Journal of Electronics and Information Science (2019) 4: 1-7
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2019.41011
ISSN 2371-9524

Water Conservancy Data Acquisition and Big Data Service Based on Multi-data Sources

Xu Zhu

Peking University, Beijing, China
1500010673@pku.edu.cn

Keywords: multiple data sources; water conservancy data acquisition; data cleaning; ARIMS; big data services

Abstract: To reflect the application value of data development, based on the data of multiple data sources, the water conservancy and the big data service were studied. First, the acquisition of public data was studied. Computers were used to quickly and efficiently collect data into libraries, which greatly reduce the difficulty of data acquisition. Then, the method of data cleaning was determined to improve data quality and enhance the effectiveness and reliability of the data in the application process. Finally, the water conservancy prediction model was applied to the flood prevention decision-making service system based on the integrated platform. The results showed that the acquisition of public data greatly improved the efficiency of data acquisition. By cleaning the obtained data of repeated values, error values, outliers and missing values, higher quality water situation data was obtained. The water conservancy prediction model improved the accuracy of the prediction, and the flood control decision service system provided an efficient and operational integrated platform. Therefore, the water conservancy prediction model has a certain guiding role in flood control decision-making. It is the key to big data services for water conservancy.

1. Introduction

After years of construction management, various departments of the water conservancy industry have accumulated a large amount of data information. It includes meteorological data, water quality data, ecological environment data, geological disaster data and other measurement data, as well as national water resources census results and other water-related information, such as geographical information, economic information, etc. [1]. These data have accumulated over the years and eventually formed a large data set of water conservancy. More and more industries, including the Internet, finance, transportation, energy and other fields, have begun to launch big data applications. The water conservancy is the basic industry of economic development and national economy, so it is imperative to develop the big data service of the water conservancy [2].

In recent years, with the continuous improvement of China's water resources sharing mechanism and platform, as well as the formulation of relevant legal systems and regulations, the water conservancy's data sharing work is progressing steadily [3]. In addition, with the development of sensing technology and emerging media, non-traditional types of water conservancy data are

increasing. Taking urban sudden floods as an example, large amounts of data are generated at each stage of the early warning process, the disaster occurrence process, and the post-disaster recovery process. This includes not only hydrological, ecological, hydrological, and meteorological data, but also geographic data, drainage facility data, topographical topography, and river distribution data. In addition to the traditional data collected by hydrometeorological detection systems, there are still a large number of flood-related media data on the Internet [4].

Adequate water conservancy data presents new opportunities, but it also presents challenges. Fast and effective acquisition of public data, effective management of data, comprehensive data mining is the basis for improving the value of data. It is also the basis for water conservancy transformation using big data from water conservancy. Water conservancy construction and water resources management provided key business services and became the key and core value of the big data service of the water conservancy.

2. Literature review

The analysis of the research status at home and abroad shows that for data collection, the current problems are mainly concentrated in the large amount of data, diverse data sources, multiple data formats and low data density. The large amount of data requires that the method of data collection is sufficiently efficient. Data sources are diverse in traditional data collected by paper hydrological yearbooks and devices, as well as a large amount of text, images and other information on the Internet. Various types of data include not only structured data such as text and numbers, but also a large amount of unstructured data such as images, links, voice, and video. How to purify data faster through model algorithms is an urgent problem for water conservancy big data.

Varotsos et al. studied the concept, characteristics, and development of big data at home and abroad. Common basic technologies and cutting-edge technologies in data collection and sensing, data storage and processing, data analysis, data visualization, and big data security and privacy protection are analyzed. The latest research directions of these technologies are pointed out. The technical and policy challenges of big data are summarized. Its technical nature is analyzed. This has guiding significance for the research and engineering application of big data [5].

Zheng et al. analyzed the big data related to urban sudden flood disasters and their characteristics. The difficulty of dealing with big data processing related to urban sudden flood disasters is deeply analyzed. In addition, combined with the characteristics of urban sudden flood disaster data, a hybrid big data analysis tool based on Hadoop was designed. It plays a certain role in the research of big data application in the field of urban sudden flood disasters [6].

Based on the in-depth study of historical water data, Shafiee et al. proposed a combined water level prediction model that considers time correlation and spatial correlation. The model can provide effective predictions for water levels from one to six hours, so that it can provide information for flood control and disaster mitigation during urban flood seasons [7].

3. Methodology

3.1 The research route

The research route is shown in Figure 1. Under the guidance of this research route, the characteristics of multiple data sources, data acquisition and cleaning techniques were studied. At the same time, the big data service system of the water conservancy industry was explored.

