IEICE TRANS. INF. & SYST., VOL.E97–D, NO.1 JANUARY 2014

LETTER

# Problem Analyzing by Distributed, History and Trend Templates with Integrated Supporting of DCMS

**Zhiming CAI**[†a)], ***Member***, **Zhe YANG**[†], ***and*** **Menghan WANG**[††], ***Nonmembers***

**SUMMARY** In analysis of general-purpose problems which involves many factors from different viewpoints, an important challenge is to acquire different opinions and distributed modeling templates from multiple remote experts, and to aggregate these templates. In order to deal with this problem, we developed the Distributed Cooperative Modeling System (DCMS) by integrating our achievements [1]–[5]. The paper introduces how to analyze a complex problem using DCMS, with distributed templates from multiple experts, historical templates based on statistical data, and trend templates deduced from historical data, with the example of analyzing "diversification of Macao industries"

***key words:*** *problem analysis, DCMS, modeling template, decision supporting*

## 1. Introduction

In order to assist the analysis of complicated general-purpose problems and systems, we proposed the distributed modeling methodology and its networked supporting platform, the Distributed Cooperative Modeling System (DCMS) [1], [2]. DCMS provides assistance in problem analyzing, decision making and teamwork organization, featuring distributed visual modeling.

Visual modeling is a commonly used method in developing large-scale systems, especially in software engineering. The visual modeling of a complex system or problem (e.g. the diversification of Macao industries) using DCMS involves construction of the visual templates, which are the graphical representation of the various elements (objects, relationships among the objects, weights/attributes of them, etc) of the problem.

With the usage of templates to structure the problem, DCMS can perform inferences on the problem and answer questions such as optimal solution selection, reasoning on direct or indirect relationship among the objects, as well as dividing and reorganizing of the template to understand complex models of the problem in multiple views and levels [3]. The aggregation by Analytical Hierarchy Process (AHP) for single template and Ordered Weighted Geometric (OWG) for multiple templates can determine the order of a set of elements and reach the conclusion among the templates [4].

To facilitate data mining and time span analysis of the templates, DCMS has incorporated Pearson Linear Correlation, Monotonic Correlation [13], [14], Markov process and linear Gaussian [11], [12] in the system, which provides the integrated supporting to mine the appropriate relations among elements in templates, analyze the status of the problem by historical data, and predict the trend of the problem with or without relevant actions.

The distributed modeling usually involves multiple experts from many different sites. The DCMS also facilitates to overcome the difficulties of how to organize, cooperate, supervise, evaluate and optimize such distributed working of experts teams [2], [5].

The novelty of DCMS is the capability to analyze the general-purpose complex problems, by means of visual distributed modeling with mining on historical data and predicting the trend of the problem. The popular general-purpose modeling is the systematic use of the Unified Modelling Language (UML), an industry standard for modeling software-intensive systems, but not for general-purpose complex problem analyzing [19].

There exist many visual modeling tools such as PowerDesigner [8], Rational Rose [9] and VisSim [10]. PowerDesigner is the modeling tool with the strength on data modeling and database design. Rational Rose is closely designed for UML based development to produce visual models of software architectures, databases and system requirements. VisSim provides a method for constructing and simulating large-scale dynamic systems, control systems and digital signal processing, with powerful math engine for linear, nonlinear, continuous time and discrete time designs. However, the visual modeling with all of these tools is lacking the features of templates aggregation, mining on historical data and deduction of trend templates, which is worthy of analyzing problem with distributed teams.

The paper will briefly present how to analyze a problem using DCMS with distributed templates among teams, historical templates from statistical data and trend templates.

## 2. The DCMS Methodology of Analyzing Problems

The proposed methodology of analyzing problems by DCMS is:

(1) Analyzing problems using distributed templates:
Build the templates based on the historical data of the problem (e.g. the historical data of Macao Industries) with experts' own judgments; share, exchange the templates among the distributed teams; improve the distributed templates with

Manuscript received July 24, 2013.
Manuscript revised September 15, 2013.
[†]The authors are with the Faculty of Information Technology, Macao University of Science and Technology, Macao.
[††]The author is with University of Florida, USA.
a) E-mail: zmcai@must.edu.mo
DOI: 10.1587/transinf.E97.D.146

discussing and cooperating.

