Historical data analysis through data mining from an outsourcing perspective: the Three-phases model

Authors

Arjen Vleugel
Utrecht University, the Netherlands

Marco Spruit
Utrecht University, the Netherlands

Anton van Daal
In Summa, the Netherlands

Submission date of revision

November 20th , 2009

Abstract

The process of historical data analysis through data mining has proven to be of value for the industrial environment. There are many models available which describe the inhouse process of data mining. However, many companies either do not have the inhouse skills or do not wish to invest in performing inhouse data mining.

This research investigates the applicability of two well-established data mining process models in an outsourcing context. We observe that both models cannot properly accommodate several key aspects in this context. Therefore, we propose the Three-phases method, which consists of data retrieval, data mining and results implementation within an organization. Each element is presented as a visual method fragment. The model is validated through expert interviews and an extensive case study at a large Dutch staffing company. Both validation techniques substantiate our claim that our Three-phases model accurately describes the data mining process from an outsourcing perspective.

Key words: data mining, outsourcing, method engineering, data quality, implementation.

Introduction: On the need for a new data mining method

"A miner who has to work with only very current information can never detect trends and long-term patterns of behavior. Historical information is crucial to understanding the seasonality of business and the larger cycles of business to which every corporation is subject" (Inmon, 1996).

The crucial element of this quote is 'the patterns of behavior'. The main technique used to retrieve those patterns is called data mining. There are several definitions which describe data mining. We use a definition from Shaw et al. (2001): "*Data mining is the process of searching and analyzing data in order to find implicit, but potentially useful, information. It involves selecting, exploring and modeling large amounts of data to uncover previously unknown patterns, and ultimately comprehensible information, from large databases".* In the early nineties, data mining was often described as "*a blend of statistics, AI, and data base research*" and was not considered to be a field of interest for staticians, where some of them described it as "*a dirty word in Statistics*" (Pregibon, 1997). Nevertheless, the research area of data mining has increasingly become an important field of interest to both academics and practioners.

Data mining can be positioned as a corollary from business intelligence (Kudyba et al, 2001; Shmueli et al, 2006). This claim is also supported by business intelligence tool providers such as Microsoft and Oracle, who both position their data mining tool as an integral part of their overall business intelligence solution (Microsoft, 2008; Oracle, 2007). Business Intelligence (BI) can be defined as the process of turning data into information and then into knowledge (Golfarelli et al., 2004). It was first introduced in the early nineties, "*to satisfy the managers' request for efficiently and effectively analyzing the enterprise data in order to better understand the situation of their business and improving the decision process."* (Golfarelli et al., 2004). Data mining supports this by providing companies the unique ability to review historical data to help improve the managers' decision-making processes (Golfarelli et al., 2004).

Most research performed in the area of data mining is aimed at adjusting existing data mining techniques to solve a specific problem, thus creating a new data mining technique (e.g. Hui et al., 1999; Rygielski et al., 2002) . This research, on the other hand, has a different goal, which is the creation of a method concerning the whole process of data mining. Two methods (one emerged from the field of statistics, one emerged from business needs) have become the standards with regard to the description of the process. The first method was suggested by Fayyad et al. (1996) and involves five different stages. Its input is data, which eventually leads to knowledge (see Figure 1.1). The method embraces the description of the process, but does not include the use of specific tools or include a section of how to implement data mining results. Furthermore, the method does not include business needs. The business environment needs a practical model to apply data mining, one which also includes the business aspects of specific organizations.

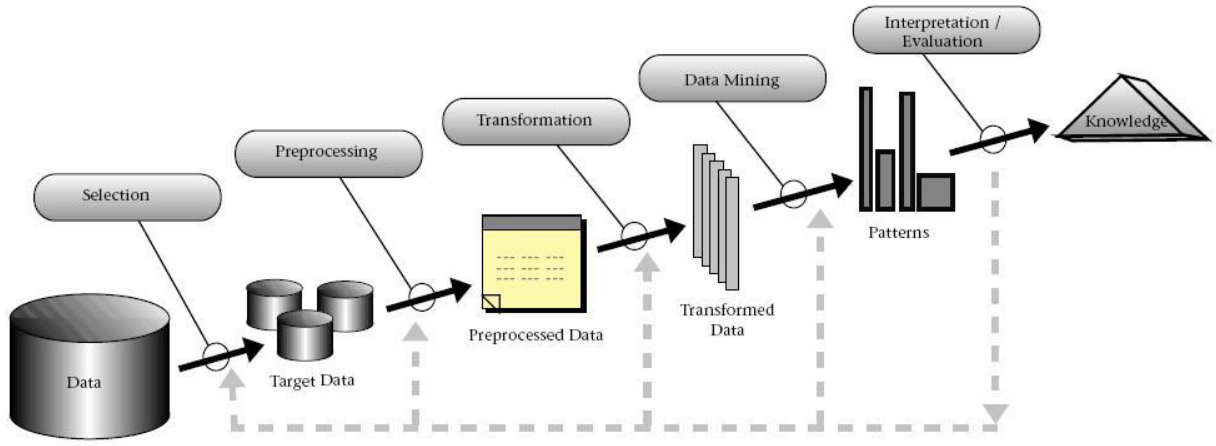


Figure 1.1: The Knowledge Discovery in Databases (KDD) process (Fayyad et al., 1996).

Because of this, the Cross Industry Standard Process for Data Mining (CRISP-DM) was developed by three major industrial players in the field of data mining: DaimlerChrysler, NCR and SPSS (CRISP-DM, 2000). Unlike the Fayyad et al. (1996) approach, the trigger was the need for a model which describes the process of implementing data mining results from a business point of view. Nowadays, the CRISP-DM method has become the standard approach concerning data mining applied in the field of business (KDNuggets, 2007).

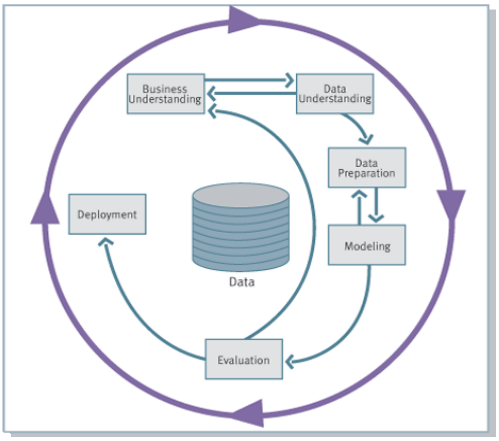


Figure 1.2: The CRISP-DM process (CRISP-DM, 2000).

It differs from Fayyad in various ways. As said above, its key focus is on organizational implementation. This translates directly into the way the CRISP model starts with creating business understanding (see Figure 1.2). The KDD model (Fayyad et al. 1996) starts with the data understanding process. Another difference is that CRISP-DM has a deployment phase, which again emphasizes the fact that it was developed for implementation purposes. The KDD model does not have a deployment phase. Both models are sufficient when describing the process of data mining. However, the recent economic trend of outsourcing calls for a different approach.

Outsourcing the process

"The financial crisis and global recession will accelerate adoption of global outsourcing and offshoring as strategic business tools as organizations respond to economic adversity with a forceful push toward cost-reduction" (Equaterra, 2008).

Outsourcing is subcontracting a process, such as product design or manufacturing, to a third party company (Bardhan et al., 2003). It is triggered by the strategic management of a company. The advantage of outsourcing is generally motivated by the drive to return to the 'core business', leaving the support business activities in the hands of the outsourcing party or third party (Loh et al. 1995). Earl (1996), states that the objectives of outsourcing are cost cutting and subcontracting responsibilities. Another advantage is expressed in a research performed by Quinn (2000), who states that strategically outsourcing innovation can put a company in a sustainable leadership position. Other advantages include risk transference to the vendor and companies becoming more flexible (Deloitte, 2005).

Despite these advantages, outsourcing in general cannot be defined as a complete success. Deloitte (2005) addressed all issues regarding the above advantages of data mining. These are additional paid costs because of services which were not addressed in the contract, vendors becoming complacent once contracts were in place and are unable to fully absorb the costs of business losses, vendors do not have the capabilities needed to provide the expected level of quality and finally, companies have mislabeled functions as non-strategic and ultimately brought these functions back inhouse. Beside these issues, the best way to outsource business processes is still in debate. For instance, one approach is to outsource parts of the inhouse process to multiple third parties. This is known as 'selective' versus 'total' outsourcing and is described by Lacity et al. (1998) and Sambamurthy et al. (2001), among others. The argument is that third parties have different core competencies, thus it is better to use multiple third parties. However, a different research came up with a contradictory conclusion; Rouse et al. (2003) report that *'the probabilities for those engaged in selective outsourcing were statistically no different*'. Based on these investigations, we shall assume that it does not matter whether the process is selectively or totally outsourced.

