Kaiserslautern, Germany, E-mail: jhansen@rhrk.uni-kl.de*

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). In this paper, we present several examples for real-time Control (RTC) tasks and decision support systems (DSS) for wastewater treatment (WWT), specifically based on case-based reasoning (CBR). Moreover, we present an approach for optimizing the prediction accuracy of these systems. The idea of this approach is to employ knowledge-intensive similarity measures instead of simple distance metrics. In order to facilitate the modeling of these measures resulting in lower deployment costs of the CBR systems, we propose a novel machine learning technique.

Keywords: Wastewater treatment, CBR, SBR, control, decision support systems, machine learning

1. Introduction

During recent years, a rising complexity of the problems in the area of WWT 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 between sewer system, WWTP, and receiving water etc. On the other hand, the technologies for measurements as well as the Computer-aided control devices (CACD) 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 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 determine 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 CBR a powerful technology, which has already proved its potentials in various industrial applications (see, e.g., [1]).

In this paper, we will present several examples for possible applications for CBR in WWT: In Section 2, we take a short look at several CBR approaches for WWT, which can be found in literature. In Section 3 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 (SBR). We will also present results of two offline CBR models, which have been developed to predict the influent flow rate and the sludge settling curves. Section 4 will focus on a DSS based on a CBR approach for

AI Communications
ISSN 0921-7126, IOS Press.

Identification and Counteraction for Harmful Microorganisms in WWTPs.

The most CBR systems applied in the WWT field nowadays employ general distance metrics to estimate the similarity between problem descriptions [9]. The only way to incorporate specific domain knowledge into such measures are feature weights. However, in order to guarantee accurate retrieval results in WWT domains, this is often not sufficient. In Section 5 we will outline methods for the optimization of the prediction accuracy of case-based WWT applications. We propose the usage of knowledge-intensive similarity measures that allow a more accurate modeling of the application specific requirements. In order to reduce the additional knowledge acquisition effort, we discuss a novel approach for learning such similarity measures from case data. Section 6 ends with some conclusions.

2. Related Work

Recently, an increasing number of publications can be found that deal with CBR and wastewater treatment, e.g.:

Krovvidy and Wee [7] developed an experimental CBR system, which was used to obtain the best treatment train for a theoretical wastewater treatment problem.

Kraslawski et al. [6] presented a case-based reasoning system for the selection of mixing devices for WWTP.

Sànchez-Marrè et al. [12] developed 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 (see Section 3) 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. [3] described an approach to put forward a Knowledge Management Methodology for EDSS.

Rodríguez-Roda et al. [11] presented a system for supervision and control of the activated sludge process of a continuous flow reactor plant (WWTP Girona, Spain), which is based on a CBR part, a knowledge based part and an adaptative controller. Rodriguez-Roda et al. [10] also presented a hybrid supervisory system to support the operation of WWTP Granollers (Spain).

Fenner and Saward [5] described a methodology to produce a performance assessment model. The model should identify changes in the internal conditions of sewer pipes. Amongst other data, they built up a case base of performance histories. The past performances are used to predict suitable management strategies in a new situation.

Comas et al. [2] developed a tool for automatic learning and reuse of knowledge in activated sludge processes, which is based on a wastewater treatment simulation model and a CBR tool.

In the next two sections, two CBR examples for RTC and EDSS will be described in detail. Both systems were developed for full-scale wastewater treatment plants.

3. Example - Real Time Control (RTC)

3.1. 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 CACD it is possible to adapt the duration of a cycle, the duration of the different steps within each cycle and the volumetric exchange ratio to the current requirements. Unfortunately, most of the SBR plants are still using fixed timer based control strategies; until now, measuring devices are predominately only used for monitoring.

3.2. Description of SBR-WWTP Messel

The WWTP Messel (Figure 2), which was put into operation in 2000, is a modern SBR plant and was designed for approx. 5,000 population equivalents. The plant was designed for biological phosphorus removal, nitrification, denitrification, and

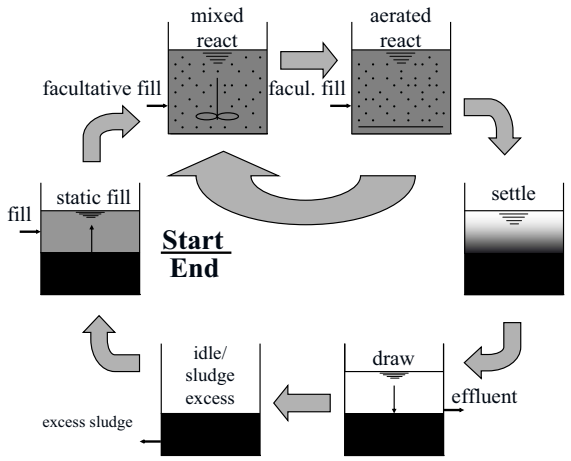

Fig. 1. The concept of SBR

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. 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. Even though the WWTP Messel is a small plant, it is a very complex technical system. Consequently, it is not quite easy to operate such a system efficiently. This particularly applies, because the WWTP is not permanently manned and is operated by only one person.