(2) Analyzing problems using historical templates:
Study direct, indirect relationships among the elements in the historical templates by the algorithm introduced in [3]; mine the correlation through Pearson linear correlation, and monotonic correlation; aggregate the historical templates to understand the status of the problem (e.g. the rank of Macao industries) by AHP/OWG stated in [2], [4].

(3) Analyzing problems using trend templates:
Work out the trend of the elements in templates by Markov/Gaussian prediction and generate trend templates of the problem to see the future status; summarize the above information to design some actions (policies) on certain purposes (e.g. apply some related policies on some elements to modify the future rank of Macao industries); re-generate the trend templates to see the effects of the actions.

The above process can be performed according to the demand of the experts, for instance, in order to look into solutions for the following questions: What future status will be caused by a given action/policy? Which elements are related in historical and trend templates? What's the difference between different experts and teams?

## 3. Analyzing Problems by Distributed Templates

The templates built in DCMS can help the distributed experts to model their understanding of a certain problem by multiple graphical views. The objects and relationships are the semantic primitives of templates. The objects would be the goals, solutions and conditions of the problem. On the other hand, the relationship lines among the objects represent the semantic network in templates, which are essential in problem analyzing [6]. The notations and definitions of the objects and relationships are introduced in [1], actually, the experts may extend the definitions of notations in DCMS to meet the requirements of their specific problem analyzing.

Using DCMS, the expert agents within a same department/team store their templates in both the local databases and the department database. The databases from different departments/teams could then be connected through internet. The templates can be distributed to the related agents and improved by knowledge sharing and synthesizing (see Fig. 1).

Figure 2 is an instance showing a template regarding the status of Macao industries. The templates can support visual considering and researching complex models in multiple views and levels, as shown in Fig. 3, in which the expert likes to see the divided views specific to "Financial" and "Employed population". Every object, line and number on the template can be double-clicked to edit/check the detailed note of unit, what, why, who, when, and other attributes.

Since analysis of large and complex problems usually calls for cooperative work of one or several teams, the system facilitates communication and work organization among distributed users [2], [5]. By assigning software-agents to supervise the corresponding expert-agents, the system records and organizes the working track of participants and provides statistical information to managers, helping in project management and supervision [5], [7].

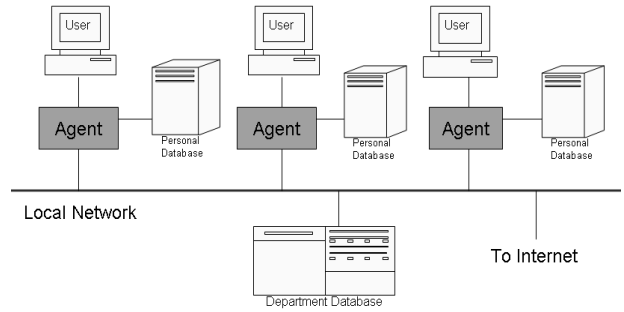


**Fig. 1** The Network of DCMS.

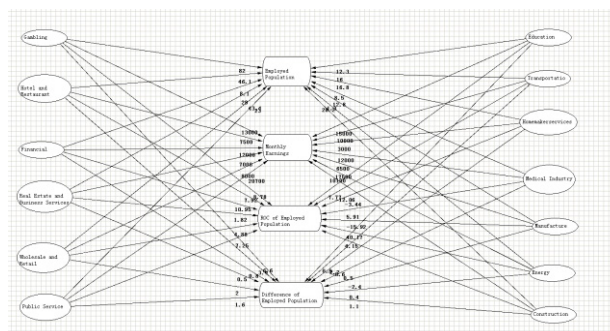

**Fig. 2** The Instance of the Template (2011 Historical Template).

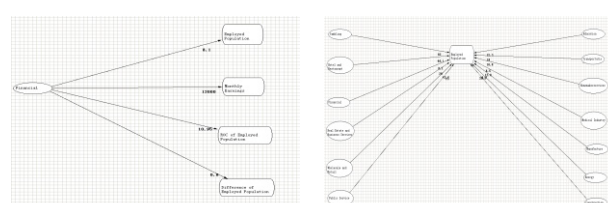

**Fig. 3** Divided Templates from Fig. 2.

## 4. Analyzing Problems by Historical Templates

The historical templates are built based on the historical data of the problem (e.g. the statistical data of Macao industries). For example, Fig. 2 is actually the historical template of Macao industries of year 2011. With analyzing a set of the historical templates, the experts usually want to look into the following questions: which elements are closely related with each other, which elements influences the elements of interest the most, how will changing of some elements affect others, and what conclusion can be drawn from these historical templates (for example, what's the rank of Macao industries according to the historical templates), etc.

