PROCESS DESIGN AND CONTROL

Automatic Knowledge Acquisition from Complex Processes for the Development of Knowledge-Based Systems

Ignasi R-Roda,*,† Joaquim Comas,† and Manel Poch†

Chemical and Environmental Engineering Laboratory, University of Girona, Campus Montilivi s/n, E-17071 Girona, Spain

Miquel Sànchez-Marrè‡ and Ulises Cortés‡

Software Department, Technical University of Catalonia, C/Jordi Girona 1-3, E-08034 Barcelona, Spain

A knowledge-based system for the supervision of a wastewater treatment plant was successfully applied to a full-scale facility. The key factor of this supporting tool development was the two-phase methodology used to acquire and fix the knowledge into the knowledge base. Both phases of the methodology are presented in the paper; the first consists of literature reviews and site interviews with domain experts, while the second is based on machine learning tools and is subdivided into four steps: data handling, classification, interpretation, and codification. The aim of this two-phase methodology is meant to ease the knowledge acquisition process. The main objective is to find the relevant issues and then reduce the space of search according to the target facility. Also, this methodology allowed the user to explore the data space to discover, if any exist, new pieces of knowledge. This methodology can be generalized to acquire specific knowledge from any (bio)chemical process, improving the development process and the efficiency of the supervisory knowledge-based system.

Introduction

The multifaceted nature of many (bio)chemical processes suggests that their efficient management and control cannot be based on a single classical perspective. These processes usually behave dynamically over time; their accurate management strategies involve interactions between different aspects, e.g., kinetics, catalysis, transport phenomena, separations, etc., and require the use of heuristic reasoning and previous experience to be solved at a multiparameter level. Although progress in control engineering, computer technology, and process sensors has enabled automatic control improvement, approaches other than a straightforward application of classical control theory are needed to efficiently manage complex processes. To provide engineering tools that solve complex problems, one needs to integrate an array of specific intelligent systems and numerical computations, allocating the detailed engineering to numerical computations, while delegating the logical analysis and reasoning to supervisory intelligent systems.[1] In this sense, knowledge-based systems (KBS) have shown promising results as supporting tools for complex process management, because of their capabilities in representing heuristic reasoning and in working with uncertain and qualitative information.

A KBS is loosely defined as an interactive computer program that attempts to emulate the reasoning process of experts in a given domain—a group of processes over which the expert makes decisions. The KBS has two main modules: the knowledge base (KB) and the inference engine. The KB includes the overall knowledge of the process as a collection of facts, methods, and heuristics, which are usually codified by means of heuristic rules. The inference engine is the software that controls the reasoning operation of the KBS, chaining optimally the knowledge contained in the KB.[2] The most successful KBS have been those that are specific to a narrowly defined problem, which are scalable to a continuously expanding KB and which integrate diverse sources of requisite knowledge.[1] The acquisition of the knowledge included in the KB is the core and also the bottleneck of the KBS development. It involves eliciting, analyzing, and interpreting the knowledge that experts use to solve a particular problem (by solve we mean to diagnose the situation, to detect the cause, and to propose the suitable actuation).

Management of biological wastewater treatment plants (WWTP) is a good example of a complex biotechnological process where KBS can play an important role as the supporting tool. The objective of WWTP is to provide a regulated water effluent with a low contaminant load to cause the least impact on the quality of the receiving water ecosystem. The basis of the treatment lies on the biological oxidation of biodegradable organic matter and nutrients from influent wastewater into stabilized, low-

* To whom all correspondence should be addressed.

† E-mail: {ignasi;quim;manel}@lequia.udg.es. Telephone: +34 972 418281. Fax: +34 972 418150.

‡ E-mail: {miquel;ia}@lsi.upc.es. Telephone: +34 93 4017016. Fax: +34 93 4017014.

10.1021/ie000528c CCC: $20.00
Published on Web 06/23/2001

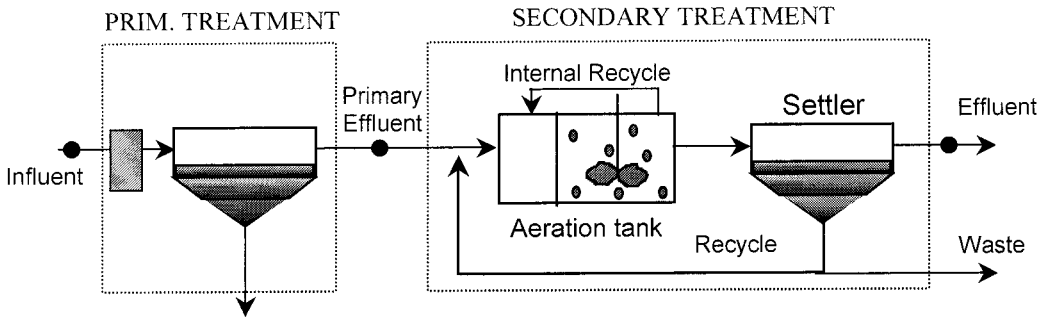

Figure 1. Flow sheet of the WWTP water line.