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 RTC approaches [19]. In the first part of the research project, very detailed computer models of the combined sewer system and the WWTP were used to develop several control strategies. These strategies are based on ammonia and nitrate sensors, as well as sludge blanket and suspended solids probes. The results of the WWTP simulation and full-scale operation show that it is possible to reduce the cycle duration during combined sewage flow 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

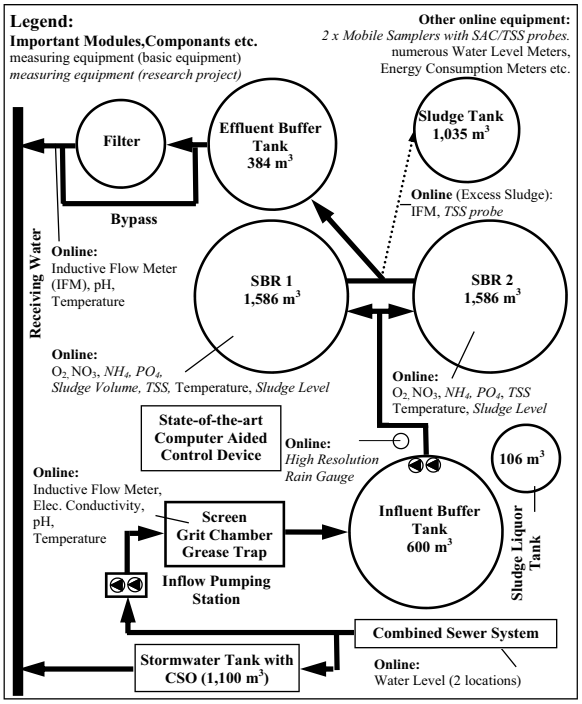

Fig. 2. Scheme of WWTP Messel

4 hours. With the help of the control strategies it was also 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 the cylce duration as early as possible.
- The optimization potential depends on several factors, e.g., influent load and wastewater temperature, but these parameters can vary significantly and sometimes rapidly.
- 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.

3.3. 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 steps of the cycle, which helps to 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.
- The practise shows that AI systems for WWTP must deliver clear and understandable results, because otherwise the system will not be accepted by the operators. Consequently, CBR has a clear advantage, because the methodology is easy to understand.
- Numerous of online monitoring data are available. With cycle durations between 3 and 9 h the database will grow and learn 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 [14]). Another possibility is to employ efficient retrieval approaches [8,13].

3.3.1. 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 step, which would have been sufficient to reach predefined effluent limit values. The opportunities for a case-based predictive SBR controller resulting from these circumstances are promising. 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 3 shows a part of our proposed system architecture. The different control strategies 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.

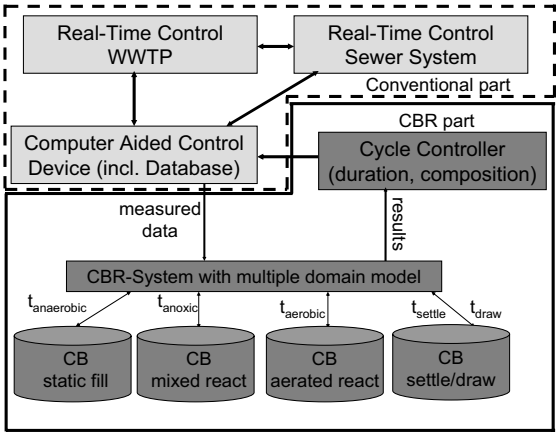

Fig. 3. Principle of the predictive SBR controller

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 a quasi-independent

series of process steps in the regarded cycle (i.e. a new step can only be started after its predecessor having finished), we can optimize (predict) the processing time of each individual step and add the predicted duration of each single step in order to obtain the overall cycle duration. 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 treatment step 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. 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.

3.3.2. Example "Influent Flow Rate"

For specific tasks and questions, 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 an equalization basin etc. But due to several reasons (e.g., infiltration water), even during phases of dry weather flow, the influent curve can vary significantly (Figure 4).

