Fashions in Data Mining and Hidden Knowledge Innovation from Clinical Database

Gunasekar Thangarasu[1], P.D.D.Dominic[2] M.C.Johnwiselin[3] and S.P. Pradeep Kumar

[1]Department of Post Graduate and Professional Studies, Limkokwing University of Creative Technology, Malaysia

[2]Department of Computer and Information Sciences, University Technology PETRONAS, Malaysia

[3]Principal, [4]Department of Computer Science and Engineering, Immanuel Arasar JJ College of Engineering, Marthandam - 629 195, Kanyakumari District, Tamil Nadu, India

E-mail : Rajni.maheshwari@gmail

(Received on 01 August 2012 and accepted on 12 October 2012)

***Abstract* – Data Mining and Hidden Knowledge Innovation (DMHKI) is one of the fast growing computer science fields. Its reputation is caused by an increased demand for tools that help with the analysis and understanding of huge amounts of data in Clinical Database. Such data are generated on a daily basis by Hospitals. This explosion came into being through the ever increasing use of computers, scanners, digital cameras, bar codes, etc. We are in a situation when rich sources of data, stored in databases, warehouses, and other data repositories, are readily available. This in turn causes big interest of medicalsocieties in the field of DMHKI. What is needed is a clear and simple methodology for extracting the knowledge that is hidden in the database. In this chapter, an integrated DMHKI process model based on the emerging technologies like XML, PMML, SOAP, UDDI, and OLE BD-DM is introduced. These technologies help designing flexible, semi-automated, and easy to use DMHKI model. They enable the building of knowledge repositories. They allow for communication between several data mining tools, databases and knowledge repositories. They also enable integration and automation of DMHKI tasks. The Journal describes a Seven-step DMHKI process model, the above mentioned technologies, and their implementation details.**

Keywords: **Data Mining, Hidden Knowledge Innovation, Clinical Database**

I. Introduction

Hidden Knowledge Innovation (HKI) is a nontrivial process of detecting valid, innovative, possibly useful, and eventually understandable patterns from large collections of Clinical database. One of the HKI steps is Data Mining (DM). DM is the step that is concerned with the actual extraction of knowledge from data, in contrast to the HKI process that is concerned with many other things like understanding and preparation of the Clinical database, verification and application of the Innovative knowledge.

The design of a framework for a Hidden knowledge Innovation process is an important issue. Several researchers described a series of steps that constitute the HKI process. They range from very simple models, incorporating few steps that usually include data collection and understanding, data mining, and implementation, to more sophisticated models like the nine-step model proposed by Fayyad. In this paper we describe the Seven-step DMHKI process model. The advantage of this model is that it is based on the Medical industry-initiated study that led to thedevelopment of a Medical industry- and tool-independent DM process model. It has been successfully applied to several medical problem domains.

II. The Seven-Steps Hidden Knowledge Innovation and Datamining Process

The goal of designing a DMHKI process model is to come up with a set of processing steps to be followed by practitioners when they execute their DMKD projects. Such process model should help to plan, work through, and reduce the cost of any given project by detailing procedures to be performed in each of the steps. The DMHKI process model should provide a complete description of all the steps from problem specification to deployment of the results.

A useful DMHKI process model must be validated in real-life Medical applications. One such initiative was taken by the CRISP-DM (Cross-Industry Standard Process for Data Mining) group. The goal of the project was to develop a DMHKI process that would help to save project costs, shorten project time, and adopt DM as a core part of the Clinical database. As a result, the Seven-step DM process was developed: Symptoms understanding, data understanding, data preparation, modeling, evaluation, and deployment. They called the entire process a data mining process which was different from the term (DM) understanding in the Asia.

A. Understanding the Problem Domain

In this step one works closely with domain experts to define the problem and define the project goals, identifies key people, and learns about current solutions to the problem. It involves learning domain-specific terminology. Adescription of the problem including its restrictions is done. The project goals then need to be translated into the DMHKI goals, and may include initial selection of potential DM tools.

B. Understanding the Data

This step includes collection of sample data, and deciding which data will be needed including its format and size. If background knowledge does exist some attributes may be ranked as more important. Next, we need to verifyusefulness of the data in respect to the DMHKI goals. Data needs to be checked for completeness, redundancy, missing values, plausibility of attribute values, etc.

