ISSN : 0974 - 7435 Volume 8 Issue 2

BioTechnology
An Indian Journal

FULL PAPER

BTAIJ, 8(2), 2013 [233-237]

An algorithm in quality inspection of large marine data based on block-nested-loops

Hu Guang-ming*, Cao Nan-ya
Modern information and education technology center Shanghai Ocean University, SHOU NO. 999 Huchenghuan Rd. Lingang new city, Shanghai 201306, (CHINA)
E-mail: gmhu@shou.edu.cn; 407060634@qq.com

ABSTRACT

Large marine data possesses several typical characteristics, such as large amount, multisource, multiple dimensions, multi-type and so on. How to design an optimal quality inspection plan and control the ocean data timely becomes more and more important for the application of large marine data. Based on skyline, it proposed a method to select the optimal quality inspection plan for the quality inspection of large marine data. Firstly, the residual of acceptance quality probability of each quality inspection plans for ocean big data were calculated by Hyper-geometric distribution model. And then, the optimal quality inspection plan was selected based on the algorithm of block-nested-loops (BNL), which compared the residual of acceptance quality probability of each quality inspection plans one by one. Finally, the proposed method is verified by inspecting the quality of the large marine data, which is collected by monitoring sites in a certain sea area.

KEYWORDS

Large marine data;
Quality inspection;
Block-nested-loops algorithm;
Residuals.

INTRODUCTION

With the development of the marine industry, information technology becomes an important way of comprehensive understanding marine and maritime research. Currently, there are a wide variety of marine data acquisition means, oceanographic data "quantity" rapid growth, while oceanographic data "classes" of diversification, oceanographic data has become big data model. Ocean monitoring data provides important information resource for marine environment, marine resources exploration, marine disaster forecasting and other studies, but the "quality" of this data really becomes a main concern.

Taking monitoring a marine aquaculture zone for example. The basic data including longitude, latitude and bathymetry; marine environmental data elements include: temperature, salinity, wave, current, tidal, etc., elements of data acquisition cycle is 10 minutes; marine aquaculture area attribute data includes: type of farming, breeding area, farming units. In the ocean data life cycle, from the collection, transmission, processing to the application, are likely to make the data produced quality problems. Because the data prior to use, the need for a large batch of marine data quality inspection. But the traditional method of data quality test data can not be di-

rectly applied to marine big quality inspection, the reason is: (a) marine data belongs to a class of spatial data, the spatial position data and attribute data with the corresponding; (b) Ocean Data Acquisition period of 10 minutes, so the dynamic characteristics of marine data, and its sharply accumulation; (c) as a result of the acquisition of means of different environmental factors, the data format, the accuracy requirements vary.

The main contribution of this paper are: (1) use the hyper-geometric distribution model to give a different set of quality inspection programs residuals; (2) propose a block nested loops algorithm based on skyline to select the optimal quality inspection programs; (3) for multi-source, multi-class, multi-dimensional and dynamic nature of maritime data, quickly determine its quality inspection optimization.

RELATED WORK

Quality inspection is to extract certain data from a batch of marine data to estimate whether the data meet the demand accuracy. As for data quality, paper[2] proposes data quality standard to build closed cycle for data quality management. In paper[3], it suggests controlling data quality in terms of data veracity, data integrity, data representativeness and comparability. And test on existed data was carried out by statistical sampling. Data quality measurement index was divided into objectively data quality indicator and subjectively data quality index[4]. User can choose different index according to need to measure data. While in paper[5], it classified data quality as interior quality, addressable quality, context quality and delivery quality. Each class is divided into specific dimension for estimation and widely cognition. Paper[6] present sampling calculation method to quantize the two important dimensions (accuracy & integrity) of data quality. And it shows concrete analysis on the influence of data quality to the four common relation algebra operations (selection, projection, Cartesian product, linkage). The above methods are data quality inspection based on traditional data. But marine data is different from traditional data and has its own properties. (1) Most marine data quality inspection is irreversibility due to its difficult to obtain and cost much. (2) Marine data are wide reach cover and asymmetry spatial and temporal distribution. Given prioritization scheme based on different batches and areas of marine data is key problem for marine data quality inspection. As few marine data quality inspection reported, paper[7] introduced methods including extreme control method, detection method, Dixon inspection detection to control data quality. According to discontinuous phenomenon caused by GPS buoy side, the researcher adopted interpolation method and continuation value to control marine data quality. Take velocity information obtained by LADCP for example, results on how velocity influence quality data controls were presented to strengthen the importance of marine data quality control. Various studied on quality inspection of different marine data have been reported, however, few studies focus on how to build quality inspection and how to control its quality.

Recently skyline calculation is favored domestic and overseas. Study on using skyline calculation in static environment and using spatial index for quick skyline query is reported. Paper[12] put forward a new skyline-based cluster stricter and applies this method in wireless sensor network. Skyline query is multiple goal program problems and it balances several factors for better decision making.

In this paper, traditional percentage method is used to propose data quality inspection method for marine data inspection. Residual collections of all quality inspection method are calculated based on hyper-geometric distribution model. We use skyline block-nesting circulation method to optimize the existing quality inspection methods. By making balance between inspection accuracy and cost, top-notch quality inspection method for the marine data was provided.

