10

Federated Knowledge Integration and Machine Learning in Water Distribution Networks

H. Afsarmanesh(1), L.M. Camarinha-Matos(2), F.J. Martinelli(2)
(1) University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands.
email: hamideh@wins.uva.nl
(2) New University of Lisbon and Uninova,
Faculty of Sciences and Technology
Quinta da Torre, 2825 Monte Caparica, Portugal
email: cam@uninova.pt; fjm@uninova.pt

Abstract

Water distribution networks are geographically distributed systems, with greater heterogeneity in terms of control structures, management strategies, and varying geometry with continuous expansion and changes in demand along their life. Due to these characteristics, water distribution companies face the problem of data and knowledge integration related with control and optimal exploitation. A few European and International RTD projects are currently focused on the design and development of a next generation system to support the control, optimal operation and decision support of drinking water distribution networks. In the context of a two year RTD project - WATERNET*, an evolutionary knowledge capture for advanced supervision of water distribution network is being developed. The WATERNET system, assists the distributed control of water management network to minimize the costs of exploitation, guarantee the continuous supply of water with a better quality monitoring, save energy consumption and minimize natural resources waste. This system comprises of several subsystems: a distributed information management subsystem, a machine learning subsystem, an optimization subsystem, a water quality monitoring subsystem, a simulation subsystem, and a supervision system that integrates these subsystems in order to assist the decision making and optimal operation of the network. This paper first provides a high level global description of this on going research on water management system. Then, it concentrates on the current stage of development on two specific subsystems in the project; the distributed information management and the machine learning subsystems.

* The work described here has been partially supported by the ESPRIT project 22.186 - WATERNET, whose partners are: ESTEC, UNINOVA, SEBETIA, WBE, Univ. of Amsterdam, ADASA Sistemas, Univ. Politecnica de Catalunya, ALFAMICRO, and Univ. of Naples.

L. M. Camarinha-Matos (ed.), *Re-engineering for Sustainable Industrial Production*

Keywords
Federated cooperative databases, integrated information sharing, knowledge integration, machine learning systems, water distribution networks, intelligent supervision

1 INTRODUCTION

Most water supply industries nowadays lack a global "overview" of the status of the production and the distribution system. It is difficult, costly, and takes a long time to recognize failures in the system, to identify the non-optimized operations and resource wastes, and to take the proper recovery and maintenance actions. Control of the system is often carried out locally, based on the operators experience. Typically, there is none or little coordinated control in order to assure a continuous supply, meet the quality standards, save energy, optimize pipeline sizes and reduce wastes. Existing systems for control and/or monitoring of the water supply and distribution are heterogeneous and of different levels of automation and reliability. Furthermore, these systems cannot be easily linked together in order to extract the needed integrated information.

However, today the consumers demand higher performance for the water supply. Recent research (Stroomloos 1994) shows that the society tends to become more and more dependent on public supplies, e.g. the water and power, while it becomes less capable to deal with interruptions (The vulnerability paradox). The same research points to the need of water companies to improve continuity and quality of water supply. The water industry will gain extensive economical benefits by: minimization of energy-consumption, the rational and effective failure diagnosis and maintenance, minimization of mixing zones of water of different qualities, minimization of detention time in the distribution system, and the rational expansion / modification of the network structures.

A network for water distribution presents a wide set of specific problems and requirements. Some of the important problems identified and addressed in the WATERNET project includes the following:

- The network is distributed through a wide geographical area; e.g., an average normal system may comprise of close to 900 Km of main distribution pipes.
- Because of the geographical structure of the covered area and the available sources and consumption requirements, each system is different, and as a consequence it employs specific mechanisms for control and management.
- In order to guarantee the continuous supply of water, several pumping stations and reservoirs must be controlled and managed in an integrated way.
- The quality of the water must be guaranteed from the capture points to the consumers.
- The network is in permanent expansion because of the increase in the number of big consumers: factories, industrial sites, etc.
- Other sources of uncertainty that must be considered are due to the disruption (total or partial) of pipes, or even pirate derivations.
- Criteria for regular and repair equipment maintenance must be met in order to optimize the exploitation of the service.

Based on the above requirements, WATERNET is developing a modular Supervision System featuring a Distributed Information Management Subsystem, an Optimization Control Subsystem, a Learning Subsystem, a Simulation Subsystem, and a Water Quality Monitoring Subsystem distributed control architecture.

An improvement in any of these problematic areas through the use of an effective drinking water distribution system produces a huge economic benefit for both the water industry and the consumers. As an illustrative example, with the mere application of a distribution optimization software, specially developed for the drinking water network of Barcelona city in Spain (Quevedo 1988), a 15% reduction of the previous distribution costs was achieved. This resulted a 2 K ecu per day cost reduction for that industry.

This paper first provides a high level global approach to the knowledge capture for advanced supervision of water distribution networks in WATERNET project. It further focuses on two specific subsystems of distributed information management and machine learning. The paper also provides some examples from the environment describing the applicability of these subsystems to distributed water control problems.

The remaining of this paper is organized as follows. Section 2 provides an integrated control approach to production and distribution of water. In this section a high level description of the work planned in the WATERNET project is presented. Section 3 describes a distributed/federated information management framework for the water control network. A summary of the PEER federated integration architecture is also provided in this section. Section 4 describes the learning system for the water management application. And finally, Section 5 concludes the paper.

