What Data Is Necessary To Data Mine For Knowledge?

Paul Alpar

School of Business and Economics, Philipps University Marburg,
Universitätsstr. 24 35032 Marburg, Germany
Phone: ++49-6421-2823703, Fax: ++49-6421-2826554
alpar@wiwi.uni-marburg.de

ABSTRACT

The data that most organizations possess is considered a valuable asset. To really benefit from it, organizations must be able to analyze these data efficiently and effectively. This activity is nowadays referred to as business intelligence and one class of algorithms used within this activity are called data mining methods. There are some famous success stories about knowledge discovered via data mining but there is also a lot of disappointment so far. The paper argues that one of the reasons for some failures to produce better results with data mining is the reliance on transactional and master data only. It points to other types of data that can enrich transactional data and help in this way to produce more interesting and more rewarding data patterns.

Keywords: knowledge discovery, data mining, data warehouse, business intelligence

1. INTRODUCTION

The computerization of most (business) processes has led to a dramatic increase in transactional data. One of the latest reasons for a revolutionary increase in transactional data is the wide-spread use of the world-wide web where each request for information is recorded in a log file. Transactional data always contain information on *what* happened and often but not always *who* the actor was. For example, in web log mining it is often not known who the web site visitor was. If we know the actor, we can combine transactional data with master data if they exist. Practitioners often misinterpret the most cited process model (see Figure 1) in such a way that transactional (and master) data are sufficient to perform knowledge discovery. A typical statement is "Data mining can help companies extract additional value from piles of accumulated customer transaction data" [7]. This is not wrong but transactional data are often not sufficient to perform successful knowledge discovery as will be shown below. Many companies first build a data warehouse, primarily for regular reporting and ad-hoc queries, and then they start thinking which more complex data analyses to perform. Building a data warehouse usually comprises merging transactional and master data from several applications and sources, a tedious task that takes a lot of resources and time [3]. If additional data are needed later for data mining they may not be available anymore or the management is reluctant to restart the process of building the data warehouse. So data mining is only performed with the data that have been collected before independently of any knowledge discovery objectives.

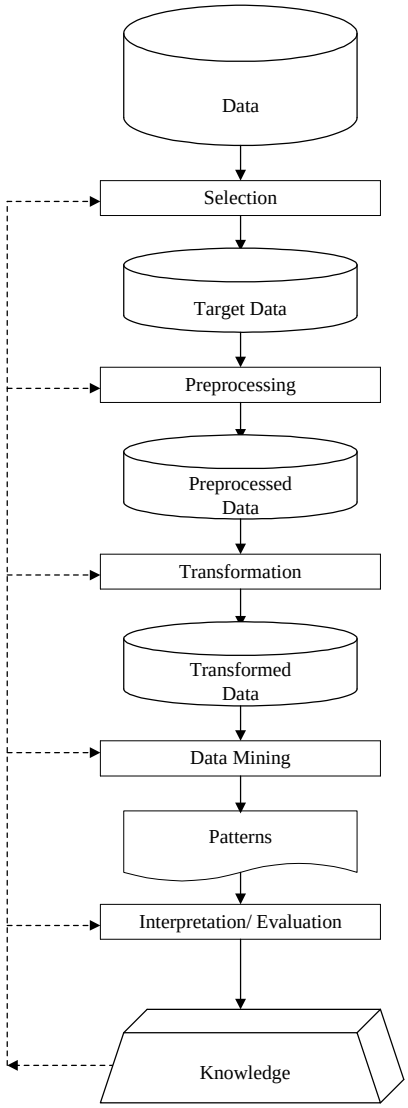

Figure 1: Knowledge Discovery Process [2]

Following such an approach, the results of the data mining process are often disappointing; they are not

interpretable, not interesting, or not actionable. Other authors suggest, therefore, to enrich data during the preprocessing step [7] or they propose an alternative process model (see Figure 2).