OCEAN DATA QUALITY INSPECTION PROGRAM AND PROGRAM RESIDUALS

Ocean data quality inspection plan

Quality inspection of marine data denoted S (N, n, c), where N is the volume, that is to be tested for the total number of marine data; n is the sample volume, the volume extracted from marine data samples to check quantity; c to receive the number of samples that appear to allow the maximum number of ocean data failed. Oceanographic data from the inspection lot to be tested

in N n samples, one by one check their quality; remember ocean sample data number of nonconforming items is d, if the number of ocean data failed to receive less number c, the batch data reaches the ocean accuracy requirements are considered not to be found in ocean data quality problems, and vice versa explanation batch ocean data quality problems.

This article uses the inspection lot reject rate to measure the level of ocean data quality standards, with the average level of quality marine data used to measure the average quality of the data. Oceanographic data which reject rate is calculated as (1) shown in the average quality level of marine data is calculated as (2) shown below:

$$p = \frac{D}{N} \times 100\% \quad \textbf{(1)}$$

$$\overline{p} = 100\% \times \frac{\sum_{i=1}^{m} d_i}{\sum_{i=1}^{m} n_i} \quad \textbf{(2)}$$

Among them, for the i-th sample batch individually check oceanographic data, we found that the number of nonconforming data; n is the i-inspection lot oceanographic data on a sample volume; m represents ocean data for the batch to be tested.

Ocean data quality solution residuals

For each batch oceanographic data to be tested, there is an acceptance quality limit its AQL (acceptance quality level) and the ultimate quality limit LQL (limit quality level). Acceptance quality limit (AQL)[13] is to be submitted when the number of data sequences acceptance testing, the process allowed the worst average quality level, it is possible to receive and reject the process average limit value. Ocean data on a number of quality inspection before the required data quality requirements according to the times given AQL inspection process value, namely inspection lot permissible nonconforming rate p. Limiting quality limit (LQL)[14] refers to sampling, probability of acceptance is limited to a low level of quality, it is in the sampling inspection for not receiving the minimum batch quality.

Based on the hyper geometric distribution model, the probability of receiving quality inspection program referred to as follow:

$$L(p) = \sum_{d=0}^{c} \frac{\binom{N-D}{n-d}\binom{D}{d}}{\binom{N}{n}} \quad \textbf{(3)}$$

$$D = N \times p \quad \textbf{(4)}$$

Therefore, based on the reception quality limit AQL residual probability of reception Ea, and the ultimate quality limit LQL residual probability of reception Eb, is given by the following formula:

$$E_a = \left| \sum_{d=0}^{c} \frac{\binom{N-round(N\cdot p_a)}{n-d}\binom{round(N\cdot p_a)}{d}}{\binom{N}{n}} - (1-\alpha) \right| = \left| L(p_a) - (1-\alpha) \right| \quad \textbf{(5)}$$

$$E_b = \sum_{d=0}^{c} \frac{\binom{N-round(N\cdot p_b)}{n-d}\binom{round(N\cdot p_b)}{d}}{\binom{N}{n}} - \beta = L(p_b) - \beta \quad \textbf{(6)}$$

Where α for the production of risk, when the quality of data to meet quality marine receiving limit AQL, its probability of acceptance $L(p_a)$ should be (1 -α) nearby, E_a for the acceptance quality limit acceptance probability residuals; β for the use of risk, when the quality level inferior limit quality limit LQL, its probability of acceptance $L(p_b)$ should β nearby, E_b is the ultimate quality limit acceptance probability residuals.

OCEAN DATA QUALITY INSPECTION SCHEME OPTIMIZATION SELECTION ALGORITHM

Block nested loop (block-nested-loops, BNL)[16] is a property value has two data points pairwise comparison method is an optimization algorithm, by its very nature is a multi-objective decision-making algorithms. In this paper, the percentage of quality inspection programs on ocean data quality inspection solution S (N, n, c) data inspection, quality inspection program for the use of BNL in the acceptance quality limit and the limit probability of acceptance quality limit residuals Ea residual probability of acceptance Eb optimized selection, in both the producer and the consumer's risk exposure under conditions selected to optimize the qual-

FULL PAPER

ity of the inspection program.

Input: marine data sets to be tested O, | O | = N;

Output: optimal maritime data quality inspection program S (N, n, c).

Step 1 Find maritime data quality verification scheme set Q, | Q | = N2

Step 2 for (i = 1; i <= N; i + +) {

Using the formula (5) Find the residuals ai, and put it in the residuals set Ea;

Using the formula (6) request residuals bi, and put it in the residuals set Eb;

/ / Using the formula (5) and (6) request residuals set Ea and Eb;} Step 3 sets Ea and Eb residuals as input, called skyline block nested loop algorithm; through the block nested loop algorithm to calculate the optimal solution (ak, bk) (0 <k | Ea |);

Step 4 (ak, bk) Q from the program chooses the optimal solution set S (N, n, c).

