1

A Data Mining & Knowledge Discovery Process Model

Óscar Marbán[1], Gonzalo Mariscal[2] and Javier Segovia[1]
[1] *Facultad de Informática (Universidad Politécnica de Madrid)*
[2] *Universidad Europea de Madrid*
Spain

1. Introduction

The number of applied in the data mining and knowledge discovery (DM & KD) projects has increased enormously over the past few years (Jaffarian et al., 2008) (Kdnuggets.com, 2007c). As DM & KD development projects became more complex, a number of problems emerged: continuous project planning delays, low productivity and failure to meet user expectations. Neither all the project results are useful (Kdnuggets.com, 2008) (Eisenfeld et al., 2003a) (Eisenfeld et al., 2003b) (Zornes, 2003), nor do all projects end successfully (McMurchy, 2008) (Kdnuggets.com, 2008) (Strand, 2000) (Edelstein & Edelstein, 1997). Today's failure rate is over 50% (Kdnuggets.com, 2008) (Gartner, 2005) (Gondar, 2005).

This situation is in a sense comparable to the circumstances surrounding the software industry in the late 1960s. This was what led to the 'software crisis' (Naur & Randell, 1969). Software development improved considerably as a result of the new methodologies. This solved some of its earlier problems, and little by little software development grew to be a branch of engineering. This shift has meant that project management and quality assurance problems are being solved. Additionally, it is helping to increase productivity and improve software maintenance.

The history of DM & KD is not much different. In the early 1990s, when the KDD (Knowledge Discovery in Databases) processing term was first coined (Piatetsky-Shapiro & Frawley, 1991), there was a rush to develop DM algorithms that were capable of solving all the problems of searching for knowledge in data. Apart from developing algorithms, tools were also developed to simplify the application of DM algorithms. From the viewpoint of DM & KD process models, the year 2000 marked the most important milestone: CRISP-DM (CRoss-Industry Standard Process for DM) was published (Chapman et al., 2003). CRISP-DM is the most used methodology for developing DM & KD projects. It is actually a "de facto" standard.

Looking at the KDD process and how it has progressed, we find that there is some parallelism with the advancement of software. From this viewpoint, DM project development entails defining development methodologies to be able to cope with the new project types, domains and applications that organizations have to come to terms with. Nowadays, SE (software engineering) pay special attention to organizational, management or other parallel activities not directly related to development, such as project completeness

Source: Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca,
ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

and quality assurance. The most used DM & KD process models at the moment, i.e. CRISP-DM, SEMMA, has not yet been sized for these tasks, as it is very much focused on pure development activities and tasks (Marbán et al., 2008). In (Yang & Wu, 2006) one of the 10 challenging problems to be solved in DM research is considered to be the need to build a new methodology to help users avoid many data mining mistakes.
This chapter is moved by the idea that DM & KD problems are taking on the dimensions of engineering problems. Hence, the processes to be applied should include all the activities and tasks required in an engineering process, tasks that CRISP-DM might not cover. The proposal is inspired by the work done in SE derived from other branches of engineering. It borrows ideas to establish a comprehensive process model for DM that improves and adds to CRISP-DM. Further research will be needed to define methodologies and life cycles, but the basis of a well-defined process model will be there.
In section 2 we describe existing DM & KD process models and methodologies, focusing on CRISP-DM. Then, section 3 shows the most used SE process models. In section 4, we propose a new DM & KD process model. And, finally, we discuss the conclusions about the new approach and future work in section 5.

2. DM & KD process models

Authors tend to use the terms process model, life cycle and methodology to refer to the same thing. This has led to some confusion in the field.
A process model is the set of tasks to be performed to develop a particular element, as well as the elements that are produced in each task (outputs) and the elements that are necessary to do a task (inputs) (Pressman, 2005). The goal of a process model is to make the process repeatable, manageable and measurable (to be able to get metrics).
Methodology can be defined as the instance of a process model that lists tasks, inputs and outputs and specifies how to do the tasks (Pressman, 2005). Tasks are performed using techniques that stipulate how they should be done. After selecting a technique to do the specified tasks, tools can be used to improve task performance.
Finally, the life cycle determines the order in which each activity is to be done (Moore, 1998). A life cycle model is the description of the different ways of developing a project.
From the viewpoint of the above definitions, what do we have in the DM & KD area? Does DM & KD have process models and/or methodologies?

2.1 Review of DM & KD process models and methodologies

In the early 1990s, when the KDD process term was first coined (Piatetsky-Shapiro & Frawley, 1991), there was a rush to develop DM algorithms that were capable of solving all problems of searching for knowledge in data. The KDD process (Piatetsky-Shapiro, 1994) (Fayyad et al., 1996) has a process model component because it establishes all the steps to be taken to develop a DM project, but it is not a methodology because its definition does not set out how to do each of the proposed tasks. It is also a life cycle.
The 5 A's (Martínez de Pisón, 2003) is a process model that proposes the tasks that should be performed to develop a DM project and was one of CRISP-DM's forerunners. Therefore, they share the same philosophy: 5 A's proposes the tasks but does not suggest how they should be performed. Its life cycle is similar to the one proposed in CRISP-DM.
A people-focused DM proposal is presented in (Brachman & Anand, 1996): Human-Centered Approach to Data Mining. This proposal describes the processes to be enacted to

carry out a DM project, considering people's involvement in each process and taking into account that the target user is the data engineer.
SEMMA (SAS, 2008) is the methodology that SAS proposed for developing DM products. Although it is a methodology, it is based on the technical part of the project only. Like the above approaches, SEMMA also sets out a waterfall life cycle, as the project is developed right through to the end.
The two models by (Cabena et al., 1998) and (Anand & Buchner, 1998) are based on KDD with few changes and have similar features.
Like the KDD process, Two Crows (Two Crows, 1999) is a process model and waterfall life cycle. At no point does it set out how to do the established DM project development tasks.
CRISP-DM (Chapman et al., 2003) states which tasks have to be carried out to successfully complete a DM project. It is therefore a process model. It is also a waterfall life cycle. CRISP-DM also has a methodological component, as it gives recommendations on how to do some tasks. Even so these recommendations are confined to proposing other tasks and give no guidance about how to do them. Therefore, we class CRISP-DM as a process model.