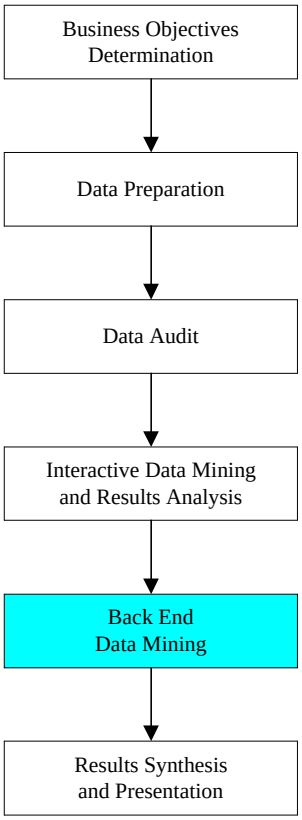

Figure 2: Knowledge Discovery Steps [4]

It is especially the step of "back end data mining" that gives rise to the hope that this process will lead to better results. According to Hirji [4] "Back end data mining involves data enrichment and additional data mining algorithm execution (...) in order to increase the dimensionality of the data mining data set with third-party demographic data (...)."

Our view, based on carrying out a number of data mining projects, is that additional data, not just demographic data, may be needed at collection time, in the selection step, and in the preprocessing step. The collection or creation of these data is based on knowledge that is sometimes loosely referred to as background, domain, or application-specific knowledge. It enables an analyst to collect, create, or select data that are relevant to the analysis task without being an explicit part of a transaction. To be able to initiate the necessary activities in due time, the analyst should roughly know ahead what s/he will be mining for. A process model proposed by SPSS, a statistical software vendor, in collaboration with some industry representatives, called CRISP-DM (Cross Industry Standard Process for Data Mining), explicitly requires to collect data only after the business and data mining objectives have been defined (see Figure 3). Further, it suggests constructing and integrating data, if needed, during the phase of data preparation. This model is, however, not specific about why and what type of new data should be generated via construction or integration.

We try to determine various types of data that may play a role in a data mining task without being part of transactional or master data. The importance of such data is demonstrated through examples from specific projects that we have conducted in the last years. The data types have been determined inductively through action research [6] so they are biased by the kind of projects undertaken. The projects cover several industries and almost all widely used data mining methods (decision trees, artificial neural networks, association rules, case-based reasoning, various clustering other statistical methods, scoring). However, it may be that mining in scientific data would reveal that additional or other data types are important.

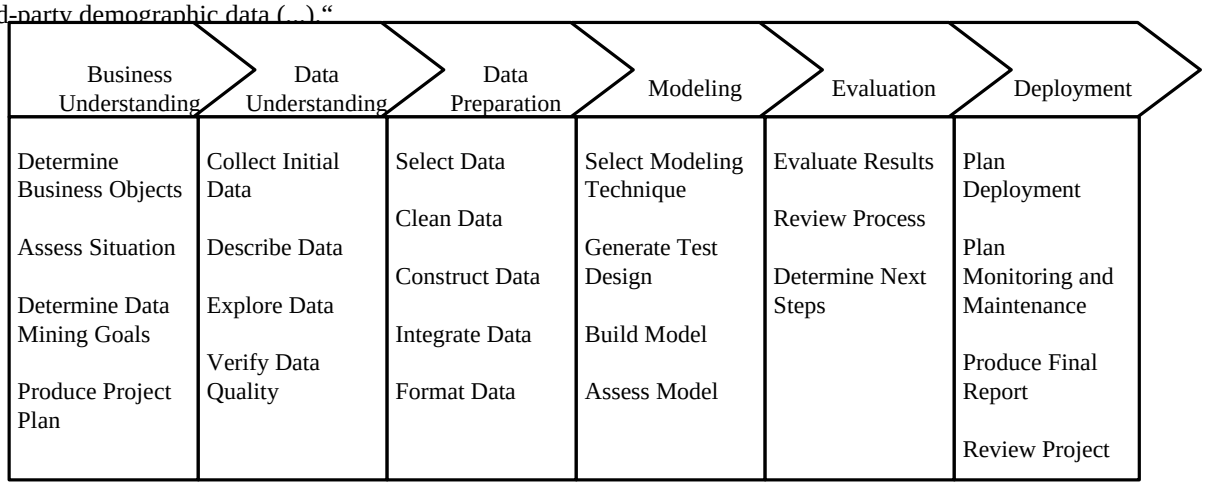

Figure 3: CRISP-DM Reference Model [1]

2. CIRCUMSTANTIAL DATA

Circumstantial data relate to the conditions when the transaction occurred. It is usually easier to create or collect such data at the time that transactional data are collected than later, perhaps after months or years, when data mining takes place. For example, to get detailed regional weather data by the hour and minute for each transaction after several months or years might be possible but it is certainly more difficult and expensive than recording the data when a transaction occurs. Table 1 shows an example where the task was to segment customers of a telecommunication product.