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 were 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 (Figure 5): In this figure the measured and the predicted influent flow rate curve for a 24 hour interval are almost identical.

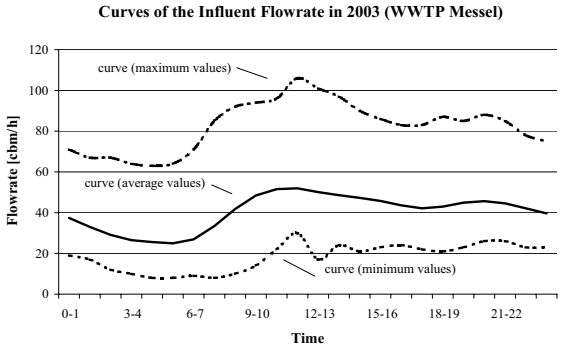

Fig. 4. Bandwith of influent flow rate curves during phases of dry weather flow in 2003 (WWTP Messel)

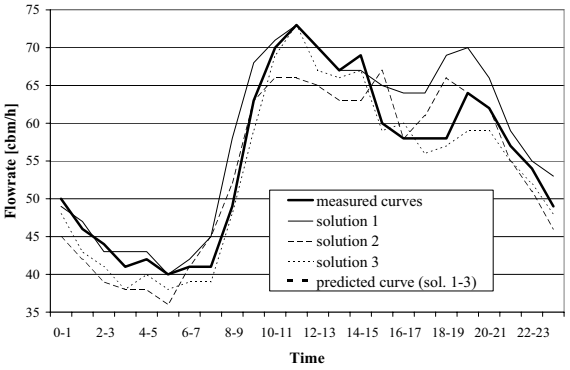

Fig. 5. Example for a good curve prediction

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.

3.3.3. Example "Settle/Decant"

During the settle and decant 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 official effluent limit values, 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 6. 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, the operational values are usually much better than the comparable design values. Therefore, sludge level 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 6); this could further increase the hydraulic capacity. Due to the high optimization potential of the settle and decant phase, it was decided to develop the CBR subsystem "Settle/Decant" first.

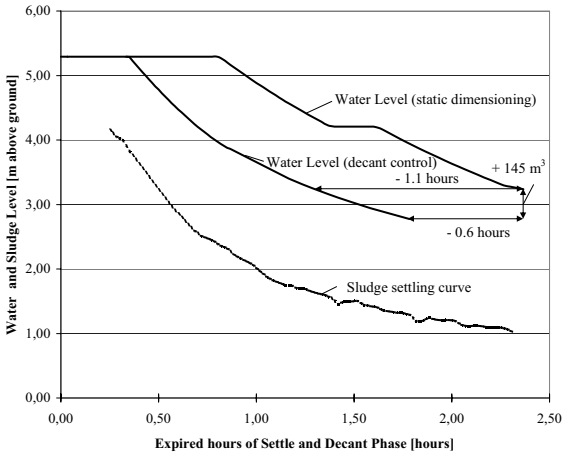

Fig. 6. Potential for optimization of settle and decant phase

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: The initial settling velocity mainly depends on the sludge volume at the beginning of the settle phase. 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. 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).

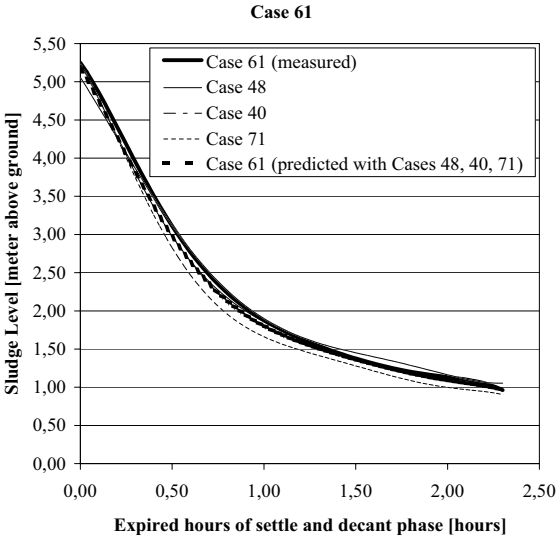

Fig. 7. Good prediction of the sludge settling curve

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 7 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 7. Figure 8 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 in case of a sludge displacement danger immediately.