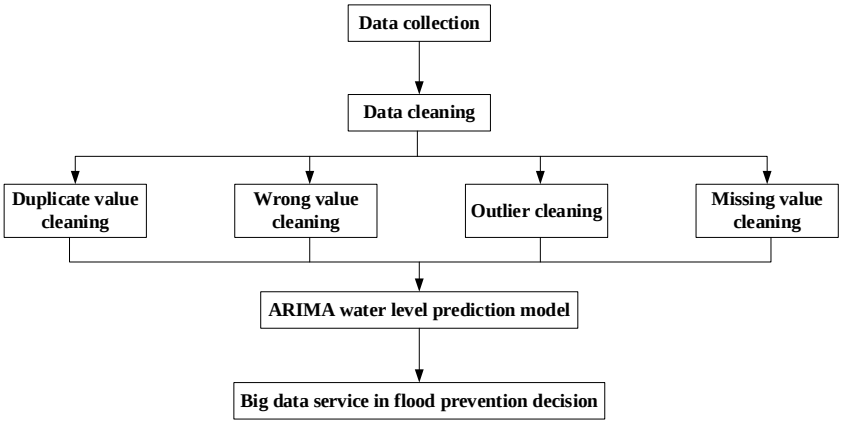

Figure. 1 The research route

3.2 Acquisition of water conservancy data

The government portal is an important part of e-government construction. This is also an indispensable channel for government interaction with the public. Government portal is of great significance for promoting the openness of government affairs and administration according to law, accepting public supervision, improving administrative management and fully performing government functions. The portals of the water conservancy departments at all levels disclose information on water conditions and rainfall conditions to a certain extent. For example, the “Online Services” section of the website of the water conservancy office in Shaanxi Province published information on the province's water situation bulletin and rainwater briefing from June 2010 to the present. The “Convenient Information” section of the website of the Beijing Water Affairs Bureau provides eight data public services, such as urban rivers and lakes, surface water quality, and large and medium-sized reservoirs. The National Environmental Protection Department data center published tens of millions of data on eighteen major items and seventy-six small items such as weekly water quality, hourly air quality, and solid waste management.

According to their respective characteristics, local news websites have also formed a unique water conservancy information carrier, and the public can obtain official water conservancy information in the first time. For example, the water conservancy website of the water conservancy hall in Hebei Province includes water conservancy news, local water affairs, world micro-text, water conservancy video, and graphic description. The water conservancy news is presented in a combination of text, images and videos. The focus is highlighted and the characteristics are distinct.

Compared with traditional news media websites, social networking platforms are also time-sensitive and more focused on randomness, and everyone can participate in the discussion of water events. Take Sina Weibo as an example, individuals can browse and comment on interesting information on Weibo as viewers, or publish content on Sina Weibo as publishers for others to browse. Similarly, content posted on Weibo also supports various forms of text, images, and videos. For example, after the incident, the masses will learn the truth through various channels and conduct extensive discussions on Weibo. Therefore, it is very important to understand the public opinion. In a three-dimensional, multi-level and objective way, the emotional trend of the public is understood. This can provide scientific and comprehensive reference for decision-making and research.

3.3 Data cleaning method

Data cleaning refers to the conversion of noise data into data that meets data quality

requirements by using mathematical statistics, data mining, and other methods. The conditions to be met in the data cleaning process are shown in Table 1.

Table 1 Conditions required for the data cleaning process

Number	Conditions
1	Both single data sources and multiple data sources need to be able to detect and remove data with significant errors and inconsistencies.
2	Factor intervention and user programming effort are reduced as much as possible. The program has some scalability, which is easier to apply to the cleaning of other data sources.

Data cleaning can be divided into the following four types according to different implementation methods and scopes:

First, manual implementation. Noise data is found and processed by manual inspection. This method is less efficient. In the case of a large amount of data, it is difficult to complete data cleaning.

Second, the preparation of applications. This method is relatively rigid and not flexible enough. Although some specific problems can be solved, when cleaning multiple different data sets, the amount of programming increases and the program becomes complicated.

Third, the problem of a specific application domain. For example, according to statistical principles, records of numerical anomalies are looked up. This approach has been implemented in mature commercial systems.

Fourth, data cleaning independent of a specific application domain.