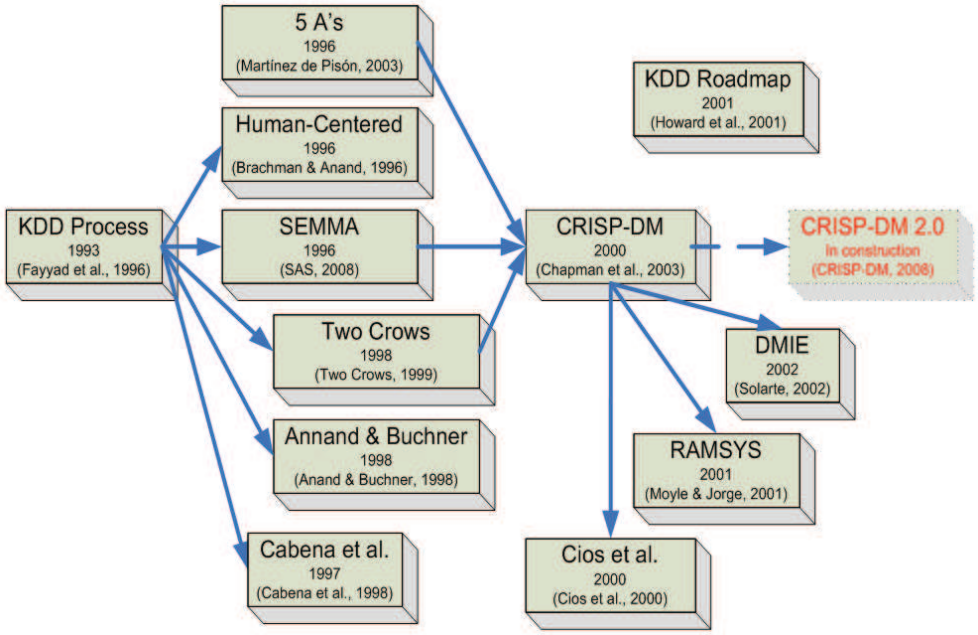

Fig. 1. Evolution of DM & KD process models and methodologies

Figure 1 shows a diagram of how the different DM & KD process models and methodologies have evolved. It is clear from Figure 1 that CRISP-DM is the standard model. It borrowed ideas from the most important pre-2000 models and is the groundwork for many later proposals.
The CRISP-DM 2.0 Special Interest Group (SIG) was set up with the aim of upgrading the CRISP-DM model to a new version better suited to the changes that have taken place in the business arena since the current version was formulated. This group is working on the new methodology (CRISP-DM, 2008). The firms that developed CRISP-DM 1.0 have been joined

by other institutions that intend to input their expertise in the field to develop CRISP-DM 2.0. Changes such as adding new phases, renaming existing phases and/or eliminating the odd phase are being considered for the new version of the methodology.

Cios et al.'s model was first proposed in 2000 (Cios et al., 2000). This model adapted the CRISP-DM model to the needs of the academic research community, providing a more general, research-oriented description of the steps.

The KDD Roadmap (Howard et al., 2001) is a DM methodology used in the DM Witness Miner tool (Lanner Group, 2008). This methodology describes the available processes and algorithms and incorporates experience derived from successfully completed commercial projects. The focus is on the decisions to be made and the options available at each stage to achieve the best results for a given task.

The RAMSYS (RApid collaborative data Mining SYStem) methodology is described in (Moyle & Jorge, 2001) as a methodology for developing DM & KD projects where several geographically diverse groups work remotely and collaboratively to solve the same problem. This methodology is based on CRISP-DM and maintains the same phases and generic tasks.

<table>
<tr><th>Model</th><th>Fayyad et al.</th><th>Cabena et al.</th><th>Anand & Buchner</th><th>CRISP-DM</th><th>Cios et al.</th></tr>
<tr><td>No of steps</td><td>9</td><td>5</td><td>8</td><td>6</td><td>6</td></tr>
<tr><td rowspan="10">Steps</td><td rowspan="2">Developing and Understanding of the Application Domain</td><td rowspan="2">Business Objectives Determination</td><td>Human Resource Identification</td><td rowspan="2">Business Understanding</td><td rowspan="2">Understanding the Data</td></tr>
<tr><td>Problem Specification</td></tr>
<tr><td>Creating a Target Data Set</td><td rowspan="5">Data Preparation</td><td>Data Prospecting</td><td rowspan="2">Data Understanding</td><td rowspan="2">Understanding the Data</td></tr>
<tr><td>Data Cleaning and Pre-processing</td><td>Domain Knowledge Elicitation</td></tr>
<tr><td>Data Reduction and Projection</td><td>Methodology Identification</td><td rowspan="3">Data Preparation</td><td rowspan="3">Preparation of the data</td></tr>
<tr><td>Choosing the DM Task</td><td rowspan="2">Data Pre-processing</td></tr>
<tr><td>Choosing the DM Algorithm</td></tr>
<tr><td>DM</td><td>DM</td><td>Pattern Discovery</td><td>Modeling</td><td>DM</td></tr>
<tr><td>Interpreting Mined Patterns</td><td>Domain Knowledge Elicitation</td><td>Knowledge Post-processing</td><td>Evaluation</td><td>Evaluation of the Discovered Knowledge</td></tr>
<tr><td>Consolidating Discovered Knowledge</td><td>Assimilation of Knowledge</td><td></td><td>Deployment</td><td>Using the Discovered Knowledge</td></tr>
</table>

