Research on abnormal data processing method in intelligent data adapter based on bayesian network

HAO Ping, LU Bin-li

College of computer science and technology, Zhejiang university of technology, Hangzhou, China

College of computer science and technology, Zhejiang university of technology, Hangzhou, China

forseone@163.com, lukelu1988@gmail.com

Keywords: Bayesian Network, Junction Tree Reasoning, Heterogeneous data processing, Data Mining, Probability Inference

Abstract. This paper put forward a bayesian network junction tree reasoning and rule reasoning hybrid algorithm to solve the adaptation problem in Public Data Center of heterogeneous system, based on the research in intelligent data processing method. Using adapter's dynamic data monitoring function, first, pick out all abnormal data. then, apply the hybrid algorithm on the abnormal data. finally, recover the abnormal data and report abnormal processing result. This method has been applied to many domestic universities's intelligent data exchange system in the Public Data Center. Through practice, this algorithm can effectively improve the reliability and integrity in heterogeneous data exchange system, and obtained good application effect.

Introduction

As the development of internet's cloud computing information technology, the networking and platform changing of information resource has become a trend. Because of the complex business relationship between enterprise, government and other social organizations, if we want to construct the public service platform to develop the network storage, information sharing, statistical analysis, cloud computing and other public services, we need to integrate information effectively in different information system that distributed in different distinct, so the data adapter and data exchange component is a key part, which play a role of bridge. But then the problem is that once the exchanging data is abnormal or missing, will bring big influence to the normal operator for the whole platform service, so data reliability is particularly important to platform changing system.

Have access to relevant paper, these are some description of the heterogeneous data integration, but most of it are data exchanging between single system. There are little research on abnormal data processing method between multisystem.

This paper based on digital campus's public data service center, studying on the data adapter's reliability and integrity in the process of heterogeneous data exchanging. and put forward a bayesian network reasoning and rule reasoning hybrid algorithm, and applied the algorithm to the data monitoring and intelligent processing in data adapter of public data center service system.

Bayesian network

Bayesian network[2] is a directed acyclic graph(DAG) composed of a set of random variables and directed edge which present the influence between random variables, it can reason conditional probability between random variables effectively.It usually consists of a set of random variables set $V = (v_1,...,v_i)$, and using the network structure of directed acyclic graph expressing the conditional dependence between random variable set V. Bayesian network get a lot of application in reasoning problem with intrinsic uncertainty, such as automobile fault diagnosis system, medical diagnosis system and data mining medium.

Bayesian network foundation

Beyasin network defined as a binary group of random variable probability distribution P and directed acyclic graph(DAG):

Published by Atlantis Press, Paris, France.

$$B = (G, P) \tag{1}$$

The directed acyclic graph expressed as random variables set V and variable dependence directed edge set E :

$$G = (V, E) \tag{2}$$

Bayesian network must meet the Markov condition, described as follows:

$$I_p(X, ND_x \mid PA_x) \tag{3}$$

ND_x representing non-descendant variable set of variable X , PA_x representing all father node of variable X . I_p representing probability independent under probability distribution P , variable sets's independence described as follows:

$$P(X \mid Y) = P(X) \text{ equals to } P(Y \mid X) = P(Y) \tag{4}$$

Joint probability distribution of all random variables can be described as follows:

$$P(V_1, ..., V_n) = \prod_1^n P(V_i \mid PA_i) \tag{5}$$

Bayesian network constructing

Bayesian network construction is divided into three stages:

(1)Determine the random variable set V and value range of each random variable.Through the knowledge of domain expert, determined the random variable set V and value range showed in Table 1:

Table 1: Abnormal data processing random variable set and discretization value

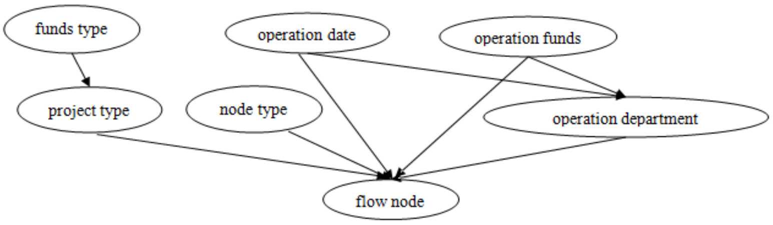
[Random Variable]	[Description]	[Discretization Value]
[project type]	[major]	[1]
	[campus]	[2]
	[provincial]	[3]
[funds type]	[high]	[1]
	[middle]	[2]
	[low]	[3]
[operation department]	[application]	[1]
	[review]	[2]
	[approval]	[3]
	[monitoring]	[4]
[flow node]	[application]	[1]
	[review]	[2]
	[approval]	[3]
[node type]	[time]	[1]
	[funds]	[2]
	[procedure]	[3]
	[sequence]	[4]
[operation date]	[miss]	[1]
	[normal]	[2]
	[abnormal]	[3]
[operation funds]	[miss]	[1]
	[normal]	[2]
	[abnormal]	[3]

(2)Constructing network, after determining the random variable set, we can use knowledge learn, or the domain experts knowledge to construct the network. This paer used the method of K2 search, K2 algorithm is a greedy heuristic search method, it selects the most suitable father node set through score function, repeat the selection for each node, and output full bayesian network finally.

(3)Bayesian network constructed using K2 algorithm is shown in Fig. 1, consists of selected seven random variables in Table 1 and some directed edge. As shown in Fig. 1, the value of $V_{flownode}$ is influence by $(V_{projecttype}, V_{no \det ype}, V_{operationdate}, V_{operationfunds}, V_{operationdepartment})$ directly and $(V_{fundstype})$

indirectly. The result is very obvious, because the flow node is key data of work flow data extraction. Once the flow node data is lost, we need directly influenced data to reasoning the probability distribution.

Fig. 1: Abnormal data processing bayesian network



(3) Determining the node conditional probability. Bayesian network's prior condition probability can be determined by parameter learning or structure learning from training data set, or from expert knowledge directly. Based on bayesian network in Fiag. 1 and Eq. 5, we need to determine the conditional probability of $P(V_{fundstype})$, $P(V_{projecttype}|V_{fundstype})$, $P(V_{no\det ype})$, $P(V_{operationdepartment}|V_{operationdate},V_{operationfunds})$ etc. after that we can determined the probability distribution of all seven variables in Table. 1.

Junction tree reasoning in bayesian network

Bayesian network inference using directed acyclic graph as the structure foundation, compute posterior probability on the basis of the prior probability. Bayesian network accurate reasoning[3~4] is a kind of practical in bayesian network reasoning, accurate reasoning is consists of group tree reasoning, combinatorial optimization reasoning, diagram code reasoning, combined tree reasoning and so on.

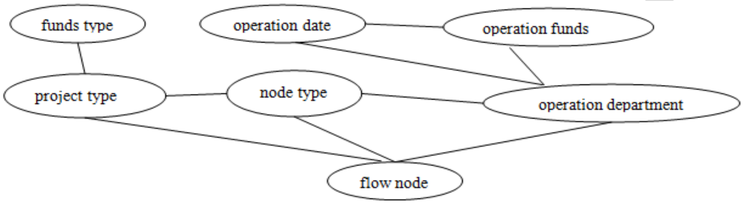
Constructing Moral graph

Algorithm steps described as follows:

1)defined G as constructed bayesian network, V as node set in network
2)remove all direct in G, get undirected graph G'
3)extract v_i from node set V
4)find father node set PA_{vi} of v_i
5)for each pair of nodes in PA_{vi}, if there is no edge, then add an edge
6)repeats steps 3-5, until node set V is empty

The moral graph of the bayesian network in Fig. 1, shown as Fig. 2:

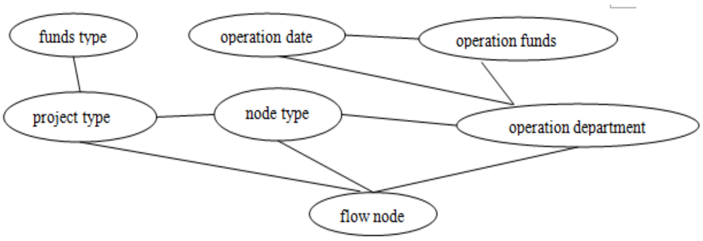
Fig. 2: Abnormal data processing moral graph