Table 1: Enriched call data

Col No	Column Name	Format	Description
...	...	...	...
5	from_connect_tm	...	...
6	from_weekend_ind	numeric1	Set if day called at origin = Sunday or Saturday
7	to_country_cd	...	...
8	to_connect_dt	...	...
9	to_connect_tm	...	...
10	to_weekend_ind	numeric1	Set if day called at destination = Sunday or Saturday
11	vacation_ind	numeric1	Set if day called at origin in {24.12 - 31.12 or 1.7 -31.8}
12	days_since_last_call_qty	numeric	No of days since customer called last
13	destination_typ	...	...
...	...	...	...

The basic call data (numbers of the caller and the person called as well as the time) were immediately augmented by data that indicate whether the call took place on the weekend (column 6), during vacation time (column 11), and a number of other indicators. Some of these played an important role during the clustering of customer groups. As can be seen from column 11, such data can be soft in the sense that they reflect an assumption or perception. In this case, it is assumed that the caller is on vacation if the call was made in the indicated time period.

Circumstantial data can be also important in the selection step. In another application, the task was to develop a scoring model with which customers of a financial institution can be targeted for specific investment products [5]. One of the variables that played an important role was the value of the customer's stock portfolio. This obviously can fluctuate considerably. To control for the fluctuations of the stock market, data about the German stocks index DAX were used in such a way that customer data were selected for training purposes from a period of time when the DAX was relatively stable (see Figure 4). In this case, this was the period between January 2, 2001 and June 29, 2001.

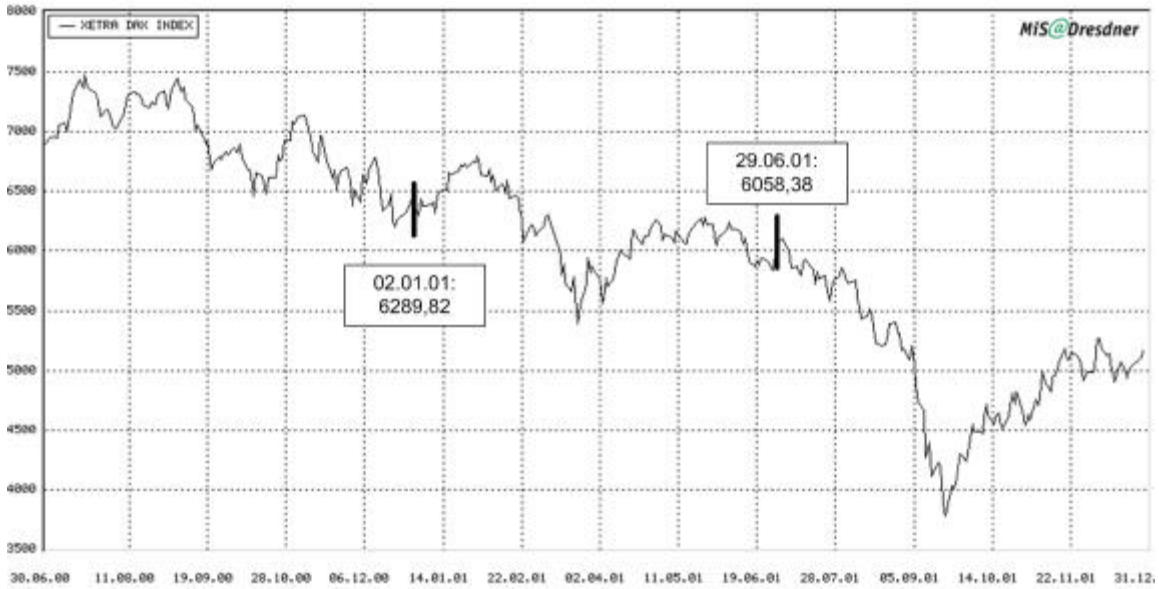

Figure 4: Run of the DAX between April 20, 2000 and December 31, 2001

3. STRUCTURAL DATA

While in some cases a simple categorization of data can be helpful, in many cases the knowledge of more complex relationships among objects involved in transactions is needed. For example, in the analysis of purchase baskets in a supermarket the scanner data allow to mine for associations at the level of universal product codes. However, given that many products from the same brand and kind come in different packages, it may be meaningful to summarize them as a pre-processing step. Sometimes, further aggregations along product categories (across brands) are needed to discover association rules. Therefore, the structure of product groups needs to be applied to transactional data before data mining can be performed.

