Exploiting Heterogeneous Monitoring Data for Spatiotemporal Algal Bloom Prediction

Taewhi Lee, Miyoung Jang, Jang-Ho Choi, Jongho Won, and Jiyong Kim
Smart Data Research Section, AI Research Lab.
ETRI (Electronics and Telecommunications Research Institute)
Daejeon, Republic of Korea
{taewhi, myjang, janghochoi, jhwon, kjy}@etri.re.kr

***Abstract*—Harmful algal blooms need to be mitigated because they can cause significant negative effects to humans and other organisms. If such algal blooms can be predicted in advance by monitoring water quality, they can be suppressed at an early stage by making decisions to take actions. We describe our ongoing work on integrating the heterogeneous water quality monitoring data and on recovering the missing data using tensor completion techniques. We also discuss the challenges in carrying out this study.**

***Index Terms*—algal bloom, data integration, tensor completion**

I. Introduction

Algal blooms are natural phenomena where the population of photosynthetic organisms rapidly increases in aquatic ecosystems [1]. Harmful algal blooms need to be mitigated because they can cause significant negative effects to humans and other organisms. When they occur in water sources, those effects may be exacerbated. If such algal blooms can be predicted in advance by monitoring water quality, they can be suppressed at an early stage by making decisions to take actions, such as dispatching algae harvesting ship, water surface aerator, or ultrasonic algae controller, opening floodgates, and spraying yellow soil.

The accuracy of algal bloom prediction depends on the quality of the monitoring data. In order to collect water quality data more densely in near real-time, attempts are being made to collect data through various types of device as shown in Figure 1.

- **Fixed sensor data.** Water quality data collected from fixed sensors that are installed on the pontoons at specific points.
- **Moving sensor data.** Water quality data collected from an unmanned surface vehicle (USV) equipped with water quality sensors. The USV collects the water quality data as it travels the target area along the predefined route.
- **Hyperspectral image data.** Water quality data collected from an aerial drone equipped with a hyperspectral sensor camera. The concentration of chlorophyll-a (Chl-a) and phycocyanin (PC), which are indicators for algal biomass, can be extracted from the hyperspectral images.

However, these heterogeneous data cannot be directly used to algal bloom prediction. In order to predict algal blooms

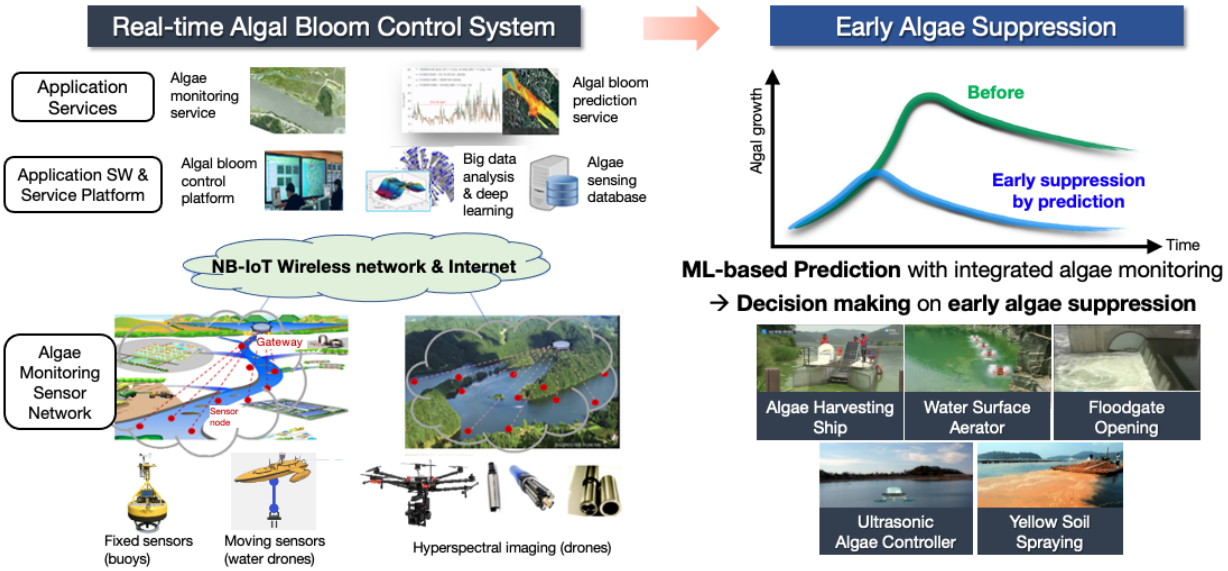

Fig. 1. algae monitoring data collection for algal bloom prediction

based on machine learning, it is necessary to preprocess these heterogeneous data to improve the quality of the data as follows.

- **Data integration.** Those monitoring data can be fed to machine learning by integrating data measured by different devices at different times and different locations.
- **Missing data recovery.** Data omissions occur frequently due to various reasons, such as weather conditions, hardware failure, and budget limitations.

