A DATA-DRIVEN APPROACH: URBAN WATER QUALITY PREDICTION THROUGH UBIQUITOUS DATA

[1]Mr.V. RAJA SEKHAR, [2]PURIMETLA GURU SAI
[1]Assistant Professor, [2]MCA Student
Department of Master of Computer Application,
Rajeev Gandhi Memorial College of Engineering and Technology
Nandyal, 518501, Andhra Pradesh, India.

ABSTRACT

Urban water quality is of great importance to our daily lives. Prediction of urban water quality help control water pollution and protect human health. However, predicting the urban water quality is a challenging task since the water quality varies in urban spaces non-linearly and depends on multiple factors, such as meteorology, water usage patterns, and land uses. In this work, we forecast the water quality of a station over the next few hours from a data-driven perspective, using the water quality data and water hydraulic data reported by existing monitor stations and a variety of data sources we observed in the city, such as meteorology, pipe networks, structure of road networks, and point of interests (POIs). First, we identify the influential factors that affect the urban water quality via extensive experiments. Second, we present a multi-task multi-view learning method to fuse those multiple datasets from different domains into an unified learning model. We evaluate our method with real-world datasets, and the extensive experiments verify the advantages of our method over other baselines and demonstrate the effectiveness of our approach.

1. INTRODUCTION

Urban water is a vital resource that affects various aspects of human, health and urban lives. People living in major cities are increasingly concerned about the urban water quality, calling for technology that can monitor and predict the water quality in real time throughout the city. Urban water quality, which serves as “a powerful environmental determinant” and “a foundation for the prevention and control of waterborne diseases” [1], refers to the physical, chemical and biological characteristics of a water body, and several chemical indexes (such as residual chlorine, turbidity and pH) can be used as effective measurements for the water quality in current urban water distribution systems [2].

With the increasing demand for water quality information, several water quality monitoring stations have been deployed throughout the city’s water distribution system to provide the real-time water quality reports in a city. Figure 1 illustrates the water quality monitor stations that have been deployed in Shenzhen, China. Besides water quality monitoring, predicting the urban water quality plays an essential

role in many urban aquatic projects, such as informing waterworks' decision making (e.g., pre-adjustment of chlorine from the waterworks), affecting governments' policy making (e.g., issuing pollution alerts or performing a pollution control), and providing maintenance suggestions (e.g., suggestions for replacements of certain pipelines).

Predicting urban water quality, however, is very challenging due to the following reasons. First, urban water quality varies by locations non-linearly and depends on multiple factors, such as meteorology, water usage patterns, land use, and urban structures. As depicted in Figure 1, the water quality indexes (RC) reported by the three stations demonstrate different patterns. Exiting hydraulic model-based approaches try to model water quality from physical and chemical perspective, but such hydraulic model can hardly capture all of those complex factors. Moreover, the parameters I model are hard to get, which make it difficult to extend to other water distribution systems. Second, as all the stations are connected through the pipeline system, the water quality among different stations are mutually correlated by several complex factors, such as attributes in pipe networks and distribution of POIs. Traditional hydraulic model-based approaches build hydraulic model for each station and ignore their spatial correlations, and thus their performance is far from satisfactory. Hence, besides identifying the influential factors, how to efficiently characterize and incorporate such relatedness poses another challenge.

Fortunately, in the era of big data [3] [4] [5], unprecedented data in urban areas (e.g., meteorology, POIs, and road networks) can provide complementary information to help predict the urban water quality. For example, temperature can be an indicator of water quality, with higher temperature indicating better water quality. The possible reason is that the water consumption tends to grow when temperature is high since most people may choose to take a shower, and the increased water consumption is one major cause that prevents the water quality's deterioration in the distribution systems.

To benefit from the unprecedented data in urban areas, in this paper, we predict the water quality of a station through a data-driven perspective using a variety of data sets, including water quality data, hydraulic data, meteorology data, pipe networks data, road networks data, and POIs. First, we perform extensive experiments and data analytics between the water quality and multiple potential factors, and identify the most influential ones that have an effect on the urban water quality. Second, we present a novel spatio-temporal multi-task multi-view learning (stMTMV) framework to fuse the heterogeneous data from multiple domains and jointly capture each station's local information as well as their global information into an unified learning model [6].

We summarize the contributions as follows:

_ Data-driven Perspective: We present a novel data-driven approach to co-predict the

future water quality among different stations with data from multiple domains. Additionally, the approach is not restricted to urban water quality prediction, but also can be applied to other multi-locations based coprediction problem in many other urban applications.

