DATA QUALITY IN THE CONTEXT OF CUSTOMER SEGMENTATION

Dirk Arndt, Norman Langbein
DaimlerChrysler AG, Research & Technology, Information Mining (RIC/AM)
P.O. BOX 2360, 89013 Ulm, Germany
{dirk.arndt, norman.langbein}@daimlerchrysler.com

Abstract: There's no such thing as data quality in general: it all depends on the particular business application [10]. With this in mind, we begin by introducing customer segmentation as a basic data mining application in analytical CRM. We describe the structure of the data mining environment and outline data selection as an essential, but difficult sub-problem. From there, we go on to develop three perspectives for approaching the issue: marketing, data mining, and data quality. Then we introduce a process model for the systematic identification of input variables for customer segmentation. Additionally, we offer practical suggestions and set out difficulties and requirements for further development.

Key Words: customer segmentation, clustering, data quality, CRM

1 INTRODUCTION

Today, customer relationship management (CRM) is one of the key issues in the field of marketing. CRM is a customer-focused approach aimed at acquiring new customers, linking them to the enterprise, and recovering them if they should become disloyal. Thus, we may organize CRM along the customer lifecycle (CLC), distinguishing between acquisition, loyalty, and recovery programs [1]. Furthermore, CRM activities may be grouped into three categories: (1) strategic CRM (sCRM), which comprises all the actions linked to long-term strategies in terms of marketing approaches and the internal CRM infrastructure, (2) operational CRM (oCRM), which includes all the activities occupied with customer contact and related internal business processes, and (3) analytical CRM (aCRM), which provides all the components needed to gather, store, and analyze customer-related data. Thus, aCRM creates the information base for well-aimed oCRM activities.

Using the CRM concept, enterprises are dedicated to delivering personalized products and services to their customers. For companies such as DaimlerChrysler which are faced with huge mass markets and, for the most part, standardized products, this mission is difficult to accomplish. Yet modern information technology enables such organizations to gather knowledge about their customers which can then be exploited to integrate customized elements into what are basically uniform products and services [14]. Therefore, a burning issue in aCRM is to find and describe homogenous customer segments which can later be treated with fine-tuned oCRM instruments. In this way, customer segmentation forms the basis of customer insights in the loyalty program [7]: it builds the core organizational structure for the development of mass customizing approaches.

There are two ways to approach customer segmentation. The first is business-driven with cluster segments described prior to clustering (e.g. based on product categories or sales areas). The second approach is data-driven customer segmentation, which can be performed by using various statistical or data mining techniques. In the latter, the literature sometimes distinguishes between algorithms geared for predictive segmentation and algorithms targeting clustering [5]. We, however, do not share this view. Assuming that previously defined classes exist and decision tree algorithms are employed to assign customers to these classes, we consider such tasks as belonging to classification and, therefore, to supervised learning. For us, segmentation targets grouping customers who show similarities in multiple dimensions. So we run a cluster analysis. Clustering is unsupervised in nature. And, for unsupervised systems such as k-means or Kohonen networks, there is no need to train or construct a classifier since the information is organized into groups based on common characteristics. There are a wide variety of clustering algorithms available on the market [13]. Nevertheless, it is common practice to use clustering algorithms in combination with other techniques such as regression and factor analysis [19], which in this context, are largely used for data preparation and definition of clustering parameters.

If customer segmentation is performed in a real business environment, it often needs to be applied to central or virtual data warehouses or to a variety of distributed databases. However, an enormous amount of intra-organizational and outside customer-related data tends to be available, and this may differ greatly both in quantity and quality [1, 11]. The challenge in practice is to find the most promising input data for the application and to apply suitable clustering algorithm(s). In this article we introduce an approach that does just this.

The approach was developed and tested in the Information Mining Department at DaimlerChrysler Research and Technology. We start by describing the overall task in section 2. Next, we explain the different perspectives for viewing clustering. The different kinds of customer data which can be used as input data for clustering are briefly set out in section 4. Section 5 presents a process model for identifying clustering variables from large databases. We conclude by summarizing the key points and outlining open issues.

2 THE TASK

Cluster analysis and, as we follow, customer segmentation may therefore be viewed as a data mining project. The data mining process consists of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. With respect to the detailed process model, we refer to the CRISP-DM approach [6]. According to CRISP-DM, the initial step in any data mining project is to define the detailed business objectives. This means defining the criteria for a successful outcome from the business perspective. After this first step, the process becomes highly iterative because there are many factors that are strongly correlated, which influences the final performance of the data mining project (cluster analysis). And this means that we need to go back and forth between the single phases and tasks of the data mining process.