4. Results and discussion

4.1 Establishment of data cleaning method

Due to the diversity, complexity, and various uncertain factors in the data collection, transmission, and storage processes, noise data that does not meet the quality requirements is introduced in the data acquisition process. Abnormalities, errors, repetitions, and missing data, as shown in Figure 2, are all noise data. These data mainly include missing values, outliers and error values. Analysis of these data will lead to inaccurate or even wrong conclusions. Therefore, by cleaning the noise data, the data quality can be improved, and the accuracy of the prediction analysis result can be improved.

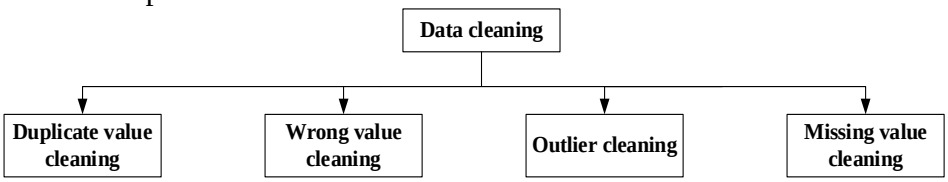

Figure. 2 Data cleaning

Most statistical methods assume that the input data is complete and does not contain missing values, but most data sets in the real world contain missing values. Therefore, missing values need to be removed or replaced with reasonable replacement values before analysis. Inference can be used to recover missing values when there is redundant information in the data or when external information is available. The specific cleaning method is shown in Table 2.

Table 2 Data cleaning method

Cleaning object	Cleaning content
Duplicate value	The duplicate value is due to the presence of duplicate data in the data source, or the collection process has repeatedly collected the same set of data. For the cleaning of duplicate data, it is necessary to use the uniqueness limit of primary key and union primary key when importing the database to complete the cleaning.
Error value	The tuple with the error data is first detected and then processed. Knowledge of the nature of the data is used to find error values. In general, when defining a data dictionary, there is a basic rule for the data.
Abnormal value	Abnormal values refer to data with a relatively large relative error in the obtained data, and also refer to data that is significantly inconsistent with other data in a batch of data. Commonly used outlier detection methods are: distance-based outlier detection, density-based outlier detection, and cluster-based outlier detection.
Missing value	There are many reasons for the lack of data. Taking water conservancy data collection as an example, it may be due to data loss caused by manual reasons, or it may be caused by data logging problems, network connection failures, etc. There are two ways to eliminate the missing data: the instance containing the missing data is eliminated, and the reasonable data is used to replace the missing data.

4.2 Selection of water level prediction model

The water level prediction model is mainly prepared for urban flood control. By predicting the change of water level, the water level information is provided to decision maker. The forecast period is short. At the same time, the requirements for prediction accuracy are relatively high. The model can transform the non-stationary time series into stationary time series by differential processing. The water level prediction model is constructed by considering the influence of the past water level on the current water level. It is a model that is more suitable for water level prediction.

In table 3, the model has a complete set of modeling methods and steps, which are mainly divided into four parts: In the first part, the premise of model modeling is to test the stationary time series of data stationarity. Therefore, the stationarity of data was first tested. The unstable data are smoothed by differential processing. The stationarity test adopts the augmented unit root test in the statistical method, namely ADF method. In the second part, the model parameters are determined and the values of the parameters p, d, q in the model are determined. Among them, d refers to the difference times after the data is stabilized. The p value is determined by analyzing the partial autocorrelation coefficient (PACF) of the water level sequence. The q value is determined by analysis of the correlation coefficient (ACF). In the third part, the water level data is used to establish a model to predict the water level in the short term. In the fourth part, the prediction results are analyzed. At the same time, the relationship between the predicted value and the observed value is plotted. The relative error between the predicted value and the observed value is calculated and analyzed, and the predicted result is evaluated. Relationship between three models and ACF and PACF is shown in Figure 3.

Table 3 Relationship between three models and ACF and PACF

	ACF	PACF
AR	trailing	p step truncation
MA	q step truncation	trailing
ARMA	trailing	trailing

4.3 Big data service in flood prevention decision

Data acquisition and data cleaning techniques are applied based on an integrated platform for knowledge visualization. Combined with the water level prediction model, the flood control decision support system was established to provide big data services for flood prevention decisions. The system can provide scientific and reasonable guidance to decision makers in case of urban flood. The damage was minimized. The urban flood control decision support system is divided into four functional modules: basic information management module, water condition warning module, historical water statistics analysis module and system management module. The functional structure diagram is shown in Figure 3.