In this paper, we describe our ongoing work on integrating the heterogeneous water quality monitoring data and on recovering the missing data using tensor completion techniques. Then, we conclude by discussing the challenges in carrying out this study.

II. Heterogeneous Monitoring Data Integration

A. Data Collection

Water quality monitoring data are being collected for our study area through the various devices mentioned in Section I. Our study area is the So-ok-cheon stream near to the Chu-so-ri region, a branch of the Geum river which is one of the six main rivers in South Korea, as shown in Figure 2. It is the area where green algae often occur due to topographic reasons. The study area is split into about 2000 grid cells, which are unit areas for algal bloom prediction.

The format of monitoring data is different depending on the device being measured. Fixed sensor data collected from the pontoon and moving sensor data collected from the USV include various kinds of features, including water temperature, pH, dissolved oxygen, electrical conductivity, total organic carbon, total nitrogen, total phosphorous, and chlorophyll-a. Such features cannot be extracted from hyperspectral images taken by aerial drones, except chlorophyll-a and phycocyanin.

The cycle of data collection is also different. While fixed sensor data are reliably collected on a hourly basis, moving sensor data and hyperspectral image data are collected on a daily basis. Also, there are many omissions in moving sensor data and hyperspectral image data because USVs and aerial drones cannot operate in bad weather conditions, or they may also have to operate within budget.

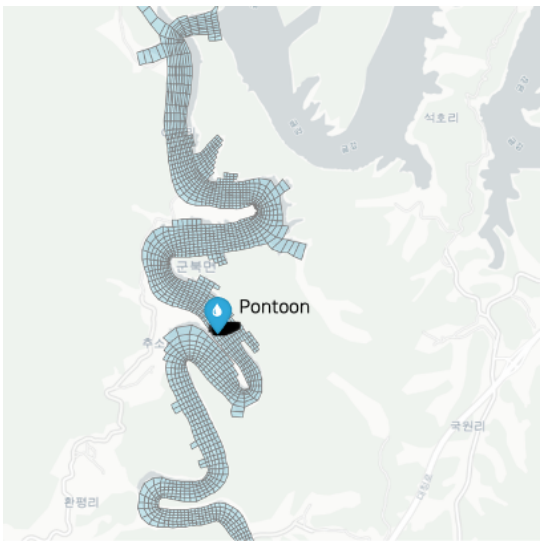

Fig. 2. Study area

B. Data Integration

We integrate heterogeneous monitoring data by mapping into grid cells. The data integration process can be summarized as follows.

1) **Grid cell construction.** For each cell, polygon objects are created using the coordinates of its boundary points.
2) **Data-to-cell mapping.** Each monitoring data record is mapped to the corresponding cell polygon object, which contains the location coordinate of the data record.
   - **Fixed sensor data.** A pontoon is installed in a specific location, so fixed sensor data can be directly mapped into a specific cell.
   - **Moving sensor data.** Moving sensor data are collected every certain distance along the moving path of USVs. Multiple values may be mapped to in the same cell for the same datetime.
   - **Hyperspectral image data mapping.** Hyperspectral image data are spatially continuous because the data values are extracted from the image pixels. Therefore, it is required to sample the data by a certain distance.
3) **Representative value selection.** Since two or more values may be mapped to one cell for the same datetime, representative values have to be selected. It may be used to apply some statistics function like max(), avg(), and median() or to pick a certain record by policy.
4) **Missing data recovery.** Data omissions occur frequently due to various reasons, so it is necessary to recover missing data for more accurate prediction. It will be described in Section III.

III. Missing Data Recovery

There are a large number of missing data in the collected data. In order to train machine learning models for algal bloom prediction, the missing data have to be handled in some ways. Naive methods, such as deleting data records with missing values or filling them with the average of surrounding values in the time or space dimension, would give inaccurate prediction results. Deep learning-based imputation techniques like DataWig [2] have been developed to impute missing values in tabular data. Many studies have been conducted to fill in the missing data [3], [4], but the larger the fraction of missing data, the more difficult it is to recover the data.

We have been trying to apply tensor data completion techniques using auxiliary information, Auxiliary Information Regularized CP model (AirCP) [5], which can be applied even in cases where the fraction of missing data is large. The authors applied the AirCP method to recover the spatiotemporal dynamics of hashtags. We apply this method to recover our