The influencing factors can be arranged in a four-dimensional environment. All four dimensions and their respective subcategories interwork closely and the various combinations result in a wide range of the modeling performance. Figure 1 below depicts the overall system.

The *target variable dimension* exists solely for supervised modeling approaches where it is most difficult to define the best output variable (e.g. classifier) from the outset [11]. Since we concentrate on unsupervised systems here (see section 1), we do not address the issue further.

Another dimension is the *problem type dimension*. This refers to the business goals of the analysis on the one hand and includes the related data mining problem types on the other. There are various data mining problem types cited in the literature [see e.g. 3, 22]. Chiefly, we differentiate between clustering, classification, prediction, association, deviation recognition, and data description. The same business problem may frequently be solved through the application of different data mining problem types. They may be applied separately or in combination. As we have set out in section one, we focus on cluster analysis as the basic approach for customer segmentation. If other problem types and corresponding algorithms are employed, this is done in order to optimize and understand the final cluster results.

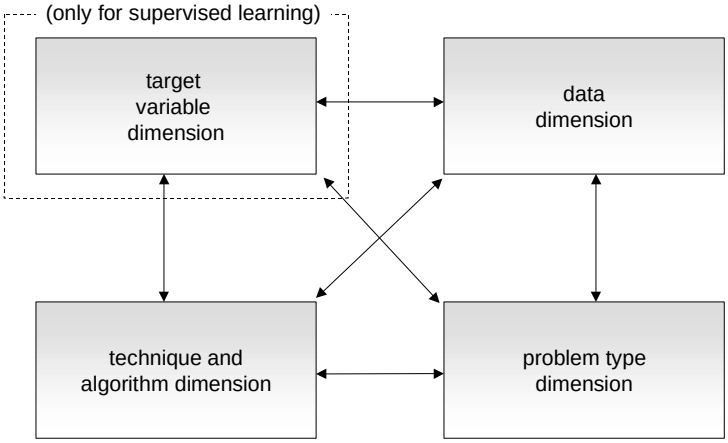

Figure 1: The General Data Mining Environment.

There are several data mining techniques and algorithms available for each of the data mining problem types. This is reflected in the *technique and algorithm dimension*. In fact, there are a multitude of mathematical approaches for finding clusters in data, and entire tomes are dedicated to the subject [19, 3]. The three main clustering techniques are divisive methods, agglomerative methods, and self-organizing maps. Numerous algorithms are available for these techniques. For example, using the agglomerative methods alone we are able to choose between single linkage, average linkage, complete linkage, centroid, median, and ward-based algorithms.

Finally, the fourth dimension, *data dimension*, is used to find the "right" internal and external data for a certain application. Most difficult to accomplish, this is done in three steps: (1) making selections from the different kinds of data (e.g. communication data, demographic data, lifestyle data), (2) deciding in which of the different databases these data can be found, and (3) choosing single attributes and records within these databases [2].

For the business problem at hand, we concentrated on two dimensions. Since the target variable dimension does not exist for unsupervised data mining techniques and the problem type dimension in our case was set as cluster analysis, this left the technique and algorithm dimension and the data dimension for fine-tuning. Most data mining algorithms are designed with specific data mining goals and specific

types of input data in mind. This is also true for clustering algorithms. Modern research focuses on scalability of algorithms, high-dimensional clustering techniques, the effectiveness of clustering complex shapes of data, proper data preparation and selection, and methods for clustering mixed numerical and categorical data [13]. Under these circumstances it is clear that there is more than one way to effectively harmonize the two dimensions. For practical reasons, we must build a heuristic in order to find the most useful data and apply the best algorithms available.

Again, there are two general approaches: we can either search for or develop special algorithms for relatively fixed data or look for data which best fit a given set of algorithms. How well the data fit the algorithms and how the fit can be measured typically depends on the business goals (see subsection 3.1). In practice we often lack the time and resources to develop or modify complicated data mining algorithms. But we normally deploy standard, commercially available algorithms as a matter of course [23]. Thus, for the development of our approach we preferred the second alternative. And when attempting to find the data with the best fit we are again confronted with two options. We may choose to select the best fitting variables (so-called *forward selection*) or exclude the most unpromising ones (*backward elimination*). For our process model, we chose to go in both directions.

Typically, decisions concerning the fine-tuning the data and the technique and algorithm dimensions are made under certain constraints resulting from limited time and resources. For this reason, we aimed to develop a standard process which would help make the procedure in future projects more reliable, less expensive, replicable, and faster. This can be accomplished, for example, through the development of an experience base and the implementation of tools such as checklists, software modules or user guides. All of these tools must, of course, be connected to the relevant steps of the process model.