Table 1. Quantitative Data Measured in the WWTP

source	variable	sampling location	frequency of sampling
analytical	COD, BOD, TSS, NH_4^+, TKN, NO_2^-/NO_3^-, Cl^-, and P	influent, primary effluent, and effluent	daily
	MLSS, MLVSS, and V30	aeration tank and recycle	daily
sensors	pH and Cond	influent and effluent	online
	DO	aeration tank	online
	flow rates (flow)	influent, primary effluent, effluent, recycle, internal recycle, and waste	online
global	SRT, SVI, and F/M		daily

energy compounds, maintaining a mixture of microorganisms and supplying oxygen by aerators.

Mechanisms to accurately control WWTP have been developed over the last years, yet the problem is still far from being solved. This is due to several features of the process such as (a) the changes in both the quantity and quality of the influent; (b) the variation over time of the living catalyst (the microorganisms) in both the quantity and the number of species; (c) limited detailed knowledge of the biological treatment mechanism; (d) scarce and often unreliable use of online analyzers ; (e) delays in obtaining the analytical results of some parameters; and (f) data relative to the process that are subjective and cannot be numerically quantified. To guarantee control of such a complex process, it seems reasonable to link classical control techniques to KBS, which handle the particular characteristics of the process. Literature quotes several examples of KBS that have been applied to improve the management of chemical processes and WWTP in particular.[3−8]

However, most of these KBS never fully succeeded for two main reasons: they were too complex, and the available knowledge could not be captured in reliable models and advisory systems.[9] In fact, the performance of KBS as the supporting tool to enhance the control of WWTP has been conditioned to the quality of knowledge included in the KB. For a given complex process such as WWTP, the KB should include two kinds of knowledge: (a) general knowledge, i.e., theoretical, already existing information from a set of similar processes, and (b) specific knowledge, i.e., practical, information from the particular study process. The stepping stone that involves processing the information to acquire and fix the knowledge has led to the development of different supporting tools that avoid the use of heuristic knowledge. These techniques belong to different areas of the artificial intelligence, such as case-based reasoning[10,11] and soft computing.[12−15] Other authors suggest the integration of different techniques to develop hybrid systems as the most optimal solution.[16−18]

This paper presents a particular case in which the application of a modular KBS has contributed enhancement of the control of a full-scale WWTP. The real value of this KBS rests with the size and the quality of its KB, which integrates general and specific knowledge. The KB was built according to a two-phase methodology: the first phase is based on the traditional way to acquire and fix knowledge from a complex process (which consists of literature reviews and site interviews with domain experts), while the second phase is based on machine learning tools[19,20] to acquire specific knowledge from the database of the target facility.

The aim of this two-phase methodology is meant to ease the knowledge acquisition process. The main objective is to find the relevant issues, reducing the information-processing requirements and the space of search according to the target WWTP (ensuring a compact KB). Also, this methodology will help to explore the data space to discover, if any exist, new pieces of knowledge.

The first part of the paper describes the facility where the KBS has been applied, together with its database. Then, both phases of the knowledge acquisition methodology are detailed, analyzing in particular the results of the second phase, while the modular structure of the developed KBS is outlined. Finally, conclusions and remarks of the paper are listed.

Description of the WWTP and Its Database

The facility includes primary and secondary treatment (Figure 1) and uses the modified Ludzack−Ettinger process[21] to remove the organic matter, suspended solids, and nitrogen from municipal and industrial wastewater of about 100 000 inhabitants equivalents.

Plant operators carry out an exhaustive characterization of the influent and effluent water and sludge quality, including both quantitative and qualitative information. Quantitative data are provided by online sensors and by the analytical determinations realized in samples collected daily from different locations of the facility. Combinations of these measurements allow the calculation of global variables: sludge residence time, sludge volume index, or food-to-microorganism ratio. Qualitative data include daily in situ observations by the operators and microscopic determinations of the biomass, which is measured twice a week and consists

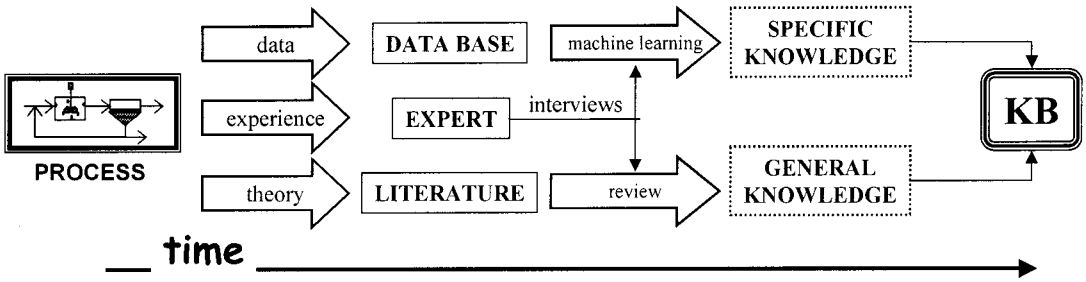

Figure 2. Scheme of the sources and methods followed to acquire knowledge from the WWTP.

