Exploration on Big Data Architecture and its Correlation Analysis for Safety Monitoring in Operation of Basin's Hydro-junctions

Shuangping Li¹,², Zuqiang Liu¹,²*, Min Zheng¹,² and Huawei Wang¹,²
¹ Changjiang Spatial Information Technology Engineering Co., Ltd. Wuhan, Hubei, 430010, China
² Hubei Province Engineering Technology Research Center for Water Resources Information Perception and Big Data, Wuhan, Hubei, 430010, China
*Corresponding author’s e-mail: 120664509@sina.com

Abstract. In order to promote the application of big data analysis technology in the safety management in operation of basin's hydro-junctions, the concept of big data and the "5V" characteristics are briefly described in this paper, which explains all the basic characteristics of monitoring data with big data. At the first time, it is proposed that the big data for safety monitoring in operation of basin's hydro-junctions is composed of four dimensions: monitoring data of survey, design and construction, real-time monitoring data during operation and maintenance period, management data in operation and maintenance of dams and social resource data. The general relationship between the various factors of monitoring data, and the classification, process and mining method of correlation analysis for monitoring big data are analyzed. The method of analyzing the safety monitoring data in operation of basin's hydro-junctions has been broadened, which meets the needs of the information and intelligence in management and operation of basin's hydro-junctions.

1. Introduction

1.1 Concepts and characteristics of big data
The development of data science has experienced three stages: the generation of data, the formation of data and the birth of big data. The continuous development of each stage has also deepened the concept of people using data to see the world. With the development of modern measurement technology and smart devices, the form of data has evolved from traditional structured data to complex data forms consisting of structured data, semi-structured data and unstructured data. With the rapid development of Internet and mobile internet, people can share the platform of these data. The massive data generated by these platforms are discovered and utilized by people, and a new concept of big data emerges. Big data analysis is an important technical means for big data applications. Big data analysis adopts the latest data analysis model from technical means. Through the unique correlation between data, many related and valuable conclusions can be drawn.

Big data refers to a collection of data that cannot be captured, managed, and processed by conventional software tools within a certain time frame. It is a huge, high growth rate and diversified information asset that requires a new processing mode to have stronger decision-making power, insight and process optimization ability. This is the definition of "Big Data" given by Baidu. The first
The core point of the book *The Era of Big Data* [1] by Victor Meyer-Schonberg and Kenneth Cooker is that big data is full data. In the age of big data, “not random samples, but all data”, understanding things is no longer from random samples, but from all data. Big data has the following five characteristics [2].

- **Volume**: Huge volume of data, including the amount of collection, storage and calculation. A large number of TB, PB, EB, and ZB data are waiting to be processed.
- **Variety**: There is a wide variety of data sources and types. Data can be obtained from multi-measures, multi-platforms, multi-sensors, etc. Data types include structured and unstructured data such as text, images and videos etc.
- **Velocity**: Data grows with high velocity, and it need fast processing and high timeliness.
- **Veracity**: Data is not veracity due to noise, missing, inconsistency, ambiguity and so on.
- **Value**: Big data makes people quantify and understand the world in an unprecedented dimension, which contains tremendous value. The ultimate goal of big data is to mine value from data.

Academician Li Deren believes that spatio-temporal data mining is the process of automatically discovering and extracting implicit, non-obvious patterns, rules and knowledge from massive, multi-source spatio-temporal big data. Obviously, data mining is more difficult than data processing and information extraction, and requires intelligent reasoning based on big data and knowledge base [2]. In fact, safety monitoring data is also typical spatio-temporal data, and it is also necessary to use the big data analysis technology to perform correlation mining.

### 1.2 Characteristics of safety monitoring data

For a single large-scale water conservancy project, the dam safety monitoring data also has the characteristics of "5V". For example, the monitoring points for direct observation of the dam in the Three Gorges Project are more than 12,000. At present, there are more than 5,300 automatic monitoring points connected. Furthermore, there are environmental monitoring and geological disasters monitoring in the dam and reservoir area, so the number of sensor devices is large, and there is a huge amount of automatic acquisition data and manual reading data. In the aspect of "Variety", there are more than a dozen types of dam monitoring projects, including deformation monitoring, seepage and pressure monitoring, stress and strain monitoring, water level monitoring, inspection, video surveillance and other data, which of data types include structured and unstructured data such as numerical, text, pictures, images and so on. In the aspect of "Velocity", the Three Gorges Project contains a large number of data such as automatic measurement points and video surveillance, which grows fast, and the processing speed and timeliness are demanded. In the aspect of "Veracity", there are a lot of noise data, vacancy data and inconsistent data in monitoring data due to faults of acquisition equipment, data transmission problems and manual input errors. In the aspect of "Value", because dam safety monitoring is the ear and eye of monitoring dam safety operation, the data itself has great value. For the centralized safe operation and management of reservoirs and dams at basin level, the "5V" characteristics of monitoring big data are more obvious.