The correlation analysis among multiple elements in templates can reveal their hidden relationship, enabling DCMS to assist experts to understand the above questions within templates more clearly. Most importantly, experts

can use this information to adjust actions which are relatively easy to control and finally achieve regulation of certain objectives of interest of the problem. This is exactly the case in "the diversification of Macao industries" problem where no readily-made action collections are available. What's more, the number of possible actions is so huge that we probably can never traverse the solution space to test each solution and find the best policy. Besides, the industry diversification is a complicate problem which can hardly be handled without the above supporting.

One of the most widely used correlation analysis method is Pearson correlation coefficient [14]. It represents the linear correlation between two random variables, and could be easily extended to represent linear correlation of multiple variables and partial correlation of two variables with elimination of effects from other relative variables.

For multiple variables, after calculating correlation coefficient between each pair of variables, multi-variant correlation coefficient can be calculated between the dependent variable and multiple independent variables, and then partial correlation between the dependent variable and each of the independent variables will be computed, which allows us to select independent variables that have significant effect on the dependent variable step by step.

Further considering the function $y = x^2$, $x > 0$. It is obvious that x and y are totally correlated, but if we compute linear coefficient on a random sampling of points on the function curve, the result could be much smaller than what we expect. Monotonic correlation [14] measures the extent to which that two variables change correspondingly without requiring a linear change rate. This is also called rank correlation. Two commonly used rank correlation coefficients are used here: Spearman's rank correlation coefficient and Kendall's tau rank correlation coefficient.

DCMS has implemented Pearson correlation coefficient, Kendall's tau correlation coefficient and Spearman rank to analyze correlation between any two elements in a set of the templates (e.g. indices of Macao of many years in a set of historical templates). On basis of these correlation analysis and historical data, experts can mine the right related information among the historical templates, which will definitely benefit the design of the effective and quantitative actions/policies on certain purpose.

For example, as shown in Fig. 4, DCMS figured out the correlations between gambling revenue, gambling employee and number of free tourist following the historical templates from 2000 to 2007, so the experts can try policies "drop-down the free tourist from mainland of China" to control the growing of gambling. As per the regulation from right chart, the gambling employee will be about 52.5 thousands if the "free tourist" narrows around 11985.6 thousands, causing gambling revenue of 20747.6 million Macao dollars from left chart (the data in Chinese in the Fig. 4. are directly from official original databases of Macao government).

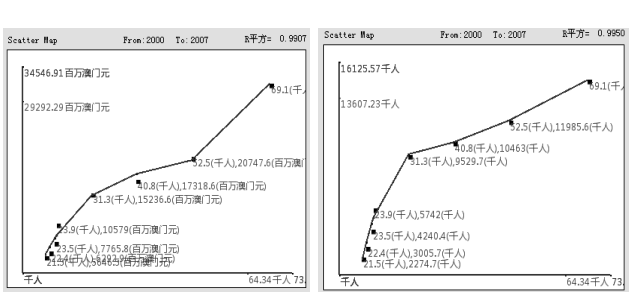


**Fig. 4** Relations of gambling revenue (left chart.) and employee (right chart.) with number of free tourist.

## 5. Analyzing Problems by Trend Templates

The trend templates can be generated in DCMS by computing future trend of elements in the templates based on historical data. In the practical problem of Macao industry diversification, by predicting the possible values of important industrial indices supposing actions or no actions are taken, the expert can recognize the existing problems in current industry status and come up with corresponding solutions/actions/policies with trends of related elements on the templates.