2 PROJECT SUMMARY

2.1 Project Objectives

The WATERNET project approaches the problems addressed in Section 1, and introduces an architecture and necessary mechanisms to support these requirements. In specific the design and development of WATERNET environment is based on the support for the following key technical issues:

a) Distributed multi-agent control system
Due to the geographical distribution and open/dynamic structure of the water distribution systems, emerging results from the area of Distributed Information Management Systems (Federated Cooperative databases, Integrated Information Sharing, etc.) are applied in the proposed system.

b) Functionality control and management system
Optimization, forecasting, diagnosis, prognosis, and recovery are obvious components for an advanced supervision system. Real-time situation assessment of the dynamic, hard to measure systems is being explored here.

c) Multi-paradigm learning system

In water distribution networks, domain models are not enough to ensure all supervision requirements. As in other areas, the codification of knowledge can be limited. For certain sub-problems, a database of solved cases can be the sole source of knowledge. In practice, the paradigm of Programming by Human Demonstration, namely a set of programmed predefined complex systems that show the particular examples of desired human behavior, have proven to help in certain complicated problems (Camarinha-Matos 1996). If traditional expert systems are often criticized for being limited in their abilities to surpass the level of existing experts, learning systems have the potential to discover new relationships among concepts, therefore exceeding the performance of experts. This does not mean that expert knowledge is being ignored. Rather, the technical approach in the WATERNET project consists of both traditional programming and programming by demonstration.

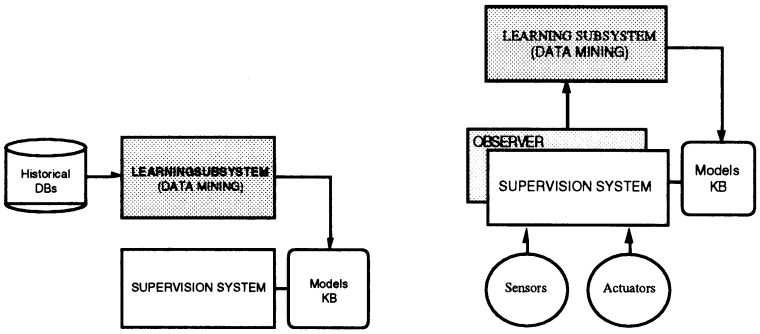


Figure 1 a) Learning from historical data b) Continuous learning

2.2 Current situation

A typical modern control and management system for water supply industry is a distributed network of stations computationally linked to a central control station. In the SMAS Sintra, a Portuguese water industry, remote stations are all points in this network that serve both as the local controller and the observation unit. According to their characteristics, a local computer system is configured and equipped with the necessary interfaces in order to control and manage every node locally. This local system applies a proper control algorithm and monitors all the sensorial data relevant for its task. A log of commands, alarms, sensorial values e.g. pressures, levels, flows, etc. is gathered locally, and transmitted periodically to the control unit database. Sensorial data is collected at short intervals - usually once every few minutes. Because of the functions offered by the communications network, any anomaly detected by this system is automatically reported to the control unit, and a proper recovering procedure is executed by the human supervisor if necessary and if available.

A control unit is localized in a control room of the water distribution services and permits the supervision of an area by the operators. It consists of a set of workstations connected by a local area network, the communications front-end, and the data storage equipments. The operator

uses a graphical interface to dialogue with the system. A diagram of the network and related stations is displayed on the workstation screen. Using the mouse, the operator can enter direct commands to a remote station in order to visualize its status of operation and perform any necessary action. The main job of the operator at a control unit, is to monitor the network and to observe the incoming alarm messages. According to the messages received, he can start the necessary action in order to recover, and if necessary he can be helped with a mobile brigade.

At each remote station, the latest data produced locally is maintained in a local database. This database can later be used by several users related to maintenance, planning, etc. Each user scans the data in order to find the information relevant for its needs. At intervals specified by the operator, the control unit polls all the remote stations for collecting their data gathered locally and stores it in its database. This scenario corresponds to a state of the art installation. However, in some cases, the situation can be less automated. At present, some water distribution systems do not even include an integrated control system, and only rely on the local human operators and spoken communications with a control unit. In the case of SMAS Sintra, at present there is only one central control unit that supports the supervision of the entire network. On the other end of the spectrum, there are water companies, e.g. the WBE, a Dutch water company, that require a more complicated system.

2.3 Need for advanced control and management systems

a) The problem of optimal operation

This problem consists of deriving strategies for the control elements in a water system (valves, pumps, turbines) so as to achieve the service objectives (meeting consumer demands with appropriate pressure levels) with the minimum possible cost. These strategies must be found ahead of time based on demand predictions. The mathematical formulation of this problem requires an optimization of a cost function which is generally non-linear and includes several constraints. Thus, the mathematical formulation is difficult and needs further research and even the evaluation of alternative techniques, e.g. the machine learning.

b) Controlling large water distribution systems

While the hardware system for telecontrol from a central dispatch is reasonably efficient at this stage, the problems of management of large amounts of information, scheduling the control of water transfer from the sources to the consumers with appropriate pressure levels ahead of time, estimating the state of the network from the readings of a limited number of sensors and dealing with breakdown alarms are still largely unsolved. Currently, the main emphasis of research in this area is on solving the optimal scheduling problem. Even for relatively small networks, this is an extremely complex problem, due to the nonlinearities involved and the computationally intensive process.

c) Information management

In addition to the problems of scheduling and control, the problems of information management have become more apparent with the accumulation of experience in telecommunication and telecontrol in the 90's. In particular, the availability of large amounts of both numeric and symbolic information, the complex structured knowledge that needs to be shared to support the distributed control and management of water supply units, and the heterogeneity of the nodes

in the network and their distinct and various tasks, has made it necessary to use more advanced data management technology than have previously been employed.