Triangularization

Triangularization is aimed at ring in graph that consists more than four nodes, cut it to triangular ring. There are four rings in Fig. 2, $(V_{fundstype},V_{projecttype})$, $(V_{projecttype},V_{no\det ype},V_{flownode})$, $(V_{no\det ype},V_{flownode},V_{operationdepartment})$ etc, they all meet the triangularization requirements. so the triangularion graph is shown in Fig. 3 as same as Fig. 2:

Fig. 3: Abnormal data processing triangularized moral graph



Identifying Cliques

The Clique of graph is defined as the largest complete subgraph, we can identify all cliques of Fig. 3 in triangularization. we get all cliques as follows: $(V_{projecttype},V_{no\det ype},V_{flownode})$, $(V_{no\det ype},V_{flownode},V_{operationdepartment})$, $(V_{operationdate},V_{operationfunds},V_{operationdepartment})$ and $(V_{fundstype},V_{projecttype})$.

Constructing Junction Tree

After identify all cliques in moral graph, wo can combine clique to build the junction tree. This tree is the optimal clique tree, and can help to shorten the time of reasoning algorithm. Junction tree constructing algorithm[1] described as follows:

1)defined empty set S as candidate separation set, take every clique node as a separated tree, we can obtain tree set $T=(t_1,...,t_i)$, and the clique set $C=(c_1,...,c_i)$

2)to every pair of clique node c_i,c_j , construct candidate separation set s_{ij} and put it into set S

3)pick up separation subset S_i which contains most nodes

4)when S_i contains more than one element, take the sum degree of all node in as standard, defined filtered set as S_i'

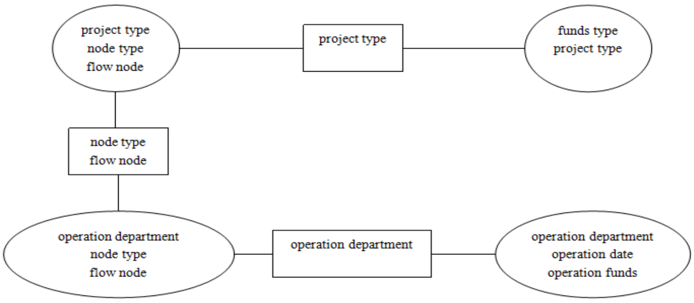
5)when S_i' contains more than one element, we can use domain expert knowledge as standard to choose subset separation set s_{ij}

6)when the separation set s_{ij} corresponding to cliques node c_i,c_j , insert s_{ij} between c_i,c_j , connected s_{ij} and c_i , s_{ij} and c_j .

7)repeat steps 3-6 until n-1 separation sets has inserted

The constructed junction tree is shown in Fig. 4

Fig. 4: Abnormal data processing junction tree



Junction Tree Message Transfer Reasoning

Using junction tree to fill abnormal data described as follows:

1)find the abnormal random variable $v_{abnormal}$

2)position and find the $v_{abnormal}$ related clique node $c_{abnormal}$

3)determine all conditional probability e

4)if e is empty, using $P(v_{abnormal})=\sum_{\mathrm{C}abnormal}\phi c_{abnormal}$ to compute the value probability distribution

5)if e is not empty, using $P(v_{abnormal}\,|\,\mathrm{e})=P(v_{abnormal},e)/P(e)=P(v_{abnormal},e)/\sum_{vabnormal}P(v_{abnormal},e)$ to compute the probability distribution

6)pick the highest probability value to repair the abnormal value

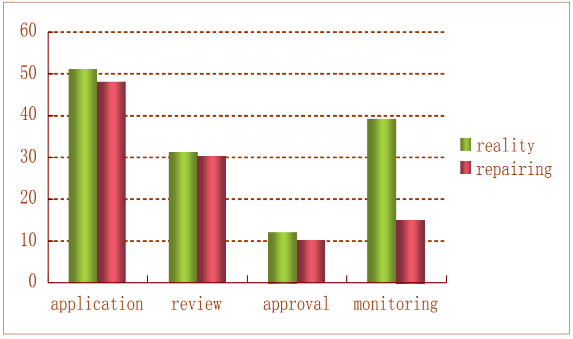
Experiment and Analysis

In order to validate the effectiveness of the algorithm, we place an intelligent data adapter between Electronic Monitoring Platform and Construction Management System in some schools, to intelligent fill the key data loss problem. The intelligent data adapter using the junction tree to reason abnormal data in Fig. 4. On the analysis of experiment data set, we conclude that this abnormal data mostly happened on operation department, operation date and flow node.

