A New Methodology for Web-Knowledge-Based System Using Systematic Thinking, KM Process and Data & Knowledge Engineering Technology:FBR-GAs-CBR-C5.0-CART

Patcharaporn Paokanta
Department of Knowledge Management, Chiang Mai University
Chiang Mai, Thailand, 50200
[1] Patcharaporn.p@cmu.ac.th

***Abstract*—In Knowledge Management perspective, Organization Learning and the selection of Knowledge Management tool affects the Knowledge Management strategy planning. Among the various KM theorem such as Learning method, organization knowledge creation, Cognitive theory, Intangible assets and knowledge capital, Measuring knowledge theory etc., Systematic Thinking plays an important role in Knowledge Management activities especially, the creation of Knowledge Management strategy, KM process and Knowledge Management system. DKET is one of several approaches for implementing the Knowledge Management tools based on the KM strategies. They are not only implemented in forms of standalone system but the web-online system also. Generally, DKET namely Ensemble Learning is well known as the technique of using different training data sets or learning algorithms. Currently, a popular learning algorithm is Fuzzy-Based Reasoning (FBR) which the concept of this theory is "each item is not matched to a given cluster but it has a degree of belonging to a certain cluster". According to these reasons, in this paper, a new methodology for Web-Knowledge-Based System by using Systematic Thinking, Knowledge Process and DKET (FBR-GAs-C5.0-CART) is proposed in terms of KM perspective. The algorithm performance comparisons of Fuzzy C-Means-CBR-GAs-C5.0-CART in several data sets are presented. The satisfied clustering results of Fuzzy-C Means-GAs-CBR-C5.0-CART attain RMSE at 5.10 for the case that full data set, on the other hand the best result of using Fuzzy-C Means-CBR-C5.0-CART attain RMSE at 12.03 in the case that unrecoded variables and CBR-C5.0-CART without symptoms variables. In the future, the other KM theories and DKET will be applied to improve the performance of this system.**

Keyword- Biomedical Computing, Knowledge-Based System, Fuzzy System, Organization Learning, Knowledge Management, Systematic Thinking, Knowledge Discovery, Medical Expert System

I. INTRODUCTION

Recently, almost organizations have their own Knowledge Management strategies. These KM strategies lead to the development of software projects. One of the well known Data and Knowledge Engineering Technology (DKET) for implement these systems is Artificial Intelligent (AI). Among the serious competition in the business organizations and other organizations, the important data, information, knowledge and wisdom are required to support the business activities such as Customer Management, Research and Development, Supply Chain Management, Enterprise Resource Planning, Production Life Cycle management process etc to be the Business intelligent organization. Before constructing these systems, Knowledge Management strategies are planed and informed to the related people for operating following the defined Knowledge Management Processes. The process to transform data to information or information to knowledge or knowledge to wisdom needs Systematic Thinking and DKET to design framework or methodology and generate the outputs also especially, Web-Knowledge-Based System.

According to the previous related published papers of author, the novel DKET, KM process and Systematic Thinking were used to improve the performance of system and methodology. This section, the literature reviews are presented below.

The study of Patcharaporn Paokanta et al. [1] presented the efficiency of data types for classification performance of Machine Learning Techniques for screening β-Thalassemia. In this paper, they proposed the classification performance of KNN, MLP, NaiveBayes, BNs and MLR on Interval scale is better than Nominal scale, with the accuracy percentage 88.98, 87.40, 84.25, 83.46 and 81.89, respectively.

The second paper of Patcharaporn Paokanta et. al [2]. They proposed the rule Induction for screening Thalassemia using Machine Learning Techniques: C5.0 and CART. The objective of this study is to fine out rules and suitable algorithms for implementing Thalassemia KBS. The results of this research presented the different rules and the classification performance of using C5.0 is better than CART with the accuracy percentage 84.25 and 77.17, respectively.

The third paper of Patcharaporn Paokanta et. al [3]. The Knowledge and Data Engineering: Fuzzy approach and Genetic Algorithms for clustering β-Thalassemia Knowledge-Based Diagnosis Decision Support System was proposed. The aim of this study is to improve the quality of data and methodology by using K-Mean clustering, Fuzzy C-Mean and Fuzzy C-Mean-GAs. The results of this paper revealed that K-Mean cluster obtained the clustering better than Fuzzy C-Mean and Fuzzy C-Mean-GAs with RMSE 13.0077, 13.0235 and 14.3527, respectively.

