Applying Parallel Programming and High Performance Computing To Speed up Data Mining Processing

Ruijian Zhang
Purdue University, USA
Zhang45@purdue.edu

***Abstract*—** Water quality assessment and prediction of Lake Michigan are becoming major challenges in Northwest Indiana, USA. Traditionally, mechanistic simulation models are employed for water quality modeling and prediction. However, given the complicate nature of Lake Michigan in Northwestern Indiana, the detailed simulation model is extremely simple in comparison and, at some point, additional detail exceeds our ability to simulate and predict with reasonable error levels. In this regard, my project applied data mining technologies, as an innovative alternative, to develop an easy and more accurate approach for water quality assessment and prediction. The drawback of the data mining modeling is that the execution takes quite long time, especially when we employ a better accuracy but more time consuming algorithm in clustering. Therefore, we applied the High Performance Computing System of the Northwest Indiana Computational Grid to deal with this problem. Up to now, the pilot experiments have achieved very promising preliminary results. The visualized water quality assessment and prediction obtained from this project would be published in an interactive website so that the public and the environmental managers could use the information for their decision making.

***Keywords*—** *Water Quality; Data Mining; Parallel Programming; High Performance Computing.*

I. INTRODUCTION

The objective of this project is to explore the consideration of improving water quality assessment and prediction by applying data mining technology for better accuracy and easier-implemented modeling, and applying parallel computing on the Northwest Indiana Computational Grid (NWICG) for reducing the execution time.

Traditionally, mechanistic simulation models are employed for water quality modeling and prediction. However, the nature in Lake Michigan of Northwest Indiana is very complex, and even the most detailed simulation model is extremely simple in comparison. At some point, additional detail exceeds our ability to simulate and predict with reasonable error levels. In light of these limitations, data mining as an innovative and attractive alternative of water quality modeling and prediction may express the complex behavior of the nature in Northwest Indiana and Lake Michigan region. This project utilized data mining technologies, such as clustering and classification, for water quality assessment and prediction. The drawback of the data mining modeling is that the execution takes quite long time, especially when we employed a better accuracy but more time consuming algorithm in clustering. But, NWICG's high performance computing (HPC) system helped us address this issue effectively [3, 4, 5]. So far, this innovative approach has obtained very promising results based on our pilot study which demonstrated that the sixteen processor high performance computation system achieved more accurate water quality prediction and reduced the execution time by more than ten times. The experiment has showed a very encouraging direction on water quality assessment and prediction through applying data mining models on high performance computing system. This project will continue to investigate the potential for data mining models to support water quality assessment and prediction in Northwest Indiana and compare the execution time between running on a normal single PC and applying parallel computing on the Northwest Indiana Computational Grid. As a result, the project developed an easy-implemented, fast and more accurate approach for water quality assessment and prediction.

II. THE SIGNIFICANT OF WATER QUALITY PROJECT

Water quality assessment and prediction are critical to protect ecosystems, human health, aquaculture production systems and species including fish and shellfish, urban development, securing food, and promoting cleaner industry. Water quality could be affected dramatically or even devastated by a variety of factors, such as water erosion, rainfall, drought, infiltration, runoff, industry accident, and leak of oil or chemistry material. However, water quality is still predictable. The prediction and forecasting of water quality hazards are vital to address the issues of the suddenly increasing water pollution or material in water potentially deleterious to human and ecosystem health (e.g. pathogens, heavy metals, organic compounds).

Great Lakes collectively represent the largest body of fresh water on the surface of the earth and the single most important source of fresh water to the contiguous United States. Especially, Northwest Indiana owns 45 miles expanse, a quite large portion of the Lake Michigan shoreline. It is part of a complex and dynamic process of the Great Lakes systems. Northwest Indiana offered industry the advantage of lake transportation, abundant water for industry processes, and scarcely populated area close to a major urban center. The region remains one of the most industrialized in the area. The water quality of Northwest Indiana and Lake Michigan region affect the human desire to live, work and play along the coast. The sustainable management of such impacted systems requires a quantitative assessment and prediction of ecosystem dynamics and services to guide decision making activities.