3 PERSPECTIVES FOR DATA ASSESSMENT

In section 1 we mentioned that there are a great variety of clustering algorithms. Even when applied to the same data set, the various algorithms may produce very different results. This is due to the fact that the algorithms use different similarity measures or optimizing methods [23]. Thus, we recommend applying more than one clustering algorithm to the same data sample. We view the cluster results as satisfactory if two or more algorithms yield outputs that look reasonably similar [19]. Now, we must find general criteria which reasonably suit all the algorithms available to us and which are considered for usage during modeling.

We must also take into account that, when clustering algorithms are applied to larger data sets, performance often declines significantly [17]. Using more than one clustering algorithm further impacts performance [19]. Therefore, it makes sense to work with samples, select the smallest attribute set, and use only two or three clustering algorithms. For this reason, we recommend reducing the input data considerably before applying the clustering algorithms.

When evaluating potential selection and deselection criteria, three perspectives should be considered: the marketing perspective, the technique and algorithm perspective, and the data quality perspective. Each perspective covers several criteria which may be unique to only one perspective or may emerge repeatedly. We explain each perspective singly in the following subsections.

3.1 The Marketing Perspective

As any data analysis must be based on the underlying business context, it is reasonable to look at the problem from a marketing perspective. As stated, we view customer segmentation as a basic framework for the development of powerful loyalty programs in CRM. We aim to treat the different segments with suitable oCRM tools. Therefore, we generally require that the individuals within a segment be as similar as possible (*intra-group homogeneity*) and that the differences between the segments be as distinct as possible (*inter-group heterogeneity*) [3]. Such segmentation enables us to recognize and describe different groups of similar customer profiles and, with this knowledge, we are capable of developing customized services and communication for the segments we are interested in. In order to tailor the oCRM instruments, we have four additional requirements concerning the segments to be built [20]. From a marketing point of view, the segments should fulfill the following prerequisites (see table 1 below):

Requirements for Segments	**Description**
Addressable	The individual segments must be reached through the use of oCRM instruments.
Describable	This allows us to understand the specifics of each segment. The segmentation criteria (variables) have to be aligned with the business context. If, for example, we plan to establish a special service program for heavy users in the car industry, we certainly should be able to describe the segments with attributes such as annual mileage, service periods, and number of warranty and ex gratia claims.
Time-stable	As a certain amount of time is required to develop and carry out marketing plans and to measure the results for chosen segments, the segments must be stable for a pre-defined time frame.
Of sufficient size	In order to treat the customer segments cost-effectively, they must contain a minimum number of individuals. The number is based on the profit and cost calculation for each segment. Naturally, this requirement is inconsistent with the requirement of intra-group homogeneity, so we must carefully balance the ratio.

Table 1: Requirements for Customer Segments from a Marketing Perspective.

3.2 The Data Mining Perspective

As set out in section 2, the challenge from the data mining perspective is to fine-tune the technique and algorithm and the data dimensions – always keeping the business objectives in mind. In the literature there are several requirements placed on clustering algorithms and the relevant input data [5, 19, 3, 13, 20]. It is quite difficult to determine whether these needs address the input data or the algorithms. Yet if we assume that several algorithms will be applied and that we have a suitable way of combining the results (see above), we may focus solely on the requirements for input data. From all the lists we found in literature, we have determined five requirements as being most practical, see table 2.

Requirements for Input Data	**Description**
Scalability	As clustering algorithms are more reliable and efficient when used on smaller data sets, we should work with small but statistically reasonable samples.
Similarity	Since combining variables of different scales is not only labor-

	intensive but also error-prone, we try to work with variables of similar scales.
Multicolinearity	In order to keep the total number of input variables low, the variables should not be correlated with each other. Thus, we try to avoid dependent variables.
Significance	As unique or nearly unbiased variables do not contain discriminating information, they may be excluded from the outset.
Absence of Noise	Most real-world data contain outliers or missing, unknown, or erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.

Table 2: Requirements for Input Data from a Data Mining Perspective.

3.3 The Data Quality Perspective

A third perspective, and one which we found most interesting, is the data quality perspective. In this context we understand data quality as "consistently meeting the information customers "expectations" [9]. In our case, these information customers are the CRM managers who develop the oCRM activities (program). Looking at the selection problem from this perspective, potential data for cluster segmentation must fulfil general quality requirements in order to achieve segmentation results of high value for oCRM activities.

The measurement of data quality can be a powerful tool for the extraction of data prior to the cluster analysis. When evaluating data quality, there are two fundamentally different approaches. One is to make an independent assessment based on characteristics of the data, whereas the other is to evaluate the data quality by applying data mining technique(s). In-depth research on this topic has been performed, for example, by the MIT Sloan School of Management [21] or DaimlerChrysler Research and Technology [12].

