A Study on the Reservoir Intelligent Inspection System Based on the Internet Things Technology

Qiang Yue\textsuperscript{1}, Fusheng Liu\textsuperscript{1*}, Changqing Song\textsuperscript{2*}, Yanmin Liu\textsuperscript{3}, Ning Cao\textsuperscript{4*}, Russell Higgs\textsuperscript{5}, Junqing Li\textsuperscript{2*}, Gregory M.P. O’Hare\textsuperscript{6}, Cuinan Yang\textsuperscript{4}, Qiang Zhou\textsuperscript{7}

College of Water Conservancy and Civil Engineering\textsuperscript{1}, Agricultural Big Data Research Center\textsuperscript{2}, Faculty of Computer Science\textsuperscript{3}, College of Information Engineering\textsuperscript{4}, School of Mathematics and Statistics\textsuperscript{5}, School of Computer Science\textsuperscript{6}, Department of Computer Science and Technology\textsuperscript{7}

Shandong Agricultural University\textsuperscript{1,2}, Dalhousie University\textsuperscript{3}, Qingdao Binhai University\textsuperscript{4}, University College Dublin\textsuperscript{5,6}, Tsinghua University\textsuperscript{7}

Taian, China\textsuperscript{1,2}, Halifax, NS, Canada\textsuperscript{3}, Qingdao, China\textsuperscript{4}, Dublin, Ireland\textsuperscript{5,6}, Beijing, China\textsuperscript{7}
yueqiang406@163.com, liufsh@sda.edu.cn, cqsong@sda.edu.cn, ning.cao2008@hotmail.com, a397858801@126.com

Abstract—The application of the Internet of things technology brings opportunity for modern dam safety intelligent inspection. Aiming at the present situation of reservoir dam safety inspection data, information island serious deficiencies and poor timeliness problem, the research applied big data thoughts into water management to solve the numerous data, demand diversifications of related interest groups, overall water difficulties and other problems confronted by hydraulic engineering. By using the historical database which contains large amount of data and feedback information, a health early warning model of reservoir was designed and established in large data environment based on decision tree C5.0 algorithm. The Dingdong Reservoirs health status was forecasted with the model as a case study. The results show that the accuracy of the early warning model is 91.92%. The reservoir is in health state, corresponding to no warning level. The early warning model is feasible and effective for utilizing abundant case resources, it can be used widely in reservoir health management. The results obtained in this paper are beneficial to the sustainable development and scientific management of reservoir.

Keywords—big data; early warning; water resources data; reservoir; decision tree

I. INTRODUCTION

With the rapid development of information technology, water information infrastructure and application systems are being increasingly applied to water conservancy projects’ construction and management and also to water administration management; as a result, the related data volume has increased vastly. In general these ‘big data’ resources are gradually changing people’s work and lives, the data as a simple processing object has begun to change into a basic resource [1], [2]. The introduction of big data technology to the water conservancy industry, as the basic technology of water conservancy construction and administration, has become an inevitable trend [3].

Water information covers the water conservancy engineering survey, planning, design, construction, operation management and maintenance. Also covered are flood control, water resources management, water administration management, and soil and water conservation. Water resources can be in multi-ownership and cover multiple categories, so that the associated data arises from dispersed sources and in different forms from varying application services. These realities can restrict the extensive use of the data and hence diminish its value. Thus the collection, storage, transmission, processing and application of water conservancy data has become a problem and challenge facing the development of water conservancy. Enhancing the value of water conservancy data to provide a scientific basis for associated project construction management is an important part of the water resources’ informatization program.

To summarize the characteristics of water conservancy big data, this paper studies water conservancy driven by data and analyzes the big data mining mode and its application in this field. It adopts the optimal decision tree method to study a reservoir health risk early warning system; it initially explores the use of big data methods on water data to enhance its value and intelligent management.

II. DATA DRIVEN AND LARGE WATER DATA

A. Data-driven principle

Data-driven originated from the computer, its basic idea is from data, not depend on the model, through the process of a large number of online and offline data processing and analysis, dig up the implied information in the data, to support decision making, implementation system of monitoring, diagnosis, decision and optimization, etc. Because of the data driven, it is not necessary to establish a global mathematical model of the controlled system, especially for non-linear and uncertain process control. In recent years, the idea of data driven has been applied in various fields, such as engineering control, fault diagnosis, multi-information integration technology [4].