Can we safely state that data mining, being a data analytical tool, can successfully be outsourced? According to a research on how to develop advanced analytical expertise performed by Henschen (2009) , most companies (48%) plan to train in-house BI experts, where 35% plan to hire either temporary or permanent experts. This roughly translates into the conclusion that a third of all companies (which were enrolled in the research) is planning to acquire outside consultants e.g. outsource (either temporarily--in the form of a project--or permanently) advanced analytical expertise. When reviewing the issues regarding outsourcing, the characteristics of the third party vendor is one of the most important factors for outsourcing failure; contractual frictions, knowledge deficiencies at the third party site, insufficient capabilities and complacency are all third party related. Regarding the mislabeling of functions in terms of data mining, we can state that the results of data mining can be of strategic importance to a company, but the process itself is not: it is a means to reach a certain strategic goal, but not a goal itself). Thus, we believe data mining to be a non-strategic function.

This research aims not only to provide a third party with a generally applicable method to perform the appropriate data mining technique, but also to provide a strategy for implementing the method's results. In other words, to improve the business processes by using historical data. This has become increasingly popular in the scientific field and is known as process mining. Process mining can be defined as "*the method of distilling a structured process description from a set of real executions"* (van der Aalst et al., 2004). It involves the use of logs; a list of actions performed by entities which can be bundled into a process. Usually companies tend to log an exhaustive list of performed actions with regard to their business processes. The research provides scientific validation (or rejection) for process mining in terms of optimization of the business processes: can these processes increase corporate performance? We adopt best practices from both data mining and process mining to improve the business processes of an organization.

Research question and approach

To create the scope of this research, we defined the following research question: how can historical data analysis be outsourced to improve corporate performance?

Corporate performance is defined as "*all of the processes, methodologies, metrics and systems needed to measure and manage the performance of an organization*" (Buytendijk, 2001). The research question above consists of the following sub questions:

- *How can we retrieve data which include companies' preferences with regard to data mining results?*
- *Given a specific case with unique characteristics, how do we choose the best data mining technique?*
- *How can data mining results be implemented in such a way that corporate performance is increased?*

Our initial aim was to adjust the currently available methods in such a way that they would provide an answer to our research question. However, we could not adapt either one of them without losing methodological integrity. Therefore, we have adopted parts of both the KDD model and the CRISP-DM model in order to build a new method which is inherently appropriate from an outsourcing perspective. We have analyzed both methods and deconstructed them for the purpose of identifying the phases of a data mining process. We have used concepts from the scientific field of method engineering to create our method.

Method engineering within an IT context can be defined as "the engineering discipline to design, construct and adapt methods, techniques and tools for the development of information systems" (Brinkkemper, 1996). What defines this field is the use of method fragments, which are coherent pieces of IS development methods, which are visualized in the form of process deliverable diagrams (PDD's; Weerd et al., 2006). A PDD combines two unified modeling techniques into a single diagram. An activity diagram is based on unified modeling language (UML 2.0, 2004) and is shown on the left side of a PDD. The right side of a PDD shows the deliverables of the activities akin to a UML-based class diagram. We have addressed each phase of the data mining process by capturing it in a method fragment. The method has been validated through numerous expert interviews on the one hand and a real-world business implementation at a large staffing company on the other hand.

The remainder of this paper is structured as follows. In the next section, we introduce the Three-phases method. Then, each phase (data retrieval, data mining and result implementation) of our method is presented through a corresponding method fragment. The fragments correspond with the phases mentioned in the overall method. Next, we validate the method by applying it to a real-world case in which a large staffing company wishes to improve its corporate performance. Finally, we present our conclusions and present several suggestions for further research.

The Three-phases method

Figure 2.1 introduces the Three-phases method. Its key characteristic is the fact that it embraces outsourcing of the data mining process by defining a clear distribution of roles and incorporating an elaborate implementation phase. Regarding the roles, there are two parties involved when applying the Three-phases method. The first party is the case company, which provides data for the purpose of data mining. The other party, which we refer to as the third party, is the party which performs the data mining activities. Usually, the case company is the trigger of the process. In each of the Three-phases, both parties are represented. The distinctive distribution of roles with regard to outsourcing is the first major difference between existing models and the Three-phases model (see table 2.2 for the exact distribution of roles).

We address three different phases when performing data mining, the first being data retrieval. In this phase, we introduce hypotheses as a new approach for the retrieval of data source independent data. Data source independent data refer to all structured data provided by a company. It should be noted that these data do not necessarily have to be stored in a database. Furthermore, we include a way of filtering out data errors.

The second phase, the mining itself, is characterized by the fact that it is applicable to any organization. We have analyzed four common data mining tools from which we have derived seven common techniques. We then described those techniques in terms of applicability. The final step of the second phase is the validation of both the technique and the results. The third phase, concerning the implementation of the results, addresses the wishes of the case company with regard to the solution; do they either want an implemented reporting tool or a simple report containing the data mining results? We use the business domains from the Business IT Alignment (BITA) model (Scheper, 2000) to describe the organization. The approach is also described in terms of these domains. The final step is the deployment of the result.

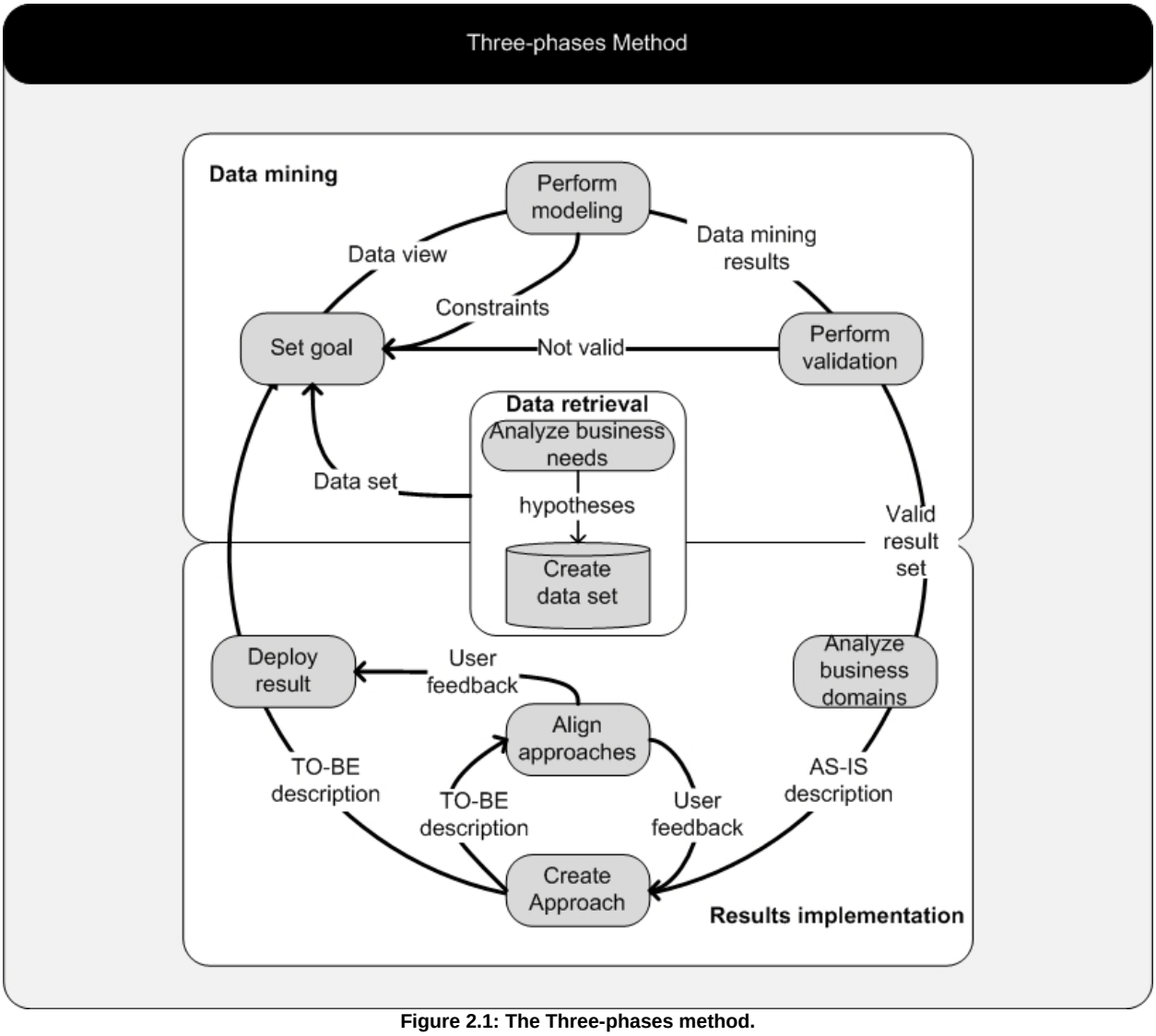


Figure 2.1: The Three-phases method.