The assessment based on characteristics of data is usually a fast and reasonable way to obtain an initial impression of the data. Literature cites many criteria for characterization of data [18, 10, 15]. Given our selection task, we consider five data quality requirements as paramount. Table 3 provides an overview of our requirements:

Data quality requirements	**Description**
Completeness	This criterion necessitates that all objects of a variable have a defined value. If segments cannot be described as complete due to missing values, errors within oCRM activities are likely.
Relevancy	This criterion requires that the information content of a variable correspond with the information needs of a request. If data used for clustering is not relevant to a specific business context, for example, it is unlikely that the segmentation will be able to help customize product and service offerings (see also 3.1).
Correctness	The material condition (specified condition) must correspond to the condition in the database (actual condition). A clustering which is carried out with invalid data will possibly result in segments which are also invalid.
Clarity	The data need to be interpreted properly. A prerequisite for clarity is the existence of sufficient documentation [11].

Timeliness	The values of a stored variable must be up-to-date in relation to the actual, real-world values. If timeliness is not taken into account, the clustering and/or the description of the segments will be based on old data, which may have a significant impact on their temporal stability.

Table 3: Requirements for Input Data from a Data Quality Perspective.

Employing a data mining technique on data for data quality purposes can be viewed as data quality mining. The goal of data quality mining is to detect, quantify, explain, and correct data quality deficiencies in large databases. There are several approaches for data quality mining [16]. When applying data quality mining, we use different evaluation techniques but largely investigate the same data quality criteria. For these reasons, we do not go into more detail here.

4 CUSTOMER-RELATED DATA

Whether we are looking at data warehouses or distributed databases, as far as the context goes, we can always divide customer-related data into four groups: identification data, communication data, descriptive data, and purchase data.

Identification data are data such as the name and address of a person. They are needed to contact the people within the generated segments and may not be considered in connection with the possible outcome quality of the cluster analysis. Nevertheless, the quality of identification data themselves is crucial for the success of CRM activities and must be monitored carefully [8].

Communication data result from the interactions between a customer and an enterprise. They include attributes (variables) such as the last or first contact date, the reason for contact, the preferred channels of communication, and the number of service inquiries per year. Within CRM (marketing) they play a significant role because they provide a close and up-to-date view of a customer on a very individual level and reflect the customer's relationship to the organization. Nowadays, the importance of the Internet as a communication channel and customer touchpoint is soaring. This is due to the opportunities for monitoring customer behavior (e.g. click-stream analysis) when clicking through the web site. All the data resulting from this click-stream can be saved and further analyzed and individual offers can be made in real time based on actual customer requests and monitored data.

In the category of *descriptive data* we summarize all the data which can be used for the general description of customers and which are not communication or identification data. This is the most voluminous category and may contain some of the data listed below [see also 1]:

- *Demographic* data (e.g. gender, age, income).
- *Psychographic* data (e.g. hobbies, interests).
- *Behavioristic* data (e.g. purchase history, purchase frequency).
- *Event-driven* data (e.g. marriage, birth of child, recent move).

Purchase data include information about the purchase history of a customer: for example, the sales revenue generated, the loan history, the preferred product categories or the number of items sold. These data are vital in describing the monetary value of customer segments and in developing tailored marketing strategies.

All subcategories may include various data fields (variables). The data of a category are collected at several customer touchpoints within the enterprise [11]. Unfortunately they often need to be bought from external data providers [1, 11], which can be expensive. These providers offer a wide variety of different data [2]. In spite of all this, descriptive data are updated most infrequently and vary much more in terms of data format and data quality than the two categories described previously.

Figure 2 illustrates the issues addressed in this section.