_ Influential Factor Identification: We identify spatially-related (such as POIs, pipe networks, and road networks) and temporally-related features (e.g., time of day, meteorology and water hydraulics), contributing to not only our application but also the general problem of water quality prediction.

_ Unified Learning Model: We present a novel spatio-temporal multi-view multi-task learning framework (stMTMV) to integrate multiple sources of spatio-temporal urban data, which provides a general framework of combining heterogeneous spatio-temporal properties for prediction, and can also be applied to other spatio-temporal based applications.

_ Real evaluation: We evaluate our method by extensive experiments that use real-world datasets in Shenzhen, China. The results demonstrate the advantages of our method beyond other baselines, such as ARMA, Kalman filter, and ANN, and reveal interesting discoveries that can bring social good to urban life.

2. LITERATURE SURVEY

L. A. Rossman, R. M. Clark, and W. M. Grayman, "Modeling chlorine residuals in drinking-water distribution systems," Journal of environmental engineering, vol. 120, no. 4, pp. 803–820, 1994.

A mass transfer-based model is developed for predicting chlorine decay in drinking-water distribution networks. The model considers first-order reactions of chlorine to occur both in the bulk flow and at the pipe wall. The overall rate of the wall reaction is a function of the rate of mass transfer of chlorine to the wall and is therefore dependent on pipe geometry and flow regime. The model can thus explain field observations that show higher chlorine decay rates associated with smaller pipe sizes and higher flow velocities. It has been incorporated into a computer program called EPANET that can perform dynamic water-quality simulations on complex pipe networks. The model is applied to chlorine measurements taken at nine locations over 53 h from a portion of the South Central Connecticut Regional Water Authority's service area. Good agreement with observed chlorine levels is obtained at locations where the hydraulics are well characterized. The model should prove to be a valuable tool for managing chlorine-disinfection practices in drinking-water distribution systems.

Y. Zheng, "Methodologies for cross-domain data fusion: An overview," IEEE Transactions on Big Data, vol. 1, no. 1, pp. 16–34, 2015.

Traditional data mining usually deals with data from a single domain. In the big data era, we face a diversity of datasets from different sources in different domains. These datasets consist of multiple modalities, each of which has a different representation, distribution, scale, and density. How to unlock the power of knowledge from multiple disparate (but potentially

connected) datasets is paramount in big data research, essentially distinguishing big data from traditional data mining tasks. This calls for advanced techniques that can fuse knowledge from various datasets organically in a machine learning and data mining task. This paper summarizes the data fusion methodologies, classifying them into three categories: stage-based, feature level-based, and semantic meaning-based data fusion methods. The last category of data fusion methods is further divided into four groups: multi-view learning-based, similarity-based, probabilistic dependency-based, and transfer learning-based methods. These methods focus on knowledge fusion rather than schema mapping and data merging, significantly distinguishing between cross-domain data fusion and traditional data fusion studied in the database community. This paper does not only introduce high-level principles of each category of methods, but also give examples in which these techniques are used to handle real big data problems. In addition, this paper positions existing works in a framework, exploring the relationship and difference between different data fusion methods. This paper will help a wide range of communities find a solution for data fusion in big data projects.

Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," ACM Transactions on Intelligent Systems and Technology, vol. 5, no. 3, pp. 38:1–38:55, 2014.

Urbanization's rapid progress has modernized many people's lives but also engendered big issues, such as traffic congestion, energy consumption, and pollution. Urban computing aims to tackle these issues by using the data that has been generated in cities (e.g., traffic flow, human mobility, and geographical data). Urban computing connects urban sensing, data management, data analytics, and service providing into a recurrent process for an unobtrusive and continuous improvement of people's lives, city operation systems, and the environment. Urban computing is an interdisciplinary field where computer sciences meet conventional city-related fields, like transportation, civil engineering, environment, economy, ecology, and sociology in the context of urban spaces. This article first introduces the concept of urban computing, discussing its general framework and key challenges from the perspective of computer sciences. Second, we classify the applications of urban computing into seven categories, consisting of urban planning, transportation, the environment, energy, social, economy, and public safety and security, presenting representative scenarios in each category. Third, we summarize the typical technologies that are needed in urban computing into four folds, which are about urban sensing, urban data management, knowledge fusion across heterogeneous data, and urban data visualization. Finally, we give an outlook on the future of urban computing, suggesting a few research topics that are somehow missing in the community.