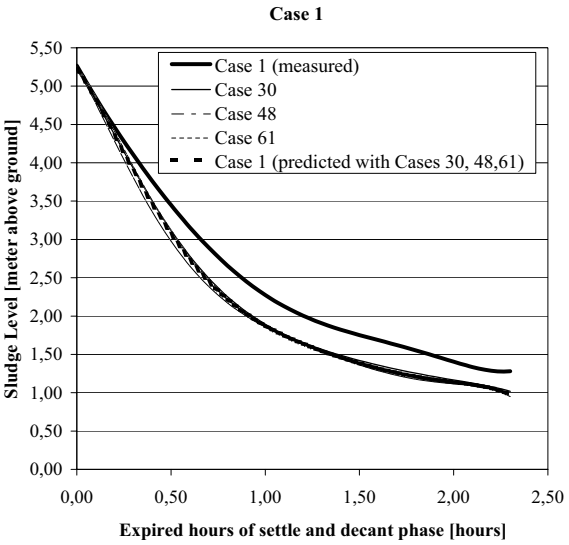

Fig. 8. Example for a worse prediction

3.3.4. Future Work

As a consequence of the good results reached with the 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.

4. DSS Harmful Microorganisms

4.1. 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 [4]. 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 problems consider this one of the most important problems to be solved. Nowadays, various approaches for counteractions exist to eliminate the problem-generating microorganisms [4], e.g., deployment of lime, polymers or pulverized lignite, installation of selectors, in/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 DSS 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 4.2 and 4.3.

4.2. 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. 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 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 (e.g., wastewater temperature, seasonal occurance), 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.

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 problem. The value range of these attributes is the interval of real values. These values correspond to the results of a microscopic examination of activated sludge samples. 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. The complete representation consists of 40 attributes describing the problem part and 11 attributes describing the solution part.

4.3. Project Summary

The approach presented in this example was developed as a part of 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*).

5. 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 some distance metric.

In many existing CBR applications in the WWT field, however, the employed distance metrics are often quite general and do not encode much knowledge about the underlying application domain. Examples of such metrics are the specific forms of the *Minkowski* metric or the *Heterogeneous Value Difference Metric (HVDM)*. Here, the only possibility to consider specific domain knowledge is the usage of feature weights. A comparative study on the use of these kinds of similarity metrics in environmental domains is given by Núñez et al. [9].

However, as typical for any heuristics, its quality can be increased significantly if it is possible to incorporate as much as possible meaningful domain knowledge. In particular when dealing with symbolic data, the use of feature weights is often not sufficient in order to obtain a good approximation of the underlying utility function. Knowledge about the relationship between the valid values of symbolic attributes is an important aspect to be considered when estimating the utility of given cases. Also for numeric attributes a domain-specific model of the similarities between the possible values can significantly increase the utility approximation. In order to be able to encode such knowledge efficiently, commercial CBR tools provide the possibility to model so-called *local similarity measures* for each individual attribute. Depending on the data type of the attribute, local similarity measures can be represented by similarity tables or special similarity functions (for more details see [15]). The similarity values computed by these local similarity measures finally have to be aggregated, e.g. by using a weighted sum.

Experiments in other CBR domains have shown that the additional use of local similarity measures can improve the retrieval quality significantly [15]. Also for the WWT applications described in the previous sections, such knowledge-intensive similarity measures have led to good results. However, in order to be able to model such measures accurately, one has to acquire a lot of specific domain knowledge.

This knowledge acquisition can be realized in two different ways. On the one hand, one may interview a domain expert in order to obtain the knowledge to be encoded into the similarity measure manually. On the other hand, one may also apply machine learning techniques in order to extract knowledge from some given training data 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 lo-

cal similarity measures, is a very difficult job that often can only be done intuitively.
- In particular in the WWT domains, each CBR application has its one characteristics due to the individual configurations and boundary conditions of the underlying WWT plants. Therefore, one usually cannot use exactly the same similarity model for each application but needs measures that have been optimized for the individual applications and plants.
- By learning the similarity measure more or less automatically one may decrease the deployment costs of a CBR application significantly. This is a very important aspect to be considered when thinking about a commercial deployment of the CBR technology in the WWT field.

A lot of approaches to learn one important part of the similarity measure, namely the feature weights, have been developed up to now [18]. Núñez et al. [9] 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. This approach is only applicable when the occurring classes are quite simple (e.g., only described by a simple class identifier) 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.

5.1. 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. The advantage of this alternative learning approach (see [15] 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 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 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 then is to find a similarity measure leading to these optimal retrieval results as close as possible. Here, genetic algorithms have been applied successfully [16].