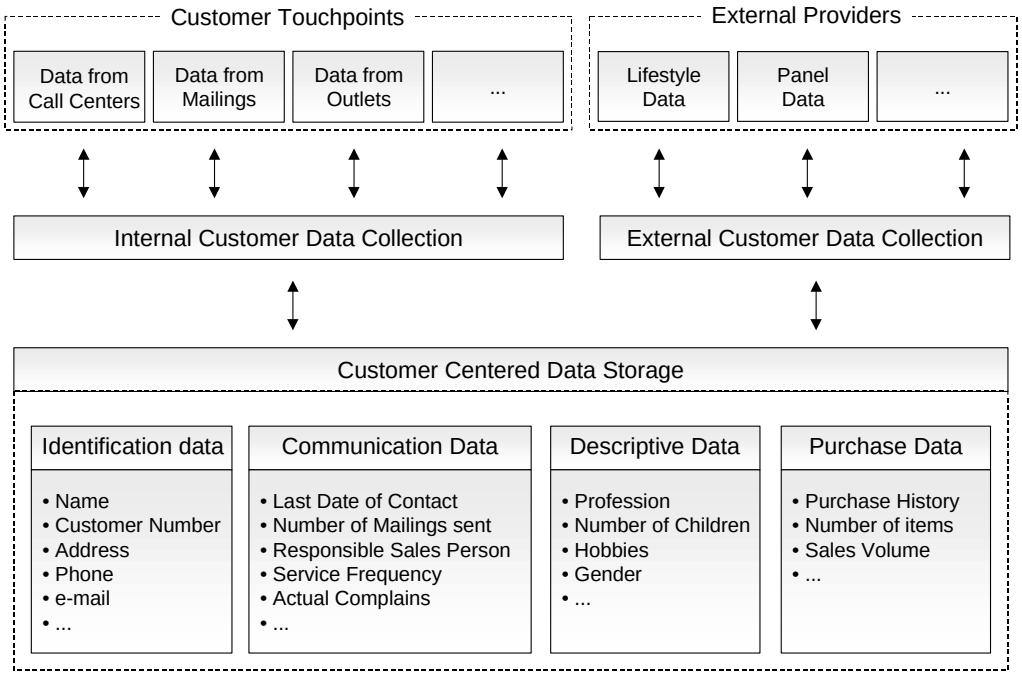

Figure 2: Customer Data Systematic.

5 PROCESS MODEL FOR IDENTIFICATION OF ACTIVE VARIABLES FOR CLUSTERING

As outlined in section 2, it can be difficult to find the appropriate variables for customer segmentation in large or distributed databases. Enormous quantities of internal and external data, which differ in terms of quantity and quality, are usually available. On the other hand, the purpose of data collection may sometimes vary from the purpose of data usage. For example, data that were originally collected for use with controlling applications oftentimes end up being used for marketing purposes. As a result, identifying appropriate variables for customer segmentation is frequently time consuming and costly, and there is a high risk of selecting ill-suited variables. Faced with these challenges in practice, we have developed an approach based on the different perspectives as outlined in section 3 and the customer data as set out in section 4. The goal of the approach is to make the identification process more reliable, less expensive, and replicable. But before introducing the process model, we consider two further issues.

In cluster analysis we distinguish between so-called active variables and passive variables. Active variables are the input variables for the clustering algorithm. Passive variables are also connected to the individuals to be clustered but they are not actively used during modeling. Instead, we try to use these variables after modeling to describe the customer segments that have been built. There is no general rule

of how to decide whether a variable should be active or passive. Normally, the basic decision is made from a business point of view. If, for example, we run the segmentation in order to treat certain customer segments with individualized communication instruments, we select communication-related variables for active use. But if we lack the necessary data quality (e.g. in terms of completeness or timeliness) we might not be able to utilize some variables as input variables for clustering. Hence, we apply them to the clusters after the analysis has been carried out. Nevertheless, since we aim to use the most suitable data in the clustering algorithm, we focus here on the selection of active variables.

In section 4 we distinguished between four categories of data: identification data, communication data, descriptive data, and purchase data. In the following we make the assumption that communication and purchase data may also be viewed as descriptive data. The chief difference between "normal" descriptive data and the other two data categories is that the latter result from the exchange between the customer and the organization (see section 4). This is important during data selection for predictive modeling in the acquisition program [11] and for working with the customer segments built. However, the difference does not play a role when choosing features prior to modeling.

When developing a process model for systematic feature (variable) selection (prior to modeling), the main objective is reliable, fast, and cost-effective selection of variables. Hence, we need to group and rank the criteria from all three perspectives with respect to these basic demands. In doing so, we developed a three-step process model. The three steps are (1) evaluation of the business context, (2) evaluation of time aspects, and (3) data profiling. As a result of the process, the number of active variables to be used for clustering is identified from a large number of available variables. Before explaining the details in the subsections below, we introduce the overall model in figure 3.

Within the process model we start by using the backward elimination approach (see section 2) since, when dealing with data from various customer touchpoints, there is a great probability that key variables will be missed if we employ forward selection [11].