Due to the previous paper of author, the performance of algorithms and methodology is improved, moreover, the quality of data is considered also. The novel ensemble method called FBR-C5.0-CART is classified after the dimension of data is increased from 10 to 17 variables and the number of classes of the output is reduced from 5 to 3 classes as the result of the third paper of Patcharaporn Paokanta et.al that Fuzzy C-Mean and Fuzzy C-Mean-GAs can not detect 2 classes of the outcome.

For this reason, in this paper Systematic Thinking, Knowledge Management Processes and DKET are applied to Web-Knowledge-Based System based on different Thalassemia data sets. Moreover, the results of using DKET: FBR-C5.0-CART for Medical Web-Knowledge-Based System are proposed in term of the performance comparison of methodology.

The organization of this paper after the introduction is the second section which DKET in Web-Knowledge-Based System are illustrated as the definition and relation between both approaches. Then the case study of DKET which is the ensemble method called FBR-CBR-C5.0-CART is presented in term of performance comparison of different data sets. In the fourth section, the results of this experiment is demonstrated. Finally, the conclusion and discussion are presented.

II. DKET AND WEB-KNOWLEDGE-BASED SYSTEM

Nowadays, Knowledge Management is required in almost organizations as the reason that those enterprises consume the important data, information, knowledge and wisdom for operating their business activities which lead to the increasing of benefit. Therefore, Knowledge Management is necessary to the recent global business competition. The appropriate KM strategy planning for the natural of each firm is the important key for the business success which cannot disregard. Among various KM strategies, software systems are needed to manage these required information and knowledge. The popular system for managing, storing and sharing these is well known as Web-Knowledge-Based System. Generally, this system are developed based on three components including rules, fact and inference engine.

In an inference engine development, the methodology is selected from DKET which is the technology for discovering problems, solutions and solving these problems by using the discovered solutions. There are several DKET algorithms which can be separated to qualitative and quantitative DKET algorithms. Moreover, the quantitative DKET algorithms can be classified to various approaches for example Artificial Neural Network, Bayesian, Evolutionary, Fuzzy, Statistics, and Case-Based Reasoning etc. These DKET algorithms are usually used for mining text in World Wide Web implementation especially search engine applications, Decision Support System, Expert System or Knowledge-Based System etc. Moreover, the hybrid approach called Ensemble methods which is the using multiple method combination to improve the better performance than using only one method.

For the rules component in Web-KBS, it is the rules which obtain from the knowledge elicitation process. These obtained rules can be generated by tacit and explicit knowledge form related documents or experts. Sometimes, before obtaining rules or knowledge, the collecting and analyzing data are required to transform or extract the essential information which these DKETs are well known as Data Mining or Knowledge Discovery in Database. There are several association rules algorithms for inducing rules such as C5.0, CART, Apiori, FP-growth and Eclat algorithms etc. The elicited rules are compared to the obtained facts which are a component of three parts of Web-KBS.

The final part of Web-KBS is the fact components. In this part, facts are defined and collected from documents and Experts as same as the rules component, the obtained facts can be tacit and explicit knowledge. The different between two components is their functions that are facts play an important role as the input. On the other hand, rules act as the standard outputs for comparing with facts. In the next section, the novel DKET called FBR-CBR-C5.0-CART Web-KBS and its results will be presented as the result comparison study.

III. Methodology of FBR-CBR-C5.0-CART Web-KBS Using Systematic Thinking and DKET

According to the review of the components of Web-KBS, in this section the novel methodology of FBR-CBR-C5.0-CART Web-KBS is revealed in Fig. 1.

The procedure of FBR-CBR-C5.0-CART Web-KBS starts at the first step which is the design process of Knowledge Management methodology. In this step, KM process which is the iteration process will be defined including, Knowledge Identification, Knowledge Acquisition, Knowledge Storage and Retrieval, Knowledge Creation, Knowledge Codification and Refinement, Knowledge Transfer and Utilization, Knowledge Sharing and Knowledge Retention

Knowledge Management Process is related to data, information, knowledge, wisdom and intelligence. In KM perspective, knowledge can be categorized as two main types including, tacit and explicit knowledge which the first one means the knowledge form experts and focuses on 2P (Process and People) on the other hand explicit knowledge is the knowledge from documents and focuses on 2T (Tool and Technology).

In the next step, tacit and explicit knowledge will be captured from related documents and experts through using Systematic Thinking (ST) which is one approach of Organization Learning. Generally, ST has three components including input, process and output.

