KNOWLEDGE DISCOVERY IN ENVIRONMENTAL DATA

JOAQUÍN IZQUIERDO*, JOSÉ L. DÍAZ, RAFAEL PÉREZ,
P. AMPARO LÓPEZ AND JOSÉ J. MORA
Centro Multidisciplinar de Modelación de Fluidos (CMMF)
Universidad Politécnica de Valencia, Spain

Abstract. An approach to tackling management problems of water resources is the introduction of the advanced ICTs (Information and Communication Technologies). In IWM (Integrated Water Management) and, in general, in the environmental field the power of these technologies has allowed for large-scale data collection campaigns. The number of parameters that must be measured to monitor an ecosystem is potentially high. Systematic measurement of those parameters generates huge amounts of data that should be suitably interpreted and used. A pragmatic approach has to be used to get the best information = knowledge from all this data. Within such amounts of data there is a lot of hidden information, in terms of models, patterns and trends. But information is difficult to be extracted since data is of varying quantity and quality. As a consequence, semi-automatic knowledge extraction from data has gained great importance within the economic and scientific community. Knowledge Discovery from Databases (KDD) has emerged as a framework where a plethora of techniques for identifying useful and understandable patterns in data have flourished. Most of those techniques can be used with success in the environmental field and, in particular, in IWM. In this paper we give an overview of what KDD is and mention some applications in these areas.

Keywords: Integrated Water Management, Data Mining, Environmental Databases

* Joaquín Izquierdo Sebastián, Centro Multidisciplinar de Modelación de Fluidos, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022, Valencia, Spain; e-mail: jizquier@gmmf.upv.es, Tel. +34 963879890. Fax. +34 963877981

P. Meire et al. (eds.), Integrated Water Management: Practical Experiences and Case Studies, 51–68.
© 2008 *Springer.*

1. Motivation: Overwhelming Amounts of Data!

According to the guidelines for the implementation of the EU Water Framework Directive, WFD, (Annex V), the list of water quality elements contains biological (priority), hydromorphological (supporting) and physico-chemical (supporting) parameters. The Directives give recommendations on the parameters that must be measured, especially on priority substances and "relevant" substances, for instance, such as the substances in List 1 of Directive 76/464EEC and List 2 of priority substances according to Article 16 of WFD. Obviously the number of parameters that must be measured is potentially high and that is why a pragmatic approach has to be used when the parameters to be measured are selected (Lavkulich, 2004). Systematic measurement of all those parameters will generate huge amounts of data that under the favourable hypothesis of being well organized will produce databases of considerable size that will have to be suitably digested and interpreted if their content is to be useful.

This is only one example of what happens in many companies, industries and research groups in almost any field. Large amounts of data are nowadays available because gathering data is easy and usually inexpensive. As environmental concern grows and information technology develops, more and more data on many different aspects (physical, chemical, biological…) of the environment are gathered. Environmental monitoring is an important source of data. Typically, samples are taken from air, soil, water and analyzed for their properties. Then, laboratory tests designed to estimate the influence of some substances on the environment and the humans themselves add new information. Remote sensing is another source of data: different satellites continuously provide images of the Earth containing information about atmosphere, geology, vegetation and land use dynamics. In addition, such advances as Geographical Information Systems (GISs), which has spread rapidly, are a valuable source of environmental relevant information, for such applications as digital elevation models, insolation assessments and land uses changes.

Most people make use of this information to observe the evolution of certain variables, control abnormal values and write simplified reports. This is an objective in itself. But, without doubt, within such amounts of data there is a lot of hidden information, in terms of models, patterns, trends, etc. Being able to extract this information would give more global perspectives, more interpreted elements, more efficient and reliable environmental and human health indicators, and increase our capability to predict. Such information is difficult to extract, since data is of varying quantity and quality. For one thing, data is abundant and of varying spatial and temporal distribution. In many instances it is too large for a human to process and make useful for the problem at hand. On the other hand, raw data, as collected by people or by sensors, typically include erroneous data.