Traditionally, mechanistic simulation models are employed for water quality modeling and prediction. For example, over the past decade the Computational Aquatic Ecosystem Dynamics Model (CAEDYM) has been developed as a generic and customizable tool for scientists and managers

978-1-5090-5507-4/17/$31.00 ©2017 IEEE
ICIS 2017, May 24-26, 2017, Wuhan, China

to simulate ecological and other water quality variables of interest [6, 10, 11, 12]. CAEDYM is a process-based freeware model of the major biogeochemical processes influencing water quality. It optionally models inorganic particles, oxygen, organic and inorganic nutrients (C, N, P and Si), multiple phytoplankton and zooplankton groups, fish and bacteria. Recent developments also include optional modules for benthic organisms (e.g. clams, macroalgae), pathogens and microbial indicator organisms, and a generic geochemical module capable of simulating pH, aqueous speciation (including metals), precipitation/dissolution reactions and sediment digenesis. Configuration is flexible so that the modeler can focus on the processes of interest. CAEDYM has been applied to a variety of aquatic systems including wetlands, lakes/reservoirs, rivers, estuaries and the coastal ocean.

It is a common strategy in water quality modeling and prediction to attempt to remedy predictive inadequacies by incorporating additional mechanistic detail into the model. This approach reflects the reasonable belief that enhanced scientific understanding of basic processes can be used to improve predictive modeling. However, the nature in Northwest Indiana and Lake Michigan region is too complicate to be predicted accurately by this way. This leads to the consideration of applying data mining technologies, such as clustering and classification, for water quality modeling and prediction, and applying the NWICG's high performance computing system to speed up the execution of the very time consuming enumerative algorithm [2, 7, 8, 9]. This project compared our data mining modeling with CAEDYM mechanistic simulation for water quality assessment and prediction in Northwest Indiana and Lake Michigan region. It has been shown that our project could improve the water quality assessment and prediction in the terms of both accuracy and execution time.

III. APPLY DATA MINING

Data mining is the process of extracting patterns from data. The process consists of the following steps: data collecting and data preparation, clustering, classification and prediction. The framework of this project is to design and analyze the data mining methods for water quality modeling. To conduct the examination of this modeling, we designed and implemented an optimal algorithm using enumerative method for better water quality prediction accuracy. The decision tree classification tool C5 was also employed in this project [1, 2].

We obtained water quality data from a collaboration partner in the local industry community, *Save the Dunes Conservation Fund's* Water Program. They took the responsibility of monitoring the water quality and provided daily values of various attributes of water quality at 17 sites in the Salt Creek area in Northwest Indiana. These raw data are incomplete and inconsistent in terms of both time periods and water quality attributes. Some values are missing; some values are invalid. We have to perform data cleaning, data format and normalization, missing or invalid data treatment and other data preprocessing and preparation work. In the preliminary experiments, for the purpose of simplicity we chose only five water quality attributes. They are: water temperature, dissolved oxygen, pH value, specific conductivity and turbidity.

We applied clustering to assign cases of a dataset into subsets (called clusters) so that data in same cluster are similar in some sense. The similarity can be measured by the distance of each other in terms of the values of the attributes. In preliminary experiments, the algorithm we designed assigned the water quality data into 3 clusters of quality levels: good, fair and poor. We will expand the clusters to five later: Excellent, Good, Fair, Poor, and Very Bad later.

The k-clustering problem is NP hard which means there is no efficient algorithm to solve the problem optimally. Traditionally, *k*-means algorithm is often applied because of its simplicity and efficiency which allow it to run on large datasets. It is an iterative heuristic method which does not guarantee the convergence to the global optimum. The results depend on the initiation of clusters and it usually converges to a local optimum. We designed and implemented an optimal algorithm based on enumeration. The complexity of the algorithm is in the order of n to the power of k (n represents the number of cases in the dataset and k represents the number of clusters). It usually takes quite long time to execute, but it could guarantee the convergence to the global optimum. Taking advantage of the NWICG's high performance computing system, it is practical for us to use complexity to exchange the accuracy. Applying parallel programming and the high performance computing system, the execution time for the enumerative algorithm could be reduced to an acceptable level. Our algorithm was implemented in C++. We run our program on the Purdue University's High Performance Cluster – *Falcon*, which has 8 nodes, each with 4 CPUs, and/or *Miner*, which has 512 nodes, each with 4 CPUs.