3. EXISTING SYSTEM

Several studies in the environmental science have been tried to analyze the water quality problems via data-driven based approaches,

and those studies covers a range of topics, from the physical process analysis in the river basin, to the analysis of concurrent input and output time series [64] [65]. The approaches adopted in these studies include instance-based learning models (e.g., kNN) as well as neural network models (e.g., ANN). In general, those data-driven approaches in the environmental science can fall into the following three major categories: Instance-based Learning models (IBL), Artificial Neural Network models (ANN) and Support Vector Machine models (SVM).

- Instance-based learning models (IBL) is a family of learning algorithms that model a decision problem with instances or examples of training data that are deemed important to test model [66]. As a typical example of IBL, k-Nearest Neighbors
- (k-NN) is widely used due to its simplicity and incredibly good performance in practice.
- For example, the work introduced by Karlsson et al. [67] addressed the classical rainfall-runoff forecasting problem by k-NN algorithm, and demonstrated promising results. Toth et al. [68] used k-NN to predict the rainfall depths from the history data, and showed the persistent outperformance of k-NN over other time series prediction methods.
- As another example, Ostfeld et al. [69] developed a hybrid genetic k-Nearest Neighbor algorithm to calibrate the two-dimensional surface quantity and water quality model. Artificial Neural Network (ANN) is a network inspired by biological neural networks (in particular the human brain), which consists of multiple layers of nodes (neurons) in a directed graph with each layer fully connected to the next one [65]. Neural networks have been widely employed to solve a wide variety of tasks, and can acheive good results. For instance, Moradkhani et al. [70] proposed an hourly streamflow forecasting method based on a radial-basis function (RBF) network and demonstrated its advantages over other numerical prediction methods. Also, the work introduced by Kalin [44] predicted the water quality indexes in watersheds through ANN.
- Support Vector Machines (SVMs) are typical supervised learning models that analyze data used for classification and regression [71].
- In aquatic studies, it was also extended to solving prediction problems [64]. For instance, Liong et al. [72] addressed the issue of flood forecasting using Support Vector Regression (SVR) which is an extension of SVM. Another work by Xiang et al. [73] utilized a LS-SVM model to deal with the water quality prediction problem in Liuxi River in Guangzhou.
- However, none of these approaches is applied into urban scenarios, which is quite different from our applications. Moreover, those existing approaches process the data

from a single source, and can hardly integrate the data from different sources. Thus, their applications in the urban scenarios are restricted.

DISADVANTAGES

- The system is implemented only Multi-task Multi-view Learning Approaches.
- Instance-based learning models (IBL) is a family of learning algorithms that model a decision problem with instances or examples of training data that are deemed important to the model.

4. PROPOSED SYSTEM

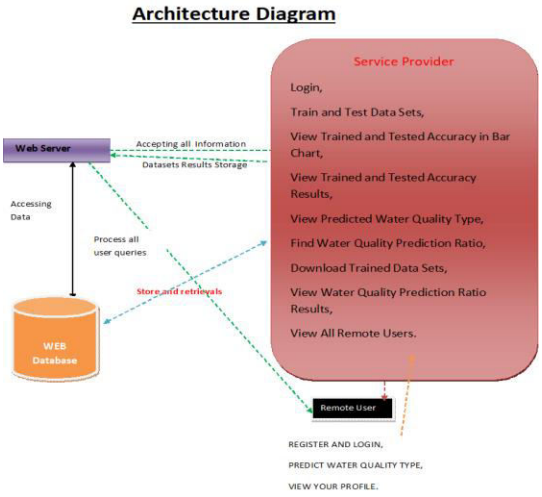
- Data-driven Perspective: We present a novel data-driven approach to co-predict the future water quality among different stations with data from multiple domains. Additionally, the approach is not restricted to urban water quality prediction, but also can be applied to other multi-locations based coprediction problem in many other urban applications.
- Influential Factor Identification: We identify spatially-related (such as POIs, pipe networks, and road networks) and temporally-related features (e.g., time of day, meteorology and water hydraulics), contributing to not only our application but also the general problem of water quality prediction.
- Unified Learning Model: We present a novel spatio-temporal multi-view multi-task learning framework (stMTMV) to integrate multiple sources of spatio-temporal urban data, which provides a general framework of combining heterogeneous spatio-temporal properties for prediction, and can also be applied to other spatio-temporal based applications.
- Real evaluation: We evaluate our method by extensive experiments that use real-world datasets in Shenzhen, China. The results demonstrate the advantages of our method beyond other baselines, such as ARMA, Kalman filter, and ANN, and reveal interesting discoveries that can bring social good to urban life.

ADVANTAGES

1) Water quality data: We collect water quality data every five minutes from 15 water quality monitoring stations in Shenzhen City. It comprises residual chlorine (RC), turbidity (TU) and pH. In this paper, we only use RC as the indexfor water quality, since RC is the most important and effective measurement for water quality in current urban water distribution system.