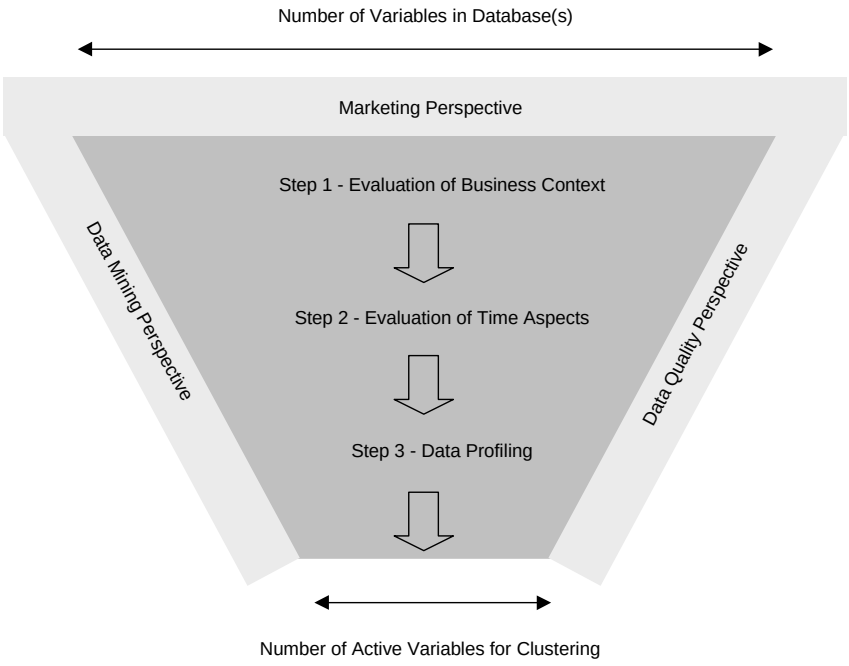

Figure 3: Process Model for Identification of Input Variables.

5.1 Evaluation of the Business Context

The initial step of our process model is the examination of the business context. Looking at the business context first enables us to eliminate a large number of irrelevant attributes from scratch. We therefore save resources as fewer variables need to be treated in the subsequent steps, which are more time consuming and costly. When determining the context that could be important for a particular project, we recommend organizing a joint workshop with the marketing and IT departments. Here, the specific fields of contextual interest are identified and verified. What are the criteria and where do they come from?

An active variable must have a logical connection to the underlying business problem, and it must be relevant for the achievement of the defined objectives of the planned oCRM activities (e.g. the foundation of tailored customer clubs). This is required not only from the marketing perspective but also from the data quality perspective (see section 3). From the marketing perspective, relevance is reflected in the requirement that the individual segments be describable. Therefore, the CRM experts at the workshop must state the intended marketing objectives and determine the resulting context requirements. From the data quality perspective, the relevance of a variable indicates its importance and necessity. The IT department must describe the actual status of the data and its understanding of the data context. Here, we often detect principal misunderstandings, misuses, and misinterpretations of data fields.

In this context, clarity is another aspect of the data quality perspective which needs to be considered. Clarity of a variable indicates that the underlying metadata, which specify the semantics of the variable, are of high quality, allowing the variable to be clearly interpreted. If the meaning and content of a variable are not clear then there is danger of information being interpreted incorrectly. Consequently, the efficiency of accompanying CRM activities is jeopardized. In practice we often do not have sufficient metadata. If this is the case, we must select a reasonable number of promising variables (using the best context descriptions available), compare the description with the real-world conditions and, in case of considerable mismatches, develop new metadata. While this might be very time consuming, it is necessary for the reason mentioned above.

The proper evaluation of the contextual criteria is relatively difficult and cannot be done automatically. Contextual fit is not really scalable and there are no easy-to-use or simple measures such as those that exist for completeness of a data field, for example. To tackle the task, we recommend generating a nominal scale (e.g. unimportant, important, and most important). Then, the experts assign these values to the variables (independently from each other) and the combination of the results leads to the final ranking. Next, we select the best variables and check if the meta-information is clear and reliable.

If a variable does not fulfil the requirements of contextual fit to the business problem and clarity, it has to be excluded as an active input variable, since in this case it makes no contribution to the achievement of the business goals. As a result of step 1, a large number of variables may usually be excluded with a reasonable amount of time and workpower.

5.2 Evaluation of Time Aspects

If a variable is logically related to the business context, we next propose looking at time aspects. Out-of-date or unstable information is generally not valuable, independent of all other data characteristics. Similarly, time aspects are typically obtainable from the metadata we have already inspected or gathered

in step one. The time aspects in question are *time stability* and *timeliness.* An active variable should be temporally stable and show a high degree of timeliness (see sections 3.1 and 3.3).

From the marketing perspective, the segments need to be temporally stable for a pre-defined time frame, because it takes a certain amount of time to plan and carry out oCRM activities. This means that the contents of active variables used for clustering should not change over a defined minimum period of time. This is of particular importance when dealing with large customer databases, since it is not economical in most cases to continuously carry out customer segmentations in short time intervals. For measurement of time stability we compare the pre-defined period with the update frequencies and the frequency of changes in reality. If there are database updates during the pre-defined period, a clear violation of the required time stability is established and the corresponding variables are excluded or must be transformed (see next paragraph). Second, we check the real-world conditions. This is also necessary because the real world values sometimes change faster than the data is updated within the enterprise.