Table 2. Qualitative Data Measured in the WWTP

source	variable	sampling location	frequency of sampling
floc characterization	morphology, size, and filament effect on structure	aeration tank	twice a week
microfauna identification and abundance	*Aspidisca*, *Euplotes*, *Vorticella*, *Epistylis*, *Opercularia*, Carnivorous cilliates, dominant protozoa keygroup, biodiversity of the microfauna, flagellates >20 μm, flagellates <20 μm, *Amoebae*, *Testate amoebae*, rotifer	aeration tank	twice a week
filamentous bacteria identification and abundance	zooglea, *Nocardia*, 021N/*Thiothrix*, type OO41, *Microthrix parvicella*, filamentous bacteria dominant, number of different filamentous bacteria	aeration tank	twice a week
observations	presence of foam, floating sludge, settling test	aeration tank and settler	daily

Table 3. List of Troubleshooting Included on the KB after the First Knowledge Acquisition Phase

	secondary treatment	
primary treatment	nonbiological origin	biological origin
old sludge	tensoactives scum	filamentous bulking
septic sludge	storm	slime viscous bulking
sludge removal systems break	hydraulic shock	foaming (*Actynomicetes*)
clogged pumps or pipes	overloading	deflocculation (pinpoint)
low efficiency of grit removal	underloading	deflocculation (dispersed growth)
high density of primary sludge	aeration problems	toxic shock
inadequate sludge purges	imbalanced flow rates	N/D (rising sludge)
hydraulic shock	mechanical and electrical problems	N/D (loss of nitrification)
high solids loading	clarifier problems	transition state
other mechanical problems	transition state	

of floc characterization, microfauna identification and abundance, and filamentous bacteria identification and abundance. Tables 1 and 2 summarize all of the available information, distinguishing the source of measurement for each variable and the sampling location.

Knowledge Acquisition

Figure 2 shows the main sources and methods that were used to process the information in order to acquire the knowledge included in the KB of the KBS. The knowledge was structured in decision-tree fashion. A decision tree is a set of nodes and arcs, where each node is a question related to a particular set of information, while each arc is a possible value for that information. Heuristic rules can then be easily extracted from these decision trees, represented as IF a set of conditions is true, THEN certain conclusions can be drawn. The knowledge acquisition process was subdivided into two different phases.

First Phase. The general knowledge of the process was mainly acquired from literature reviews. It was synthesized in a set of decision trees, which included diagnosis, cause identification, and actuation strategies, for the typical troubleshooting that interferes with the performance of the biological treatment of wastewater. Literature reviews offer a wide range of problems related to the domain, along with a correspondingly wide range of potential solutions to these problems.

To modify and adapt these general decision trees to the particularities and data of this particular facility, it was necessary to schedule a series of interviews with the experts of the WWTP (manager and operators). This step, which was long and iterative, implied drawing, modifying, and checking every inference path considered in the KB, according to the characteristics of the facility (i.e., rejecting those paths that could not be evaluated in the target WWTP because of the lack of some sensor or analytical device to determine certain variables).

The result of this first phase was a group of decision trees that were more specific and compact, containing some particularities of the facility. However, interviews with experts have, in general, a poor productivity, from 2 to 5 units of information/session. This is especially due to the subjective and incomplete point of view expressed by the experts, who do not find it easy to formulate how to solve the day-to-day problems. Experts tend to emphasize recent, first-hand experience, providing a limited and potentially biased view of the domain.[2]

Table 3 shows the list of all operational troubleshooting that was considered. The list covers primary and secondary treatment, distinguishing between the nonbiological and the biological origin of the problems. The latter origin causes a decrease in the biological reactor performance or dysfunctions in the secondary settlement.

Figure 3 shows an example of a decision tree synthesized in this first knowledge acquisition phase. The tree, part of the overall tree concerning deflocculation troubleshooting, reflects the main paths to diagnose the dispersed growth, to detect the cause(s), and to propose