The problems of management and control of water networks continue to be a challenging one for most European cities. From one side, no generally applicable techniques have been yet derived for the problems tackled in the past. On the other side, the significant progress in the computing power and data management techniques can now be applied to the water industry. It is possible to envisage a solution for the whole new range of management and control problems, which had not yet been dealt with in previous telecontrol systems. The application of new technology, in turn improves the ability of water authorities to provide a good service to the consumers taking maximum advantage of the limited natural resources available.

Extracting knowledge from available or being collected data, by application of machine learning / data mining techniques, seems to be a promising approach.

3 DISTRIBUTED INFORMATION MANAGEMENT SUBSYSTEM (DIMS)

Modern Water supply and water distribution management heavily depends on the support of a strong information management system. A distributed/federated information management framework needs to be designed and developed to support the cooperation and information exchange among different sites and their activities involved in intelligent water distribution control network. This network of sites must be open and dynamic to support its expansion, as the number of sites in the water distribution network increases. Consequently, it must support the management of the dynamic and incremental evolution of the information within the integrated network. The sites in the network can be autonomous since they may pre-exist to the development of the water distribution system, serving a specific purpose, and must be able to retain their local control and autonomy. For instance, there are sites that provide certain information and/or services, e.g. maintenance, complaint handling, demographic information, maps and geographic information, etc.

The sites are also heterogeneous due to the variety of their functionalities ranging from the production, distribution, and maintenance to data gathering, control, simulation, diagnosis, prognosis, learning, optimization, and planning. In the current design of the DIMS subsystem, it is assumed that the information of each site in the network is represented using a common information-representation model and can be accessed using a common access-language. Existing standards, such as common interface with the IDL (of ODMG) and common exchange architecture with CORBA (of OMG), may need to be used to support possible database model heterogeneity, which is outside the scope of the current research and this paper.

Each site in the network is being extended with a layer that supports the federated information sharing and exchange among different sites. In the development of the federated architecture, an existing software prototype, the PEER system, developed at the University of Amsterdam (UvA) is applied. PEER is a federated object-oriented information management system (Afsarmanesh 1993), (Tuijnman 1993), (Afsarmanesh 1994) that is designed and prototyped at UvA; its development is based on the AIM (Agent Information Management) software developed by UvA in ESPRIT - ARCHON project (Wittig 1992).

3.1 Management of information in water control network

In water supply and distribution network, typically the information is gathered, processed, and stored in geographically distributed sites that are called units in this section. Every unit serves a specific function in the system and thus units are of different kinds. In order to perform its tasks, every unit relies on accurate and proper integration of remotely collected data and the availability of up-to-date shared information, at the time when it is needed.

Three main specific kinds of units can be identified. Every production site comprises a "remote unit" that collects the local data and executes some local level of control. A "control unit" performs the control supervision and control of the water supply and distribution system. for a number of remote units. There is a third kind of unit, the "auxiliary unit" that assists the tuning and optimizing tasks of different functionalities of the system. Examples of auxiliary units include the machine learning unit, the simulation and optimization unit, the water quality monitoring unit, etc..

The units mentioned above are interconnected by a communication intranet network. Some of these units are tightly and some loosely coupled. For instance, the remote units are tightly-coupled with the control units. All the information gathered and stored at any remote unit is also transfered and stored in a control unit. In some existing water management networks, the remote units are only connected to the control units, while in some others they are also connected with each other and with the auxiliary units in a network. On the other hand, some units are only loosely-coupled and they only share and exchange a part of their information with each other as required. In order to properly perform their tasks, for instance, the water quality monitoring units need to access a part of the information stored in the control units.

3.2 Need for information sharing and exchange among nodes

The proper functionality of all units involved in the water control systems depends on the sharing and exchange of data with other units. The autonomy of some units need to be preserved, for instance, the supervision unit is an autonomous unit, while the remote unit has only partial control over its functionality and takes orders from the control units. Similarly, the heterogeneity of information representation and classification needs to be supported. In general, the same piece of information is viewed differently by two units, and different level of details are associated with it. For instance, to the water quality monitoring unit, all the details of every water quality sample are very important piece of information, while to the alarm handling and maintenance units, only a part of certain water quality samples are of interest and relevance if they are the cause of alarm situations.

3.2.1 Example cooperation scenario

Typically auxiliary units require to share and access information from the control units, and the remote units (either directly or through the control unit) in the network. Consider the learning subsystem defined in Section 4 of this paper that performs several types of tasks. For instance it supports the production planning, alarm handling, preventive maintenance, etc. For every one of these tasks, learning algorithms need to perform a number of subtasks, for which they require access to different data from certain other units in the network.

In more details, the task of providing preventive maintenance information, involves the two major subtasks of: (1) Characterization of devices life cycle, and (2) identification of behavioral changes in devices. Now, subtask (1) requires the access to maintenance history of devices, that is stored in the control unit, while subtask (2) requires the access to up-to-date current device status, flow, etc. that is coming from the remote unit. Once the data of every unit is classified and handled through its federated PEER layer, their interoperability and information exchange are properly supported by the system and the logical and physical distribution of the information over the units is completely transparent to any user in the network.

3.3 PEER federated integration architecture

The PEER federated information management system supports the management, and sharing and exchange of information in a network of loosely/tightly coupled nodes. Each node in the federation network can autonomously decide about the information that it locally manages, how it structures and represents this information, and which part of its local information it wishes to export and share with other nodes. Each node can import information that is exported by other nodes and then transform, derive and integrate (a part of) the imported information to fit its interest and corresponds to the local interpretation. PEER is a pure federated system; namely there is no single global schema defined on the information to be shared by different nodes in the network, and there is no global control among nodes. Clearly, this definition can be made more restrictive based on the configuration of an application environment. The PEER information integration infrastructure helps the human users in a cooperative team, by supporting the information integration at different levels of granularity, e.g. for the global task, or among different activities and sub-activities. PEER supports both the loose coupling and the tight coupling between nodes.