Table 1. Comparison of DM & KD process models and methodologies (Kurgan & Musilek, 2006)

DMIE or Data Mining for Industrial Engineering (Solarte, 2002) is a methodology because it specifies how to do the tasks to develop a DM project in the field of industrial engineering. It is an instance of CRISP-DM, which makes it a methodology, and it shares CRISP-DM's associated life cycle.
Table 1 compares the phases into which the DM & KD process is decomposed according some of the above proposals. As Table 1 shows, most of the proposals cover all the tasks in CRISP-DM, although they do not all decompose the KD process into the same phases or attach the same importance to the same tasks.
However, some proposals described in this section and omitted the study by (Kurgan & Musilek, 2006), like 5 A's and DMIE, propose additional phases not covered by CRISP-DM that are potentially very useful in KD & DM projects. 5 A's proposes the "Automate" phase. This phase entails more than just using the model. It focuses on generating a tool to help non-experts in the area to perform DM & KD tasks. On the other hand, DMIE proposes the "On-going support" phase. It is very important to take this phase into account, as DM & KD projects require a support and maintenance phase. This maintenance ranges from creating and maintaining backups of the data used in the project to the regular reconstruction of DM models. The reason is that the behavior of the DM models may change as new data emerge, and they may not work properly. Similarly, if other tools have been used to implement the DM models, the created programs may need maintenance, e.g. to upgrade the behavior of the user application models.

2.2 CRISP-DM

We focus on CRISP-DM as a process model because it is the "de facto standard" for developing DM & KD projects. In addition, CRISP-DM is the most used methodology for developing DM projects (KdNuggets.com, 2002; KdNuggets.com, 2004; KdNuggets.com, 2007a).
Analyzing the problems of DM & KD projects, a group of prominent enterprises (Teradata, SPSS – ISL, Daimler-Chrysler and OHRA) developing DM projects, proposed a reference guide to develop DM & KD projects. This guide is called CRISP-DM (CRoss Industry Standard Process for Data Mining) (Chapman et al., 2000). CRISP-DM is vendor-independent so it can be used with any DM tool and it can be applied to solve any DM problem.
CRISP-DM defines the phases to be carried out in a DM project. CRISP-DM also defines for each phase the tasks and the deliverables for each task. CRISP-DM is divided into six phases (see Figure 2). The phases are described in the following.

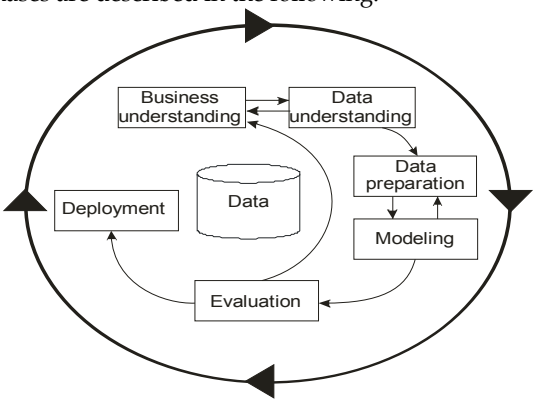

Fig. 2. CRISP-DM process model (Chapman et al., 2000)

Business understanding	Data understanding	Data preparation	Modeling	Evaluation	Deployment
Determine business objectives	Collect initial data	Select data	Select modeling techniques	Evaluate results	Plan deployment
Assess situation	Describe data	Clean data	Generate test design	Review process	Plan monitoring & maintenance
Determine DM objectives	Explore data	Construct data	Build model	Determine next steps	Produce final report
Produce project plan	Verify data quality	Integrate data	Assess model		Review project
		Format data			

Table 2. CRISP-DM phases and tasks.

- **Business understanding**: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives.
- **Data understanding**: The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.
- **Data preparation**: The data preparation phase covers all the activities required to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed repeatedly and not in any prescribed order.
- **Modeling**: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same DM problem type. Some techniques have specific requirements on the form of data. Therefore, it is often necessary to step back to the data preparation phase
- **Evaluation**: What are, from a data analysis perspective, seemingly high quality models will have been built by this stage of the project. Before proceeding to final model deployment, it is important to evaluate the model more thoroughly and review the steps taken to build it to be certain that it properly achieves the business objectives. At the end of this phase, a decision should be reached on how to use of the DM results.
- **Deployment**: Model construction is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

Table 2 outlines the phases and generic tasks that CRISP-DM proposes to develop a DM project.

3. Software engineering process models

The two most used process models in SE are the IEEE 1074 (IEEE, 1997) and ISO 12207 (ISO, 1995) standards. These models have been successfully deployed to develop software. For this reason, our work is founded on these standards. We intend to exploit the benefits of this experience for application to the field of DM & KD.

Figure 3 compares the two models. As Figure 3 shows, most of the processes proposed in IEEE 1074 are equivalent to ISO 12207 processes and vice versa. To select processes as optimally as possible, the IEEE 1074 and ISO 12207 processes should be mixed. The selection criterion was to choose the processes that either IEEE 1074 or ISO 12207 defined in more detail. We tried to not to mix processes from different groups in either process model. In compliance with the above criteria, we selected IEEE 1074 processes as the groundwork, because they are more detailed. As IEEE 1074 states that it is necessary to acquire or supply software but not how to do this, we added the ISO 12207 (ISO, 1995) acquisition and supply processes.