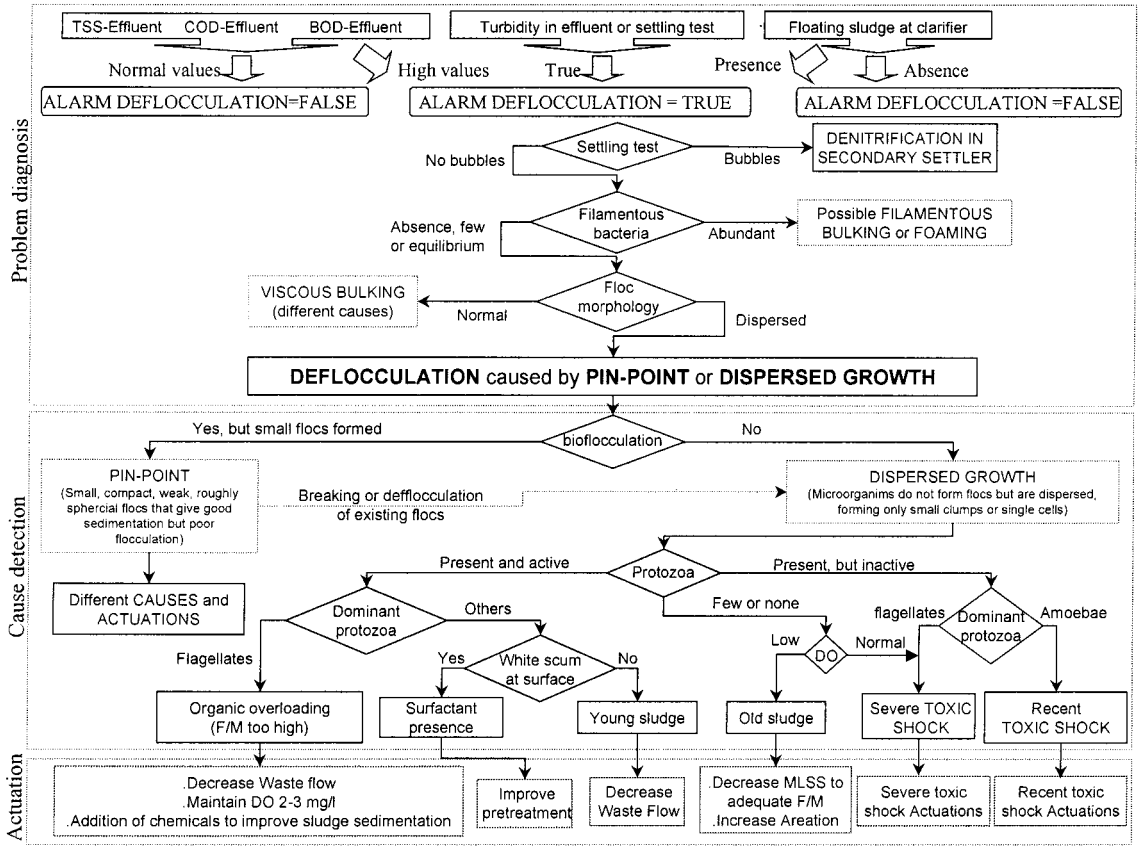

Figure 3. Example of a simplified decision tree to deal with the deflocculation problem.

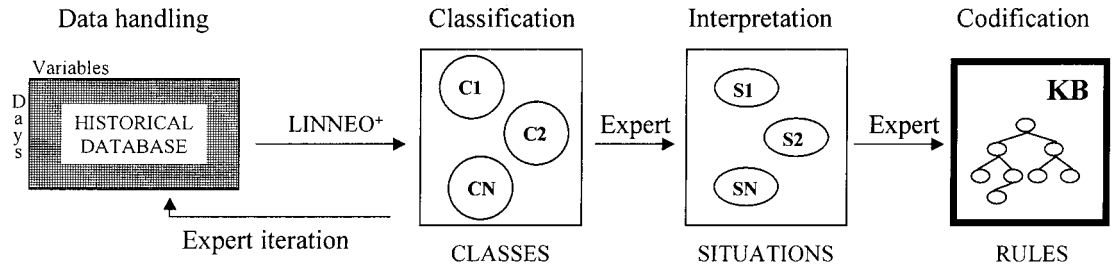

Figure 4. Scheme of the second-phase knowledge acquisition process, outlining the main developmental steps.

the suitable actuation. Note that the inference can start from all kinds of data (qualitative or quantitative, chemical or biological), with it being possible to explore different branches to reach the same conclusion. This feature of the KB allows the evaluation of the process despite the lack of some relevant data.

Every path of the decision tree was codified as a heuristic rule into G2,[22] a commercial shell to develop real-time expert systems that include the inference engine. The collection of those trees is in itself an operational memory of the process, and each tree shows clearly the decision-making process performed at each step for any situation lived in the plant. This is a very important piece of knowledge that has to be compiled and automated to speed the KBS's response. The final structure of the KB was modular, using meta rules to determine which other module/rules are to be used to solve the problem.

Second Phase. To overcome the caveats and limitations of the traditional way to acquire knowledge from the process, a second phase was proposed. As mentioned, it is based on machine learning tools. The aims are to identify the operational situations of the facility from the data as a class, then to create heuristic rules to describe those objects, and, finally, to integrate this knowledge in the KB. This second phase of the methodology is divided into four main steps: data handling, classification, interpretation, and rule codification (Figure 4).

Data Handling Step. In this study, the historical database covered a representative period of 241 consecutive days. Data from each day were considered as a particular data set. Variables considered in the database correspond to those listed in Tables 2 (quantitative information) and 3 (qualitative information). The most relevant feature of the database was the high presence of missing values, especially for qualitative variables. Nevertheless, consideration of qualitative information was of critical relevance in this study because of its great influence on the activated sludge characterization and behavioral understanding. A previous statistical analysis was carried out to filter some signals and to remove redundant variables of the database. Then, a careful selection of the most relevant variables was done according to the plant manager criteria (i.e., the expert of the process). Table 4 shows

Table 4. List of the Relevant Variables That Were Classified To Acquire Knowledge in the Second Phase

sampling location	variables
influent	flow, pH, COD, BOD, TSS, TKN, NH_4^+, Cond, Cl^-
primary effluent	flow, COD, BOD, TSS
aeration tank	MLSS, MLVSS, SVI, F/M, SRT, qualitative information (including floc characterization, biodiversity, microfauna and filamentous identification and abundance, and operator observations)
secondary effluent	flow, pH, COD, BOD, TSS, TKN, NH_4^+, Cond, observations