If a variable is of high contextual relevance but violates the requirement of time stability, we found the solution of determining a new variable which ultimately meets the requirement useful in practice. The number of overall purchases a customer makes may, for example, change on a daily basis. But if we determine the number of purchases up to a certain date, the new variable is generally valid for all possible periods of time. On the other hand, because the context of the information is changed, this solution may yield lower business relevance.

From the data quality perspective, timeliness is required. This means that the values of a stored variable are up to date relative to the actual real-world values. Here we view not the update frequency but the general validity at a certain point in time. If a variable has expired it cannot be used for customer segmentation. The information may be obtained from the metadata, reasoned from the business environment or verified by using sampling techniques.

Problems can occur if only individual values of a variable change at different points in time. In this case we recommend executing a data profiling on the corresponding timestamps (if they exist). As a result of the data profiling, measures such as average, minimum, maximum, and standard deviation can be documented. These results enable the quality of timeliness to be estimated. A slight variation of the standard deviation of the timestamps suggests, for example, that most of the records were updated within a small period of time. If the timeliness of a variable seems to be very uncertain, we normally try to avoid using it as an input variable.

If time stability and timeliness cannot be achieved, a variable should be excluded as an active variable for clustering. However, it might still be used after the analysis as a passive variable for describing the different clusters.

5.3 Data Profiling

In the last step we recommend executing a data profiling. Data profiling focuses on the instant analysis of individual attributes within a database and provides an overview of data type, length, value range, means, variance, standard derivation, frequency of discrete values, frequency of null values, etc. [18]. These basic measures can help to form a general impression of the data. For example, they are employed to estimate. the discrimination potential of a variable for clustering.

From both the data mining and the data quality perspectives, it is required that a variable be significant and contain discriminating information for clustering. The significance can be estimated, for example,

through an analysis of dispersion and position parameters depending on the scale level of a variable. If a variable possesses interval scale level with many different values, a small standard deviation suggests that the values differ only slightly from the average value and the variable thus has no or only a small discrimination potential for the formation of customer segments.

A further criterion from the data mining perspective is that the scales of variables used for clustering should be similar. Scale similarity can be measured by looking at the data type (numeric, string, etc.). This examination is of particular importance, since some clustering algorithms such as k-means work best if the input data is primarily numeric in nature [4].

Furthermore, from the data mining perspective, the absence of noise is required. Therefore a data profiling should focus on the detection of missing, unknown or false values within the variable being examined. Unknown values can be identified by looking at the metadata, which should define the range of values for a variable. False values can be identified by looking at outliers, because outliers may indicate false data. In order to detect outliers, we can look at the most infrequent values determined during data profiling. The identification of missing values corresponds with the requirement of completeness outlined in the data quality perspective. An active variable should be largely complete, i.e. its objects should have a value. The null value can be used as an indicator of missing values, if it is not defined otherwise. Another way to identify missing values is to search for the most frequently occurring values across all fields. Improvements of completeness can be achieved by application of business rules. For example, if such a rule specifies that with value x in attribute D inevitably value y must be present in attribute E, then the value y can be entered automatically if E is incomplete.

When doing a data profiling, correctness is another criterion to consider. Correctness of a variable value can be judged by a comparison of the real-world value with the value stored in the database. In order to determine correctness, we recommend looking at the most frequent values or the general distribution for detecting irregularities from a business point of view. Sometimes it is quite an effort to look at all variables and to gather the necessary business knowledge. In this case, we can again employ sampling as a heuristic approach.

In practice there are numerous tools for automatic data profiling. Based on our experiences with data quality projects at DaimlerChrysler, we developed a special tool for data profiling. There are many criteria which can be measured or estimated in a data profiling, but from our point of view it is not possible to develop a ranking of criteria or a standard process for data profiling. When choosing the criteria, we recommend considering the requirements of the actual project, experiences with previous projects, and the knowledge of experts.

At the end of this process the number of input variables for clustering should have been reduced significantly and we should be able to continue only with those variables which are highly relevant from a business point of view, are time stable, and have been profiled as promising and useful. Yet, it might be necessary to apply techniques such as factor analysis in order to combine or eliminate dependent features before clustering takes place.

