Chapter 4
Knowledge Discovery Process Models:
From Traditional to Agile Modeling

Mouhib Alnoukari
Arab International University, Syria

Asim El Sheikh
Arab Academy for Banking and Financial Sciences, Jordan

ABSTRACT

Knowledge Discovery (KD) process model was first discussed in 1989. Different models were suggested starting with Fayyad's et al (1996) process model. The common factor of all data-driven discovery process is that knowledge is the final outcome of this process. In this chapter, the authors will analyze most of the KD process models suggested in the literature. The chapter will have a detailed discussion on the KD process models that have innovative life cycle steps. It will propose a categorization of the existing KD models. The chapter deeply analyzes the strengths and weaknesses of the leading KD process models, with the supported commercial systems and reported applications, and their matrix characteristics.

INTRODUCTION

The term 'Knowledge Discovery' (KD) or Knowledge Discovery in Data (KDD) was first coined in 1989. Fayyad defined knowledge discovery as it concerns with "the entire knowledge extraction process, including how data are stored and accessed, how to use efficient and scalable algorithms to analyze massive datasets, how to interpret and visualize the results, and how to model and support the interaction between human and machine. It also concerns support for learning and analyzing the application domain" (Fayyad et al. 1996).

This means that data mining is simply one of the KD process's steps. Piatetsky-Shapiro explained the difference between knowledge discovery and data mining: "…data mining was used more by database and business folks. The term

DOI: 10.4018/978-1-61350-050-7.ch004

"knowledge discovery" (which I coined in 1989) was more popular among researchers in Artificial Intelligence. Both terms are used to describe the process of searching for useful knowledge in data, but [the term] data mining is much more popular" (Piatetsky-Shapiro 2000).

According to Fayyad et al. (1996), KDP is "the process of using the database along with any required selection, preprocessing, subsampling, and transformations of it; to apply data mining methods (algorithms) to enumerate patterns from it; and to evaluate the products of data mining to identify the subset of the enumerated patterns deemed knowledge".

The various models discussed in this paper are related to data mining and knowledge discovery. They vary in the number, iterations, activities, and structures of their stages. The paper includes analysis of the strengths and weaknesses of each of these methodologies. This paper's survey is different from two older surveys done by Kurgan and Musilek (2006) and Hofmann (2003) in the way it considers the leading KD process models. Our paper completes these two surveys with many of new KD process models presenting the evolutions of these models, and provides a characteristics matrix that summarizes the main differences among the considered models.

KNOWLEDGE DISCOVERY PROCESS MODELING CATEGORIZATION

The following are the proposed categories for Knowledge Discovery Process (KDP) modeling:

1. *Traditional KDP Approach*. This approach is widely used by most of KDP modeling innovators. Starting with Fayyad's et al. (1996) KDD process modeling, many of KDP modeling used the same process flow including most of the following steps: business understanding, data understanding, data processing, data mining/modeling, model evaluation, and deployment/visualization.
2. *Ontology-based KDP Approach*. This approach is the integration of ontology engineering and traditional KDP approach steps. Three directions were identified in this approach: Ontology for KDP, KDP for Ontology, and the integration of both previous directions (Gottgtroy 2007).
3. *Web-based KDP Approach.* This approach mainly deals with web log analysis. It is mainly similar to traditional KDP approach, but it has some unique steps to deal with log web data, see (Pabarskaite and Raudys 2007) and (Buchner et al. 1999).
4. *Agile-based KDP Approach.* This approach is the integration between agile methodologies and KDP traditional methodologies (Alnoukari et al. 2008).

THE LEADING KDP MODELS

The following leading KDP models have been chosen by the authors based on their innovation steps, and their applications in both academia and industry:

1. Knowledge Discovery in Databases (KDD) Process by Fayyad et al. (1996).
2. Information Flow in a Data Mining Life Cycle by Ganesh et al. (1996).
3. SEMMA by SAS Institute (1997).
4. Refined KDD paradigm by Collier et al. (1998).
5. Knowledge Discovery Life Cycle (KDLC) Model by Lee and Kerschberg (1998).
6. CRoss-Industry-Standard Process for Data Mining (CRISP-DM) by CRISP-DM (2000).
7. Generic Data Mining Life Cycle by (DMLC) by Hofmann (2003).
8. Ontology Driven Knowledge Discovery Process (ODKD) by Gottgtroy (2007).