A PEER layer is defined and augmented to every node that needs to share and exchange information with any other node in the network. In the PEER layer of a node, the information is then structured and defined by several kinds of `schemas', as shown in Figure 2. The user needs to identify (1) the information that is available locally in the node (LOC schema), (2) the information that the node needs to access remotely and so the node imports it (IMP schemas), (3) the information that is available locally and other nodes need to access, so the node exports it (EXP schemas), and (4) the integration of the LOC information with the IMP information to create a coherent pool of information (INT schema) that is needed to be accessed by this node. At the development stage, the design of these schemas, makes both the information managed by a node and the interfaces to other nodes explicit. However, once the PEER layer is developed, the integration facility of PEER, its distributed schema management, and its distributed query processing (Afsarmanesh 1993), (Tuijnman 1993), (Afsarmanesh 1994) make the distribution of information and the heterogeneous information representations among different nodes totally transparent to the user.

Information integration in PEER is supported by a declarative specification using the PEER Schema Definition and Derivation Language SDDL (Afsarmanesh 1993). The SDDL language supports the integration, derivation, interrelation, and transformation of types, attributes and relationships from their sources.

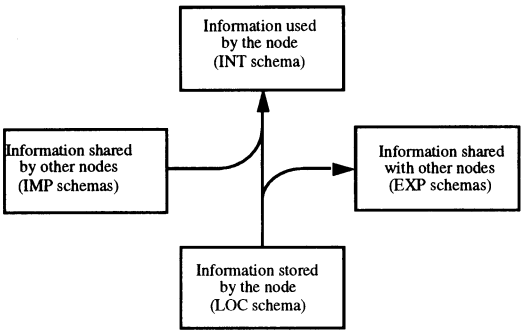


Figure 2 The architecture of the information managed by a node

A prototype implementation of the PEER system is developed in the C language and includes two user interface tools (Afsarmanesh 1994), a Schema Manipulation Tool (SMT) and a Database Browsing Tool (DBT).

3.4 Information analysis for building a PEER layer

One difficulty in building a DIMS for water management systems is due to the fact that the typical existing subsystems lack a recorded semantic description for the data that they work with. Namely, there are no schemas (meta-data) or textual description for the stored data. The gathered information is usually dumped in files and the knowledge about their semantics is only carried by the expert operators who need to interpret the stored data and issue commands, both at the local site and at the control unit.

In order to build a PEER layer for any unit in the network, first step is to identify all the information that is handled by the unit locally. The next step is to pin point the data that this unit needs to access (import) from the other units in the network, and then to identify the operations applied to the data (local or imported) that in turn supports the functionality of the unit.

So far, a preliminary analysis of some information management requirements at the SMAS-Sintra water industry is achieved. The goal of this analysis is to focus on the design of the PEER layer for the remote units. Once this process is completed, the information will be classified/reclassified by the object-oriented database model of PEER and will be represented through corresponding schemas (LOC, IMP, EXP, and INT) at the PEER layer of such units. The analysis results identifies the static and dynamic characteristics of information handled in a remote unit in SMAS industry. Namely, a complete set of all the gathered and handled data; data structures that are currently defined on the data and a textual description for the data items; operations that are currently applied to the data; and other system functionalities is being identified. Following are preliminary the results of this analysis.

a) Variety of data gathered:
Data sensored on the level, flow, pressure, motor power, energy and valve reading, the alarm data etc. represent some of the various kinds of information collected at the unit.

b) Frequency and the amount of gathered data:
The status of all the sensors is read and gathered once every few minutes during the normal operation, generating huge amount of data. In irregular and exceptional situations, e.g. the maintenance situation, requires even more data collection than the periodic reading. This data constitutes the historic database.

c) Simple Data structures currently available:
All information is stored in files, so the record formats are the only structures used to classify and store the information.

d) Operations applied on the data at present:
The alarm handling, historic data search, and data transfer requests are example operations at the unit.

e) Some logged System functionalities:
Logs of actions taken, and control commands are kept at present in the remote unit.

All data gathered at this stage belongs in the LOC schema of the remote units. At present, in the case of SMAS-Sintra there is no cooperation and information sharing among the remote units themselves. Also, there is no communication line between the remote units and other auxiliary units. Therefore, the remote units do not import any information from outside and their main purpose is to collect relevant information and to export them to other units. Thus, the EXP schema of a remote unit is the same as its LOC schema.

4 LEARNING SUBSYSTEM

In most advanced water distribution management systems, the main implemented aspects deal with the problem of control and daily exploitation in which the main objective is the guarantee of a continuous supply. The experience shows that, even with such systems, there are frequent requests to change the control algorithms and control strategies of the remote stations in order to refine the entire system's exploitation. Mainly these requests are based on the analysis of the collected data, for example when the distribution services find evidences such as: if they change the current pumping strategy they can pump the same amount of water in a less expensive way.