Markov Process has been proven very successful in many predictions [11], [12]. By Markov, if $\mathbf{P}(\mathbf{X}_{t+1}|\mathbf{x}_t)$ denotes the probability distribution of state at time t+1 defined for each value in $\mathbf{x}_t$, which is the collection of possible state values at time t, the predicted distribution at time k starting from state $x_0$ will be

$$P(\mathbf{X}_k) = \sum_{t=0}^{k} \mathbf{P}(\mathbf{X}_{t+1}|\mathbf{x}_t)$$

If the state variable is continuous rather than discrete, the summation on the right-hand side should be replaced by an integral. In DCMS, prediction for discrete variable and continuous variable are respectively implemented.

Handling the prediction for discrete state variable can actually be: the summation mentioned above can be turned into arithmetic of matrices. We use $\mathbf{P}$ to denote the state transition probability matrix. $\mathbf{P}$ consists of n rows and n columns, where n is equal to the number of possible states. A cell of $\mathbf{P}$, $p_{ij}$, is $P(X_{t+1} = j \mid X_t = i)$, that is, the probability of transition from state i to state j. One thing to be noticed is that if a certain state did not appear in the historical data, its relative cells will be zero and the Markov process will never transit to that state using the computed transition matrix. If the training data is large enough, we can plus one to each count to avoid this problem. After obtaining the transition matrix, the probability distribution of time t+1 is the probability distribution of time t multiplied by the transition matrix.

For continuous state variable, the transition matrix could no longer be used, since the value of continuous variable cannot be iterated. In fact, an approximate distribution for the state variable and an appropriate function form to model relationship between neighboring states should be chosen. Hence, the popular linear Gaussian model [14], [15] is used to analyze a single state variable to demonstrate

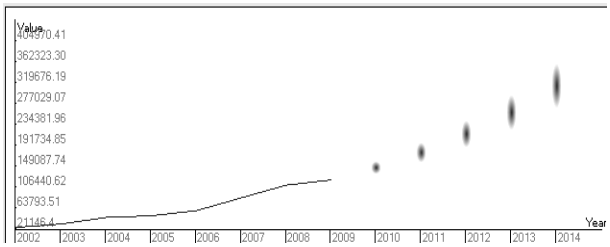


**Fig. 5** Historical & predicted trend of gambling revenue from 2002 to 2014.

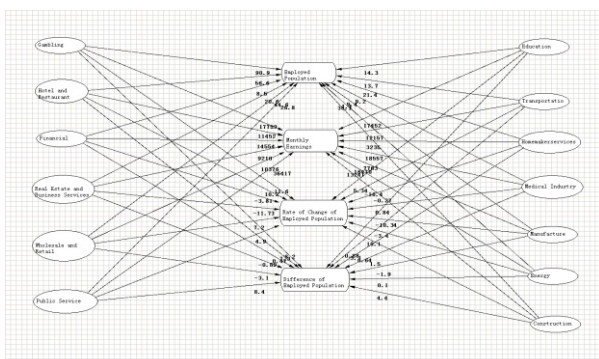


**Fig. 6** The Trend Template of 2014.

learning and prediction for continuous state variable. For more than one variable, joint distribution to substitute single variable distribution could be used. The linear Gaussian model could also be substituted by other function and distribution models and the prediction process will be similar.

The Fig. 5 is a demonstration of prediction of trend of Macao gambling revenue from 2010 to 2014 in DCMS. The historical data is from 2002 to 2009.The historical data points are connected with solid line, while predicted distributions are shown as gradient bars: the central point is the expectation value and the bar covers the value range $[\mu - 3\sigma, \mu + 3\sigma]$. The predicted results are quite close to the real data from 2009 to 2012.

In this way, the trend of each element in templates is worked out and the trend templates can be generated by DCMS. The Fig. 6 is the trend template of Macao industries in 2014 generated on the basis of historical templates from 2002 to 2012.

By adjusting historical track of some relevant elements as per the correlation analysis on historical templates, and re-computing the trend templates with re-studying them, the experts can see what will happen to trend if some actions/policies are employed. For example, in the Sect. 3, we knew that the gambling revenue will be going down to 20747.6 million Macao dollars when the "free tourist" narrows around 11985.6 thousands. What will be the overall status if the "free tourist" is narrowing for several years? The expert may repeatedly deduce the trend templates to look into the effects of different actions/policies.