2) Hydraulic data: Hydraulic data consists of flow and pressure, which are collected every five minutes from 13 flow sites and 14 pressure sites, respectively.

5. SYSTEM ARCHITECTURE



6. MODULES

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Train and Test Data Sets, View Trained and Tested Accuracy in Bar Chart, View Trained and Tested Accuracy Results, View Predicted Water Quality Type, Find Water Quality Prediction Ratio, Download Trained Data Sets, View Water Quality Prediction Ratio Results, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user's details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like REGISTER AND LOGIN, PREDICT WATER QUALITY TYPE, VIEW YOUR PROFILE.

7. CONCLUSION AND FUTURE ENHANCEMENT

This paper presents a novel data-driven approach to forecast the water quality of a station by fusing multiple sources of urban data. We evaluate our approach based on Shenzhen's water quality and various urban data. The experimental results demonstrate the effectiveness and efficiency of our approach. Specifically, our approach outperforms the traditional RC decay model [2] and other classical time series predictive models (ARMA, Kalman) in terms of RMSE metric. Meanwhile, as our approach consists of two components, each of the components demonstrates its effectiveness through extensive experiments and analysis. In particular, the first component is the influential factors identification, which explores the factors that affect the urban water quality via extensive experiments and analysis in Section 3 and 4. The second one is a spatiotemporal multi-view multi-task learning (STMTMV) framework that consists of multi-view learning and multi-task learning. The experiments have shown that STMTMV has a predictive accuracy of around 85% for forecasting next 1-4 hours, which outperforms the single-task methods (LR) by approximately 11% and the single-view methods (t-view and s-view) by

approximately 11% and 12%, respectively. The code has been released at: https://www.microsoft.com/en-us/research/publication/urbanwater- quality-prediction-based-multi-task-multi-view-learning-2/ In future, we plan to deal with the water quality inference problems in the urban water distribution systems through a limited number of water quality monitor stations.

REFERENCES

[1] W. H. Organization, Guidelines for drinking-water quality, 2004, vol. 3.

[2] L. A. Rossman, R. M. Clark, and W. M. Grayman, "Modeling chlorine residuals in drinking-water distribution systems," Journal of environmental engineering, vol. 120, no. 4, pp. 803–820, 1994.

[3] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," IEEE Transactions on Big Data, vol. 1, no. 1, pp. 16–34, 2015.

[4] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," ACM Transactions on Intelligent Systems and Technology, vol. 5, no. 3, pp. 38:1–38:55, 2014.

[5] Y. Zheng, H. Zhang, and Y. Yu, "Detecting collective anomalies from multiple spatio- temporal datasets across different domains," 2015.

[6] Y. Liu, Y. Zheng, Y. Liang, S. Liu, and D. S. Rosenblum, "Urban water quality prediction based on multi-task multi-view learning," in Proceedings of the International Joint Conference on Artificial Intelligence, 2016.

[7] H. Cohen, "Free chlorine testing," http://www.cdc.gov/safewater/chlorineresidual- testing.html, 2014, accessed on 5 August 2016.

[8] B. D. Barkdoll and H. Didigam, "Effect of user demand on water quality and hydraulics of distribution systems," in Proceedings of the World Water and Environmental Resources Congress, 2003.

[9] P. Castro and M. Neves, "Chlorine decay in water distribution systems case study–lousada network," Electronic Journal of Environmental, Agricultural and Food Chemistry, vol. 2, no. 2, pp. 261–266, 2003.

[10] L. W. Mays, Water distribution system handbook, 1999.

[11] L. A. Rossman and P. F. Boulos, "Numerical methods for modeling water quality in distribution systems: A comparison," Journal of Water Resources planning and management, vol. 122, no. 2, pp. 137–146, 1996.

[12] W. M. Grayman, R. M. Clark, and R. M. Males, "Modeling distributionsystem water quality: dynamic approach," Journal of Water Resources Planning and Management, vol. 114, no. 3, pp. 295–312, 1988.

[13] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2003, pp. 2–11.

[14] G. Luo, K. Yi, S.-W. Cheng, Z. Li, W. Fan, C. He, and Y. Mu, "Piecewise linear approximation of streaming time series data with max-error guarantees," in Proceedings

of the IEEE International Conference on Data Engineering, 2015, pp. 173–184.
[15] E. O. Brigham and E. O. Brigham, The fast Fourier transform. Prentice- Hall Englewood Cliffs, NJ, 1974, vol. 7.