Activities of the Three- phases model								
Data retrieval		Data mining			Results implementation			
Analyze business needs	**Create data set**	**Set goal**	**Perform modeling**	**Perform validation**	**Analyze business domains**	**Create approach**	Optional: **Align approaches**	**Deploy result**
Perform business interviews (CT)	Retrieve entities and attributes (T)	Set purpose (CT)	Choose technique (T)	Validate result (CT)	Perform business interviews (CT)	Select data mining result (T)	Interview end-users / stakeholders (CT)	Business preference; either:
Review business documents (T)	Create raw data set (T)	Define goals (CT)	Perform modeling (T)	Choice (C)	Define business preference (T)	Analyze AS-IS description (T)	Create alignment approach (T)	- Create change report (T)
Define hypothesis (CT)	Apply data filters (T)	Create data view (T)	Interpret result (T)		Create AS-IS description (T)	Create TO-BE description (CT)		- Implement result (CT)
	Apply data enrichment (CT)		document result (T)					

Table 2.2: tasks of the Three-phases method. Note that C is performed by the case company; T is performed by the third party

The accomplishment of the method is standardized, but allows flexibility; each step can be adjusted according to the specific situational factors of a case company. See table 2.2 for a complete list of steps. The iterative part of the method only applies to the data mining phase and the implementation phase; if performed correctly, the data retrieval should only be performed once. Note that one step is not obligatory; depending on the preferences of the case company, the approach should either be aligned between data mining result and company (in the case of a reporting tool) or not (in the case of a report). This is the second major difference between existing models and the Three-phases method; the implementation phase with regard to outsourcing is characterized to be lengthy and extensive, and thus needs to be addressed extensively.

We did not only address a current scientific issue (how to outsource data mining), but we also assured our method to be applicable for the industry. We believe that using method fragments, which are specifically designed to capture processes and its deliverables, serve our purpose rather well (Brinkkemper, 1996). Note that the output of each method fragment is the input for the next method fragment. Therefore, combining all fragments leads to a method which can directly be translated into the Three-phases method.

Phase one: data retrieval

The first steps in the process of creating a usable dataset are unmistakingly obvious: one must identify the goal from a customer's viewpoint (Fayyad et al., 1996) and create a greater understanding of the business domain (Yu, 2007). Hypotheses are used to achieve these goals. Hypotheses can be defined as "*a scientific proposition which can be falsified*" (Webster, 2007). By creating hypotheses, the third party develops a greater understanding of the customer's wishes and needs to delve deeper into the company structure in order to create usable hypotheses. The main goal of a hypothesis is "*making the search for finding relevant information possible*" (Everink, 1999). This makes use of hypotheses valid for the process of data mining (as mentioned earlier, data mining is about finding patterns in data). Note that it is not necessary to know which data are stored. The trigger in the process of data mining is usually the case company; by performing business interviews and reviewing business documents, one can create hypotheses. Note that feedback from the case company (in the form of decision makers and /or end-users) is invaluable if one wants to cover all aspects in the desired dataset. Business interviews can either be structured or unstructured (depending on the desired procedure by the case company).

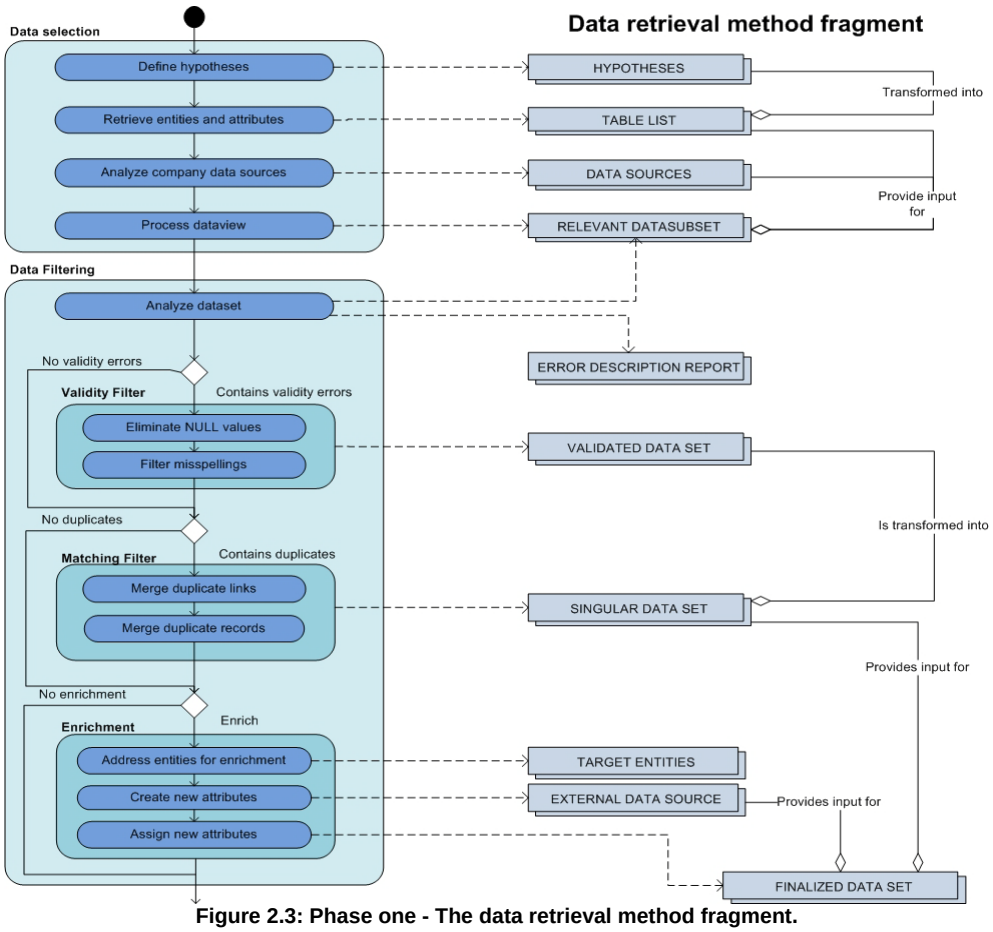


Figure 2.3: Phase one - The data retrieval method fragment.

Once a list of hypotheses is constructed, the next step is to retrieve entities and attributes from this list. For example, consider a hypothesis we used as a key hypothesis regarding the case study:

A successful allocation [=noun] depends on the type [=attribute] of actions [=noun] performed on it.

By identifying the nouns, we derive entities. Each entity is translated to a (database) table. Then, we need to assign attributes to these entities. The fact that no research has been done in this area makes it hard to validate the creation of entities as a step, as pointed out by Duntsch et al. (1998): "*the researcher's choice of attributes and measurements are part of the model building process and of the data analyses*". Basically, Duntsch et al. (1998) state that the choice of attributes cannot be defined by a general approach, since each situation has unique characteristics and therefore a general situation cannot be designed. Although it is true that each company has its own unique data, this does not imply that they use unique data structures (for instance, a relational database structure). By using the most common dividers (tables, entities and attributes), we propose a general approach for capturing data (Figure 2.3). The last step regarding data retrieval is to apply data quality filters.

This process, known as data filtering (cleaning), should remove all errors and inconsistencies in the data view(s) (Rahm et al., 2000). Because we have created a method which focuses on outsourcing, we can only address the special class of data quality issues named data entry errors. Data entry errors can be defined as false data which have been entered into the data source. In the process of performing data mining it is not possible to prevent errors from entering the data source; the third party cannot modify the way the information is stored. Clearly, for the purpose of using a valid dataset in the process of outsourcing data mining, a solution should be devised to clean the database after the data have entered it. The Three-phases method provides guidelines to ensure this, by handling misspellings and 'null' values (which are empty values), referring to the completeness of data (Engels, 1999). Furthermore, we should address 'matching errors' and ensure that duplicate tuples are eliminated, both contradicting as equivalent (Lee et al., 1999). The last step is to apply data enrichment. According to Neun et al. (2004), data enrichment is necessary "*to equip the 'raw' spatial data with information about the objects and their relationships"*. Therefore, we include data filtering in our method fragment by including independent external data sources (figure 2.3)..

Phase two: data mining

There is a whole array of different data mining techniques. Depending on the situational aspects of a project, one can decide to employ a certain technique. On top of that, each data mining technique can be performed in several different ways by using different algorithms. Having said that, it's clear that a classification of techniques needs to be constructed in order to answer the second sub-research question. To answer this, consider the following highlights of an internet poll among data mining professionals (KDnuggets, 2009). The summary of the poll results in Table 2.4 shows the main four data mining tool providers: the two most frequently used single tool providers (SPSS and SAS) and the two most frequently used overall database solution providers which include a data mining tool (Oracle and Microsoft). We have opted to not include non-commercial tools, because we feel that commercial support for a data mining tool is essential due to its inherent complexity.