6 SUMMARY AND A LOOK AHEAD

Customer segmentation carried out on large databases often faces the challenge of a great number of potential input variables. In order to create segments which are highly useful for CRM activities, the number of variables has to be reduced significantly before clustering. The reduction should follow a consistent process which is reliable, replicable, and fast, since decisions are usually made under time

pressure and cost constraints in practice. In this paper we present a first look at a process model for the identification of input variables for clustering. For customer segmentation in the context of CRM, we identified three different perspectives which should be considered. One of the perspectives is the data quality perspective, which influences each step of the process model. This leads to the conclusion that data quality is an essential factor and should always be taken into account when performing customer segmentation.

With the process model derived, we are able to generate experiences solely in the automotive industry. For this reason, a pivotal research requirement is the implementation and testing of the model in other industries with different data and business backgrounds. Furthermore, we should look for new, sophisticated ways of navigating in data generated during extensive data profiling. Since the reliable exploration of metadata is most important for the selection process, we are working on approaches to generate, store, and explore metadata of high quality.

Acknowledgement

The authors would like to thank Ms. Amy Weissenberg for her support.

References

[1] Arndt, D. and Gersten, W.: Data Management in Analytical Customer Relationship Management. In: Workshop Data Mining for Marketing Applications, In: Proceedings of the ECML/PKDD, Freiburg (2001), pp. 25 – 38.

[2] Arndt, D. and Gersten, W.: External Data Selection for Data Mining in Direct Marketing. In: Proceedings of the International Conference on Information Quality, MIT, Boston (2001), pp. 44 – 61.

[3] Berry, M. J. A. and Linoff, G. S.: Data Mining Techniques, New York et al. (2000).

[4] Berry, M.J.A., Linhoff, G.S.: Mastering Data Mining, The Art and Science of Customer Relationship Management, New York (2000).

[5] Berson, A., Smith, S., and Thearling, K.: Building Data Mining applications for CRM, New York et al. (1999).

[6] CRISP-DM: Cross-Industry Standard Process Model for Data Mining. In: http://www.crisp-dm.org/home.html (2001).

[7] Crockett, B. K. and Reed, K. L.: Three Approaches to Customer-Centric Understanding, http://crockett.CRMproject.com.

[8] Doyle, M.: CRM starts with quality data., White Paper, http://www.dedupe.com.

[9] English, L. P.: Information Quality Improvements: Principles, Methods, and Management, Seminar 5th Ed., Brentwood, TN: Information Impact International, Inc., 1996.

[10] English, L.P.: Improving Data Warehouses and Business information Quality, Methods for reducing costs and increasing profits, New York (1999).

[11] Gersten, W. and Arndt, D.: Effective Target Variable Selection from Multiple Customer Touchpoints. In: Workshop on Mining Data Across Multiple Customer Touchpoints for CRM (MDCRM02) at the Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-02), pp. 1 – 13.

[12] Grimmer, U. and Hinrichs, H.: A Methodological Approach to Data Quality Management Supported by Data Mining. In: Proceedings of the International Conference on Information Quality, MIT, Boston (2001), pp. 217 – 232.

[13] Han, J. and Kamber, M.: Data Mining. Concepts and Techniques, San Francisco et al. (2001).

[14] Henning-Thurau, T. and Hansen, U.: Relationship Marketing – Some Reflections on the State-of-the-Art of the Relational Concept. In: Henning-Thurau, T., Hansen, U. (Hrsg.): Relationship Marketing, Berlin Heidelberg New York (2000), pp. 3 – 27.

[15] Hinrichs, H.: Datenqualitätsmanagement in Data Warehouse-Systemen, Dissertation, University of Oldenburg (2002).

[16] Hipp, J., Güntzer, U., and Grimmer, U.: Data Quality Mining – Making a Virtue of Necessity, In Proceedings of the 6th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2001), pp. 52 – 57, Santa Barbara, California, May 20 2001a.

[17] Peterson, H.: Assessment of Clusteranalysis and Self-organizing-maps. In: International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, Vol. 6 (1998), No. 2, pp. 139 – 149.

[18] Rahm, E. and Do, H.H.: Data Cleaning: Problems and Current Approaches. In: IEEE Techn. Bulletin on Data Engineering (2000).

[19] Shepard, D. et al.: The New Direct Marketing. How to implement a Profit-Driven Database Marketing Strategy, McGraw-Hill, New York (1991).

[20] Stecking, R.: Market Segmentation with Neural Networks, Wiesbaden (2000).

[21] Wang, R.: A Product Perspective on Total Quality Management. Communications of the ACM, Vol. 41., No. 2 (1998).

[22] Weiss, S. H. and Indurkhya, N.: Predictive Data Mining, San Francisco (1998).

[23] Witten, I. and Frank, E.: Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations, San Francisco (2002).