Due to the geological conditions, water level, temperature, aging and so on, there are many factors affecting the dam deformation, which are interrelated and very complex. The difficulty of monitoring work lies in mining the relationship between effect sizes, effect sizes and cause sizes, cause sizes, measurement data and construction process from a large number of complex and different data [3]. From these massive data, valuable information or knowledge can be extracted, the effects and rules of various factors on dam safety can be explored, and the dam safety status can be analyzed and predicted. The technology of big data analysis is to find correlative data from massive data, and then to analyze and verify them by establishing complex models, and finally to draw a conclusion that meets the needs. This is the essential feature of big data analysis, which emphasizes correlation but neglects causality. It is also an important aspect of monitoring data correlation analysis. Therefore, we believe that the dam safety monitoring system is a typical example of big data application. The "5V" characteristics of big data naturally fit in with the dam safety monitoring.
2. Big data architecture for safety monitoring of basin's hydro-junctions

Over the past ten years, dam safety monitoring technology has developed rapidly in China. The development of dam safety monitoring information management system or decision support system has attracted more and more attention, especially the preliminary development and application of dam safety monitoring information system for basin/regional reservoirs and dams. In order to analyze and process a large amount of data generated by the safety monitoring system, the large hydropower enterprise group began to build big data center to centralize the management of the monitoring system of the basin/regional reservoir and dam group, which laying a good foundation for the analysis and early warning of the operation safety of the reservoir and dam group, and providing decision support for the responsible persons and departments.

From the Volume characteristics of big data, the amount of monitoring data is small. From the core point of view of "big data is full data"[1], we can explore the structure of monitoring big data. There are four main dimensions for the big data structure of the reservoir and dam group (Figure.1).

![Figure 1. Structure of monitoring big data of the reservoir and dam group](image)

2.1 Monitoring data of survey, design and construction

The geological structure data of dam foundation, dam design data and monitoring data during construction period are all important components of monitoring data. Among them, the geological structure of dam foundation and dam design data are unstructured data, and the monitoring data during construction period can be stored in the form of relational database, which belongs to structured data. At present, there are many analytical tools and means to analyze structured data, which belong to traditional data analysis, but it still needs to be compared and analyzed by using big data analysis software [4].

2.2 Real-time monitoring data during operation and maintenance period

During the operation and maintenance period of reservoirs and dams, the monitoring automation system collects data by various sensors installed on the dam and its adjacent rock mass, and then transmits, stores, collates, processes and analyses the data. As the design life of reservoirs and dams is generally over 100 years, with the passage of time, the monitoring automation system will produce huge amounts of structured data.

In the stage of external loads such as rainstorm, flood, earthquake and disaster of dam itself, under the background of big data, monitoring data processing originates from real-time data acquisition, of which the goal is to quickly analyze and process monitoring big data, and to form text, voice and pictures to report to the competent department.
2.3 Management data in operation and maintenance of dams

2.3.1 Operation and scheduling data of dams. Cascade hydropower station group system is a dynamic multi-objective complex system with huge scale, numerous units, complex structure and intricate relationship. In the specific process of reservoir dispatching, there are various sources of information, such as hydrological and rainfall data for hydrological forecast, reservoir operation information, grid operation information, and meteorological information, etc. [5] In order to analyze the relationship between dam operation and dam deformation by using big data, it is necessary to acquire information automatically through the established "cascade reservoir dispatching automation system".

2.3.2 Operation and maintenance data of hydraulic gates. Hydraulic gate is the main equipment for regulating and controlling water flow in hydropower station, and it is an important part of hydraulic structure of hydropower station [6]. Whether the operation of hydraulic gate is normal or not is directly related not only to the power generation benefits of hydropower stations, but also to the safety of dams and people's lives and property downstream. The operation and maintenance of hydraulic gates are closely related to pier deformation, so it is necessary to obtain relevant information.

2.3.3 Monitoring data of hydraulic turbine. The vibration of the hydraulic turbine is often caused by mechanical, hydraulic and electrical reasons. The emergence of vibration has a direct impact on the safe and stable operation of hydraulic turbine, affects the economic benefits of hydropower stations, and threatens the safety of hydropower stations. Especially at present, hydraulic turbine is developing in the direction of high specific speed and large capacity, which makes the structure of the unit increase, the stiffness decrease relatively, and the vibration problem will be more prominent [7]. Therefore, it is of great significance to study the deformation and dynamic response monitoring of spiral case structure change during hydraulic turbine operation.