In some countries there are even various levels for cost of consumed energy according to the phases of the day. Of course, these conclusions can only be reached a posteriori; namely after the system is running for some time. Originally, and without necessary collected information on a running system, it is not possible to achieve such optimizations. In some other situations a priori decisions can be made to support the change in the water consumption, e.g. to support the seasonal behaviors. These aspects support the idea of developing learning tools for exploitation of the collected data. With such tools the water distribution control technicians can concentrate on the optimization and planning aspects, and to explore the gathered data in order to build a realistic model of the system. Some of the relevant aspects to be learned are:

- Models describing the behavior of specific nodes, mainly those closer to the high concentration of consumers.
- Models to relate the alarms as symptoms of future breaks (prognosis): in pipes, pumps, etc.
- Models for daily forecast of water demands, permitting optimal management of production.
- Consumer trends models, helping future planning of expansion of the network.
- Diagnosis models that help to identify and locate problem areas in the network.

4.1 Examples of learning tasks

There is therefore a large number of potential opportunities for application of learning techniques in the water distribution system. Some examples of learning tasks are:

4.1.1 Support for Production Planning

In this area a number of problems can be considered, all of them related to the planning of water production and distribution or the optimization of this process:

a) Identification and characterization of pattern repetitions along the time (seasonal patterns)
This task is intended to detect temporal repetitions of patterns in collected data:

- Patterns at specific time periods along the day.
- Patterns along the week.
- Seasonal patterns along the year.

Example:

During weekdays, between time H1 and H2, the average consumption reaches the level C.

Or, in a rule form:

IF weekday(D) and T >=H1 and T<=H2 THEN expected_consumption = C.

A characterization of such patterns may be an important input for the production planning activity. This knowledge can help in:

- Avoiding periods of lack of supply.
- Optimizing the production costs, by anticipation of customer needs.

b). Identification of tendencies (changes) in consumption
The objective here is to detect any monotonic changes in the average values of consumption for equivalent time periods. This information gives an idea on how the network is evolving (in terms of demand) and may be a good tool for the network expansion / restructuring activities. Superimposing (comparing) the seasonal patterns with average values for different (but equivalent) time periods can give an idea on how the network is evolving.

Example:

Average flow C in subnet N is steadily increasing X% per year.

c) Identification of working period of devices
Objective: Identification and evaluation of the working periods of devices (groups/pumps) in order to reduce the operation costs. This is specially useful in regions with different energy prices along the day.
Example:
Detection of all devices operating in periods of high price energy.

4.1.2 Identification of the behavior of human operators

Learning the operators' behavior can be useful for:

- Training of new operators.
- Giving advise in critical situations to less experienced operators.
- Contributing to the optimization of the algorithms of local controllers.

The existence of a reference model of good behavior may be necessary in order to validate the examples. Or at least some a-priori qualitative model should be used.
Examples of such learning tasks are:

a) Actions as a function of physical system parameters
To learn which actions are typically performed by operators as a function of read values of system parameters. For this purpose, the values of system variables in a time window before realization of each action have to be analyzed. Similarities of values of some variables in all time windows before a certain kind of action may induce the conclusion that such values were the reason for the decision of applying that action. Nevertheless, such conclusion needs an assessment by a human expert or a validation against an a-priori qualitative model of the system.
Example:
IF level of reservoir R reaches the value L AND group P is on
THEN switch pump P off.

b) Actions as a function of time
Detect / identify actions that are performed in specific times.
Example:
Every weekday group B should be switched on at 18:00.
Or:
IF weekday(D) AND time(D) = 18:00 THEN switch_on(B).

4.1.3 Monitoring and alarm handling

In this area, a set of tasks related to the acquisition of supervision knowledge are considered. For instance:

a) Alarm detection
a.1) Alarm detection based on readings of system variables
Detect primary causes of alarms, i.e., identify explanations for alarms based on the observation of variables evolution before the alarm situation.
Example:
Relationship between an alarm and low pressure level, high flows or pipe break.

a.2) Alarm detection function of time
To detect alarms that occur with some timely repetitive pattern. For instance, alarms due to insufficiencies of the distribution network or due to a bad production plan.
Example:
Every Saturday, near time H, the minimum level alarm of reservoir R goes on.

a.3) Consequences of alarms not processed
Identify what can happen to the system if some alarms are not handled by the operators. The operators may ignore some alarms by assuming these alarms have no consequences. One approach is to verify if any action was done in a time interval after the occurrence of the alarm. If not, we can consider that the alarm was discarded and then a careful analysis of the system behavior, before and after the alarm, is necessary.
Example:
If some maximum level alarms are not handled, other level alarms may be generated and an overflow of reservoirs may occur.

a.4) Inter-relationships between alarms
To detect inter-relationships between alarms that occur in a short time interval. Such alarms may represent several failures or a single failure that provoked various alarms.
Example:
IF a reservoir R is too full THEN the maximum level alarm goes on
AND pressure on pipe D alarm ALSO goes on.

b) Identification of actions related to alarms
b.1) Actions that cause alarms
To identify actions that may cause the system to enter (even temporarily) in an abnormal state. This might be useful in order to create protection or advise mechanisms.
Example:
A few minutes after the start of pomp P, the minimum level of reservoir R goes on.

b.2) Actions after alarms
To identify which actions, i.e., which recovery plan, should be performed for each alarm situation. For this purpose it is necessary to analyze all actions performed after an alarm. On the other hand, it may happen that different operators have different strategies for problem solving after an alarm occurs.
Example:
After a maximum level alarm for reservoir R, pumps P1 and P2 must be switched off.

4.1.4 Preventive maintenance
For instance:

a) Characterization of devices life cycle
Identify the influence of the operating conditions on the life cycle of devices, namely in terms of malfunctioning situations, maintenance, etc. This knowledge might help identifying persistent

problems in a specific region (which need particular attention), problems with a given equipment supplier, the average life time of each class of equipment, etc.

Example:

Valve V, which normally operates at pressure P1, can have a useful life X times larger if operating at pressure P2.

b) Identification of behavioral changes in devices

To identify the evolution of devices behavior with age / time of use.