Name	Times used	Stand alone tool	Techniques
SPSS	142	Yes	Predictive modeling (decision trees, neural networks, statistical models as linear regression and logistic regression, time series). Clustering models. (clustering and sequence clustering). Association rules (decision trees) and Screening models (Neural network, Naïve Bayes)
Excel & MSSQL	81	No	Clustering, Association rules, Decision trees, Linear regression, Logistic regression, Naïve Bayes, Neural network, Sequence clustering, Time series.
SAS	79	Yes	(Non-) Linear and logistic regression, Decision tree, Neural networks, Memory based reasoning (k-means clustering), Ensembly nodes (association) and user defined models, two-stage model
Oracle DM	7	No	Classification (Logistic regression, Naïve Bayes, decision tree), Regression (Multiple regression, time series). Attribute Importance (Minimum descriptive Length), Anomaly detection (One-Class support vector machine), Clustering, Association (market basket analyses), (Factorization), Time series.

Table 2.4: Adapted from KDD nugget poll May 2009: Commercial software data mining tools used

We then derived the most common data mining techniques which were supported by all the tools. These were Clustering, Neural networks, Regression, Association rules, Decision tree, Time series (not covered by SAS, they include Memory-based reasoning to forecast) and Naïve Bayes. Naïve Bayes is in fact mentioned by all, but with different names. The technique is used to determine the impact of an attribute on other attributes; SAS refers to it as two-stage, ODM calls it attribute importance and SPSS uses screening models.

We are aware of the fact that there are more techniques available for data mining purposes. However, these other techniques are usually designed for a specific case and contain constraints on when to use a specific case (e.g. Hui et al., 1999; Rygielski et al., 2002). Therefore, we did not include them. Table 2.5 classifies the seven common techniques according to three different purposes.

Purpose	Techniques
Rule-based data mining	Association Rules and Decision Trees
Descriptive data mining	Naïve Bayes and Clustering
Predictive data mining	Time-series and Regression

Table 2.5 Data mining classification

We wanted to include a general approach concerning a choice for the right mining technique. Therefore, we included a classification in terms of purposes about the most common techniques. We did not include neural networks in our classification. Neural networks learn from a training set, generalizing patterns inside it for classification (Berry et al. 1997). Although an output is guaranteed to be found,. "*Neural networks are not commonly used for data-mining tasks, because they often produce incomprehensible models and require long training times*". (Craven et al. 1997). We consider neural networks useful in certain cases, but only if the other available techniques are not appropriate.

Rule-based data mining

The well-known division between predictive and descriptive data mining (Fayyad et al., 1996) is not exhaustive; some techniques can be used to serve both purposes. Therefore, we propose a new category: rule-based data mining. Rule-based data mining can either be used as input for either predictive or descriptive purposes or as stand-alone data mining output. It is independent of either descriptive or predictive, since it covers both categories but is not specifically bound to either one of them. Rule-based data mining is characterized by the fact that it is purely based on rules. It consists of decision trees or association rules. A decision tree is a form of inductive reasoning, which can be defined as creating rules by extracting patterns from data (Quinlan, 1985). Another definition is provided by Apte et al. (1997), which identifies the use of decision trees; to find the odds of an outcome based on values in a training set. At the root node (root of the tree), the database is examined and a splitting criterion is selected. At a non-root node N, the family of N is examined. From this family, a splitting criterion is selected (Gehrke et al, 2000).

The association rules data mining technique is used to show an association between attributes (or entities containing attributes). Normally, results are presented in the format "X→Y", where X is a set of attributes and Y is a set of attributes (Srikant, 1997). Rule based data mining helps users (i.e. data miners) to find support for certain claims. For example: "Is it likely that a male employee is allocated in a technological branch?". Whether to use either association rules or a decision tree is determined by the third party. The choice depends on the question whether one prefers to compare a subset of attributes to generate a rule (rule based) OR one wants to retrieve a successful decision path (decision tree).

Predictive data mining

The first general remark concerning predictive data mining from our research perspective is the fact that, although valuable, we are analyzing internal case company data; in other words, the prediction is based on historical internal trends. In order to make accurate predictions, external market information should be taken into account as well. Predictive data mining tends to become more accurate when the size of the used dataset grows (Keogh et al., 2003). The results are validated by using a small fraction of the total dataset for testing the algorithm. Regarding the modeling techniques, predictive data mining includes the use of time series and regression analysis.

The time series algorithm uses a sequence of data points (measured in time intervals). The model is used to forecast future events based on known past events. (Keogh et al., 2003). The key attribute for using this technique is a 'date' variable (or another form of timestamp). A distinction can be made between two different types of predictions. One of them is aimed at short-term predictions, while the other is aimed at long-term trend indications (Box et al., 1994).
The other technique regarding predictive analyses is the regression technique. This technique does not require a 'date variable'. The output is a formula, which differs depending on the type of regression (either linear, multiple or non-linear) (Kachigan, 2005).

Descriptive data mining

Descriptive models describe patterns in existing data, and are generally used to create meaningful subgroups such as demographic clusters (Fayyad et al., 1996). Unlike predictive models, it is not clear what the outcomes of descriptive data mining will be; it depends on the technique and the used dataset. Furthermore, the results will not always be applicable. Since descriptive data mining is analyzing so-called business rules, we need to understand those rules before we can make valid conclusions. Therefore, when validating these outcomes, the case company should always be involved. Descriptive data mining consists of clustering and naïve bayes. The latter technique measures the impact of a variable.

Clustering methods partition a set of objects into clusters such that objects in the same cluster are more similar to each other then objects in different clusters according to predefined criteria (Huang et al., 1997). Clustering consists of hierarchical and partitional clustering. Partitional clustering involves creating a set of centroids equal to the amount of clusters. Each object is then assigned to a centroid (Jain et al., 1999). An example is the K-means algorithm, which has been embedded in the Microsoft, Oracle and SPSS tools. Hierarchical clustering treats every single object as a cluster and iteratively merges them to reach a desired subset of clusters.

Data mining method fragment

The process of performing the modeling activities consists of setting a goal, selecting the dataset needed to achieve this goal and deciding on the most appropriate data mining technique (Figure 2.6). This choice is based on the characteristics of the different purposes. The "State goals" sub-phase consists of settling on a purpose (predictive, descriptive or rule-based), define specific goals regarding this purpose and eventually a retrieval of a data selection based on this purpose plus goals. Next, the first choice is whether rules need to be created (e.g. perform rule-based data mining) or not; this is because rules can provide input for either descriptive or predictive data mining, and therefore the first option is to perform rule-based data mining. Then, the choice between descriptive and predictive data mining needs to be made. This "State goals" and "Rule-based data mining" sub-phases are deliberately unaligned in the model as shown in Figure 2.6 to emphasize the fact that it is a choice, not merely to state a sequence order.

The result of any data mining technique is referred to as a data mining model. Such a model is validated by both consulting the case company (to verify the applicability of the results) as well as through a standard data mining procedure. Usually, a fixed percentage (75%) of the historical data is used to build the model. The remaining part of the data is used to validate the model. The result of these validation steps is stored in a validation report. This is visualized as a method fragment in figure 2.6.