Algorithm Analysis: In this algorithm, seeking maritime data quality verification scheme set Q is the time complexity is O (N2); seeking residuals set time complexity is O (N2); block nested loop algorithm time complexity is O (N2); from the program chooses the optimal solution set Q the time complexity is O (N). Therefore, the time complexity of the algorithm is O (N2).

EXPERIMENTAL ANALYSIS

Experimental data

Taking a portion of farmed sea area monitoring site data, for example, which includes monitoring points within the study area N is 1392 bits of data, each data point bits include three categories, namely, spatial location data, marine and aquaculture information data element data

Four different sampling ratio f of marine data for quality testing, ocean data for batch N, sample size n, N batch were taken at 5%, 10%, 15% and 20%, the number of c take different receiver values listed in TABLE 1 as a percentage of ocean data quality inspection program.

TABLE 1 : List of the percentage sampling plan

f	*n*	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}
5%	69	0	1	2	3	4	5	6	7	/	/	/	/	/	/	/	/
10%	139	0	1	2	3	4	5	6	7	8	9	10	/	/	/	/	/
15%	208	0	1	2	3	4	5	6	7	8	9	10	11	12	13	/	/
20%	278	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

Optimization of quality inspection program

According to the above four different percentages sampling method, using the hyper-geometric distribution model, based on AQL and LQL value corresponding reject rate, reject rate were taken Pa = 0.02, Pb = 0.1, the calculated probability value received L (Pa) and L (Pb), and the corresponding residual value Ea and Eb.

BNL algorithms use different sampling fractions generated ocean data quality inspection program to choose. First of marine data sampling plan is defined as a collection of residual handicap, quality inspection program for all residues interact almost set twenty-two comparison to filter out the balance between the optimal solution residuals, After repeated comparisonsÿand the optimal quality inspection program will turn out.

CONCLUSION

This article introduces the idea of skyline to optimize ocean data quality inspection scheme selection. Using the hyper-geometric distribution model to calculate residuals, and then chooses the optimum ocean data quality inspection program via a block nested loop algorithm, the experimental results the feasibility of the method. This paper demonstrates how to choose the best quality inspection programs of oceanographic data rapidly, improves the theoretical system of marine data quality inspection.

REFERENCES

[1] Han Jing-yu, Xu Li-zhen, Dong Yi-sheng; Data quality survey[J].Computer Science, **35(2)**, 1-12

(2008).

[2] Bao Yang, Qi Xuan; Large software systems data quality issues[J].Computer Engineering and Design, **32**, 963 **(2011)**.

[3] Xu Zi-zhou, Song De-rui; The Control method of marine environmental monitoring data quality[J]. Marine Environmental Science (in Chinese), **28(3)**, 329-334 **(2009)**.

[4] R.Y.Wang, H.B.Kon, S.E.Madnick; Data quality requirements analysis and modeling [C]. In:Proc. of Ninth ICDE, **(1993)**.

[5] E.Rahm; Do Hong-hai. Data cleaning: problems and current approaches [J].IEEE Data Engineering Bulletin, **23(4)**, 3~13 **(2000)**.

[6] A.Parssian, S.Sarkar, V.S.Jacob; Assessing information qualityfor the composite relational operation joins [C]. In:Proc. of Seventh International Conference on Information Quality, **(2002)**.

[7] Shi Jing-tao, Zhou Zhi-hai; Ocean station data quality control technology[J]. Marine Technology, (in Chinese), **30**, 11-30 **(2011)**.

[8] Zhang Suo-ping; Single point GPS wave research methods and data quality control[J].Marine Technology, (in Chinese), **27(3)**, 15-18 **(2008)**.

[9] Xie Ling-ling, Xiong Xue-jun, Yang Qing-xuan; LADCP configuration files and data quality control parameter settings[J] Marine Technology, (in Chinese), **28(1)**, 19-23 **(2009)**.

[10] Zhu Lin, Zhou Shui-geng; Skyline computation:survey [J].Computer Engineering and Applications, (in Chinese), **44(6)**, 160-165 **(2008)**.

[11] I.Bartolini, P.Ciaccia, M.Patella; Efficient sort-based skyline evaluation[J].ACM Transactions on Database Systems (TODS), **33(4)**, 1~49 **(2008)**.

[12] Wang Yan-jie; Research on Skyline Computation and Application Based on Data Stream [D]. Jiangsu: Jiangsu University, (in Chinese), **(2011)**.

[13] Wang Zhen-hua; Principle, methods and application of sampling inspection for quality control of geospatial data [D]. Shanghai: Tongji University, (in Chinese), **(2011)**.

[14] V.Kuralmani, K.Govindaraju; Modified Tables for the Selection of Double Sampling Attribute Plan Indexed by AQL and LQL[C]. Communications in Statistics. Part A: Theory and Methods, **24(7)**, 1897 **(1995)**.

[15] Yu Shanqi; Sampling Inspection and Quality Control[M].Beijing:Peiking University Press, **(1991)**.

[16] S.Borzsonyi, D.Kossmann, K.Stocker; The Skyline Operator[C] Proceedings of the 17th International Conference on Data Engineering (ICDE), 2001. Washington, DC, USA: IEEE Computer Society, 421-430 **(2001)**.