Table 5. Values of the Most Relevant Variables for Classes 1, 5 and 8 from the Classification Step

sampling location	variables (units)	class 1	class 5	class 8
influent	flow (m^{-3}·day^{-1})	7650	13563	11270
	COD ($g·m^{-3}$)	725	755	812
	TSS ($g·m^{-3}$)	290	293	314
	Cl^- ($g·m^{-3}$)	307	3360	312
	Cond ($\mu S·cm^{-1}$)	1350	9806	1170
primary effluent	COD ($g·m^{-3}$)	381	460	509
	TSS ($g·m^{-3}$)	81	109	92
aeration tank	MLSS ($g·m^{-3}$)	1750	2030	1900
	dominant protozoa	sessile ciliates	sessile cilliates	small flagellate
	zooglea	none	abundant	none
	dominant filamentous	none	*Zooglea*	*Nocardia*
	biodiversity	4	5	3
effluent	COD ($g·m^{-3}$)	49.9	65	91
	TSS ($g·m^{-3}$)	8.9	13	37
global	SVI ($g·mL^{-1}$)	394	433	250
	SRT (days)	11.8	8.3	6
	F/M (kg of BOD·kg of $MLSS^{-1}$·day^{-1})	0.37	0.3	0.32

some of the 65 variables that were finally selected. The final database was then structured in a heterogeneous matrix, which contained the days in rows and the variables in columns.

Classification Step. Once the historical database of the WWTP was structured in a summary matrix, it was fed to a clustering tool in order to conduct the classification step. Each cluster should represent a group of days characterized by a particular situation of the facility. In this case we used the software Linneo$^+$,[23] a semiautomated machine learning tool based on classification methods for ill-structured domains. It is an unsupervised iterative clustering method, which determines useful subsets of data, assuming that observations vary in their membership degree with regard to each possible class. Linneo$^+$ assumes that the initial number of classes is not known. It uses the conventional concept of distance as a fuzzy similarity value to build classes that can change (i.e., augment or diminish) during a learning process. The authors had previous experience with Linneo$^+$, which had been successfully compared to a statistical method (*K* means) in a clustering study of quantitative data from a full-scale WWTP.[23]

Linneo$^+$ works by defining a space of *n* dimensions, where *n* is the number of variables included in the database (65 in this study). Within this 65-dimension space, all days are grouped in different hyperspherical classes, which are specified by a center and a radius (the size of the class). To establish the case for a day to belong to a particular class, the software uses a "distance" criterion, which also determines the shape of the clusters. At the beginning of the classification process, the classes are still undefined; thus, the first day is taken and placed within the space to form the first class. The center of this class is calculated based on the values of the variables that describe the day. Then, the second day is placed within the same space. When it is close enough to the first center, then the 2 days will belong to the same class, with its new center being recalculated with the mean values of the variables of both days. If the distance between the 2 days is too large, a new class is formed. Linneo$^+$ offers a set of tools to perform previous data analysis to iteratively estimate a radius, which let the user obtain a certain number of classes. Also the user is able to seed the space using relevant objects (days) to build upon the classification.

In this study, Linneo$^+$ used a radius equal to 10 and the generalized Hamming[24] as distance criteria. On the basis of this radius value, 12 classes were obtained from the WWTP database. These classes corresponded to a reasonable number of differentiated situations (according to the number of expected situations contained in the database). All of the 12 situations were clearly identified and labeled by the experts. Some examples of the results from the classification step are shown in Table 5, with the values of the most relevant variables corresponding to the center of classes 1, 5, and 8.

Interpretation Step. Once the classification process is done, it is necessary to examine and interpret all of the classes identified with the clustering tool. The plant manager performed the interpretation and labeling. On the basis of his experience, he related the characteristics of each class to a particular episode of the WWTP during the period of time considered in the database. Table 6 shows the list and interpretation of the 12 classes, some of them containing also a set of substates, and the number of days pertaining to each class.

Class 1 is the most common situation (i.e., highest number of days), and the values of the variables of the center do not differ significantly from the mean values. This class was labeled as normal or optimal plant operation. Classes 2−4 have abnormal loading rates. Class 2 contains days with a high influent rate combined with a dilution of the contaminants (rainy and stormy days). Class 3 is characterized by a low loading rate together with a normal influent rate (underloading). Class 4 makes reference to those days with higher loading rates than expected (organic overloading). Class 5 was interpreted as a chloride shock leading to a very high conductivity. Classes 6 and 7 correspond to periods of full N/D episodes and rising sludge (uncontrolled denitrification in the secondary settler), respectively. Class 8 (deflocculation) also describes a situation with poorly settling and compacting activated sludge. This

Table 6. List of Classes Obtained by Linneo$^+$ and Their Corresponding Interpretation According to the Criteria of the Plant Manager

class no.	no. of days	interpretation
		Normal WWTP Operation Days (Two Substates)
1	102	days with normal WWTP operation
	22	days with optimal WWTP operation
		High Influent with Dilution of Contaminants (Two Substates)
2	2	rainy days
	4	stormy days
3	21	organic underloading
4	2	organic overloading
5	2	chloride shock
6	2	nitrification/denitrification (N/D)
7	5	rising sludge
8	5	deflocculation (effluent turbidity due to small flagellates)
		Filamentous Bulking Sludge (Two Substates)
9	3	due to *Thiothrix* (affecting the effluent)
	2	beginning of a strong bulking sludge episode due to *Thiothrix* (transition to bulking)
10	9	viscous bulking sludge due to *Zooglea* and foaming sludge due to *Microthrix*
		Foaming Sludge (Four Substates)
11	23	*M. Parvicella* (with normal microfauna biodiversity)
	20	*M. Parvicella* (with low microfauna biodiversity and small flagellates dominant)
	8	*Nocardia* with abundant small flagellates
	6	severe episode of foaming sludge due to *Nocardia* affecting the process (loss of solids)
12	3	WWTP configuration change

situation is caused by the total absence of filamentous organisms leading to small flocs (pinpoint floc) or because of a very low production of exopolymer (bioflocculation does not occur).