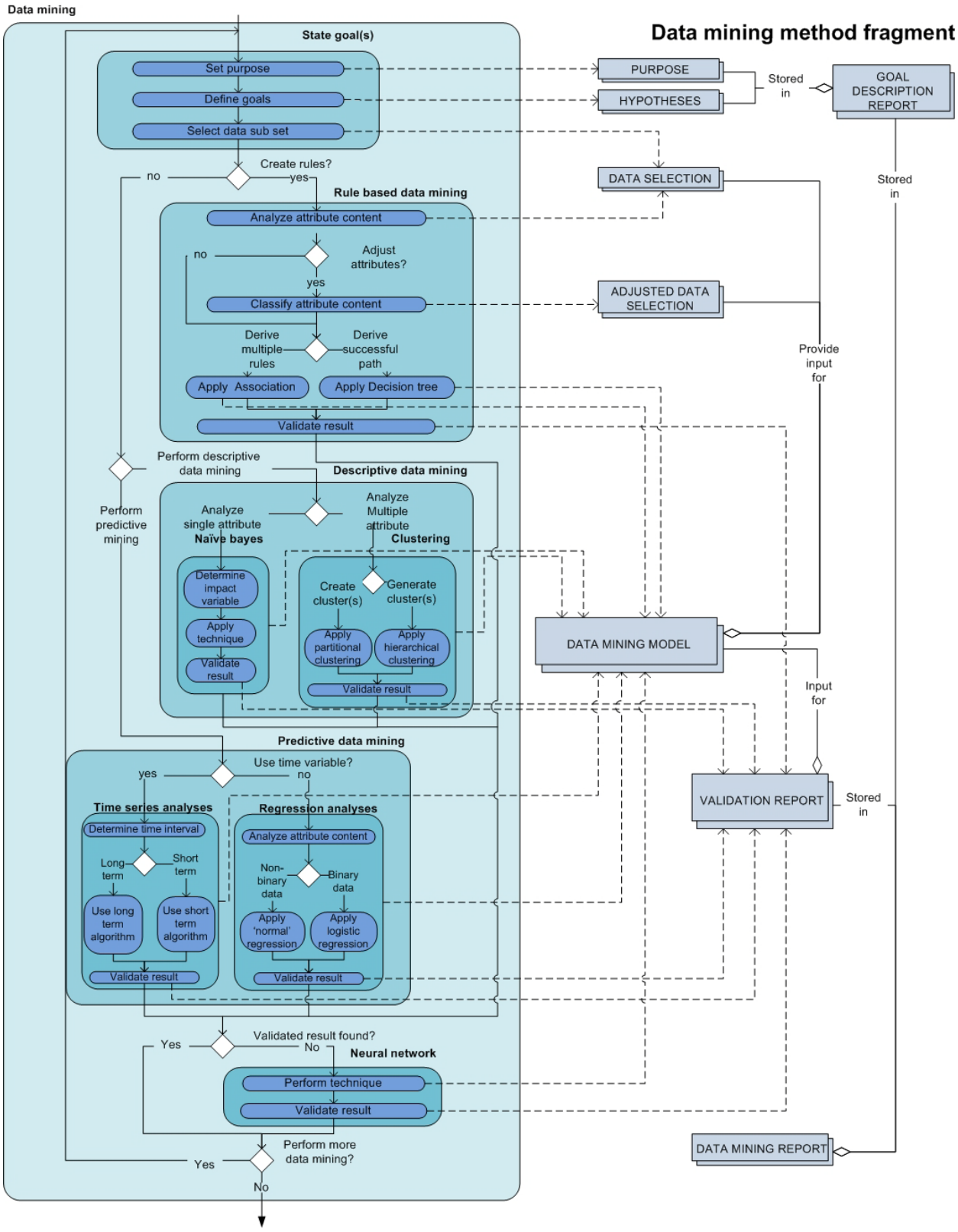


Figure 2.6: Phase two - The data mining method fragment.

Phase three: results implementation

The final phase of the Three-phases model focuses on the implementation of the validated data mining results. The incorporation of this phase is what distinguishes the Three-phases model the most other data mining models. Furthermore, it provides scientifically validated approaches in the context of data mining and data retrieval to the field of industry. We use one existing model to measure the current company performance and one model to measure to impact of change with regard to implemented results. This latter model is known as the Strategic Grid model (McFarlan et al., 1988). It offers a high-level positioning model to describe the current situation (see Figure 2.7). We use it to decide whether the potential strategic impact of future information systems (e.g. the implementation of data mining) is either high or low.

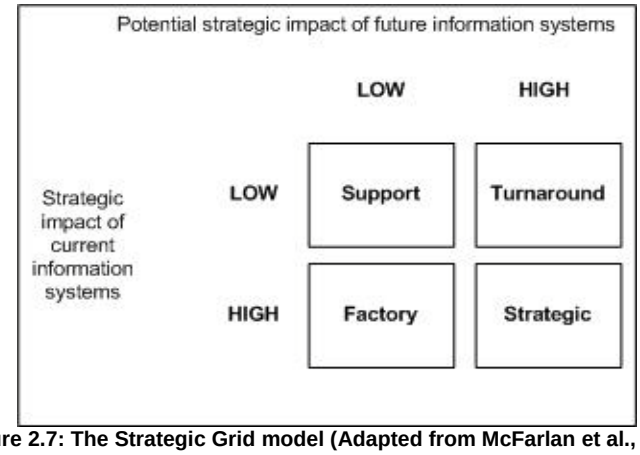


Figure 2.7: The Strategic Grid model (Adapted from McFarlan et al., 1988)

With respect to measuring company performance, numerous models are available (e.g. Maul et al., 2003; Rosemann et al., 2006). We choose to use the Business IT alignment (BITA) model from Scheper (2002), depicted in Figure 2.8. The BITA model is designed for optimal alignment of IT implementation projects, which applies very well to implementation of data mining results. The model uses five business domains (factors) to measure company performance.

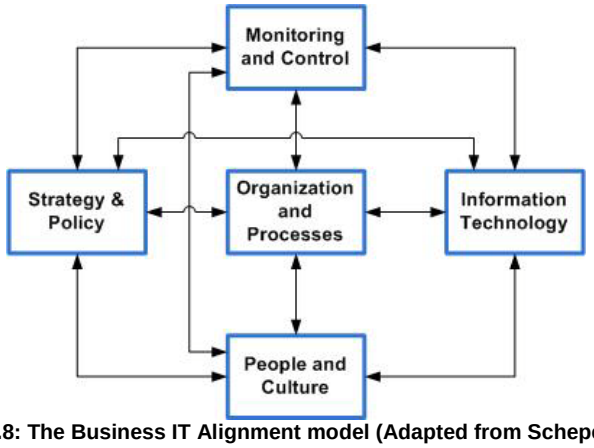


Figure 2.8: The Business IT Alignment model (Adapted from Scheper, 2002)

Strategy and policy. Three levels of organizational decision making can be identified, each dealing with a different executing horizon (Schmidt et al., 2000). The strategic level deals with a relatively long planning horizon (2-5 years), the tactical level deals with a mid-range horizon (approximately 6-24 months) and the operational level deals with the short term horizon. The decision whether to implement the data mining results is made at the tactical or even strategic level but directly effects the operational level. In other words, the different types of data mining results also affect the operational level, since this deals with the environment created by decisions made at the tactical level.

Data mining result	Decision level
-	Operational level
Descriptive, Rule-based and Predictive data mining results	Tactical level
Predictive data mining results	Strategic level

Table 2.9: Affected decision levels when implementing data mining results

Organization and processes. We distinguish between process, procedure and action/rule. A procedure is a unique way of working (Webster, 2008), which consists of one or more processes, a sequence and rules. A process consists of steps, which itself contains an order and content. A sequence determines the sequential order of the processes in the procedure. Finally, rules provide the input for either a process or a sequence of steps. However, rules are not affected by either one of them. Rules, or business rules, can be defined as a statement that defines or constrains some aspect of the business (Business rules group, 2000). Business rules are usually expressed either as constraints or in the form '*if conditions then action'* (Halle, 2001). Note that the different types of data mining results (i.e. predictive, descriptive or rule-based) each uniquely affect business rules, processes and execution sequence.

People and culture. Once implemented, the success of an implementation can be measured by its usage. High usage in the end can only be achieved when people have a positive attitude towards the implementation. One model which explains the possible patterns people follow when confronted with new approaches towards their way of working is the ARIA model (van Driel 1999). It describes a path were people are first Amazed by the introduction of the new approach. Afterwards, they may experience the approach as threatening and will Resist it. People tend to fit the old approach in the new approach, i.e. Imitate. Finally, given ample time, the approach may be Accepted. The lesson learned here is that in order to successfully implement an IT-solution, one cannot ignore the importance of company culture.

Information technology. We would like to emphasize the fact that we do not have to implement a tool to perform data mining, as it is exactly these activities which are outsourced in our approach. What we do need is a visualization tool, but only if the company is willing to embrace the data mining process in such a way that it will become an integral part of its business.

Data retrieval method fragment
By reviewing the business domains, we identify strategy and policy, organization and processes and monitoring and control as the key domains. We bundled them into a so-called process description (table 2.10), which uses the elements of these key domains.

Process	**Impact level**	**Monitoring**
Description of a process.	Level of impact, either operational, tactical or strategic.	Description about how the process is being monitored.

Table 2.10: A process description template.

This leaves only information technology and people and culture unaddressed. Regarding information technology, we do not include it in our method fragment since we don't implement a tool for performing data mining (again, we outsource the process).

People and culture, on the other hand, are only to be considered when the case company requires a full implementation of the result, not just a report. This is because a report allows the case company to execute change itself; a fully implemented approach is proposed by the third party (and "executed"). This is another major difference compared with existing models; since the process is outsourced, we first have to decide who performs the implementation.

First, the current situation is analyzed through interviews, and results in a document in the format of Table 2.10 with respect to the AS-IS situation. Then, the case company faces a choice; do they want to implement a result reporting tool or a one-time report containing the results? This is captured in a "company wish".
Second, we create a TO-BE situation by selecting the data mining results and describe the situation in the format of the AS-IS situation. If the results need to be fully implemented, we should interview stakeholders (in other words, incorporate the people and culture domain) before we can start implementing the results.

Note that the logical order of steps is to first perform an analysis of the current situation before the company decides on the desired result (report or tool). This is because the AS-IS description with regard to the data mining results is used in this process. It is not included in our model (only as a choice) since it only affects the case company and not the data mining process of the third party.