We employed C5, a decision tree based data mining tool, to perform the further classification (assessment) and prediction [1]. In the preliminary experiments, we were using two previous days' or three previous days' water quality information to predict the following day's water quality classification level. Through repeatedly working backward and forward adjusting of the data format, the water quality prediction accuracy was improved greatly while executing in an acceptable short time period. The assessment accuracy rates by C5 reached 99 percent. The prediction accuracy rates by C5 reached 82 percent currently and are anticipated to be close to ninety percent if we apply more attributes later (22 versus 5). Comparing with the average 80 percent prediction accuracy by applying k-means algorithm for clustering and with the average 80 percent prediction accuracy by the mechanistic simulation models, this was a significant improvement.

IV. HIGH PERFORMANCE COMPUTING SPEED UP THE PROCESSING

In this project, we were running data mining program on Purdue University's High Performance Cluster – *Falcon*, which has 8 nodes, each with 4 CPUs, and/or *Miner*, which has 512 nodes, each with 4 CPUs. This High Performance

Computation system is supported by the Northwest Indiana Computational Grid program funded by the U.S. Department of Energy grant ($4.9 million). Taking advantage of the NWICG's high performance computing system, it is practical for us to use complexity to exchange the water quality prediction accuracy. HPC dramatically decreases the running time of the enumeration based algorithm, which is time consuming but more accurate.

Facilitated perfectly by the current existing Northwest Indiana Computational Grid, the project has applied the high performance computing system which dramatically decreases the running time of the enumeration based algorithm, so that makes this very time consuming algorithm for water quality prediction becoming practical.

Up to now, the experiments of this innovative approach obtained very promising results. The sixteen processor high performance computation system achieved more accurate water quality prediction and reduced the execution time by more than ten times. The experiment results showed applying data mining models on high performance computing system is a very encourage direction for Northwest Indiana's water quality assessment and prediction.

In light of almost 3000 cases, the optimal enumerative algorithm is very time-consuming (running in days) when executing on a single machine. In order to improve the performance, we applied parallel programming on the enumerative algorithm. The program run on multi-processors in parallelism. The parallel computing assigned one process as the master process and others as slave process. The master process read in the data, broadcasted the data and other information to the slave processes. The parallel programming split the clustering calculation into threads for slaves. Each slave process worked on calculations for clustering. Finally the master processor received the results from each slave process and chose the global optimized clustering [4].

Our experiments compared the running time of the enumerative algorithm using a single processor with that using High Performance Computing with 8 processors and 16 processors. The number of the cases was 500, 1000, 1500, 2000, 2500, and 2700 respectively. Figure 1 and Figure 2 below show the execution time of our preliminary experiments. It exposed the critical role of HPC for the time consuming enumerative algorithm to process large amount of data. For example, for 1500 cases, running on a single processor, it took more than eight hours to finish the clustering; whereas running on HPC with 16 processors, it only took about 32 minutes. This improvement of execution time is critical, because it is important to have a reasonable amount of indicative and representative data samples (cases) in order to discover patterns in data samples using data mining technologies. It is not unusual to have a data set of 1500 cases by this approach.

The results of these experiments produced a solid fundamental for future developments in our water quality assessment and prediction project. We created decision tree employed C5 for the water quality assessment and prediction. Using the produced decision trees to predict unseen cases, the prediction accuracies reach 82 percent, two percent better than applying k-means algorithm and about the same improvement than applying the mechanistic simulation models. It is anticipated that if we use more attributes of the water quality in the future, the accuracy rate could be further improved.