Example:

Relationship between the time of operation of a group and the volume of pumped water.

4.1.5 Improvement of user satisfaction

To characterize special anomalies, like frequent ruptures in pipes, from the analysis of users' complains and information. Example: Frequent ruptures in the same pipe / zone may suggest its replacement.

4.1.6 Other possibilities

Some additional learning tasks could be:

a) Identify the range of values for system variables in normal operation

To learn which is the range of values, in normal operation, for each system variable. This information contributes to the characterization of what is a normal state and is useful for the supervision activity. These values may also be used for simulation and optimization purposes.

Example:

The pressure in pipe P should be between P1 and P2.

b) Detection of abnormal variations of some variable

To detect and explain (if possible) unexpected variations of some variable, by comparison to the expected pattern.

Example:

A pressure was almost steady for a while and it had a sadden change in spite of no actions have been done.

The above list is a starting point, which needs to be assessed by the end users in terms of usefulness, priorities, availability of data and a-priori knowledge, etc. On the other hand, other learning tasks may be missing in this list, but the approach was to use this list as a first step in an iterative process.

4.2 Preliminary experiments

This section describes some preliminary experiments on machine learning applied to historical data from one of the SMAS-Sintra stations (Pedra Furada). The objective was mainly of a preparatory nature rather than an attempt to generate useful knowledge out of the data. Therefore, the motivation was to:

- Understand the structure and contents of the historic data repositories;

- Identify basic data transformations needed before these data can be applied to standard learning algorithms;
- Start testing the behavior of some algorithms with a selected subset of data.

Training Data:

```
**ATTRIBUTE AND EXAMPLE FILE**

LN-11: (FLOAT)   -> Level
LC-6-1: (FLOAT)  -> Flow
LC-6-2: (FLOAT)  -> Flow
LP-1: (FLOAT)    -> Pressure
Action: "AG_18(Off)" "AG_18(On)" "None";
@

1.5  4.2  0.4  2.8  "None";
1.3  ?    ?    ?    "None";
?    4.2  0.1  2.7  "AG_18(On)";
?    4.5  0    2.8  "None";
1.3  13.5 0.7  2.8  "None";
1.5  13.1 1.1  2.8  "None";
1.5  12.4 1.1  2.9  "None";
1.5  12.6 1.1  2.9  "None";
1.8  13.1 1.1  2.7  "None";
1.8  13.1 1.1  3.6  "None";
2    ?    ?    ?    "None";
?    13   0.9  3.2  "None";
?    1.9  0.1  3.4  "AG_18(Off)";
2    4.7  0    3.8  "None";
2    4.4  0.4  2.9  "None";
2    4.4  0.3  2.8  "None";
2    4.4  0.4  2.8  "None";
2    4.4  0.4  2.9  "None";
1.8  4.2  0.4  2.8  "None";
1.8  4.4  0.4  2.8  "None";
1.5  4.5  0.4  2.8  "None";
1.5  3.7  0.4  2.8  "None";
1.5  4.5  0.4  3    "None";
1.3  ?    ?    ?    "None";
?    4.4  0.3  2.9  "AG_18(On)";
?    4    0    2.9  "None";
1.3  13.8 0.1  3    "None";
1.3  13.1 1.1  3    "None";
1.3  13.6 1.1  2.8  "None";
1.3  13.1 1.1  2.8  "None";
1.5  13   1.1  2.7  "None";
1.5  13.3 1.1  2.8  "None";
1.5  13.3 1.1  2.8  "None";
1.5  13   1.1  3    "None";
1.5  13   1.1  3    "None";
1.5  13.1 1.1  2.9  "None";
2    ?    ?    ?    "None";
?    12.8 1.1  2.9  "None";
1.8  13.6 0    2.9  "None";
?    1.4  0.1  2.9  "AG_18(Off)";
1.8  4.2  0.3  2.9  "None";
1.5  4.4  0.4  3    "None";
...
...
```

Figure 3a

Generated rules:

```
**RULE FILE**
@
Time: [ Mon Sep 16 19:47:07 1996 ]
Examples: g18p_cn2.txt
Algorithm: UNORDERED
Error_Estimate: LAPLACIAN
Threshold: 0.00
Star: 5
@

*UNORDERED-RULE-LIST*

IF    LC-6-1 < 3.40
  AND 0.05 < LC-6-2 < 0.15
  AND LP-1 > 1.75
THEN  Action = "AG_18(Off)" [7 0 1.25]

IF    LN-11 > 1.90
  AND LC-6-1 < 2.70
THEN  Action = "AG_18(Off)" [4 0 3.75]

IF    LN-11 > 1.90
  AND 2.30 < LP-1 < 2.75
THEN  Action = "AG_18(On)" [0 1 1.75]

IF    4.20 < LC-6-1 < 4.60
  AND 0.05 < LC-6-2 < 0.25
  AND LP-1 > 3.45
THEN  Action = "AG_18(On)" [0 2 0.62 ]

IF    LN-11 < 1.90
  AND LC-6-1 > 2.70
  AND 0.05 < LC-6-2 < 0.15
  AND LP-1 < 1.55
THEN  Action = "AG_18(On)" [0 1.50 0.62]

IF    LN-11 > 1.90
  AND LC-6-1 < 8.60
  AND LP-1 < 1.55
THEN  Action = "AG_18(On)" [0 0.50 2.25]

...
...
```

Figure 3b

In this project the intention is not to create new learning algorithms but rather to adopt, assess and, if necessary, adapt standard learning systems to this particular application domain. It is our belief that more than new algorithms, the main challenges are to:

- Identify techniques for pre-processing of raw data in order to extract high level features;
- Integrate multiple learning tasks;
- Integrate the learning system with other functionalities (Supervision Systems Optimization System,etc.).