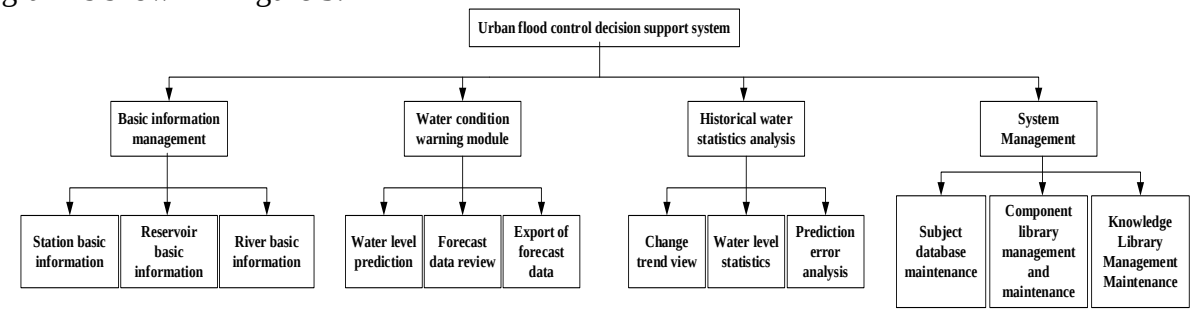

Figure. 3 The functional structure diagram

The basic information management module can modify, delete, add, and query basic information of reservoirs, stations, and rivers through the basic information management module. The water condition warning module and water level prediction are the core content of this system. It can predict the water level in the next one to seven days. If the predicted value of a station exceeds the warning value, the data will be marked with red, which provides a reference for flood prevention. The historical water statistics analysis module can view the water level value of the specified site and the specified time period, and display the results in the form of a line chart. At the same time, the water level data distribution within the specified date range is counted, including the average, maximum, minimum, and mode of the water level data. The predicted value and the measured value in the specified time range are extracted and compared, and the error statistical table and the error statistical graph of the two are obtained. By analyzing the error statistics, the accuracy of the water level prediction and the error variation law in the time range are summarized, so that it can evaluate the previous prediction results and provide reference for the improvement of water level prediction in the future. The system management module maintains and expands the database, component library, and knowledge library to improve the system defects and rich system functions. The flood control decision service system can quickly generate water level prediction results according to different stations and time periods. It has strong versatility.

5. Conclusion

From the perspective of technological development, the emergence of big data and its diversity is inevitable, which leads to many problems in data acquisition, collection and analysis. Using data

cleaning, the model was selected for water level prediction. This model is applied to the flood prevention decision service system based on the integrated platform. This has a certain guiding role in flood control decisions. The main conclusions include the following aspects:

First, the water conservancy data published on the government website was obtained. Some time-sensitive recent data and a large amount of historical data are quickly acquired. The efficiency of data acquisition has been greatly improved, which makes the acquired data more accurate.

Second, by cleaning the acquired data with repeated values, error values, outliers, and missing values, higher quality water data can be obtained. At the same time, the accuracy of subsequent model predictions is indirectly improved, which makes the prediction of the model more accurate.

Third, the selected water conservancy prediction model can provide an efficient and operational integrated system for the flood control decision process to a certain extent.

References

[1] Yang Y, Zhao R, Zhou S, et al. Integrating multi-source data to improve water erosion mapping in Tibet, China. Catena, 2018, 169, pp. 31-45.

[2] Wang J, Kang S, Sun J, et al. Spatial prediction of crop water requirement based on Bayesian maximum entropy and multi-source data. Transactions of the Chinese Society of Agricultural Engineering, 2017, 33(9), pp. 99-106.

[3] Bai Y, Bai X, Meng-Ting H U. Research on Io T-based Water Environment Benchmark Data Acquisition Management. Journal of Anhui Radio & Tv University, 2017, 94(1), pp. 012145.

[4] Mcdonald S, Groff C, Low M, et al. How Boynton Beach Florida Is Pioneering Optimization of Water Utilities Through Integration of Big Data Analytics to Create Gis-Centric Decision Support Dashboards. Proceedings of the Water Environment Federation, 2017, 2017(6), pp. 4864-4869.

[5] Varotsos C A, Krapivin V F. A new big data approach based on geoecological information-modeling system. Big Earth Data, 2017, 1(1-2), pp. 47-63.

[6] Zheng H, Hong Y, Long D, et al. Monitoring surface water quality using social media in the context of citizen science. Hydrology & Earth System Sciences, 2017, 21(2), pp. 949-961.

[7] Shafiee M E, Barker Z, Rasekh A. Enhancing water system models by integrating big data. Sustainable Cities & Society, 2018, 37, pp. 485-491.