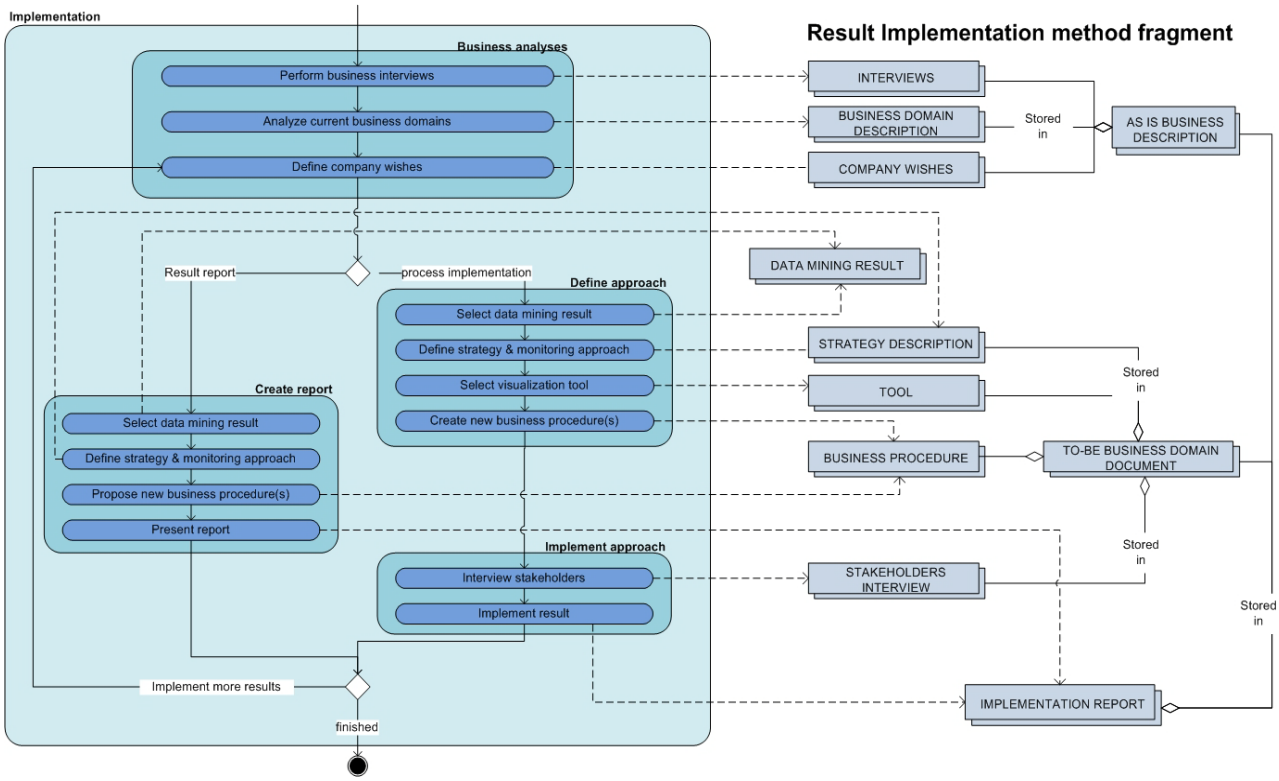


Figure 2.11: Phase three - The result implementation method fragment.

Positioning the model

We have used a method engineering approach as described by Brinkkemper (1996). The method fragments proposed in the section above are usually built to create an implementation method. However, our method cannot be classified as such a method. An implementation method consists of unique project situations, feature groups and a method base (Harmsen et al., 1994). The Three-phases method does not include unique project situations, although it allows for the use of situational factors. Furthermore, we did not specifically include a method base. Instead, we derived elements from the KDD model (Fayyad et al., 1996) and the CRISP-DM model (Chapman et al., 2000) and created a new approach, partially based on these reference models for each phase.

Unlike a situational method, the final result of our approach is a generic method to apply the process of data mining to an organization. It consists of a situational approach concerning the use of a certain data mining technique and a generic process description for the implementation of the data mining results. This last element indirectly refers to a process, which describes a partially ordered set of activities intended to reach a goal (Kueng et al., 1997). This emphasizes the fact that there is no need for a method base. The generic process already describes all the project situations in itself.

When comparing the method with both the CRISP-DM approach and the KDD process (Fayyad, 1996), the differences between the KDD method and the Three-phases method are obvious. Contrastingly, one might argue that the proposed method is merely a revision of the CRISP-DM method with additional refinements in process areas relevant to outsourcing. However, unlike the CRISP-DM method, the general hub of the Three-phases method is the business need, not the business data (compare Figures 1.2. and 2.1). A third party cannot achieve successful data mining without knowing the case company; thus, without reviewing the business needs. The CRISP-DM method, as being more focused on internal data mining, does not unambiguously state business needs as being the general hub, which may at least partly be due to the fact that internal data retrieval specialists already have a greater understanding of the business data. Furthermore, one major problem regarding outsourcing is often the trust between the parties. The CRISP-DM model includes the creation of a project team; the project team can consist of both internal and external employees (CRISP-DM, 2000). However, the specific issue of trust between those employees is not dealt with. On the other hand, The Three-phases model distinguishes between third party tasks and case company tasks. It allows both parties a great deal a transparency such that vendors cannot be complacent (since they are constantly being monitored) and since the full extent of the data mining process is captured from the beginning, additional costs are reduced (“case companies implicitly know what they are paying for”). To summarize,

the Three-phases method can be classified as a generic process which performs a wide variety of implementation situations of an outsourced knowledge discovery process.

Case study

In order to validate whether the Three-phases model serves its purpose, we applied it to one of the largest Dutch staffing companies. In this case, the trigger was the marketing department of the staffing company, who were indicating a need for improving corporate performance. To be more specific, they stated there was a lot of historical data, and they intuitively presumed it could contain valuable information. Using data mining to retrieve patterns from this historical data to improve the corporate performance was exactly what they needed. They assigned one person to oversee the project. We, as researchers, acted as the third party. The staffing company did not really have specific goals in advance.

Data retrieval

We first analyzed the business needs by reviewing various business documents. These documents consists of system documentation, procedural descriptions about the marketing department and annual reports. Based on the information obtained here, we performed one semi-structured interview with the marketing department representative. We found that the case company divided the Netherlands in five regions, each region having between 50-100 amount of locations. The goal of each location is to acquire as many vacancies as possible. Next, locations need to fulfill as many of these vacancies as possible by allocating employees to these vacancies. A list of actions is defined for the allocation process. Finally, the vacancies are derived from the customers of the staffing company, which are organizations. We created an initial hypothesis. This initial hypothesis can be defined as the starting point regarding the rest of the hypotheses and is used as a basic for a second interview with the marketing department representative:

"A successful allocation depends on the type of actions performed on it".

We were able to retrieve the following list of hypotheses from this second interview, having the key hypothesis in mind:

- ***Vacancies** can be acquired through performing the right **actions**.*
- *There is a large difference between **locations** in the amount of successful **allocations**..*
- *The success of an allocation depends on the **employee** who is allocated.*
- *The success of an allocation depends on the characteristics of an **organization**.*

We retrieved four entities from the hypotheses. We then assigned attributes to these entities. The entities could be matched with corresponding tables in the available database from the case company. Thus, we were able to use the attributes from this database. We then filtered out misspellings and validity errors. The last step regarding the Data Filtering phase was to apply enrichment. We enriched the data set with extra information from the Dutch chamber of commerce regarding organizational branches, which was stored in Excel sheets. This was done to allow for a comparison between the case company and the market. The end result is captured in Table 3.1.

-	*Vacancies*	*(ID, type, date, status)*
-	*Actions*	*(ID, description)*
-	*Locations*	*(ID, name, region)*
-	*Employee*	*(ID, name, education)*
-	*Organization*	*(ID, name, branch)*
-	*Allocation*	*(ID, company ID, employee ID, branch, date)*

Table 3.1: Finalized data subset structure at the end of phase one.

Data mining

We performed a third interview with the staffing company representative in order to capture the specific goals with regard to data mining. From this unstructured interview, we retrieved the following goals:

Goal 1 *(predictive*) Which branches differ in terms of trends compared with the market (when we predict on the terms)

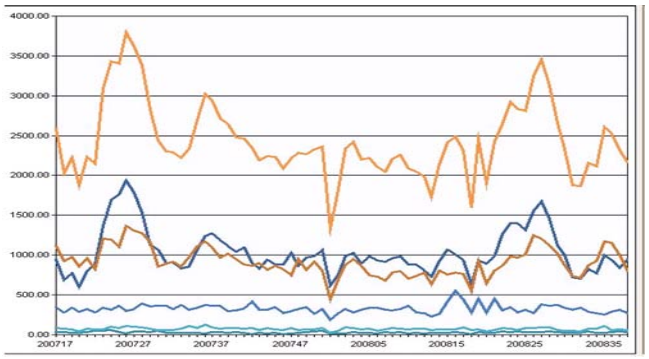
Goal 2 (*descriptive*) Do the actions differ when we categorize between market dependent and independent branches? Market dependent branches are branches which show the same trend as the total market.