2.3.4 Maintenance and reinforcement data of dams. Safe operation of dams throughout their life cycle is accompanied by hydrology, hydraulics, geology, earthquake, aging and other risk factors, which inevitably lead to defects or problems, such as cavitation of overflowing buildings, corrosion of metal structures, abnormal leakage, cracks, landslides, etc. [8]. These problems will lead to changes in dam safety and affect the safe operation of the dam. Therefore, combined with dam safety monitoring and inspection, these potential safety hazards should be found and eliminated in time to ensure the operation of the dam under design conditions. Therefore, the process and effect of dam maintenance and reinforcement are highly correlated with dam monitoring data.

2.4 Social resources data
The safety of reservoir and dam group is a worldwide problem. Managers hope that they can comprehensively analyze and compare the accident causes of relevant dams around the world through their own monitoring system data, and give the problems existing in the operation of dams and how to solve them. For this reason, besides the data of monitoring system itself, a large number of social resources data, such as scientific and technological reports on dam monitoring, safety evaluation, monitoring, early warning and maintenance and reinforcement, are needed all over the world to construct a new monitoring big data to solve the safety problems of dam operation that people are concerned about. Social resources data are mostly unstructured or semi-structured data, which is difficult to use. Search and extraction of social resources data mainly use semantic search engine and data mining algorithm [3].

3. Correlation analysis of big data for safety monitoring

3.1 Correlation analysis
Traditional correlation analysis is based on theoretical models, with a small amount of accurate data as support, and then solved by some specific algorithms. However, for complex systems and open and semi-open systems, traditional correlation analysis methods are difficult to establish theoretical models for them. Nowadays, under the background of big data, some complex mechanisms that cannot be resolved can be characterized by some correlation law between lots of data.

Generally, the correlation analysis of dam safety monitoring data only analyzes the effect sizes such as dam deformation, crack, seepage and seepage pressure, stress and strain monitoring, and some cause sizes such as the reservoir water level, temperature, aging and other reasons. The qualitative correlation between the various effect sizes can be represented by Figure 2.

![Figure 2. Correlation analysis chart](image)

### 3.2 Correlation analysis of big data

With the rapid development of information technology, the ability of data acquisition and data storage has been greatly improved. Especially, with the extensive development of automatic monitoring in dam safety monitoring industry, huge amounts of data has been rapidly accumulated, and large-scale and super-large-scale data storage systems keep emerging. The change of any effect sizes in the monitoring system is determined by other cause sizes interacting with it, and the information that determines the law of dam variation is distributed among all variables associated with it, and it is difficult to model and characterize by any single variable.

#### 3.2.1 Analytical methods and classification

From the difference of correlation types, multivariate data correlation analysis methods can be divided into linear and nonlinear correlation analysis methods. At present, linear correlation analysis methods are mature, such as commonly used canonical correlation analysis, linear regression, principal components analysis, independent components analysis, cross-correlation functions, consistency functions, causality analysis and so on. However, in reality, the relationship between systems is not always linear, and the characteristics of non-linearity,
non-equilibrium and collinearity often appear. As a result, linear correlation analysis often fails to meet the needs of analysis, and often leads to some erroneous conclusions. For nonlinear correlation analysis, when the number of variables is small, we can select a suitable nonlinear function to fit by observing the scatter plot. However, these methods are not applicable as the number and complexity of variables far exceed the capabilities of manual processing. In recent years, the popular methods of non-linear correlation analysis are mainly based on fractal theory, recursive graph theory, mutual information theory, grey system theory, non-linear regression theory, dynamic time warping method, dynamic frequency warping method and correlation rule mining, and some achievements have been made in many fields.

From the purpose of correlation analysis, the multivariate data correlation analysis method can be divided into two categories: feature extraction method and factor selection method. Feature extraction is to extract the main information by mapping or transforming the original variable space into low-dimensional feature space; Factor selection method can be regarded as a special form of feature extraction method, which aims to ensure the performance of model prediction by removing redundant variables and irrelevant variables in the original variables. This method directly processes the input variables, so it has obvious rapidity and good universality.

At present, the main correlation analysis methods can be roughly classified according to Table 1. Each method has its own emphasis in dealing with multivariate data correlation analysis. Fractal theory, mutual information theory and grey system theory can also be used for linear correlation analysis, while mutual information theory, grey system theory, dynamic time distortion and dynamic frequency distortion can be used for feature extraction and factor selection. Therefore, Table 1 is not a strict division.

| Classification | Feature extraction method | Factor selection method |
|----------------|--------------------------|-------------------------|
| **Linear Analysis** | canonical correlation analysis, linear regression, principal components analysis, independent components analysis | cross-correlation functions, consistency functions, causality analysis |
| **Nonlinear Analysis** | fractal theory, recursive graph theory, mutual information theory, grey system theory, non-linear regression theory, dynamic time warping method | dynamic frequency warping method, correlation rule mining |

3.2.2 **Correlation analysis process.** According to the characteristics of multivariate data obtained by safety monitoring multi-source information, combined with the traditional ideas of data processing and analysis, when the correlation analysis is performed for given multivariate data, the process can be roughly divided into the following steps [9].

