SaKEM: A Semi-automatic Knowledge Engineering Methodology for Building Rule-based Knowledgebase

Maqbool Ali, Maqbool Hussain, Sungyoung Lee*
Department of Computer Science and Engineering
Kyung Hee University
Yongin, South Korea
{maqbool.ali, maqbool.hussain, sylee}@oslab.khu.ac.kr

Byeong Ho Kang
School of Engineering and ICT
University of Tasmania
Hobart, Australia
byeong.kang@utas.edu.au

***Abstract*—Knowledge engineering is one of the key research area to build knowledgebase for providing solutions to real-world problems. Due to rapidly increase of data growth rate, it is almost impossible to extract hidden knowledge with manual approach. Moreover, a number of methodologies have been proposed that focus on some specific aspect of data mining process rather than end-to-end knowledge engineering methodology. Keeping in view these facts, a *Semi-automatic Knowledge Engineering Methodology* (SaKEM) is proposed that covers all major stages that are involved in *Knowledge Discovery in Databases* (KDD) process. For realization of SaKEM, a toolset called *Data Driven Knowledge Acquisition Tool* (DDKAT) is developed. The proposed methodology is designed for *Mining Minds* project but it can be utilized by other service-enabled platforms as well.**

Keywords—features selection; data preprocessing; decision trees; model translation; production rules; knowledge acquisition

I. INTRODUCTION

Knowledge is a key to compete and success in every field of life, which can be discovered from heterogeneous sources by various knowledge discovery approaches. In current arena, almost every organization need decision support system that use discovery knowledge techniques for making better decisions. Normally knowledge is acquired either by *expert-driven*, where the expert heuristics are used or *data-driven*, where state-of-the-art data mining methods are applied. Due to rapidly increase of data growth rate, it is almost impossible to extract hidden knowledge with manual approach [1]. This paper proposed a *Semi-automatic Knowledge Engineering Methodology* (SaKEM) for building a knowledgebase using data-driven as well as expert-driven approaches. The proposed methodology consists of five important phases namely *data selection*, *data preprocessing*, *model learning*, *model translation*, and *rule conformance* that are realized through *Data Driven Knowledge Acquisition Tool* (DDKAT).

The SaKEM is designed for *Knowledge Curation Layer* of *Mining Minds* (MM) project. The MM is a platform which will take benefit from the technology of big data with respect to variety as well as volume, mapping of life events through sensory environment and reasoning and prediction to process the real-time data for providing personalized services [2]. The multilayer architecture of MM framework is depicted in Fig. 1, where SaKEM is elaborated. Each layer of MM framework has specific tasks such as the *Data Curation Layer* is responsible for obtaining data from *Multimodal Data Sources*, *Information Curation Layer* for describing the user context as well as behavior, *Knowledge Curation Layer* for developing the health and wellness knowledge, *Service Curation Layer* for creating health and wellness support services, and *Supporting Layer* for providing security as well as data visualization facility [3].

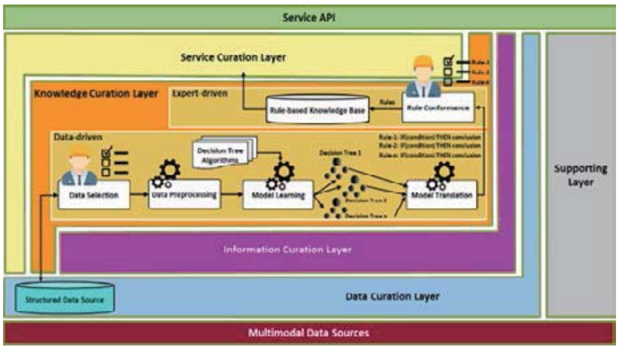

Fig. 1. SaKEM and MM framework layed architecture.