Goal 3 *(predictive)* Can we predict on incoming vacancies or allocations?

Based on these goals, we modified our data set in such a way that we could capture all the goals. Actually, we selected all data except locations and employees (we do not need them to answer the goals stated above; however, in case the staffing company needed additional data mining to be performed, then we would already have captured this in our dataset). We used the time series technique on the first goal. We clearly state this is predictive modeling. This is because descriptive modeling does not include the future; we are also interested to see whether the algorithm classifies the branches in terms of market dependency in the future. For the second goal, we used the naïve bayes technique, since we are only interested in the relation between two variables: the description attribute of actions and the market dependency attribute in branches. The third goal was researched by using time-series data mining (i.e. predictive modeling based on a time variable).

Result of goal 1

By comparing the total amount of allocations with each different branch, we derived two branches which were market independent (e.g. tends to react different compared to the market). See figure 3.2. The y-axe shows the amount of fulfilled allocations, the x-axe the timeline. Clearly, branches 4-5 are market independent, since they do not react like the market trend. Note that these branches are by far 'the smallest' compared to branches 1, 2 and 3. This means that the staffing company allocates less in these branches compared to branches 1, 2 and 3. When we zoom in, we do see that the 'normal trend (e.g. less allocations in the end of the year) is the same, but the fluctuating market trend like peaks in 200722 is different. Therefore, these branches are market independent.



Color	Branch
	Total
	Branch 1
	Branch 2
	Branch 3
	Branch 4
	Branch 5

Figure 3.2: Time series analysis: amount of allocations over time in five branches.

Result of goal 2

We then bundled branch 4 and 5 in a type called 'market independent' and bundled branches 1-3 in a type called 'market dependent'. We now performed naïve bayes on this result, to check whether the actions were different between those two types. However, we did not find any difference between the actions performed on them (Table 3.3 for results)

Action	Total performed		Total Actions		Percentage of total	
from 2005-30 till 2008-40	Independent	Dependent	Independent	Dependent	Independent	Dependent
Normal phone call	1147	7231	**28442**	**187316**	4,03	3,86
Call after sales	1190	8386	**28442**	**187316**	4,18	4,48
Mission confirmation INL	1436	12294	**28442**	**187316**	5,05	6,56
Transmission confirmation INL	1468	12148	**28442**	**187316**	5,16	6,49
Optional Phone call	1882	12835	**28442**	**187316**	6,62	6,85
Introduce by using c.v.	4724	31859	**28442**	**187316**	16,61	17,01
Follow-up vacancy	10605	56547	**28442**	**187316**	37,29	30,19

Table 3.3 Naïve Bayes analysis: market dependent versus independent allocations.

We could argue that follow-up vacancy is different for both types (market dependent, independent), since the gap is 7%. However, in both markets follow-up vacancy is performed most, around twice as much as others. We consulted this result with the staffing company, and they state that they did not find the difference applicable, since the actions did not differ that much. This is because the action is only performed after a company has reported a vacancy. It is thus not possible to influence this action (it purely depends on the customer company).

Result of goal 3

The last data mining we performed for the staffing company was to check whether we could predict the amount of incoming vacancies and fulfilled allocations. We created a model which included the total amount of vacancies and allocations per week per year. The data included ranged from week 30, 2005, until week 40, 2008. Based on these

data, we performed a predictive analysis of weeks 41-52, 2008. We then compared the predictions with the actually fulfilled allocations or acquired vacancies (see Figure 3.3). In conjunction with the case company, We choose thresholds the decide whether the prediction is good or bad. Table 3.4 shows the prediction results. If a prediction is within the threshold, it is displayed in green; otherwise, it is shown in red. With respect to allocations, the threshold was set to 95% accuracy. Since this is an analysis of an internal process and not a market comparison, the accuracy rate needs to be relatively high (we analyze the companies process). Regarding vacancies, the prediction accuracy rate was set to 90%, since this aspect is heavily dependent on the market and thus a lower accuracy rate is acceptable.

Week	Difference prediction in % allocations	Difference prediction in % vacancies
200841	1.4	1
200842	-4.7	9.3
200843	-0.5	2.1
200844	5	12.1
200845	2.6	23.8
200846	0	38.4
200847	9.8	29.9
200848	21.8	34.4
200849	19.7	41
200850	11	24.6
200851	27.6	23.7
200852	17.3	-38.8

Table 3.4: Time series analysis: predicting the amount of incoming vacancies and fulfilled allocations.

The results of these mining techniques were presented to the staffing company. They evaluated the predictions with respect to the allocations as being valuable for six weeks. Regarding vacancies, the predictions were labeled usable for four weeks, despite the fact that the last prediction was not completely accurate. This is because the case company were aware that the prediction on vacancies is based on previous year and current trends. Because of the financial crisis, the trend of previous year is heavily different then current year. Therefore, the case company reckons the algorithm simply can't predict 'good', thus they were very satisfied with the results (since it was in fact, nearly 'good').

Results implementation

One unstructured interview was held regarding the AS-IS situation of the staffing company. Furthermore, we visited three locations in order to retrieve the normal procedure regarding allocations. This provided us with information regarding the predictions and forecasting procedure, the allocation procedure and the business preference regarding the solution.

We captured this in a process description (Table 3.5). The business preference regarding the solution was stated very clear; the staffing company did not want a reporting tool, they wanted a report containing the results. Thus, we did not need to align the current situation with the TO-BE situation. We simply needed to present a report.

Process	Impact Level	Monitoring
Predict on allocations. Based on previous allocations, a trend is created. This is done in Excel. The prediction is based on experience. Results are not validated.	Operational level	Performed once a month, only used 'how we are doing'. Process is not transparent; only performed by one person.
Create market campaigns. Based on preferences of marketing department	Tactical level	Performed not based on time; Effect visible, but not measured.
Forecasting. Top level management predictions about how the staffing company is likely to do. Based on market trends, compared with previous years.	Strategic level	Performed every year; evaluated every year. No control or monitoring on this process.

Table 3.5 Current staffing company process description

We selected the results from the predictive data mining phase, which were in fact all results since the descriptive results could not be applied. The descriptive results did not affect the organization and where thus excluded. Based on business preference, the requested output was to be a report. By analyzing the current AS-IS situation regarding

the predictive data mining result, we distinguished three different processes (Figure 3.5). We proposed the following changes regarding these processes:

- Predict on allocations: This should change to an automated process, reducing time to perform the process and creating a transparent situation so that every end-user can interpret the predictions.
- Create marketing campaigns: Should be based on predictions; thus, trends are being altered pro-actively. Effect of current marketing campaigns should also be data mined.
- Forecasting: Once every quarter of a year, check current trend and historical trends. Not much research has been done in this respect within the staffing company. People are reluctant to automate the process. The operational level needs to be changed first in order for the strategic level decision makers to become convinced.

Recapitulation

This paper has introduced a method which, unlike current dominating methods, explicitly describes the process of data mining from an outsourcing perspective. Our Three-phases model provides an answer to the problem of how to outsource data mining processes to help improve corporate performance. The model's key strengths can be captured as follows:

(1) Data source independent data. We propose the use of hypotheses before any source of data has been analyzed. Thus, it is easier to communicate with a case company (since we are talking about research aims, not data structures) and it facilitates a more data source-independent approach as well.

(2) Counters data quality issues. We embed a matching and a validity filter in our method for filtering out data quality errors. However, this directly uncovers a notable disadvantage of data mining from an outsourcing perspective; errors can only be filtered out after the fact, since the data is provided by the company. Internally executed data mining processes put more pressure on companies to minimize the amount of 'false data' before it enters the database. On the other hand, companies will always thrive to keep their data clean, since 'blurred' data leads to inefficiency, which, in turn, ultimately leads to less profit. Thus, in an ideal world, it should not make a big difference. Furthermore, we do not need to filter out all errors, since data mining concentrates on 'the big picture'; one or even ten errors will not mean much when we analyze more than (for instance) 300.000 rows. On the other hand, the current approach is already robust in such a way that it filters out most errors and allows for heavy modifications in terms of data enrichment. That being said, we believe that the described approach presented in the current paper will led to a high quality data independent data set.

(3) Narrows down the choice regarding the appropriate data mining technique. As mentioned earlier, there are hundreds of data mining techniques available. However, it is our opinion that first classifying the data mining techniques and then analyzing the most common techniques according to those classes provides a solid foundation to perform data mining, even though we do not cover all techniques. We have introduced a third category (in addition to the well-known descriptive versus predictive categorization by Fayyad et al. (1996)) which we named 'rule-based data mining'. In our belief, rules are an independent category and can serve as both input for predictive and descriptive data mining as well as stand-alone data mining (in other words, we perform rule-based data mining). We believe that nearly all techniques can be captured according to those three classes of data mining. The described method fragment in Figure 2.6 includes seven common data mining techniques, which were derived by comparing four different data mining tools and then selecting the common techniques.