## 6. Conclusion

With help of the integrated methodology and support by DCMS, the experts can take advantage of the visual, distributed history/trend templates to research and analyze the following questions: What's the historical status of the problem (rank of elements, etc.)? What are the different judgments from different experts and teams? How are the elements related in templates? What are the direct and indirect relations hidden in historical and trend templates? What future status will be caused by a given action/policy? Furthermore, what is the trend of the overall and the partial status (e.g. all or part of indices of Macao) respectively? Which trend is correlated with which other trends (by correlation analysis on trend templates)? The integrated supporting from DCMS can help the experts to answer these questions, not only qualitatively, but also quantitatively in most cases.

In the future works, we will make use of the huge historical data of Macao (many indices were available from 1970's), enabling more hidden information among historical templates to be mined and trend templates with different indices over different period to be deduced. As a result, the past/future social/economic status of Macao will be addressed effectively by DCMS. We may also apply the methodology towards other problems with enough statistical data (for instance, other city's data) to verify the effectiveness of DCMS.

The project is funded by Macao Fund of Development of Science and Technology.

### References

[1] C. Zhiming, Y. Zhe, and W. Menghan, "The distributed modeling methodology and platform," Int. J. Digital Content Technology and its Applications, vol.7, no.9, pp.409–417, May 2013.
[2] C. Zhiming and Y. Jun, "The selection modeling system with grouped agents," Proc. IEEE 22nd Int'l Conf. Advanced Information Networking and Applications. Japan, 2008.
[3] C. Zhiming, Y. Jun, and Z. Yingjie, "Multi-agent/multi-goal modeling templates and The 1-1, n-1, 1-n, n-n deductions of relationship-chain," Int. J. Hybrid Information Technology. vol.3, no.1, pp.49–64, Jan. 2010.
[4] C. Zhiming, H. Liangli, and Y. Jun, "The concentration and distribution of visual group selecting templates and sorting solutions by AHP," J. MUST, vol.1, no.2, pp.13–21, 2007.
[5] C. Zhiming and Y. Jun, "The process conducting and agent audit in the distributed enterprise modeling," Proc. 2008 IEEE Asia-Pacific Services Computing Conference, Taiwan, 2008.
[6] C. Zhiming and Y. Eric, "Addressing performance requirements using a goal and scenario-oriented approach," Proc. 14th Int'l Conf. Advanced Information Systems Engineering. LNCS, pp.706–710, Canada, 2002.
[7] C. Zhiming and L. Ying, "The distributed cooperative work of multi-agent on network," Proc. Inter. Conf. CAID&CD"99, Thailand, 1999.
[8] http://www.sybase.com/products/modelingdevelopment/powerdesigner
[9] http://www-01.ibm.com/software/awdtools/developer/rose/
[10] http://www.vissim.com/products/overview.html

[11] R.A. Howard, Dynamic Programming and Markov Processes, MIT Press, Cambridge Mass, 1960.
[12] S. Proper and P. Tadepalli, "Solving multiagent assignment Markov decision processes," Proc. 8th International Conference on Autonomous Agents and Multiagent System, May 2009.
[13] G.W. Corder and D.I. Foreman, Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach, Wiley, 2009.
[14] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. Morgan Kaufmann, 2005.
[15] M. Negnevitsky, Artificial Intelligence: A Guide to Intelligent Systems, 2nd ed. Pearson Education, 2002.
[16] C. Gutwin, R. Penner, and K. Schneider, "Knowledge sharing in software engineering: Group awareness in distributed software development," Proc. 2004 ACM conference on Computer supported cooperative work (CSCW '04), pp.72–81, 2004.
[17] E. Carmel and R. Agarwal, "Tactical approaches for alleviating distance in global software development," IEEE Software, vol.18, no.2, pp.22–29, 2001.
[18] F. Lanubile, D. Damian, and H.L. Oppenheimer, "Global software development: Technical, organizational, and social challenges," ACM SIGSOFT Software Engineering Notes, vol.28, no.6, p.2, 2003.
[19] F.J. Lucas, F. Molina, and A. Toval, "A systematic review of UML model consistency management," Information and Software Technology, vol.51, no 12, pp.1631–1645, Dec. 2009.