Since there is no fixed model and no need to build the model library, the data-driven controller is relatively simple, fast and robust. For offline data analysis and statistics, it can find the trend of the development of the data quickly and easily, so sensitive to the presence of abnormal data. Data-driven not only can effectively utilize the real-time information received and stored by the sensors in the application environment, but also work with the model-driven method.
B. Water conservancy big data

The long-term business practice of water conservancy information has accumulated a large number of heterogeneous independent business data. Remote sensing, GIS, sensor network and radio frequency technology and other modern information technology development and application of comprehensive development of the water resources of the spatial scale and factor types, water data to form a stable data update and incremental mechanism [5]. Water resources data cover water and soil environment conservation, water resources protection, flood control, water conservancy construction and management and maintenance aspects, water resources data has gradually emerged multi-source, multi-dimensional, large and polymorphic large data characteristics. Water conservancy data is an important scientific basis for administrative decision-making of water conservancy industry, and it is an important resource for water conservancy.

The large value of water resources has the objective value of knowledge, development and practice. The value of large data of water conservancy needs to be realized in the dynamic environment of comprehensive analysis and service of water conservancy information. Integration of the Internet of things, cloud computing and other advanced technology to manage the massive data efficiently, find and make full use of the potential value of the data, making the value to maximize the upgrade.

III. THE KEY TECHNOLOGY OF WATER CONSERVANCY BIG DATA

A. Distributed storage and processing technology

The use of relational database and distributed file system can solve the problem of centralized storage of water conservancy data and unified management of structured data and unstructured data. The Hadoop open source software framework not only supports large data-intensive distributed storage, but also has powerful batch data processing and analysis capabilities, often used for offline data storage and analysis, as a relational database management system. Hadoop divides the application into many small portions, each of which can be executed or re-executed on any node in the cluster. Hadoop provides an HDFS distributed file system for storing data from all compute nodes for the entire cluster a very high data bandwidth, the entire framework can automatically handle node failures. Hadoop architecture uses Zookeeper to provide coordination management services within the cluster, using H Base column database storage and management of data, through Pig, Hive and Mahout to achieve data mining analysis.

B. Large data analysis technology

Traditional data analysis tools only from the simple statistics, query and management data and other aspects of processing, can not tap the potential value of information. Big data technology in the database of massive data automatically extracts the extraction of implicit information, access to the application of the law and the model [6]. The large data analysis technology is based on the traditional analysis method, and integrates the fields of multidisciplinary technology, including statistical analysis, pattern recognition, and machine learning and so on. Among them, there are Bayesian network, artificial neural network and decision tree [7].

Bayesian network is used to represent the variable connection between probabilities of complex causality graph pattern, through reflects the probability that the relationships between data, find credible potential dependency [8]. From the point of view of information processing, artificial neural network is used to simplify the abstract simulation of complex network composed of interconnected neurons, and different neural networks are constructed according to different connection methods [9].

$S$ is a collection of $n$ data samples; the decision attribute has $m$ different values, define $m$ different categories of $P_i (i = 1, 2, \cdots, m)$.

Let $n_i$ is the number of samples in the category $P_i$, and the mathematical expectation of its information quantity is information entropy:

$$I(p_1, p_2, \cdots, p_m) = -\sum_{i=1}^{m} \frac{n_i}{n} \log_2 \left( \frac{n_i}{n} \right) \quad (1)$$

Set the $\{a_1, a_2, \cdots, a_k\}$ to satisfy the condition attribute $A$. $A$ can be decomposed into a set of $k$ sub $S$ for $\{C_1, C_2, \cdots, C_k\}$, $A$ is the test attribute, and the branch of the $P$ of the class collection corresponds to the subset. Let $n_{ij}$ be the number of samples in the subset $C_j$ of $P_i$, by entropy as $A$ divided into subsets:

$$E(A) = -\sum_{j=1}^{k} \left[ \frac{n_{1j} + n_{2j} + \cdots + n_{mj}}{n} \times I(p_{1j}, p_{2j}, \cdots, p_{mj}) \right] \quad (2)$$

$$I(p_{1j}, p_{2j}, \cdots, p_{mj}) = -\sum_{i=1}^{m} \frac{n_{ij}}{n_{ij}} \log_2 \left( \frac{n_{ij}}{n_{ij}} \right) \quad (3)$$

Formula: $(n_{1j} + n_{2j} + \cdots + n_{mj})/n$ is the weight of the first $j$ subset, $I(p_{1j}, p_{2j}, \cdots, p_{mj})$ is the expected information for subset $C_j$.