- Data preprocessing: Firstly, the algorithm need to preprocess the original data, and unify the data format, then clean the data to fill the missing data and eliminate the outliers. This part of the work is often the most cumbersome and time-consuming;
- Data compression: Secondly, according to the characteristics of automated monitoring of multivariate data, data compression is carried out. It need to reduce dimensionality of high-dimensional data and intercept long-cycle data, thereby the algorithm can reduce the data size to a scale that can be processed by the computer; how to compress the monitoring data is a hot issue, and the relevant theoretical methods are also rich;
● Correlation analysis: Thirdly, it is the core algorithm part of the correlation analysis method, which uses various correlation models to calculate the correlation indicators for exploratory correlation analysis of the processed data.

● Evaluation and Interpretation: Finally, the validity of the results is checked with expert experience knowledge, and the correlated results are evaluated and interpreted, so that the correlated results can be transformed into applicable knowledge.

The complete process of correlation analysis is shown in Figure 3.

3.2.3 Data mining of correlation analysis. The general correlation relationship in the dam monitoring big data has been analysed above (Figure 2). The dam monitoring system of reservoir is a typical complex large-scale system, it has many data sources. The causes of dam deformation are affected by geological conditions, water level, temperature, aging and so on. There are many effect sizes, which are interrelated and very complex. How to mine the relationship among effect sizes, effect sizes and cause sizes, cause sizes, observation and construction process from huge complex and structurally different data is the purpose of correlation analysis for monitoring big data. That is to say, it is our mission to mine the relationship that cannot be recognized by analysis technology at this stage from the big data of monitoring.

At present, the mining method of correlation analysis for monitoring big data can only adopt the basic methods such as traditional grey correlation theory, linear regression and non-linear regression (Table 1). How to deal with unstructured data and how to use modern special software and intelligent algorithm for big data to carry out exploratory correlation analysis are the goals of correlation analysis.

4. Conclusion
With the rapid development and application of modern technologies such as artificial intelligence, big data, cloud platform, Internet of Things and mobile Internet, it has become a reality to collect massive and multi-source perceptual information in real time with automated and intelligent security monitoring system. A large amount of monitoring data has been accumulated rapidly. How to analyze and process these increasing massive monitoring data is a difficult problem we are facing. Using the method of correlation analysis, discovering rules from monitoring data, mining knowledge from monitoring data, and obtaining valuable information from monitoring data, we can better manage these monitoring data, which have great significance and function for analysing the safety behaviour of dams, exploring the effects and rules of various effect sizes on dam safety, evaluating and predict dam safety, and promoting the construction of intelligent dams and the safe maintenance of hydro-junctions.

Acknowledgments
This study is supported by the Changjiang Survey Planning Design And Research Co., Ltd. Research Project, No.CX2018Z26.

References
[1] Victor, M., Kenneth, C. (2013) The Era of Big Data. Zhejiang People’s Publishing House, Hangzhou.
[2] Li, D. (2016) Looking Into the Geospatial Informatics in the Era of Big Data. Journal of Surveying and Mapping, 45(4): 379-384.
[3] Wang, H., Zhai, W.M., Jiao, Y.Y., et al. (2014) Geotechnical Engineering Monitoring in the Big Data Era-Turning and Opportunity. Rock and Soil Mechanics, 35(9): 2634-2641.

[4] Yang, J. (2016) Application of Big Data Analysis Technology in Bridge Health Monitoring. Modern Transportation Technology, 13(3): 30-32.

[5] Guo, S.L., Chen, Y.H., Liu, Pan., et al. (2010) Research Progress and Prospects of Joint Optimization Scheduling of Reservoirs. Advances in Water Science, 21(4): 496-503.

[6] Xing, L.S., Zhou, J.H. (2012) Operation Accidents and Countermeasures for Hydraulic Gates of Hydropower Stations. Water Power, 38(8): 70-73.

[7] Yang, H., Liu, G, Liu, C., et al. (2014) Status and Development Trend of Hydropower Unit Condition Monitoring. Journal of China Institute of Water Resources and Hydropower Research, 12(3): 300-30.

[8] Feng, Y.X. (2017) Summary of Dam Operation Safety Management of Hydropower Station. Dam and Safety, (2): 1-6.

[9] Zhang, P.L. (2015) Multivariate Data Correlation Analysis Method Based on Maximum Joint Information Coefficient and its Application. National University of Defence Technology, Changsha