(4) Embraces proven concepts to measure corporate performance. We incorporate the BITA model of Scheper (2002) to measure company performance in terms of business domains. We are aware that we do not measure performance in terms of financial data or other 'hard numbers'. This is not how we interpret performance regarding data mining. We want to measure current procedures and see if data mining results can optimize them; we do not need (for instance) revenue data to achieve this objective.

(5) Standardized method, but allows for flexibility. The approach to outsource data mining is standardized in terms of the three-phases model. However, in our model, we do not force the researcher or case company to use a specific data mining technique. Furthermore, we allow for two different types of implementation. We also do not capture the use of a specific tool or a specific reporting tool.

To test our method in terms of completeness, robustness and applicability we have applied it to our case company. We found that the method serves our purpose rather well. However, with regard to implementation, we still think the method can be made more mature in terms of standardized business documents (templates) for describing the business. To conclude, further research would be desirable to better capture data mining purposes.

References

Aalst van der W.M.P, Weijters A.J.M.M (2004). Process mining: a research agenda*. Computers in Industry,* 53 (3), pp. 231-244.

Apte C, Weiss S (1997). Data mining with decision trees and decision rules. *Future generation computer science,* 13, pp. 197-210.

Bardhan AD, Kroll C., (2003). *The new wave of outsourcing.* Fisher Center Research Reports.

Berry M, Linoff G (1997). Data mining techniques; for Marketing, Sales, and Customer support. *Data mining methodology*, pp. 65-93. Wiley computer publishing.

Box G., Jenkins G. (1994). *Time Series Analysis: Forecasting & Control.* University of Wisconsin. Prentice Hall.

Brinkkemper, S. (1996). Method engineering: engineering of information systems development methods and tools. *Information and Software Technology 38*(4). pp. 275-280.

Chapman P, Clinton L, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R (1999). Crisp-DM 1.0. Step-by-step data mining guide. *Published by SPSS Inc.*

Craven M.W., Shavlik J. W. (1997). Using neural networks for data mining. *Elsevier Future generation computer systems*. ,13, 2-3 , pp. 211-229.

Deloitte, 2005. Calling a Change in the Outsourcing Market. *Deloitte Development LLC.*

Driel v H. (1999). *Oude nieuwe media.* E-view, 99 (1).

Duntsch I., Gediga G., (1998). Uncertainty measures of rough set prediction. *Artificial Intelligence*, 106(1), 109-137.

Earl J.M. (1996). The risks of Outsourcing It. *Sloan management review,* 37(3). 26-32

English, L., (Ed.) (1999). *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. Wiley computer publishing.

Everink J., (1999). Resultaat gericht onderzoek. Holistisch onderzoek en leefbaarheid (3). *Bureau Everink.*

Fayyad U, Piatestky-Shapiro G, Smyth P (1996). From data mining to knowledge discovery in databases. *American association for artificial intelligence*. 1, pp. 1-34.

Buytendijk, F (1999). 'CPM Helps Build the High-Performance Organization'*. The Gartner group* , 1-46.

Golfarelli M., Rizzi S., Cella L. (2004). Beyond data warehousing; what's next in business intelligence? *Data Warehousing and OLAP*. 1, pp. 1-6.

Halle von B (2001). Business Rules Applied: Building Better Systems using the Business Rules Approach. IBM system journals, 31 (3).

Harmsen, F., Brinkkemper, S., Oei, J.L.H. (1994). Situational method engineering for informational system project approaches. *In: Methods and Associated Tools for the Information Systems Life*. Pp. 169–194

Huang, Z. (1997) Clustering Large Data Sets with Mixed Numeric and Categorical Values. *In: Proceedings of The First Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Singapore, World Scientific.

Inmon WH (1996). The data warehouse and data mining. *Communications of the ACM.* 39, (11). pp. 49-50.

Jain A.K, Murty M.N, Flynn P.J. (1999). Data Clustering: A review. *ACM Computer Survey*, 31 (3).

Kachigan S.K. (2005). *Multivariate Statistical Analysis, A Conceptual Introduction*. New York. Radius Press.

KDnuggets (2007). *Poll: Data Mining Methodology*. Retrieved 16:50, February 23, 2009, from http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm.

KDnuggets (2009). *Poll: Data Mining Tools Used*. Retrieved 16:50, June 2, 2009, from http://www.kdnuggets.com/polls/2009/data-mining-tools-used.htm.

Keogh E., Kasetti S (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Journal data mining and knowledge discovery*. 7(4), pp. 349-371.

Kueng P., Kawalek P. (1997). Goal-based business process models: creation and evaluation. Business *process management journal,* 3 (1). pp.17-38.

Lee, M.L., Lu, H., Ling, T.W., Ko, Y.T.(1999): *Cleansing Data for Mining and Warehousing*. In proceedings of the 10th Intl. Conference on Database and Expert Systems Applications. Singapore, DEXA.

Loh, Lawrence and Venkatraman, N. (1995). *An Empirical Study of Information Technology Outsourcing: Benefits, Risks, and Performance Implications.* In proceedings of the international conference on information systems. Amsterdam. (ICIS).

Microsoft (2008). Microsoft sql 2008 server. *Microsoft product information on sql server 2008.*

Neun M., Weibel R., Burghardt D. (2004). *Data enrichment for adaptive generalization*. ICA workshop on generalisation and multiple representation. Leicester.

Oracle (2007). Oracle enterprise manager. *Oracle product information on the ODM Business suite 2007.*

Pregibon, D. (1997). *Data Mining*. Statistical Computing and Graphics, 7. pp.8)

Quinlan J.R., (1983). Induction of decision trees. .Learning efficient classification procedures. *Machine Learning: An Artificial Intelligence Approach*. 1, pp. 81–106.

Quinn J.B., Outsourcing Innovation: The New Engine of Growth. *Sloan Management review.* 41(4), 13

Rahm E, Do H H (2000). Data Cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 1, pp. 1-11.

Rouse, A.C., Corbitt, B., (2003). Revisiting IT outsourcing risks: analysis of a survey of Australia's top 1000

organizations. *In proceedings of the 14th Australasian Conference on Information Systems*. Perth.

Sambamurthy, V., Straub, D.W., Watson, R.T., 2001. Information technology managing in the digital era. *Information Technology and the Future Enterprise, New models for Managers*. Prentice Hall, 1, pp. 281–305.

Scheper, W.J. (2002). Business IT Alignment: solution for the productivity paradox (In Dutch). *Deloitte & Touche, Netherlands.*

Schmidt, G., & Wilhelm, W. (2000). Strategic, tactical and operational decisions in multi-national logistics networks. A review and discussion of modeling issues. *International Journal of Production Research*, 38.

Shaw M, Subramaniam C, Woo Tan G, Welge M (2001). Knowledge management and data mining for marketing. *Decision support systems*. 31 (1). pp. 127-137.

Srikant, R., Vu, Q., and Agrawal, R. 1997. *Mining association rules with item constraints.* In Proceedings of the Third International Conference on Knowledge Discovery in Databases and Data Mining, New York.

Yu P., Coa L, Zhang C., Zhao Y., Williams G. (2007). *Domain driven data mining*. Proceedings of the 2007 international workshop on Domain driven data mining. San Jose.

Author biographies

Arjen Vleugel, MSc, is a Business Intelligence consultant. He graduated in the area of data mining. At the moment, he is employed at In Summa B.V. in Raamsdonksveer, an organization specialized in data warehousing and other business intelligence solutions like data mining. Other scientific research include a co-authorship on a scientific paper called “Mining E-mail to leverage knowledge networks in organizations".

Dr. Marco Spruit is an Assistant Professor in the Organisation & Information research group at the Institute of Information and Computing Sciences of Utrecht University. His information systems research revolves around Knowledge Discovery processes to help achieve organizational goals through Data Mining techniques, Business intelligence methods, Linguistic Engineering techniques and Web 2.0 technologies. Additionally, he investigates Information Security models and Cloud Computing frameworks as infrastructural safeguards and enablers for Knowledge Discovery processes. Marco initiated his Knowledge Discovery research agenda while performing his PhD in Quantitative Linguistics at the University of Amsterdam. In 2005 he was awarded an ALLC Bursary Award for this work.

Ir. Anton van Daal is chief executive officer at In Summa B.V. in Raamsdonksveer. He started his career in 1996 at Jac van Ham, where he implemented various ICT systems. He then switched to JVH Gaming production, in order to set up the automatic departments of the company. In 2001, he co-founded In Summa. As CEO of the company, he oversees the development of the management information solution of In Summa, which is called the webdashboard (which itself is based on OLAP).