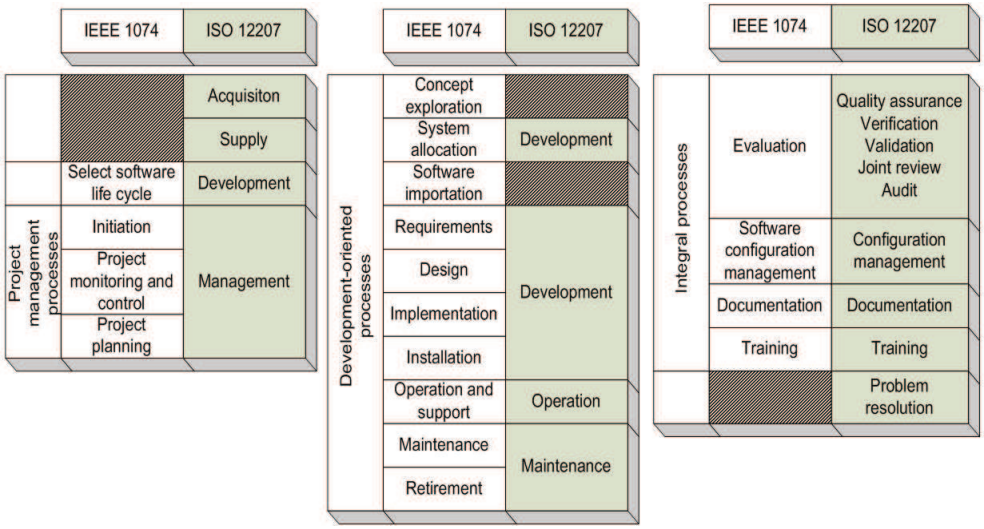

Fig. 3. Mapping ISO 12207 to IEEE 1074 (Marbán et al., 2008)

Figure 4 shows the joint process model developed after examining IEEE 1074 and ISO 12207 according to the above criteria. Figure 4 also shows the activities in each major process group according to the selected standard for the group in question.

We describe the key processes selected for this joint process model below:

- The acquisition and supply processes are taken from the Primary Life Cycle Processes set out in (ISO, 1995). These processes are part of the software development process initiation and determine the procedures and resources required to develop the project.
- The software life cycle selection process (IEEE, 1997) identifies and selects a life cycle for the software under construction.
- The project management processes (IEEE, 1997) are the set of processes that establish the project structure, and coordinate and manage project resources throughout the software life cycle.
- Development-oriented processes (IEEE, 1997) start with the identification of a need for automation. It may take a new application or a change of all or part of an existing application to satisfy this need. With the support of the integral process activities and under the project management plan, the development processes produce software (code and documentation) from the statement of the need. Finally, the activities for installing,

operating, supporting, maintaining and retiring the software product should be performed.

- Integral processes (IEEE, 1997) are necessary to successfully complete the software project activities. They are enacted at the same time as the software development-oriented activities and include activities that are not related to development. They are used to assure the completeness and quality of the project functions.

In the following section we are going to analyze which of the above activities are in CRISP-DM and which are not. The aim is to build a process model for DM projects that is as comprehensive as possible and organizes the activities systematically.

PROCESS	ACTIVITY
Acquisition	
Sypply	
Software life cycle selection	Identify available software life cycles
	Select software life cycle
Project management activities	
Initiation	Create software life cycle process
	Allocate project resources
	Perform estimations
	Define metrics
Project monitoring and control	Manage risks
	Manage the project
	Retaib records
	Identify software life cycle process improvement needs
	Collect and analyze metric data
Project planning	Plan evaluations
	Plan configuration management
	Plan system transition
	Plan installation
	Plan documentation
	Plan training
	Plan project management
	Plan integration
Deployment oriented activities	
Pre-development	
Concept exploration	Identify ideas or needs
	Formulate potencial approaches
	Conduct feasibility studies
	Refine and finalize the idea or need
System allocation	Analyze functions
	Decompose system requirements
	Develop system architecture
Software importation	Identify imported software requirements
	Evaluate software import sources
	Define software import method
	Import software
Development	
Requirements	Define and develop software requirements
	Define interface requirements
	Priorizate and integrate software requirements

PROCESS	ACTIVITY
Design	Perform architectural design
	Design data baase
	Design interface
	Perform detailed design
Implementation	Create executable code
	Create operating documentation
	Perform integration
Post-Development	
Installation	Distribuye software
	Install software
	Accept software in operational environment
Operation and support	Operate the system
	Provide technical asistanse and consulting
	Maintain support request log
Maintenance	Identify software improvement needs
	Implement problem reporting method
	Maintenance support request log
Retirement	Notify user
	Conduct parallel operations
	Retire system
Integral activities	
Evaluation	Conduct reviews
	Create traceability matrix
	Conduct audits
	Develop test procedures
	Create test data
	Execute test
	Report evaluation results
Software configuration management	Develop configuration identification
	Perform configuration control
	Perform status accounting
Documentation development	Implement documentation
	Produce and distribuye documentation
Training	Develop training materials
	Validate the training program
	Implement the training program

Fig. 4. Joint process model

4. A data mining engineering process model

A detailed comparison of CRISP-DM with the SE process model described in section 3 is presented in (Marbán et al, 2008). From this comparison, we found that many of the processes defined in SE that are very important for developing any type of DM engineering project are missing from CRISP-DM. This could be the reason why CRISP-DM is not as effective as it should be. What we proposed there was to take CRISP-DM tasks and processes and organize them by processes as SE researchers did. The activities missing from CRISP-DM are primarily project management processes, integral processes (that assure project function completeness and quality) and organizational processes (that help to achieve a more effective organization).