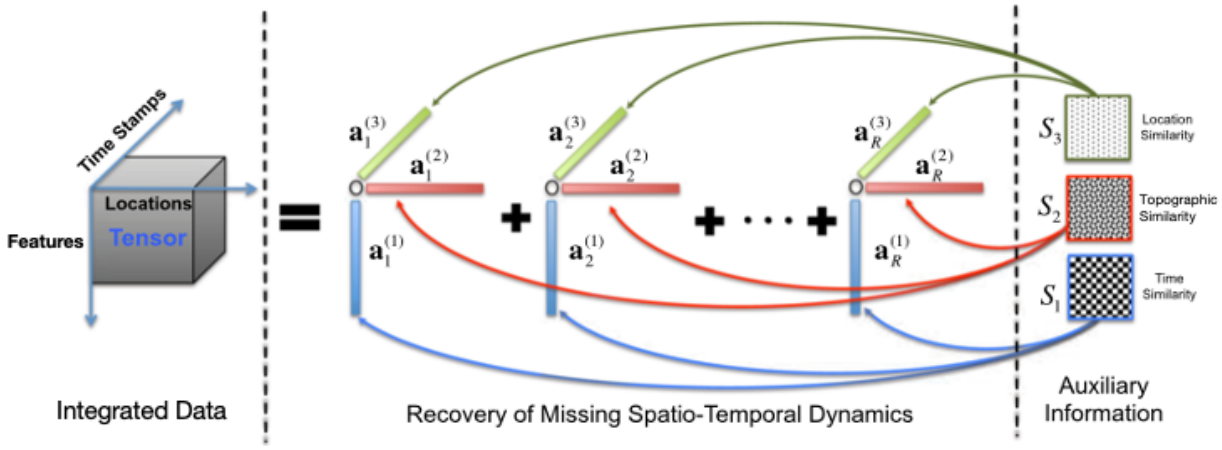

Fig. 3. Tensor data completion using Auxiliary Information Regularized CP model (adapted from [5])

integrated data by injecting topographic similarity, as well as location similarity and time similarity, as shown in Figure 3.

- **Location similarity.** A similarity matrix encoding spatial relationships. The location similarity matrix is derived under the assumption that the closer the grid cell, the more similar the concentrations of algae.
- **Topographic similarity.** A similarity matrix encoding topographic characteristics. For example, this matrix can indicate whether the grid cell is the edge or the middle of river by the distance to the nearest land. The topographic matrix is derived under the assumption that the more similar the topographical characteristics, the more similar the concentrations of algae.
- **Time similarity.** A similarity matrix encoding temporal relationships. The time similarity matrix is derived under the assumption that the closer the time of data collection, the more similar the concentrations of algae.

We continue to collect water quality data on our study area with a novel direct-readable water quality complex sensor, which is newly developed in our project. Unfortunately, enough data has not been accumulated yet. We are currently testing various topographic similarity matrices to enhance the accuracy of algal bloom prediction and are analyzing collected data for each cell.

IV. Discussion and Challenges

As we are challenging to predict the concentration of algae for areas where water quality data has not been collected, we are facing a lot of difficulties. We describe the challenging issues that must be addressed.

To build machine learning models for algal bloom prediction, we need enough data to construct training and test datasets. It is better for the prediction to collect water quality data daily or more frequently, but this is impossible when the weather condition is bad or the hardware fails. This puts us in a situation where there is no exact real data to compare with predicted values.

A limited budget is another factor that lowers the frequency of data collection because of the costs involved in operating USVs and aerial drones. The features that can be extracted are also different depends on the devices. It can be another research topic to determine the frequency or cycle of data collection under these constraints.

V. Conclusion

We presented a method to integrate the heterogeneous water quality monitoring data and to recover the missing data by applying tensor data completion techniques using auxiliary information. We used three types of auxiliary information to recover the missing data, i.e., location similarity, topographic similarity, and time similarity. We plan to compare the performance of algal bloom prediction using various auxiliary similarity matrices, or various missing data recovery methods. Also, we will address the challenging issues discussed in this paper.

Acknowledgment

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2018-0-00219, Space-time complex artificial intelligence blue-green algae prediction technology based on direct-readable water quality complex sensor and hyperspectral image).

References

[1] T. Lee, J.-H. Choi, M. Jang, J. Won, and J. Kim, "Enhancing prediction of chlorophyll-a concentration with feature extraction using higher-order partial least squares," in *Proceedings of 2020 International Conference on Information and Communication Technology Convergence*, 2020, pp. 1666–1668.

[2] F. Biessmann, T. Rukat, P. Schmidt, P. Naidu, S. Schelter, A. Taptunov, D. Lange, and D. Salinas, "DataWig: Missing value imputation for tables," *Journal of Machine Learning Research*, vol. 20, no. 175, pp. 1–6, 2019.

[3] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *Journal of Big Data*, vol. 8, pp. 140:1–37, 2021.

[4] S. Jäger, A. Allhorn, and F. Bießmann, "A benchmark for data imputation methods," *Frontiers in Big Data*, vol. 4, p. 48, 2021.

[5] H. Ge, J. Caverlee, N. Zhang, and A. Squicciarini, "Uncovering the spatio-temporal dynamics of memes in the presence of incomplete information," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, ser. CIKM '16, 2016, pp. 1493–1502.