Before a commitment to any particular learning technique or a decision about any necessary changes to classical algorithms is made it is necessary to make a first (empirical) trial evaluation.

In this way, a set of inductive algorithms, generating decision trees or rule sets, were selected and applied to a subset of data. This work is still in a very early phase but it helped in getting a more clear feeling about the problem. It has to be pursued and repeated with other classes of algorithms. Lets consider an example:

The learning task considered in the following example is the characterization of situations that lead an operator to perform some action. Figure 3a shows a partial example set selected for the operation of one particular pumping group. Figure 3b shows the set of rules learned by a CN2-style (Clark 1991) algorithm.

Training Data:

```
**ATTRIBUTE AND EXAMPLE FILE**

LN-11: (FLOAT)   -> Level
G_18_status: "On"  "Off"
Action: "AG_18(Off)"  "AG_18(On)"  "None";
@

1.5   "Off"   "None";
1.3   "Off"   "None";
 ?    "Off"   "AG_18(On)";
 ?    "On"    "None";
1.3   "On"    "None";
1.5   "On"    "None";
1.5   "On"    "None";
1.5   "On"    "None";
1.8   "On"    "None";
1.8   "On"    "None";
2     "On"    "None";
?     "On"    "None";
?     "On"    "AG_18(Off)";
2     "Off"   "None";
2     "Off"   "None";
2     "Off"   "None";
2     "Off"   "None";
2     "Off"   "None";
1.8   "Off"   "None";
1.8   "Off"   "None";
1.5   "Off"   "None";
1.5   "Off"   "None";
1.5   "Off"   "None";
1.3   "Off"   "None";
?     "Off"   "AG_18(On)";
?     "On"    "None";
1.3   "On"    "None";
1.3   "On"    "None";
1.3  13.6  1.1   2.8    "None";

...
...
```

Figure 4a

Generated rules

```
**RULE FILE**
@
Time: [ Mon Sep 16  20:04:12  1996 ]
Examples: g18a_cn2.txt
Algorithm: UNORDERED
Error_Estimate: LAPLACIAN
Threshold: 0.00
Star: 5
@

*UNORDERED-RULE-LIST*

IF      LN-11 > 1.90
  AND   G_18_status = "On"
THEN    Action = "AG_18(Off)"  [4.50  0  16.50]

IF      LN-11 < 1.40
  AND   G_18_status = "Off"
THEN    Action = "AG_18(On)"  [0  4.50  10.50]

IF      1.40 < LN-11 < 1.90
  AND   G_18_status = "Off"
THEN    Action = "None"   [0  2.25  173.75]

IF      LN-11 < 1.90
  AND   G_18_status = "On"
THEN    Action = "None"  [4.50  0  102.50 ]


...
...
```

Figure 4b

The training data was obtained from the real system and submitted to the learning algorithm after some pre-processing. This example illustrates the viability of using conventional learning

algorithms to extract operation rules. However, these rules must be assessed by the domain experts.

Considering that pressures and flows are consequence of the pumping process and not variables that directly influence the decision making, a refinement can be made. Figure 4 shows this situation. For this test the parameters used are the reservoir level and status information on the group 18 (On or Off).

In previous tests unknown values were fed directly to the learning algorithm. Considering that during the sample interval (5 min., in this case) the changes in the water level are small, a new test was performed replacing each unknown ("?") by the previous known value. In this case, for the same training set, the following rules were generated:

IF LN-11 > 1.9 AND G_18_status = "On" THEN Action = "AG_18(Off)"
IF 1.65 < LN-11 < 1.90 AND G_18_status = "On" THEN Action = "AG_18(Off)"
IF LN-11 < 1.40 AND G_18_status = "Off" THEN Action = "AG_18(On)"
IF LN-11 > 1.40 AND G_18_status = "Off" THEN Action = "None"
IF LN-11 < 1.65 AND G_18_status = "On" THEN Action = "None"
(DEFAULT) Action = "None"

As mentioned before, other tests were performed with other algorithms not shown here due to lack of space, but further work is necessary.

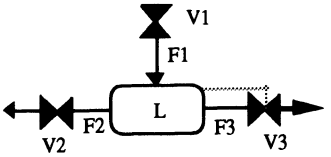
4.3 Model driven learning

It is widely recognized that applying standard inductive learning algorithms to real world problems is quite an art. Such algorithms are unable to take into account background knowledge and, therefore, it is up to the engineer to perform a set of preparatory actions / transformations on raw data in order to get useful rules or decision trees. Generated knowledge has to be assessed by a domain expert. This process is usually interactive and iterative, very time consuming and might require addition / removal of training examples, modification of the example description language or even modification of parameters of the learning algorithm. This situation has been reported by several authors.

It is also recognized that a-priori or background knowledge could be used to guide the induction process and then contribute to a more effective way of generating accurate knowledge. In our approach we plan to follow a strategy similar to that of Clark & Matwin (Clark 1993) which use Qualitative Models to guide inductive learning. A Qualitative Model (QM) (Dague 1995) tries to represent in a simplified way the "macroscopic" behavior of a physical system.A QM is a graph whose nodes represent parameters of the system and the arcs represent their relationships. Arcs have labels indicating a "qualitative" relationship between two parameters. One example of such labels is:

\+ => monotonically increases with
\- => " decreases "

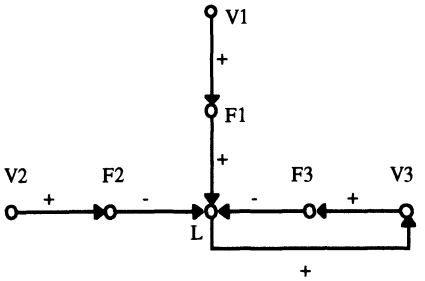
A classical illustration:



L- level of reservoir
Fi - flow in pipe i
Vi - opening of valve i

Figure 5a Hydraulic example

Causal relationships between variables:



Flow F1 increases with the opening of valve V1
Level L increases with the flow F1
The opening of V3 depends on the level L
...