Classes 9−11 correspond to abnormal situations due to the proliferation of different filamentous bacteria leading to different settling interferences depending on the causative bacterial species: bulking sludge (if the dominant filamentous bacteria was *Thiothrix*), foaming (associated with the presence of *Microthrix parvicella* or *nocardia*), and viscous or slime bulking (if *Zooglea*). Among these, class 9, labeled as filamentous bulking sludge, describes days where sedimentability of the activated sludge is significantly poor (high values of SVI), leading to a poor clarified effluent with loss of solids (high values of COD and TSS at the effluent). This effect was produced by the abundant presence of *Thiothrix*, which extends from flocs into the bulk solution and interferes with compaction, settling, and thickening processes of activated sludge. One substate of class 9 corresponds to an episode of a tendency to bulking occurring in the plant. Class 11 corresponds to days with settling dysfunctions caused by the growth of *M. parvicella* and/or *nocardia* (organisms with hydrophobic cell surfaces). These organisms form air bubble−floc aggregates less dense than water that float to the aeration tank surface to form a greasy thick brown foam. Four different substates were characterized according to the biomass state and the degree of affectation of the effluent. Class 10 is a situation that combines foaming sludge due to *M. parvicella* and viscous bulking sludge. The latter, also called zoogleal bulking, occurs when too much of the extracellular polymer that contributes to bioflocculation is produced, resulting in a viscous, poorly settling and compacting sludge.[25] Some global variables (i.e., F/M and SRT) are necessary to characterize these situations. Finally, class 12 indicates a change in the WWTP process configuration.

Codification Step. All of the knowledge and information acquired in the previous steps must be finally codified by means of heuristic rules. This process can be performed by hand or by use of automatic tools (e.g., GAR[26]). These rules, which are specific to the WWTP, replace general inference paths in the decision trees of the KB. The objective of this step is to reduce the number of essential nodes and variables that must be evaluated to reach the final conclusion. For instance, class 8 (interpreted as deflocculation) was well characterized by (a) a cloudy effluent (high value of COD or TSS at the effluent), (b) a normal value of SVI, and (c) small flagellates as the predominant keygroup of the mixed liquor microfauna. In contrast, to reach the same diagnosis using the general decision tree, it was not only necessary to have high values of TSS, COD, or BOD at the effluent, a true presence of turbidity in effluent or settling test, and the presence of floating sludge as a clarifier, but it was also required information about the presence of bubbles in the settling test analysis, the number of different filamentous bacteria, and the floc morphology (see Figure 3). Following this example, Table 7 shows the specific rule that can replace the general decision tree in the KB to diagnose deflocculation problems in the study WWTP.

Table 7. Specific Rule Resulting from Class 8

IF $COD_{effluent} > 70\ g \cdot m^{-3}$ or $TSS_{effluent} > 20\ g \cdot m^{-3}$ and SVI = normal and dominant protozoa keygroup = "small flagellates"	THEN deflocculation = true

This second phase of the knowledge acquisition process also enables the user to discover new knowledge that was not acquired from the literature because it is specific to the facility. For instance, from the interpretation of values from class 5, we were able to codify the rule shown in Table 8. This rule is a clear consequence of the experience acquired in the target WWTP and reflects the detection of chloride shock situations.

Table 8. Specific Diagnose Rule for Chloride Shock in the WWTP

IF	$COD_{influent} < 1100\ g \cdot m^{-3}$ and $Cond_{influent} > 1500\ g \cdot m^{-3}$ or $Cl^-_{influent} > 400\ g \cdot m^{-3}$	THEN chloride shock = true

The use of a machine learning tool in front of the traditional way to acquire specific knowledge of the process to build the KBS (second phase vs first phase) has an additional advantage, that is, an increase in the understanding of some cause−effect relationships between the presence of microorganism and operational problems in the WWTP. Table 9 gives some examples of specific rules based on good biological−operational correlation found in this study. These rules, only hinted in the literature, can be used as meta rules to conduct the reasoning throughout the KB.

Finally, this phase can also assist in the boundaries adjustment among the different modalities of each quantitative variable considered in the KB. Note that the set of conditions of the heuristic rules can be based on modalities, with it being necessary to determine when the value of a quantitative variable (e.g., COD at the influent) is low, normal, or high and to register any qualitative observations (e.g., microorganism presence) as few, regular, or abundant.

KBS

All of the specific knowledge acquired during this second phase was combined to rebuild the KB of the KBS. Figure 5 shows the modular structure of the KB. Each module has a specific task and consists of different sets of rules, methods, and/or procedures. The whole heuristic knowledge to carry out the management of the WWTP is grouped into three main modules: (a) the fault detection module including all of the knowledge related to mechanical equipment or electrical failures, e.g., clogged pumps or pipes, electrical fault detection, air system failure, sludge removal system break, etc.; (b) the operational problems module including the knowledge of diagnostic techniques for primary or secondary treatment troubleshooting (these problems were also divided into biological and nonbiological nature depending on the cause of the dysfunction, e.g., old sludge, scum, filamentous bulking, slime viscous bulking, deflocculation, storm, foaming, hydraulic shock, underloading, overloading, toxic shock, low pretreatment efficiency, etc.); (c) the transition states module containing all of the knowledge necessary to cope with transient states that can result in the undesirable problems contained in previous modules.