V. SUMMARY AND CONCLUSION

This project improved water quality assessment and prediction by applying data mining technology for better accuracy and easier-implemented modeling, and applying parallel computing on the Northwest Indiana Computational Grid (NWICG) for reducing the execution time. It exposed the critical role of HPC for the time consuming enumerative algorithm to process large amount of data. The experiment results showed applying data mining models on high performance computing system is a very encourage direction for Northwest Indiana's water quality assessment and prediction.

REFERENCES

[1] J. R. Quinlan, "C5: Programs for Machine Learning", Morgan Kaufmann, 1993.

[2] Helen M. Moshkovich, Alexander I Mechitov, and David L. Olson. "Rule Induction in Data Mining: Effect of Ordinal Scales", Expert Systems with Applicationsm, 2002, Vol.22 pp 303-311.

[3] Charles Severance and Kevin Dowd, "High performance computing", 2nd Edition, 1998.

[4] G. Stellner, Cocheck: "Checkpointing and Process Migration for MPI", Proc. IPPS, IEEE Computer Society Press, Silver Spring, MD, 1996

[5] P.L. Vaughan, A. Skjellum, D.S. Reese, F.C. Cheng, "Migrating from PVM to MPI, Part I: The Unify System", Proc. 5th Symp. on the Frontiers of Massively Parallel Computation, McLean, VA, IEEE Computer Society Press, Silver Spring, MD (1995), pp. 488–495 IEEE Computer Society Technical Committee on Computer Architecture.

[6] M.R. Hipsey and D.P. Hamilton, "Computational Aquatic Ecosystem Dynamics Model: CAEDYM v3", 2008.

[7] Tong, S and Chen, W. "Modeling the Relationship between Land Use and Surface Water Quality". *Environmental Management* , 2002, 66, 377-393

[8] Stow, C. A., Borsuk, M. E., and Reckhow, K. H. "TMDL Development in the Neuse River Watershed: An Imperative for Adaptive Management". *Water Resources Update*, 2002, 122, 16–26.

[9] Shrestha, S. Kazama, F. "Assessment of Surface Water Quality Using Multivariate Statistical Techniques: A Case Study of the Fuji River Basin, Japan". *Environmental Modeling & Software*, 2006, 22, 464-475

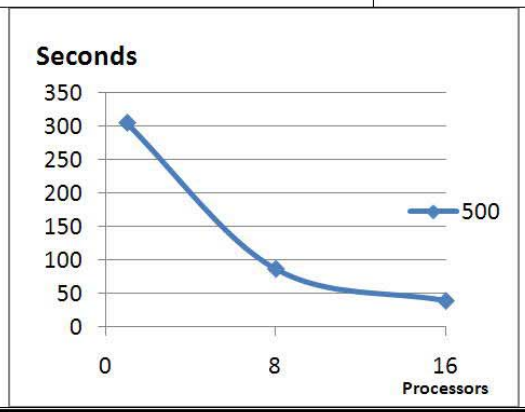
[10] Said, A. "The Implementation of a Bayesian Network for Watershed Management Decisions". *Water Resources Management*, 2005, 20, 591–605.

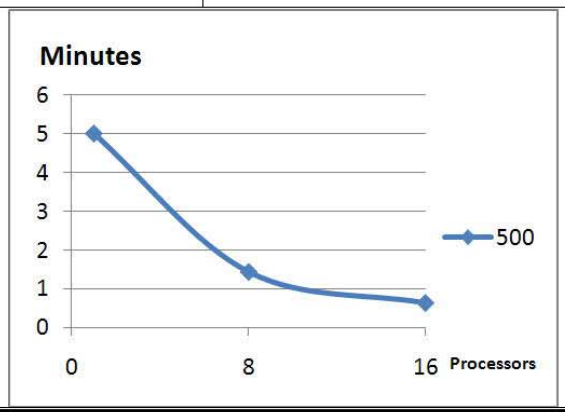
[11] Faruk, D. "A Hybrid Neural Network and ARIMA Model for Water Quality Time Series Prediction". *Engineering Applications of Artificial Intelligence, 2005,* 23, 586-594.