According to the information gain formula (4), the information gain obtained by mathematical expectation and information entropy.

$$Gain(A) = I(p_1, p_2, \cdots, p_m) - E(A) \quad (4)$$

The information gain rate is calculated as:

$$GainRatio(A) = Gain(A)/SplitI(A) \quad (5)$$

$$SplitI(A) = -\sum_{j=1}^{k} \frac{n_{ij}}{n_j} \log_2 \left( \frac{P_{ij}}{n_j} \right) \quad (6)$$

Decision tree growth to the rate of information gain as the criterion, found the best grouping variable and break point, from top to bottom in turn to construct branch structure, node attribute that test branch represents the outcome of the test, the leaf nodes represent category distribution.
IV. ENGINEERING APPLICATION

Reservoir as a channel of communication between human beings and nature, maintaining the health status is the inevitable result of the progress of Scientific Outlook on Development and social values [10], [11]. According to the connotation of health reservoir [12], [13], selection of four key dam safety, ecological environment, social function, state continuous to evaluate health Dingdong reservoir in line with the principle of scientific, systematic, gradation and maneuverability establishing easy to quantify reservoir health warning index system of the source. The influencing factor that the reservoir health involves is numerous, the information content is huge, the data randomness, uncertainty and fuzziness cause the health to forecast that is hard to adapt to the highly effective movement of reservoir, the traditional data analysis method cannot the deep level excavation in the mass data latent have the value information, needs to use under one kind of big data environment the highly effective practical big data analysis method.

The Dingdong reservoir began filling in September 1997, accumulated a large amount of observational data through surveys and statistical analysis of historical data, training data between 1998 and 2015 and when training completed then set sample, enter the reservoir comprehensive database. Factors affecting the health of the reservoir including reservoir safety, ecological environment, social function and status of sustainability in many indicators, from the long running Dingdong reservoir observation data, the main indicators affecting reservoir health including the piezometric level index C31, water quality safety index C7, underground water level index C10, sediment changes index C14.

We take C5.0 decision tree method for warning models, taking into account the characteristics of reservoir data health warning affair, choose binary tree structure analysis of the health condition, select data of 1998-2010 as the training data used to build the model using; select data of 2011-2015 as test data, used for testing model. According to the information gain maximum principle, piezometric level indicators C31 as a root node test attributes, step by step down choice node attributes and create branch. Use test data set to establish the training model conducts the test, the rate of accuracy achieves 91.92%, and the accuracy of the model is good. According to the forecast index values in the coming month, using decision tree model to predict, the state of Dingdong reservoir is “healthy”. Alarm grade is no warning, early warning signals for “green” in a month.

V. CONCLUSIONS AND RECOMMENDATIONS

Big data technology promotes the rapid development of water conservancy data acquisition, management and application. Promoting application value from huge amounts of data and making the data to lead decisions is a kind of new way for scientific decision-making. The Big data thinking from the point of view of data applications, efficient use of large data resources. With the increasing influence of big data technology on management decision making, the situation that depends on intuition to make decisions will be changed completely.

The discovery of big data value can not be separated from the value orientation and application direction of artificial identification. Manual identification is the key to make big data work. How to apply the massive data of water resources to more fields is an important task for the application of water resources.

VI. ACKNOWLEDGEMENTS

This work is supported by the National Key Technology Research and Development Program of the Ministry of Science and Technology of China (2015B010305), the Water Conservancy Scientific and Technical Program of Shandong Province (SDSLKY201305) and the significant application of agriculture technology innovation Program of Shandong province (SDNYCX1531963).

REFERENCES