Then these captured data, information and knowledge will be defined and collected using Systematic Thinking. Afterward, the obtained data will be cleaned before the rules induction process. The cleaned data will be extracted to the form of rules. Moreover, these data will be clustered to improve the quality of data. In the six step, The obtained rules by using C5.0, CART and CBR will be implemented and the obtained results of this step will be combined with the clustered results of using Fuzzy C-Means-GAs. The combination results of Fuzzy C-Means-GAs-CBR-C5.0-CART will be clustered using K-Means clustering, Fuzzy C-Means, and Fuzzy C- Means-GAs and the obtained results will be compared with the previous results in database. Finally, the best result will be stored in the database. In the next section, the results of FBR-CBR-C5.0-CART Web-KBS will be proposed by using Thalassemia data set to verify the methodology performance.

IV. Results of FBR-CBR-C5.0-CART WEB-KBS

The obtained results of using FBR-CBR-C5.0-CART Web-KBS show in Fig. 2. below.

In Fig. 2., Table 1. presents the clustering performance of using Ensemble method: Fuzzy C-Mean-CBR-C5.0-CART and Fuzzy C-Mean-GAs-CBR-C5.0-CART in different Thalassemia data sets. 60 records of Thalassemia indicators were collected from Out Patients Department cards (OPD) of hospital in Northern Thailand which includes variables obtained from Laboratory and Symptoms. The used data sets are F-cell, HbA_2, Inclusion Body and Hb typing results of children, father and mother, moreover, the symptom indicators were collected and transform trough using the proposed methodology shown in Figure 1. The total used variables are 22 indicators which separated to deferent data sets for testing the clustering performance of Fuzzy C-Means- CBR-C5.0-CART and Fuzzy C-Means-GAs-CBR-C5.0-CART. The best obtained results of using Fuzzy C-Means-CBR-C5.0-CART and Fuzzy C-Means-GAs-CBR-C5.0-CART in the first table are RMSE 5.10 in the case that unrecorded variables with no symptoms and no CBR-C5.0 and no CART. On the other hand, for unrecorded variables with no symptoms and no CBR-C5.0 and no CART obtains RMSE 12.03 in both algorithms.

The second table presents the satisfied result comes from the using Fuzzy C-Means-GAs-CBR-C5.0-CART with RMSE 5.10 and 17.91 for Fuzzy C-Means-CBR-C5.0-CART in the case that unrecorded variables and symptom indicators.

Besides in the third table, recorded variables and symptom indicators gives the satisfied result which is RMSE 12.03 in both methodology by using unrecorded variable with no CBR-C5.0-CART.

Moreover, the fourth table reveals the best result is Fuzzy C-Means-GAs-CBR-C5.0-CART with RMSE 5.10 in the case that full data sets (unrecorded variables, symptom indicators and CBR-C5.0-CART). On the other hand, Fuzzy C-Means- CBR-C5.0-CART obtains RMSE 17.91.

Finally, the fifth table reveals that the satisfied result is Fuzzy C-Means-GAs-CBR-C5.0-CART with RMSE 5.10 in the case that full data sets (unrecorded and recoded variables, symptom indicators and CBR-C5.0-CART).

In the next section, the result summarization will be discussed and compared to the other obtained results of using previous methodologies.

V. Conclusion

According to the obtained results of using Ensemble method called FBR-GAs-CBR-C5.0-CART, the clustering performance of the obtained results of Fuzzy C-Means-GAs- CBR-C5.0-CART is better than Fuzzy C-Means- CBR-C5.0-CART with RMSE 5.10 in the case that unrecorded variables with no symptoms and no