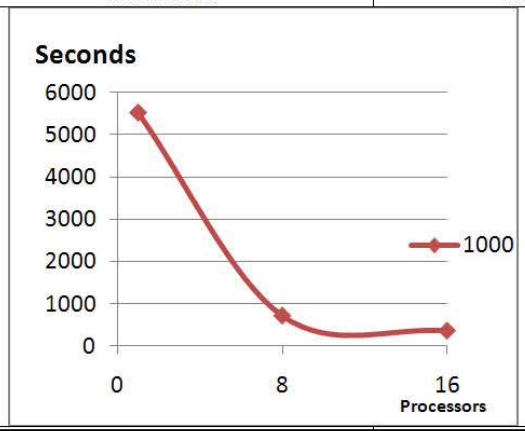
[12] He, L.M. & He, Z.L. "Water Quality Prediction of Marine Recreational Beaches Receiving Watershed Baseflow and Stormwater Runoff in Southern California, USA". *Water Research*, 2008, 42 , 2563-2573.

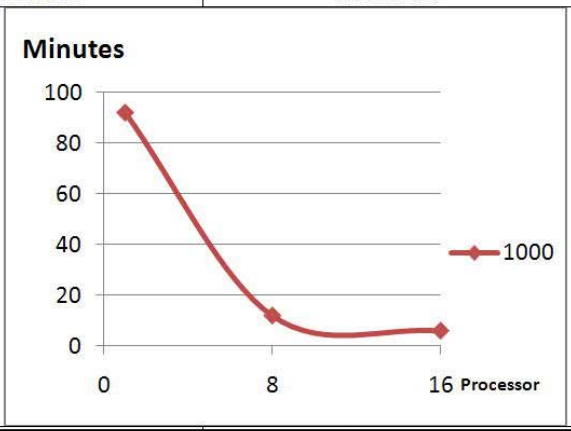
Figure 1 Experiment Results with MPI on the High Performance Computing

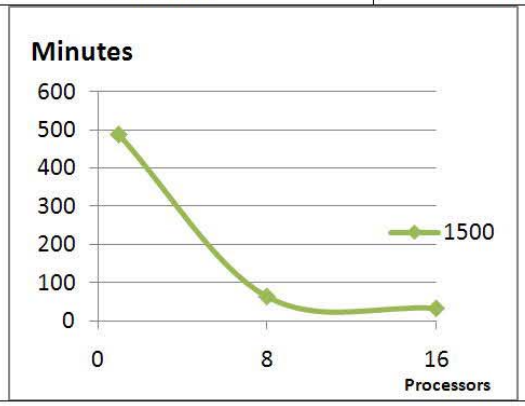
Cases	One Processor	MPI with 8 Processors	MPI with 16 Processors
500	Run time=306 Seconds =5 Minutes =0.1 Hours	Run time=87 Seconds =1.45 Minutes =0.0 Hours	Run time=39 Seconds =0.65 Minutes =0.0 Hours
1000	Run time =5533 Seconds =92 Minutes =1.5 Hours	Run time =722 Seconds =12 Minutes =0.2 Hours	Run time=363 Seconds =6 Minutes =0.1 Hours
1500	Run time=29301 Seconds =488 Minutes =8.1 Hours	Run time=3803 Seconds =63 Minutes =1.1 Hours	Run time=1918 Seconds =32 Minutes =0.5 Hours











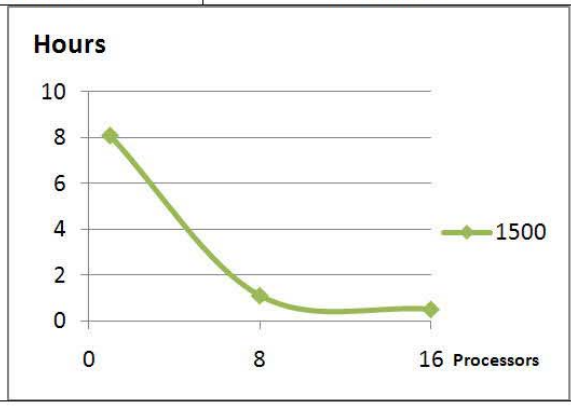


Figure 2. Execution Time in Seconds for Comparison of Various Processors in HPC