II. METHOD

A. Data Selection

The first step in any decision support system is to understand the application domain and then to identify application goal, objectives, causative factors, and their associations. All these factors help data understanding phase. Based on expertise, domain expert selects suitable parameters from available features list. In addition to that, feature selection measures also assist the domain expert in selecting informative features for decision making [4].

B. Data Preprocessing

Once a selected feature-set is obtained from a dataset, then data pre-processing is required that can play major role in improving the quality of data. This phase performs the basic data preprocessing tasks such as (1) identification and replacing of missing values with attribute's *mean/mode* mechanism, (2) detecting and replacing outlier values with *Interquartile* technique and attribute's *mean* mechanism respectively, and (3) data discretization with *Equal-Width Binning* and *Equal-Frequency Binning* methods.

C. Model Learning

The aims of this phase is to learn and build the classification model, called decision trees (DTs). In order to understand the knowledge structures, most of the health-care experts are interested in knowledge visualization or its representation [5]. The DTs are one of the popular data classification methods that represents the inductive knowledge. In this phase, DTs are built using *BFTree*, *J48*, *J48graft*, *RandomTree*, *REPTree*, and *SimpleCart* classification algorithms.

D. Model Translation

The decision trees are non-executable format and there is need to extract rules and convert them into executable format i.e. production rules. This is not an easy task and very limited information is found to translate the decision tree into production rule set [6]. To achieve this goal, this phase performs *model trimming*, *XML conversion*, *XML parsing*, and *rules conversion* tasks.

E. Rule Conformance

The aims of this phase is to build the trusted knowledgebase. To achieve this goal, all production rules are shown to domain expert through expert-driven interface, where expert conforms them one-by-one and stores them into knowledgebase.

III. Results: Case Study for Users Profile and Life-Log Mining Minds Dataset

We have applied the proposed methodology in digital health and wellness paradigms that monitors the users' daily life activities to provide healthy habits recommendations [2]. *Mining Minds Users Profile and Life-log Dataset* is selected to realize the SaKEM, where partial outcome of this methodology is illustrated in Table I.

TABLE I. Partial Outcome of SaKEM

Rule #	Production Rules
1	IF (SituationCategory ≠ LyingDown OR Sitting) THEN Recommendation = Sitting
2	IF (SituationCategory = LyingDown OR Sitting AND RiskFactor ≠ Normal) THEN Recommendation = Stretching
3	IF (SituationCategory = LyingDown OR Sitting AND RiskFactor = Normal AND Age >= 32) THEN Recommendation = Walking

IV. Conclusion

This paper demonstrated the end-to-end knowledge engineering methodology for building a knowledgebase using data-driven as well as expert-driven approaches. The methodology is demonstrated with users' profile and life-log dataset and outcome of SaKEM is used in real project of *Mining Minds*.

Acknowledgment

This work was supported by the Industrial Core Technology Development Program (10049079, Develop of mining core technology exploiting personal big data) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea).

References

[1] M. Zorrilla, D. Garcia-Saiz, "A service oriented architecture to provide data mining services for non-expert data miners", Decision Support Systems, vol. 55, no. 1, 2013, pp. 399–411.

[2] O. Banos, M.B. Amin, W.A. Khan, M. Afzel, M. Ahmad, M. Ali, T. Ali, R. Ali, et al., An innovative platform for person-centric health and wellness support, in: International Conferenceon Bioinformatics and Biomedical Engineering, Springer, 2015, pp. 131–140.

[3] O. Banos, M.B. Amin, W.A. Khan, M. Afzal, M. Hussain, B.H. Kang, S. Lee, The mining minds digital health and wellness framework.

[4] M. Dash, H. Liu, "Feature selection for classification", Intelligent data analysis, vol. 1, no. 3, 1997, pp. 131–156.

[5] M. Humphrey, S.J. Cunningham, I.H. Witten, "Knowledge visualization techniques for machine learning", Intelligent Data Analysis, vol. 2, no. 4, 1998, pp. 333–347.

[6] I.H. Witten, E. Frank, "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, 2005.