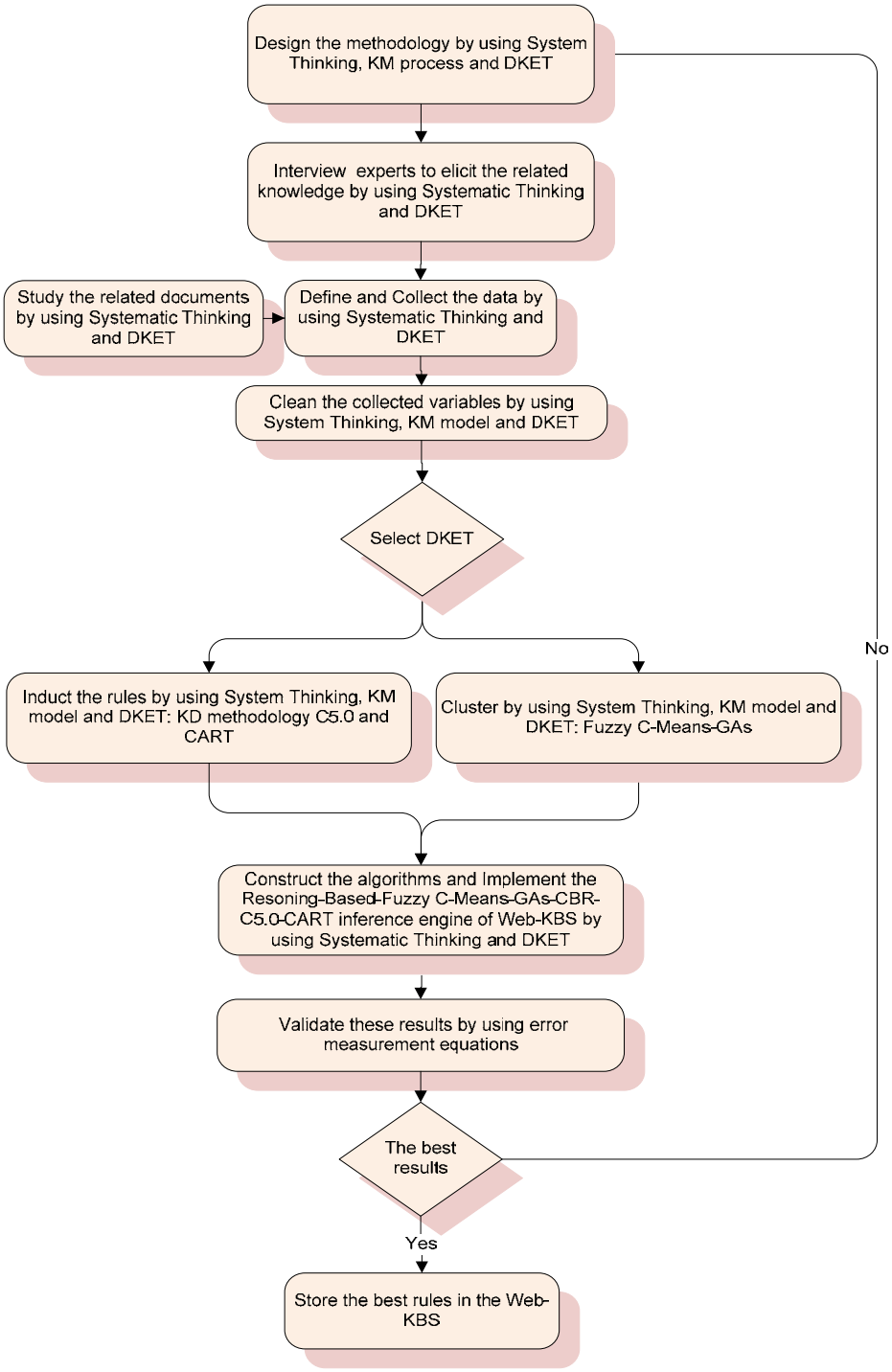

Fig. 1. Methodology of FBR-CBR-C5.0-CART Web-KBS using System

Algorithms	60 recodes 3 classes									
	Unrecoded no symptoms and no CBR-C5.0-CART					Recoded no symptoms and no CBR-C5.0-CART				
Classes	C1	C2	C3	Total	RMSE	C1	C2	C3	Total	RMSE
Fuzzy C-Means	49	3	10	62	5.10	39	10	13	62	12.03
Fuzzy C-Means-GAs	49	3	10	62	5.10	39	10	13	62	12.03
Expert	56	1	5	62	-	56	1	5	62	-

Algorithms	60 recodes 3 classes									
	Unrecoded and symptoms					Recoded and symptoms				
Classes	C1	C2	C3	Total	RMSE	C1	C2	C3	Total	RMSE
Fuzzy C-Means	31	10	21	62	17.91	26	15	21	62	21.23
Fuzzy C-Means-GAs	49	3	10	62	5.10	26	15	21	62	21.23
Expert	56	1	5	62	-	56	1	5	62	-

Algorithms	60 recodes 3 classes									
	Unrecoded and CBR-C5.0-CART					Recoded and CBR-C5.0-CART				
Classes	C1	C2	C3	Total	RMSE	C1	C2	C3	Total	RMSE
Fuzzy C-Means	39	10	13	62	12.03	30	11	21	62	18.55
Fuzzy C-Means-GAs	39	10	13	62	12.03	39	10	13	62	12.03
Expert	56	1	5	62		56	1	5	62	-