C. Use Innovative and Creative Idea

The ability to spend less time gathering data and more time applying it by using a centralized tool enables patient recruitment professionals to answer stronger and more valuable questions regarding new studies to drive reliable recommendations and enrollment projections – for example:

- Do investigators feel the enrollment goals for a particular protocol are feasible?
- What is the level of concern regarding potential challenges such as age limitations and surgery or post-hospitalization?
- Which specific aspects of safety would be problematic?

D. Preparation of the Data

This is the key step upon which the success of the entire Hidden knowledge Innovation process depends; it usually consumes about half of the entire project effort. In this step, we decide which data will be used as input for data mining tools of step 4. It may involve sampling of data, running association and implication tests, data cleaning like checking completeness of data records, removing or correcting for noise, etc. The cleaned data can be further processed by feature selection and extraction algorithms by derivation of new attributes, and by summarization of data. The result would be new data records, meeting specific input requirements for the planned to be used DM tools.

E. Data Mining

This is another key step in the Hidden knowledge Innovation process. Although it is the data mining tools that find out new information, their application usually takes less time than data preparation. This step involves usage of the planned data mining tools and selection of the new ones. Data mining tools include many types of algorithms, such as rough and fuzzy sets, Bayesian methods,evolutionary computing, machine learning, neural networks, clustering, preprocessing techniques, etc.

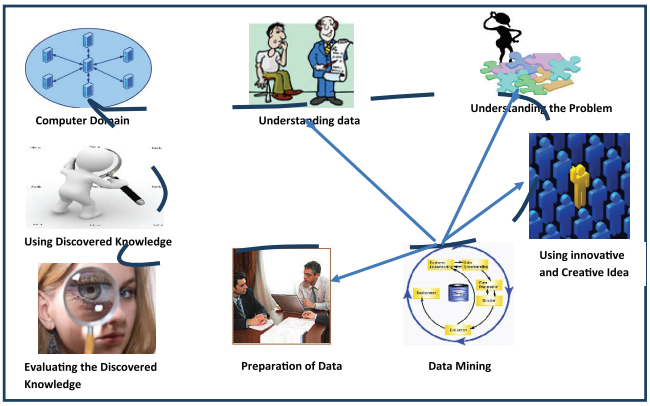

Fig.1 Seven-step of DMHKI Model

This step involves the use of several DM tools on data prepared in step 4. First, the training and testing procedures are designed and the data model is constructed using one of the chosen DM tools; the generated data model is verified by using testing procedures. One of the major difficulties in this step is that many off-the-shelf tools may not be available to the user, or that the commonly used tools may not scale up tohuge volume of data. The latter is a very important issue. Scalable DM tools are characterized by linear increase of their runtime with the increase of the number of data points within a fixed amount of available memory. Most of the DM tools are not scalable but there are examples of tools that scale well with the size of the input data; examples include clustering, machine learning, and association rules. Most recent approach for dealing with scalability of DM tools is connected with the meta-mining frame work. The meta-mining generates meta-knowledge from knowledge generated by data mining tools. It is done by dividing data into subsets, generating data models for these subsets, and generation of meta-knowledge from these data models. In this approach small data models are processed as input data instead of huge amounts of the original data, which greatly reduces computational overhead.

F. Evaluation of the Discovered Knowledge

This step includes understanding the results, checking whether the new information is unique and interesting, interpretation of the results by domain experts, and checking the impact of the Innovative knowledge. Only the approved models retained. The entire DMHKI process may be revisited to identify which alternative actions could have been taken to improve the results. A list of errors made in the process is prepared.

G. Using the Discovered Knowledge

This step is entirely in the hands of the owner of the database. It consists of planning where and how the Innovative knowledge will be used. The application area in the current domain should be extended to other domains. A plan to monitor the implementation of the innovative knowledge should be created, and the entire project documented.

The important issues are the iterative and interactive aspects of the process. Since any changes and decisions made in one of the steps can result in changes in later steps,the feed back loops are necessary. The model identifies several such feed back mechanisms:

i. From Step 2 to Step 1 because additional domain knowledge may be needed to better understand the data.

ii. From Step 4 to Step 2 because additional or more specific information about the data may be needed before choosing specific data preprocessing algorithms (for instance data transformation or discretization)

iii. From Step 4 to Step 1 when the selected DM tools do not generate satisfactory results, and thus the project goals must be modified.

iv. From Step 4 to Step 2 in a situation when data was misinterpreted causing the failure of a DM tool (e.g. data was misrecognized as continuous and discretized in Step 4. The most common scenario is when it is unclear which DM tool should be used because of poor understanding of the data.

v. From Step 4 to Step 4 to improve data preparation because of the specific requirements of the used DM tool, which may have not been not known during the data preparation step.

vi. From Step 5 to Step 1 when the discovered knowledge is not valid. There are several possible sources of such a situation: incorrect understanding or interpretation of the domain, incorrect design or understanding of problem restrictions, requirements, or goals. In these cases the entire DMHKI process needs to be repeated.

vii. From Step 5 to Step 4 when the discovered knowledge is not novel/interesting/useful. In this case, we may choose different DM tools and repeat Step 4 to extract new and potentially novel, interesting, and thus useful knowledge.