Note that the correspondence between CRISP-DM and SE process model elements is not exact. In some cases, the elements are equivalent, but the techniques are different. In other cases, the elements have the same goal but are implemented completely differently. This obviously depends on the project type. In SE the project aim is to develop software and in DM & KD it is to gather knowledge from data.

Figure 5 shows an overview of the proposed process model, including the key processes. The KDD process is the project development core. In the following we describe the processes shown in Figure 5. We also explain why we think they are necessary in a DM project.

4.1 Organizational processes

This set of processes helps to achieve a more effective organization. They also set the organization's business goals and improve the organization's process, product and resources. Neither the IEEE 1074 nor the ISO 12207 SE process models include these processes. They were introduced in ISO 15504 or SPICE ISO (ISO, 2004). These processes affect the entire organization, not just one project.

This group includes the following processes (see Figure 5):

- **Improvement**. This activity broadcasts the best practices, methods and tools that are available in one part of the organization to the rest of the organization.
- **Infrastructure**. This task builds the best environment in the organization for developing DM projects.
- **Training**. This activity is related to training the staff participating in current or ongoing DM projects.

No DM methodology considers any of these activities. We think that they could be adapted from the SPICE standard because they are all general-purpose tasks common to any kind of project.

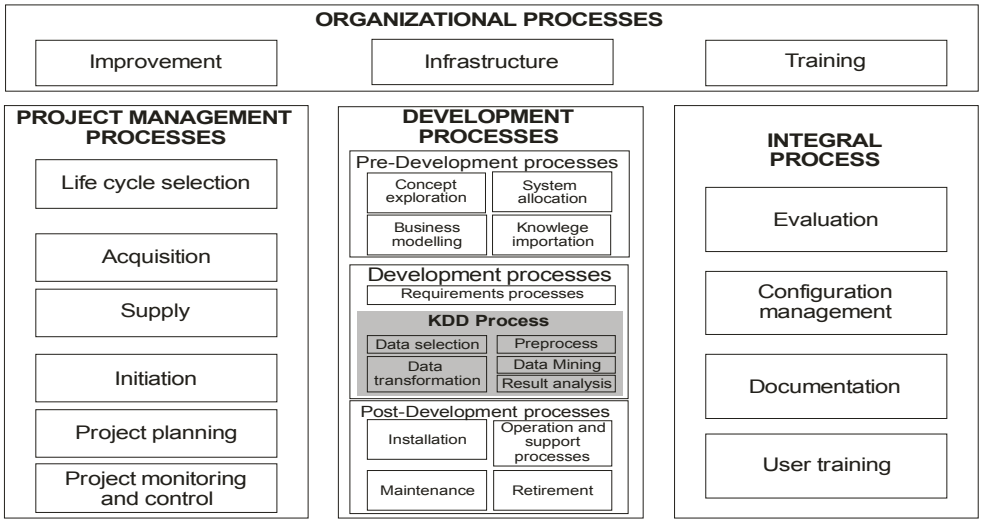

Fig. 5. DM engineering process model (Marban et al., 2008)

4.2 Project management processes

This set of processes establishes the project structure and also how to coordinate and manage project resources throughout the project life cycle. We define six main processes in the project management area. Existing DM methodologies or process models (such as CRISP-DM) take into account only a small part of project management, i.e. the project plan.

The project plan is confined to defining project deadlines and milestones. All projects need other management activities to control time, budget and resources. The project management processes are concerned with controlling these matters.

- **Life cycle selection**. This process defines the life cycle to be used in the DM project (Pressman, 2005). Until now, all DM methodologies had a similar life cycle to CRISP-DM: a waterfall life cycle with backtracking. However, there is a fair chance of new life cycles being developed to meet the needs of the different projects. This is what happens in SE from where they could be adapted.
- **Acquisition**. The acquisition process is related to the activities and tasks of the acquirer (who outsources work). Model building is one possible example of outsourcing. If there is outsourcing, the acquirer must define the acquisition management, starting from the tender and ending with the acceptance of the outsourced product. This process must be included in DM processes because DM projects now developed at non-specialized companies are often outsourced (KdNuggets.Com, 2007d). The acquisition process could be an adaptation of the process proposed in the ISO 12207 standard. It defines software development outsourcing management from requirements to software (this depends on which part is outsourced).
- **Supply**. The supply process concerns the activities and tasks that the supplier has to carry out if the company acts as the developer of an outsourcing project. This process defines the tasks the supplier has to perform to interact with the outsourcing company. It also defines the interaction management tasks. As above, this process can be adapted for DM & KD projects from ISO 12207.
- **Initiation**. The initiation process establishes project structure and how to coordinate and manage project resources throughout the project life cycle. This process could be divided into the following activities: create DM life cycle, allocate project resources, perform estimations, and define metrics. CRISP-DM's Business Understanding phase partly includes these activities, but fails to suggest how to estimate costs and benefits (Fernandez-Baizan et al., 2007). Although some DM metrics have been defined (ROI, accuracy, space/time, usefulness…) (Pai, 2004) (Shearer, 2000) (Biebler et al., 2005) (Smith & Khotanzad, 2007), they are not being used in ongoing DM developments (Marbán et al., 2008).
- **Project planning**. The project planning process covers all the tasks related to planning project management, including the contingencies plan. DM has not shown much interest in this process. DM methodologies focus primarily on technical tasks, and they overlook most of the project management activities. The set of activities considered in this process are: Plan evaluation (Biffl & Winkler, 2007), Plan configuration management (NASA, 2005), Plan system transition (JFMIP, 2001), Plan installation, Plan documentation (Hackos, 1994), Plan training (McNamara, 2007), Plan project management (Westland, 2006), and Plan integration. CRISP-DM does consider the Plan installation and Plan integration activities through its Deployment phase. Although documentation is developed in DM projects, no documentation plan is developed. CRISP-DM states that the user should be trained, but this training is not planned. The other plans are not considered in DM & KD methodologies.
- **Project monitoring and control**. The project monitoring and control process covers all tasks related to project risk and project metric management. The activities it considers within this process are: Manage risks, Manage the Project, Retain records, Identify life

cycle process improvement needs, and Collect and analyze metrics (Fenton & Pfleeger, 1998) (Shepperd, 1996). CRISP-DM considers Manage risks in the Business Understanding phase and Identify life cycle process improvement needs in the Development phase. The Manage the project activity is new to CRISP-DM, which considers project planning but not plan control. The other two activities (Retain records and Collect and analyze metrics) are new to DM & KD methodologies.