The KBS gathers all kinds of data collected in the facility: online signals, offline analytical determinations from water and sludge samples from the facility, and qualitative data corresponding to biomass microscopic examination and operator observations. For many processes (e.g., WWTP), solution features are not supplied as data but must be inferred by data abstraction.[27] Then, the data abstraction module of the KBS carries out a qualitative abstraction of the whole input data (e.g., IF TSS-effluent is $>35\ g \cdot m^{-3}$, THEN TSS-effluent is high). On the basis of this qualitative abstraction, the meta-diagnosis module determines which tree and diagnosis paths (i.e., decision rules or procedures) are explored to infer the situation. When the situation is identified, the inference engine of the KBS launches the set of rules that must determine the causes of this process state. If the right cause is determined with a certain degree of certainty, a specific actuation is recommended. In the case where the cause is not successfully determined, this causes an impasse and a nonspecific actuation must be proposed to soften the effects of the trouble. Finally, if different problems have been diagnosed, then the integration module accomplishes the integration of the different information and solutions provided by the KBS.

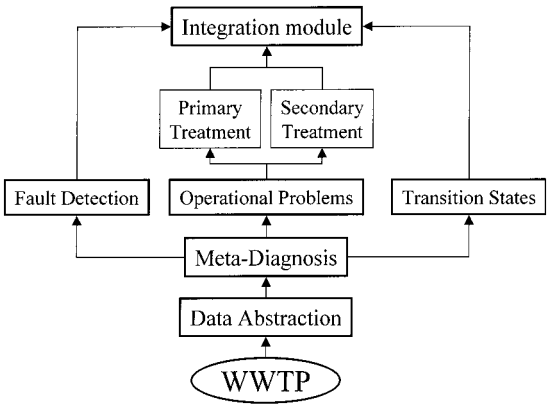

Figure 5. Modular structure of the KBS.

Conclusions

The paper presents a particular case in which the application of a KBS has contributed to successfully support the management of a full-scale WWTP. The success of this KBS rests with its compact KB, which combines general knowledge of the process and specific knowledge from the target WWTP, both acquired in a two-phase methodology. The first phase is based on literature review and site interviews with the manager and operators of the facility. The second phase, based on machine learning tools, consists of four main steps: data handling, classification, interpretation, and codification.

Among the particularities of this KB are (a) the reduced number of essential nodes and variables that must be evaluated to reach the final conclusion, (b) the discovery of new pieces of knowledge that were not acquired from the literature because they were specific to the facility, (c) the possibility of adjusting the boundaries among the different modalities of each

Table 9. Correlations Found in the Classification Process among Some Parameters

IF	microfauna abundance = "high biodiversity"	THEN	BOD and TSS at the effluent = "low values" (good quality of effluent)
IF	dominant protozoa = "sessile/crawling cilliates"	THEN	BOD and TSS at the effluent = "low values" (good quality of effluent)
IF	microfauna abundance = "low biodiversity" and dominant protozoa "small flagellates"	THEN	BOD and TSS at the effluent = "high values" (bad quality of effluent)
IF	settling test observations = "good settling" and microfauna abundance "high biodiversity"	THEN	floc characterization = "well-formed flocs"

quantitative variable considered in the KB, and (d) the increase of the understanding of some cause−effect relationships between the presence of microorganism and operational problems in the process. These facts open the door to new trends of research in the field such as, for example, the design and building of ontologies of the microorganisms present in the WWTP departing from the collection of specific experiences.

Among the benefits of using this approach to build KBS as supporting tools for WWTP, we have to mention portability and reuse. The KB created for a specific WWTP could be easily ported into another of the same kind and trained through the learning process. We think that this two-phase methodology can be generalized to acquire knowledge from any (bio)chemical process, improving the development process and the efficiency of the KBS.

Acknowledgment

This research has been partially supported by the Spanish CICyT projects TIC96-0878 and AMB97-0889. We acknowledge the support of European Union Project IST-1999-176101. The views in this paper are not necessarily those of the A-TEAM consortium.

Nomenclature

COD = chemical oxygen demand
BOD = biological oxygen demand
TSS = total suspended solids
NH_4^+ = ammonia
TKN = total Kjeldhal nitrogen
NO_2^-/NO_3^- = nitrite/nitrate
Cl^- = chloride
P = phosphorus
MLSS = mixed liquor suspended solids
MLVSS = mixed liquor volatile suspended solids
V30 = 30-min settling volume
Cond = conductivity
DO = dissolved oxygen
flow rates (flow)
SRT = sludge residence time
SVI = sludge volume index
F/M = food-to-microorganism ratio
N/D = nitrification−denitrification

Literature Cited

(1) Stephanopoulos, G.; Han, C. Intelligent Systems in process Engineering: a review. *Comput. Chem. Eng.* **1996**, *20* (6/7), 743.

(2) Jackson, P. *Introduction to Expert Systems*, 3rd ed.; Addison-Wesley: Essex, U.K., 1999.