Table I Steps in Dmhki Process

7 STEP DMHKI PROCESS	9 STEP DMHKI PROCESS	5 STEP DMHKI PROCESS
1. Understanding the domain	1. Understanding application domain, identifying the DMKD goals	1. Business objective determination
2. Understanding the data	2. Creating target data set	2. Data preparation
3. Use innovative and Creative Idea	3. Data cleaning and preprocessing	
	4. Data reduction and projection	
4. Preparation of the data	5. Matching goal to particular data mining method	
	6. Exploratory analysis, model and hypothesis selection	
5. Data mining	7. Data mining	3. Data mining
6. Evaluation of the discovered knowledge	8. Interpreting mined patterns	4. Analysis of results
7. Using the discovered knowledge	9. Consolidating discovered knowledge	5. Knowledge assimilation

The important characteristic of the DMHKI process is the relative time spent to complete each of the steps, estimates that about 20% of the effort is spent on business objective determination, about 60% on data preparation and about 10% for data mining and analysis of knowledge and knowledge assimilation steps, respectively. On the other hand, show that about 15-25% of the project time isspent on the DM step. Usually it is assumed that about 50% of the project effort is spent on data preparation. There are several reasons why this step requires so much time: data collected by enterprise companies consist of about 1-5% errors, often the data are redundant and inconsistent, also companies may not collect all the necessary data. These serious data quality problems contribute to the extensive data preprocessing step. To accommodate for the above, we propose to use time ranges rather than fixed times for the steps.

III. Conclusion

At present the DMHKI industry is fragmented. It consists of research groups and field experts which do not work closely with decision makers. This is caused by the situation where the DMHKI community generates new solutions that are not widely accessible to a broader audience; the major obstacle being that they are very complex to use. Because of the complexity and high cost of the DMHKI process, the DMHKI projects are deployed in situations where there is an urgent need for them, while many other businesses reject it because of the high costs involved. To come up with the solution to this problem may require consolidation of the DMHKI community by providing integrated DM tools and services, and making the DMHKI process easier to implement by the end-users by semi-automating it.

References

[1] S. Banerjee, V. Krishnamurthy, M. Krishnaprasad and R. Murthy, "Oracle8i – The XML Enabled Data Management System", *Proceedings of the Sixteenth International Conference on Data Engineering,* San Diego, California, 2007, pp. 561-568.

[2] R. Bourret, C. Bornhvd, and A.P. Buchmann, "A Generic Load/ Extract Utility for Data Transfer Between XML Documents and Relational Databases", *Proceeding of the 2nd International Workshop on Advanced Issues of E-Commerce and Web-based Information Systems,* San Jose, California, June, 2008, pp.134-143.

[3] L. Kurgan, K.J. Cios, M. Sontag, and F.J. Accurso, "Mining a Cystic Fibrosis Database", In: Zurada, J., and Kantardzic, M. (Eds.), Novel Applications in Data Mining, submitted, 2006.

[4] L.Kurgan and K.J. Cios, "DataSqueezer Algorithm that Generates Small Number of Short Rules", *Submitted to the IEE Proceedings: Vision, Image and Signal Processing,* 2008.

[5] K.K. Hirji, "Exploring Data Mining Implementation", *Communications of the ACM,* Vol. 44, No. 7, July 2008, pp. 87-93.

[6] DMG, The Data Mining Group, http://www.dmg.org/, 2009

[7] T. Bray, J. Paoli and E. Maler, "Extensible Markup Language (XML) 1.0 (Second Edition)", W3C Recommendation, http://www.w3.org/ TR/2000/REC-xml-20001006, October 2008.

[8] P. Buneman, M.F. Fernandez, D.Suciu, "UnQL: A Query Language and Algebra for Semistructured Data Based on Structural Recursion",*Very Large Data Bases Journal,* Vol. 9, No. 1, 2008, pp.76-110.

[9] S. Cluet and J.Simeon, YATL: A Functional and Declarative Language for XML, draft manuscript, 2007.