Figure 5b Causal graph

Other authors use other labels:

$X \xrightarrow{Q+} Y$ Y monotonically increases with X

$X \xrightarrow{I+} Y$ the rate of change of Y (dY/dt) monotonically increases with X

Q- and I- are the inverse monotonic relationships.

The purpose of establishing a QM is not to support simulation or prediction. Our main goal is to guide induction (restricting the search space, for instance). In this way, it is not mandatory that a complete qualitative model be defined. We can even think of relationships that we cannot classify in terms of Q or I labels. This means, we know (or guess) there is a relationship between the two parameters but we don't know its nature. Nevertheless this knowledge is useful, as it gives information about the parameters that must be considered in the training set.

In our current work the following examples of qualitative relationships are being used:

$$A \xrightarrow{+} C$$

$$A \xrightarrow{-} C$$

$$(A1 \mid A2) \,\&\, B \xrightarrow{+} B$$

These examples show relations between physical system variables. These relations represent influences (positive or negative) between variables; they do not represent the exact function that specifies the inter-relationship between the variables.

In this kind of relations we can use discrete or continuous variables and use logical relations to represent influences from various variables over a studied variable.

5 CONCLUSIONS

In this paper a general approach to the development of the next generation water distribution management systems was summarized. Two key components of such systems, the distributed information management and learning subsystems, were discussed and their initial functional requirements established.

A federated information management framework was described to properly support the cooperation and information exchange among different involved sites and their activities in intelligent water distribution control network. The autonomy, heterogeneity, need for loose and tight coupling, and the cooperative information sharing were identified as the characteristics of the network nodes. Three kinds of units, namely the "control unit", the "remote unit", and the "auxiliary unit" were described. Further, the results obtained in the analysis steps for building the interoperation layer among the units were presented.

Preliminary experiments with existing historic data show that control knowledge and optimization guidelines can be extracted from these data by the application of conventional machine learning techniques. However, it became also evident that transformations of raw data (extraction of high level features) are a key point in this process, although no general methodology is available. This will require extensive experimental work. The identification of the adequate data transformations required by the learning subsystem will be an important input to the specification of functionalities of the information management system. Complementarily, the use of incomplete qualitative models is foreseen as a promising approach to guide the learning process and to reduce the iterative assessment effort regarding the "quality" of the acquired knowledge.

ACKNOWLEDGEMENT

This work is funded in part by the European Commission, via the Esprit Waternet project. The authors also thank ESTEC and SMAS-Sintra for the supply of historic data and fruitful discussions.

Fernando José Martinelli also thanks CNPq (Brazilian Council of Research and Development) for his scholarship.

6 REFERENCES

Afsarmanesh, H. et al. (1993), Distributed Schema Management in a Cooperation Network of Autonomous Agents. In Proceedings of the 4th IEEE Int. Conf. on "Database and Expert Systems Applications DEXA'93", Lecture Notes in Computer Science (LNCS) 720, pages 565-576, Springer-Verlag, September 1993.

Afsarmanesh, H., Wiedijk, M., and Hertzberger, L.O.(1994) Flexible and Dynamic Integration of Multiple Information Bases, in *Proceedings of the 5th IEEE International Conference on "Database and Expert Systems Applications DEXA'94"*, Athens, Greece, Lecture Notes in Computer Science (LNCS) 856, pages 744-753. Springer Verlag, September 1994.

Camarinha-Matos, L.M., Seabra Lopes, L., Barata, J. (1996) Integration and Learning, in Supervision of Flexible Assembly Systems, *IEEE Transactions on Robotics and Automation,* Vol. 12, N. 2, April 1996, pages 202-219.

Clark, P. and Boswell, R. (1991) Rule induction with CN2: Some recent improvements, in *Machine Learning - EWSL-91* (ed. Y. Kodratoff), pages 151-163, Berlin, 1991. Springer-Verlag.

Clark, P. and Matwin, S. (1993) Using Qualitative Models to Guide Inductive Learning, in: *Proceedings of the 10th International Machine Learning Conference (ML93).* (ed. P. Utgoff), San Mateo(CA-USA): Morgan Kaufmann Publishers inc., 1993. p.49-56.

Dague, P. (1995). Qualitative Reasoning: A Survey of Techniques and Applications. *AI Communications*, v.8, nrs.3/4, p.119-192. Sept./Dec. 1995.

Quevedo J. et al. (1988) A contribution to the interactive dispatching of water distribution system, in *International Symposium on AI, Expert Systems and Languages in Modelling and Simulation*, pp41-46, Barcelona, 1987.

Stroomloos (1994) *Stroomloos*. Rathenau Institute, The Netherlands, 1994.

Tuijnman, F. and Afsarmanesh, H.(1993) Management of shared data in federated cooperative PEER environment. *International Journal of Intelligent and Cooperative Information Systems (IJICIS)*, 2(4):451-473, December 1993.

Wittig, T. (1992) *ARCHON: An architecture for Multi-Agent Systems*, Ellis Horwood, 1992.