4.3 Development processes

These are the most highly developed processes in DM. All DM methodologies focus on these processes. This is due to the fact that development processes are more related to technical matters. Consequently, they were developed at the same time as the techniques were created and started to be applied. These processes are divided into three groups: pre-development, development, and post-development processes.
The development process is the original KDD process defined in (Piatetsky-Shapiro, 1994). The pre- and post-development processes are the ones that require a greater effort.

4.3.1 Pre-development processes

These processes are related to everything to be done before the project kicks off. From the process model in Figure 5, we can identify the following Pre-development processes:

- **Concept exploration**. In CRISP-DM these activities are considered in the Business Understanding phase. However, these tasks do not completely cover the activities because they focus primarily on the terminology and background of the problem to be solved. In (Marbán et al., 2008) we conclude that some tasks adapted from SE standards need to be added to optimize project development.
- **System allocation**. CRISP-DM considers these tasks through its Business Understanding phase
- **Business modeling**. This is a completely new activity. The CRISP-DM business model is described in the Business Understanding phase, but there are no business modeling procedures or formal tools and methods as there are in SE (Gordijn et al., 2000).
- **Knowledge importation**. This process is related to the reuse of existing knowledge or DM models from other or previous projects, something which is very common in DM. CRISP-DM does not consider this process at all, and its SE counterpart is related to software and cannot be easily adapted. Consequently, the process must be created from scratch.

4.3.2 Development processes

This is the most developed phase in DM methodologies, because it has been researched since late 1980s. CRISP-DM phases include all these processes in one way or another. In SE the process is divided into requirements, analysis, design, and implementation phases. We can easily map the requirements, design and implementation phases to DM projects. The design and implementation phases match the KDD process, and we will stick with this process and its name.

- **Requirements processes**. CRISP-DM covers this set of processes, but they are incomplete. The requirements are developed in CRISP-DM's Business Understanding phase. In CRISP-DM this process produces a list of requirements, but the CRISP-DM user guide does not specify or describe any procedure or any formal notation, tool or

technique to obtain the requirements from the business models. Neither does it specify or describe how to translate requirements into DM goals and models for proper use in the subsequent design and implementation phases: the KDD process. We believe that requirements can be described formally like they are in SE (Kotonya, G. & Sommerville, 1998). For example, something like use-case models could be adapted to specify the project requirements. Further work and research, possibly inspired by SE best practices, should be put into developing a core of formal methods and tools adapted to this process in the DM area.

- **KDD process**. The KDD process matches the design and implementation phases of a software development project. This set of processes is responsible for acquiring the knowledge for the DM project. KDD includes the following activities: data selection, pre-processing, data transformation, DM, and result analysis. CRISP-DM covers the KDD process (Marbán et al., 2008).

4.3.3 Post-development processes

Post-development processes are the processes that are carried out after the knowledge is gathered. They are applicable during the later life cycle stages.

- **Installation**. This process is commended with transferring the knowledge extracted from the DM results to the users. The knowledge can be used as it is, i.e. to help managers to make decisions about a future marketing campaign, or could involve some software development, i.e. to improve an existing web-based recommender system. CRISP-DM considers planning for deploying the knowledge at the client site, but it does not regard the development of software installed and accepted in an operational environment as part of this deployment (Reifer, 2006).
- **Operation and support process**. This process is necessary to validate the results and how they are interpreted by the client, and, if software is developed, to provide the client with technical assistance. CRISP-DM only includes results monitoring. In addition, we propose tasks to validate the results (this is a new task) and to provide technical assistance if necessary (this task could be directly incorporated from IEEE 1074).
- **Maintenance**. The maintenance process has two different paths. On the one hand, if knowledge is embedded in software, this process will provide feedback information to the software life cycle and lead to changes in the software. For this path, the task can be adapted from the IEEE 1074 maintenance process. On the other hand, CRISP-DM does not include a task for knowledge used as it is, and this needs to be developed from scratch.
- **Retirement**. The knowledge gathered from data is not valid forever, and this task is in charge of retiring obsolete knowledge from the system. CRISP-DM does not cover this process, but it can be adapted from IEEE 1074.

4.4 Integral processes

Integral processes are necessary to successfully complete the project activities. These processes assure project function completeness and quality. They are carried out together with development processes to assure the quality of development deliverables. The integral processes group the four processes described below.

- **Evaluation**. This process is used to discover defects in the product or in the process used to develop the DM project. CRISP-DM covers the evaluation process through

evaluation activities spread across different phases: Evaluation, Deployment, Business Understanding and Modeling. But we think the organization of the SE process is more appropriate and covers more aspects.

- **Configuration management**. This process is designed to control system changes and maintain system coherence and traceability. The ultimate aim is to be able to audit the evolution of configurations (Buckley, 1992). We consider this to be a key process in a DM project because of the amount of information and models generated throughout the project. Surprisingly, DM methodologies do not account for this process at all.
- **Documentation**. This process is related to designing, implementing, editing, producing, distributing and maintaining the project documentation. CRISP-DM considers this process across different phases: Deployment, Deployment, Modeling, and Evaluation.
- **User training**. Current DM methodologies do not consider user training at all. This process is related to training inexperienced users to use and interpret the results of the DM project.