Algorithms	60 recodes 3 classes									
	Full unrecoded data sets					Full recoded data sets				
Classes	C1	C2	C3	Total	RMSE	C1	C2	C3	Total	RMSE
Fuzzy C-Means	31	10	21	62	17.91	25	16	21	62	21.92
Fuzzy C-Means-GAs	49	3	10	62	5.10	25	16	21	62	21.92
Expert	56	1	5	62	-	56	1	5	62	-

Algorithms	60 recodes 3 classes				
	Full data set				
Classes	C1	C2	C3	Total	RMSE
Fuzzy C-Means	31	10	21	62	17.91
Fuzzy C-Means-GAs	49	3	10	62	5.10
Expert	56	1	5	62	-

Fig. 2. Clustering performance Of FBR-CBR-C5.0-CART

CBR, C5.0 and no CART, unrecorded variables and symptom indicators, full data sets (unrecorded variables, symptom indicators and CBR-C5.0-CART). The result comparison of these methodologies and the previous proposed methodologies [4, 5, 6, 7] reveals that RMSE of Fuzzy C-Mean-GAs is 14.3527 which this result are improved by reducing the number of classes and records as the proposed results in this paper. In the future, graphical model and the other DKET will be used to discover the new knowledge and methodology.

REFERENCES

[1] P. Paokanta, M. Ceccarelli and S. Srichairatanakool, "The Effeciency of Data Types for Classification Performance of Machine Learning Techniques for Screening β-Thalassemia," in *Proc. ISABEL* 2010, pp. 1-4.

[2] P. Paokanta, M. Ceccarelli, N. Harnpornchai et al., "Rule Induction for Screening Thalassemia Using Machine Learning Techniques: C5.0 and CART," *ICIC Express Letter: An International Journal of Research and Surveys*, vol. 6, no. 2, pp. 301-306, Feb. 2012.

[3] P. Paokanta, N. Harnpornchai, N. Chakpitak et al., "Knowledge and Data Engineering: Fuzzy Approach and Genetic Algorithms for Clustering β-Thalassemia of Knowledge Based Diagnosis Decision Support System," *ICIC Express Letter: An International Journal of Research and Surveys*, vol. 7, no. 2, pp. 479-484, Feb. 2013.

[4] P. Paokanta, N. Harnpornchai, N. Chakpitak et al., "Parameter Estimation of Binomial Logistic Regression Based on Classical (Maximum Likelihood) and Bayesian (MCMC) Approach for Screening β-Thalassemia," *International Journal of Intelligent Information Processing*, vol.3, vol. 1, pp. 90-100, Mar. 2012.

[5] P. Paokanta, N. Harnpornchai, S. Srichairatanakool et al., "The Knowledge Discovery of β-Thalassemia Using Principal Components Analysis: PCA and Machine Learning Techniques," *International Journal of e-Education, e-Business, e-Management and e-Learning*, vol. 1, no. 2, pp. 175-180, Jun. 2011.

[6] P. Paokanta and N. Harnpornchai, "Risk Analysis of Thalassemia Using Knowledge Representation Model: Diagnostic Bayesian Networks" in *Proc. IEEE-EMBS BHI* 2012, pp. 61-61.

[7] P. Paokanta, "DBNs-BLR (MCMC) -GAs-KNN: A Novel Framework of Hybrid System for Thalassemia Expert System," *Lecture Notes in Computer Science*, vol. 7666, 2012, pp. 264–271.

AUTHOR PROFILE

Patcharaporn Paokanta has been a lecturer in the areas of Data Management, E-Commerce, Rapid Application and Development, System Analysis and Design, and Information Technology at the College of Arts, Media and Technology, Chiang Mai University (CMU), Thailand. She is studying for a Ph.D. in Knowledge Management and obtained her M.S. in Software Engineering in 2009from the College of Arts, Media and Technology, CMU, Thailand. In addition, she obtained a B.S. in Statistics from the Faculty of Science, Chiang Mai University, Thailand, in 2006. She was awarded an ERASMUS MUNDUS scholarship (E-Link Project) to study and performed research at the University of Sannio in Italy for 10 months. Her research interests include Data and Knowledge Engineering, Knowledge Discovery techniques, Statistics, Biomedical Engineering, Knowledge and Risk management, Artificial and Computing Intelligence, applied mathematics, Ramsey Number and Graph theory. Patcharaporn Paokanta has published articles in international journal and conference proceedings, including ICIC Express Letter: An International Journal of Research and Surveys, International Journal of Computer Theory and Engineering, International Journal of Intelligent Information Processing (IJIIP), Lecture Notes in Computer Sciences (LNCS), ISABEL 2010 and BHI 2012.