(3) Huang, Y. L.; Sundar, G.; Fan, L. T. Min-Cyanide: An expert system for cyanide waste minimization in electroplating plants. *Environ. Prog.* **1991**, *10* (2), 89.

(4) Zhu, X. X.; Simpson, A. R. Expert system for water treatment plant operation. *J. Environ. Eng.* **1996**, 822.

(5) Serra, P.; Sánchez, M.; Lafuente, J.; Cortés, U.; Poch, M. ISCWAP: A knowledge-based system for supervising activated sludge processes. *Comput. Chem. Eng.* **1997**, *21* (2), 211.

(6) Isaacs, S. H.; Thornberg, D. A Comparison Between Model and Rule-Based Control of A Periodic Activated Sludge Process. *Water Sci. Technol.* **1998**, *37* (12), 343.

(7) Furukawa, S.; Tokimori, K.; Hirotsuji, J.; Shiono, S. New Operational Support System for High Nitrogen Removal in Oxidation Ditch Process. *Water Sci. Technol.* **1998**, *37* (12), 63.

(8) Ashraf Islam, K.; Newell, B.; Lant, P. Advanced Process Control For Biological Nutrient Removal. *Water Sci. Technol.* **1999**, *39* (6), 97.

(9) Olsson, G.; Aspegren, H.; Nielsen, M. K. Operation and Control of Wastewater Treatment—A Scandinavian Perspective over 20 years. *Water Sci. Technol.* **1998**, *37* (12), 1.

(10) R-Roda, I.; Sànchez-Marrè, M.; Cortés, U.; Lafuente, J.; Poch, M. Consider a Case-Based System for Control of Complex Processes. *Chem. Eng. Prog.* **1999**, *95* (6), 39.

(11) Kraslawski, A.; Koiranen, T.; Nystrom, L. Case-Based Reasoning System For Mixing Equipment Selection. *Comput. Chem. Eng.* **1995**, *19*, S821.

(12) Belanche, Ll.; Valdés, J. J.; Comas, J.; R-Roda, I.; Poch, M. A Soft Computing Techniques Study in Wastewater Treatment Plants. *Artif. Intell. Eng.* **2001**, in press.

(13) Chen, B.-C. Back-propagation Neural Network Adaptive Control of a Continuous Wastewater Treatment Process. *Ind. Eng. Chem. Res.* **1998**, *37* (9), 3625.

(14) Wang, X. Z.; Chen, B. H.; Yang, S. H.; McGreavy, C.; Lu, M. L. Fuzzy Rule Generation from Data for Process Operational Decision Support. *Comput. Chem. Eng.* **1997**, *21*, S661.

(15) Leiviskä, K. Industrial Applications of Intelligent Methods. In *Proceedings of the 6th European Congress on Intelligent Techniques & Soft Computing (EUFIT '98)*; Verlag Mainz Wissenschaftsverlag: Aachen, 1998; Vol. 3, p 1449.

(16) R-Roda, I.; Sànchez-Marrè, M.; Comas, J.; Cortés, U.; Lafuente, J.; Poch, M. An Intelligent Integration to improve the Control and Supervision of large WWTP. In *Proceedings of the 8th IAWQ Conference on Design, Operation and Economics of Large Wastewater Treatment Plants*; Budapest University of Technology: Budapest, Hungary, 1999; p 566.

(17) Cortés, U.; Sànchez-Marrè, M.; Cecaronni, L.; R-Roda, I.; Poch, M.; Environmental Decision Support Systems. *Appl. Intell.* **2000**, *13* (1), 77.

(18) Özyurt, B.; Sunol, A. K.; Çamurdan, M. C.; Mogili, P.; Hall, L. O. Chemical plant fault diagnosis through a hybrid symbolic-connectionist machine learning approach. *Comput. Chem. Eng.* **1998**, *22* (1−2), 299.

(19) Ke, M.; Ali, M. Induction in Database Systems: a Bibliography. *Appl. Intell.* **1991**, *1* (3), 263.

(20) Mitchell, T. M. *Machine Learning*; McGraw Hill: New York, 1997.

(21) Barnard, J. L. Biological denitrification. *J. Int. Water Pollut. Control Fed.* **1973**, *72*, 6.

(22) *G2 Reference Manual*, Version 4.0.; Gensym Corp.: Cambridge, U.K., 1995.

(23) Sànchez, M.; Cortés, U.; Béjar, J.; de Gracia, J.; Lafuente, J.; Poch, M. Concept formation in WWTP by means of classification techniques: a compared study. *Appl. Intell.* **1997**, *7*, 147.

(24) Dubes, R.; Jain, A. *Algorithms for Clustering Data*; Prentice-Hall: Englewood Cliffs, NJ, 1988.

(25) Jenkins, D.; Richard, M. G.; Daigger, G. T. *Manual on the Causes and Control of Activated Sludge Bulking and Foaming*; Lewis Publishers: Chelsea, 1993.

(26) Riaño, D.; Cortés, U. Rule generation and compactation in the WWTP. *Comput. Sist.* **1997**, *1* (2), 77.

(27) Clancey, W. J. Heuristic Classification. *Artif. Intell.* **1985**, *27*, 289.

Received for review May 30, 2000
Accepted March 25, 2001

IE000528C