We accurately extract 6000 records from the management system, the missing rate in operation department is 2.1%, in operation date is 1%, in flow node is 4%.

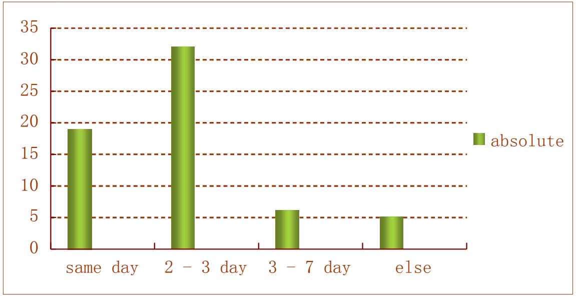
The repairing result in operation department is shown as Fig. 5, because of the human factors in monitoring department, the result is unpredictable.

Fig. 5: operation department data repair contrast



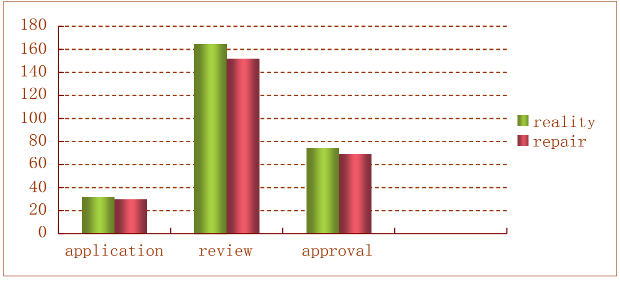
The repairing in operation date is shown as Fig. 6, because the date is continuous value, so we use the absolute value between fill data and real data to show the fill accuracy. As the Fig. 6 shown, the repair absolute value fall in the same day is 40%, and the absolute value fall in 2 -3 day is 80% up, basic conform to the monitor transaction demand.

Fig. 6: operation date data repair contrast



The repairing in flow node is shown as Fig. 7, because of the importance of flow node in transaction monitoring, Fig. 7 is shown that the algorithm is meet the abnormal data processing of monitoring platform's needs.

Fig. 7: flow node data repair contrast



Summary

To key data loss problem in the business data integration of the electronic monitoring platform, first, analysis the key random variables in system, and determined random variables set and it's value rang in abnormal data processing; Then, construct the abnormal data processing bayesian networking using expert knowledge; finally, based on constructed bayesian network, construct junction tree and reason using the junction tree. As the experiment result shows, the repaired data can meet monitor transaction needs. How to use the bayesian networks to multivariate multi-dimensional repairing will be the next step research content.

Published by Atlantis Press, Paris, France.

Reference

[1] PEAR. J. Probabilistic Reasoning in Intelligent Systems; Networks of Plausibale Inference[M]. San Francisco; Morgan Kaufmann, 1988; 62-65.

[2] HUANG C, DARWICHE A. Inference in Bayesian Networks; a Procedural Guide[J]. International Journal of Approximate Reasoning, 1994(11); 1-158.

[3] Murphy K P. A variational approximation for bayesian networks with discrete and Continuous Latent Variables[EB/OL]. htpp://www.berkeley.edu/ - murphy/publ.2005-05-14.

[4] Kjaerulff U, Hugin D. A computational system for dynamic time-sliced Bayesian networks [J]. International Journal of Forecasting, Special Issue on Probability Forcasting, 1994, 19(10);1-3.

[5] Gregory F Cooper, Edward Hemkovits. A Bayesian method for the induction of probability networks from data[J]. Machine Learning(S0885 - 6125). 1992, 9(4);309-347.

[6] Philippe Galinier, Michel Habib, Christophe Paul. Chordal graphs and ther clique graphs [EB/OL]. http://citeseer.nj.ncc.com/galinier95chordal.html 2003.

[7] B.Boerlage, Link Strengths in Bayesian networks, Master's thesis, Dept of Computer Science, U. of British Columbia 1995.

[8] Kevin Patrick Murphy et. al, The Bayes Net Toolbox for Matlab, http://www.ai.mit.edu/~murphyk/Software/BNT/bnt.html.