5.2. Exploiting Solution Similarity

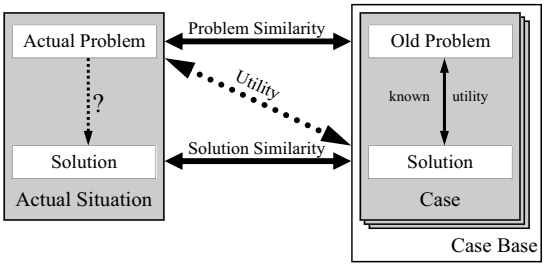

Fig. 9. The concept of solution similarity

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* [17] represented by an additional similarity measure that compares solution parts of cases instead of problem parts (see Figure 9). 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.

6. Conclusions

Despite the fact, that CBR is a powerful technology, which has already proved its potentials in different industrial applications, CBR is still 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, as discussed in this article there are some examples that show that the use of CBR in the field of wastewater treatment could be very promising, especially in case of DSS and RTC.

However, existing applications in the WWT field often achieve only suboptimal prediction accuracy due to the relative simple similarity measures typically used for retrieving relevant cases. In order to optimize the prediction accuracy of WWT applications, it seems to be necessary to employ knowledge-intensive similarity measures that take the application and domain specific characteristics into account. The drawback of such similarity measures is the additional modeling effort and the corresponding costs. In our point of view, the proposed machine learning approach allows to increase the benefit of case-based WWT applications by improving the prediction accuracy and simultaneously by reducing the deployment costs. This may help to make the CBR technology more interesting for commercial applications in the WWT field. 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 greatly indebted to all persons, partners, and institutions, which made both projects possible.

References

[1] R. Bergmann, S. Breen, M. Göker, M. Manago, and S. Wess. *Developing Industrial Case-Based Reasoning Applications: The INRECA Methodology.* State-of-the-Art-Survey, LNAI 1612. Springer, 1999.

[2] J. Comas, I. Rodríguez-Roda, M. Poch, K.V. Gernaey, C. Rosen, and U. Jeppsson. Demonstration of a Tool for Automatic Learning and Reuse of Knowledge in the Activated Sludge Process. In *Proceedings of the 2nd IWA Conference on Instrumentation, Control and Automation*, pages 237–245, 2005.

[3] U. Cortés, M. Sànchez-Marrè, J. Comas, I. Rodríguez-Roda, and M. Poch. Knowledge Management in Environmental Decision Support Systems. *AI Communications*, 14(1):3–12, 2001.

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

[5] R.A. Fenner and G. Saward. Towards Assessing Sewer Performance and Service-ability using Knowledge Based Systems. In *Proceedings of the 9th International Conference on Urban Drainage*, Portland, 2002.

[6] A. Kraslawski, T. Koiranen, and L. Nystroem. Case-Based Reasoning System for Mixing Equipment Selection. *Computers and Chemical Engineering*, 19:821–826, 1995.

[7] S. Krovvidy and W.G. Wee. Wastewater Treatment Systems from Case-Based Reasoning. *Machine Learning*, 10:341–363, 1993.

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

[9] H. Núñez, M. Sànchez-Marrè, U. Cortés, J. Comas, M Martinez, I. Rodríguez-Roda, and M. Poch. A Comparative Study on the Use of Similarity Measures in Case-Based Reasoning to Improve the Classification of Environmental System Situations. *Environmental Modelling and Software*, 19(9):809–819, 2004.

[10] I. Rodríguez-Roda, M. Sànchez-Marrè, J. Comas, J. Baeza, J. Colprim, J. Lafuente, U. Cortés, and M. Poch. A Hybrid Supervisory System to Support WWTP Operation: Implementation and Validation. *Water Science and Technology*, 45(4-5):289–297, 2002.

[11] I. Rodríguez-Roda, M. Sànchez-Marrè, U. Cortés, J. Comas, and M. Poch. Development of a Case-Based System for the Supervision of an Activated Sludge Process. *Environmental Technology*, 22:477–486, 2001.

[12] M. Sànchez-Marrè, U. Cortés, J. Lafuente, I. Rodríguez-Roda, and M. Poch. DAI-DEPUR: A Distributed Architecture for Wastewater Treatment Plants. *Artificial Intelligence in Engineering*, 10(3):275–285, 1996.

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

[14] 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.

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

[16] 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'2003)*. Springer, 2003.

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

[18] D. Wettschereck and David W. Aha. Weighting Features. In *Proc. of the 1st International Conference on Case-Based Reasoning*. Springer, 1995.

[19] J. Wiese, J. Simon, and T.G. Schmitt. Integrated Real-Time Control for a Sequencing Batch Reactor Plant and a Combined Sewer System. In *Proceedings of the 6th International Conference on Urban Drainage Modeling*, pages 325–334, Dresden, Germany, 2004.