5. Conclusions and future development

After analyzing the SE process models, we have developed a joint model based on two standards to compare, process by process and activity by activity, the modus operandi in SE and DM & KD. This comparison revealed that CRISP-DM does not cover many project management-, organization- and quality-related tasks at all or at least thoroughly enough. This is now a must due to the complexity of the projects being developed in DM & KD these days. These projects not only involve examining huge volumes of data but also managing and organizing big interdisciplinary human teams.

Consequently, we proposed a DM engineering process model that covers the above points. To do this, we made a distinction between process model, and methodology and life cycle. The proposed process model includes all the activities covered by CRISP-DM, but distributed across process groups that conform to engineering standards established by a field with over 40 years' experience, i.e. software engineering.

The model is not complete, as the need for the processes, tasks and/or activities set out in IEEE 1074 or ISO 12207 and not covered by CRISP-DM has been stated but they have yet to be adapted and specified in detail.

Additionally, this general outline needs to be further researched. First, the elements that CRISP-DM has been found not to cover at all or only in part would have to be specified and adapted from their SE counterpart. Second, the possible life cycle for DM would have to be examined and specified. Third, the process model specifies that what to do but not how to do it. A methodology is what specifies the "how to" part. Therefore, the different methodologies that are being used for each process would need to be examined and adapted to the model. Finally, a methodology is associated with a series of tools and techniques. DM has already developed many such tools (like Clementine or the neural network techniques), but tools that are well-established in SE (e.g. configuration management techniques) are missing. It remains to be seen how they can be adapted to DM and KD processes.

6. References

Anand, S.; Patrick, A.; Hughes, J. & Bell, D. (1998). A data mining methodology for cross-sales. *Knowledge Based Systems Journal*. 10, 449–461.

Biebler, K.; Wodny, M., & Jager, B. (2005). Data mining and metrics on data sets. In CIMCA '05: *Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce*. Vol-1 (CIMCA-IAWTIC'06), pages 638–641, Washington, DC, USA. IEEE Computer Society.

Biffl, S. & Winkler, D. (2007). Value-based empirical research plan evaluation. In *First International Symposium on Empirical Software Engineering and Measurement*, 2007, pages 494–494. IEEE.

Brachman, R. J. & Anand, T. (1996) *The process of knowledge discovery in databases*. pp 37–57.

Buckley, F. (1992). Configuration Management: Hardware, Software and Firmware. *IEEE Computer Society Press*, USA.

Cabena, P.; Hadjinian, P.; Stadler, R.; Verhees, J. & Zanasi, A. (1998) *Discovering Data Mining: From Concepts to Implementation*. Prentice Hall.

Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C. & Wirth, R. (2000). CRISPDM 1.0 step-by-step data mining guide. *Technical report*, CRISP-DM

CRISP-DM (2008). Crisp-2.0: Updating the methodology, www.crisp-dm.org/new.htm

Edelstein, H.A. & Edelstein, H.C. (1997). Building, Using, and Managing the Data Warehouse. In: *Data Warehousing Institute*, 1st edn., Prentice Hall PTR, Englewood Cliffs

Eisenfeld, B.; Kolsky, E. & Topolinski, T. (2003a). *42 percent of CRM software goes unused*, http://www.gartner.com

Eisenfeld, B.; Kolsky, E.; Topolinski, T.; Hagemeyer, D. & Grigg, J. (2003b). *Unused CRM software increases TCO and decreases ROI*, http://www.gartner.com

Fayyad, U.; Piatetsky-Shapiro, G.; Smith, P. & Uthurusamy, R. (1996) *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, MA

Fenton, N. E. & Pfleeger, S. L. (1998). Software Metrics: A Rigorous and Practical Approach. *Course Technology*, 2nd edition.

Fernandez-Baizan, C.; Marban, O. & Menasalvas, E. (2007). A cost model to estimate the effort of data mining projects (DMCoMo). *Information Systems*.

Gartner (2005). Gartner Says More Than 50 Percent of Data Warehouse Projects Will Have Limited Acceptance or Will Be Failures Through 2007. *Analysts to Show How To Implement a Successful Business Intelligence Program During the Gartner Business Intelligence Summit, March 7-9 in Chicago, IL*. 2005 Gartner Press Releases.

Gondar, J.E. (2005). *Metodología Del Data Mining*. Data Mining Institute, S.L

Gordijn, J.; Akkermans, H. & van Vliet, H. (2000). Business modelling is not process modelling. In *ER Workshops*, pages 40–51.

Hackos, J. T. (1994). *Managing your documentation projects*. John Wiley & Sons, Inc., New York, NY, USA.

Howard, C.M.; Debuse, J. C.W.; de la Iglesia, B. & Rayward-Smith, V.J. (2001). Building the KDD Roadmap: A Methodology for Knowledge Discovery, chapter of *Industrial Knowledge Management*. pages 179–196. Springer-Verlag.

IEEE (1997) Std. for Developing Software Life Cycle Processes. IEEE Std. 1074-1997. *IEEE*, Nueva York (EE.UU.)

ISO (1995). ISO/IEC Std. 12207:1995. Software Life Cycle Processes. *International Organization for Standardization*, Geneva (Switzerland).

ISO (2004). ISO/IEC Standard 15504:2004. Software Process Improvement and Capability determination (SPICE). *International Organization for Standardization*, Geneva (Switzerland).

Jaffarian, T.; Mok, L.; McDonald, M. P.; Blosch & M. and Stevens, S. (2006). *Growing it's contribution: The 2006 CIO agenda*. www.gartner.com

KdNuggets.Com (2002). *Data Mining Methodology*. http://www.kdnuggets.com/polls/2002/methodology.htm

KdNuggets.Com (2004). *Data Mining Methodology*. http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm

KdNuggets.Com (2007a). *Data Mining Methodology*. http://www.kdnuggets.com/polls/2007/data _mining_methodology.htm, 2007.

KdNuggets.Com (2007b). *Is data mining a mature field?* http://www.kdnuggets.com/polls/2007/data mining mature field.htm

KdNuggets.Com (2007c). *Data Mining Activity in 2007 vs 2006*. http://www.kdnuggets.com/polls/ 2007/data_mining_2007_vs_2006.htm.

KdNuggets.Com (2007d). *Outsourcing data mining*. http://www.kdnuggets.com/polls/2007/outsourcing data mining.htm

KdNuggets.Com (2008). *Data Mining ROI* http://www.kdnuggets.com/polls/ 2008/roi-data-mining.htm

JFMIP (2001). White paper: *Parallel operation of software - is it a desirable software transition technique?* Joint Financial Management Improvement Program.

Kotonya, G. & Sommerville, I. (1998). *Requirements Engineering. Processes and Techniques*. Wiley, USA.

Lanner Group (2008). *Witness Miner On-line Help System*. http://www.witnessminer.com

Marbán, O. ; Segovia, J.; Menasalvas, E. & Fernandez-Baizan, C. (2008). Towards Data Mining Engineering: a software engineering approach. *Information Systems Journal*. doi: 10.1016/j.is.2008.04.003

Martínez de Pisón Ascacibar, F.J. (2003). *Optimización Mediante Técnicas de Minería de Datos Del Ciclo de Recocido de Una Línea de Galvanizado*. PhD thesis, Universidad de La Rioja, 2003.

McMurchy, N. (2008). *Toolkit Tactical Guideline: Five Success Factors for Effective BI Initiatives*. Gartner.com

McNamara, C. (2007). *Complete guidelines to design your training plan*. http://www.managementhelp.org/trng dev/gen plan.htm.

Moore, J. (1998). Software Engineering Standards: A User's Road Map. *IEEE Computer Science Press*, Los Alamitos, California

NASA (2005). *Software Engineering Process Guidebook, Software Configuration Management Planning*. Software Engineering NASA LaRC and Analysis Lab.

Naur, P. & Randell, B. (1969). Software Engineering: Report on a conference sponsored by *the NATO science committee*.

Pai; W. C. (2004). Hierarchical analysis for discovering knowledge in large databases. *Information Systems Management*, 21:81–88.

Piatetsky-Shapiro, G. & Frawley, W. (1991) *Knowledge Discovery in Databases*. AAAI/ MIT Press, MA.

Piatetsky-Shapiro, G. (1994). *An overview of knowledge discovery in databases: Recent progress and challenges. Rough Sets, Fuzzy Sets and Knowledge Discovery*, pages 1–11.

Pressman, R. (2005). *Software Engineering: A Practitioner's Approach*. McGraw-Hill, New York.
Reifer, D. J. (2006). *Software Management*. Wiley-IEEE Computer Society Press, 7th. edition.
SAS Institute (2008). *SEMMA data mining methodology*, http://www.sas.com
Shepperd, M. (1996). *Foundations of Software Measurement*. Prentice Hall.
Smith, M. & Khotanzad, A. (2007). Quality metrics for object-based data mining applications. In ITNG '07: *Proceedings of the International Conference on Information Technology*, pages 388–392, Washington, DC, USA, 2007. IEEE Computer Society.
Solarte, J. (2002). *A proposed data mining methodology and its application to industrial engineering*. Master's thesis, University of Tennessee, Knoxville
Strand, M. (2000). *The Business Value of Data Warehouses - Opportunities, Pitfalls and Future Directions*. PhD thesis, University of Skövde
Two Crows Corp (1999). *Introduction to Data Mining and Knowledge Discovery*. 3rd edn.
Westland, J. (2006). *The Project Management Life Cycle: A Complete Step-by-Step Methodology for Initiating, Planning, Executing and Closing the Project Successfully*. Kogan.
Yang, Q. & Wu, X (2006). 10 Challenging Problems in Data Mining Research. *International Journal of Information Technology & Decision Making*. Vol. 5, No. 4 (2006) 597–604. World Scientific Publishing Company.
Zornes, A. (2003) *The top 5 global 3000 data mining trends for 2003/04*. META Group Research-Delta Summary.

Data Mining and Knowledge Discovery in Real Life Applications

Edited by Julio Ponce and Adem Karahoca

ISBN 978-3-902613-53-0
Hard cover, 436 pages
Publisher I-Tech Education and Publishing
Published online 01, January, 2009
Published in print edition January, 2009

This book presents four different ways of theoretical and practical advances and applications of data mining in different promising areas like Industrialist, Biological, and Social. Twenty six chapters cover different special topics with proposed novel ideas. Each chapter gives an overview of the subjects and some of the chapters have cases with offered data mining solutions. We hope that this book will be a useful aid in showing a right way for the students, researchers and practitioners in their studies.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Óscar Marbán, Gonzalo Mariscal and Javier Segovia (2009). A Data Mining & Knowledge Discovery Process Model, Data Mining and Knowledge Discovery in Real Life Applications, Julio Ponce and Adem Karahoca (Ed.), ISBN: 978-3-902613-53-0, InTech, Available from: http://www.intechopen.com/books/data_mining_and_knowledge_discovery_in_real_life_applications/a_data_mining__amp__knowledge_discovery_process_model

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821