4. QUALIFICATION DATA

Qualification data describe more precisely what has happened. Transactional data contain the facts, qualification data add meaning to these facts. In the above mentioned example, segmentation of customers of a telecommunication product, each transaction contains the origination and destination phone number. Through a comparison of the destination number with the customer's home and office numbers, which are usually recorded in the master data, each call was qualified to be a call *home*, to the *office*, to the *home city*, to the *home country*, or *other*. Another comparison of the origination and the destination numbers qualified the calls as being *in-country* or *out-of-country* calls. It turned out that one of these variables played a role in forming the customer clusters. Another example is visits to a web site. They can be relatively easily distinguished by the number of web pages requested, the visit length, or the action taken (e. g., a registration, an information request, or a purchase). However, if visitors usually do not take an action at a web site, it may be helpful to characterize each visit before further analyzing the visiting behavior. One simple qualification using cookie technology can be the distinction between *first visit* and *repeat visit*.

Another example relates to a set of purchase data from mail orders. Some of the customers acted as sales people in the sense that they were also placing collective orders for several customers. Their pay-off was a bonus on the accumulated sales amount. The mail orders did not contain any information whether an order was placed for one or several customers. The goal of the analysis was to search for association rules among products in a sales transaction (basket analysis). Collective orders would distort the results since they contain several transactions. Therefore, the mail orders were qualified as collective orders if the total order value exceeded a threshold value. These purchases were then excluded from further analysis.

The first example shows that transactions can be qualified based on facts while the second one shows that, sometimes, they must be based on "educated guesses."

5. NON-TRANSACTIONAL TRACKING DATA

Non-transactional tracking data record the behavior of objects of interest, often customers, even when no transactions occur. When a transaction occurs transaction processing systems carry out necessary actions including the update of stock values (e.g., customer account balance or number of seats available on a plane). However, the lack of activity of an object can also be an important information worth recording. As an example, we look at data of a direct order company. The orders come via phone, mail, or internet. In each of the cases, a quick credit assessment takes place. Most cases are decided automatically by a program that decides whether to deliver the merchandise to the customer and what kind of payment to accept. This process is referred to as *behavioral scoring* in the case of existing customers because the decision is based on their past buying and paying behavior (in the case of new customers *application scoring* takes place). The method used to derive rules for the decision can be a decision tree or a scoring scheme. Obviously, the relating data are generated and updated when a purchase or a payment takes place. But it is also important to update data when a customer with a balance due does not make a payment. Table 2 shows some of the data used in the example. The variable *number of successive months overdue* is a variable which value is updated monthly even if a customer has no transactions with the company in a given month.

Table 2: Customer data of a direct order company

CHAR005	Average of monthly balance
CHAR007	Balance of the actual month
CHAR008	Percent of used debt limit
CHAR013	Number of successive months overdue
CHAR022	Longest time of overdue in the last 12 month
...	...
CHAR516	Average of all payments
...	...
CHAR553	Percent of returns and refusals
CHAR555	Number of months since last return or refusal
...	...

6. SUMMARY

Table 3 gives an overview of the described data types with respect to some attributes. The values of the attributes should be understood as typical values.

Data type	Time of generation with respect to transactional data	Transaction dependent?	Ownership (ass.: company is owner of transaction data)
Circumstantial	at the same time	yes	others
Structural	before or after	no	company
Qualification	after	yes	company
Non-transactional tracking data	at any time	no	company or others

Table 3: Additional Data Types for Data Mining

Our findings are more specific with respect to types of data that can be used to enrich transactional and master data than general knowledge discovery process models but they are still not "cooking recipes" for data collection for data mining. Each case requires its own business analysis. Our findings may help to make management more aware of the data needs in data mining while data miners may find them helpful to conceptualize their thoughts.

REFERENCES

[1] Chapman, P. et al., "CRISP-DM 1.0 - Step-by-step Data Mining Guide", CRISP-DM consortium, 2000.

[2] Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, pp37-51, Fall 1996.

[3] Fayyad, U and R. Uthurasamy, "Evolving Data Mining into Solutions for Insights", Communi- cations of ACM, Vol. 45, No. 8, pp28-31, 2002.

[4] Hirji, K. K., "Exploring Data Mining Implementation", Communications of the ACM, Vol. 44, No. 7, pp87-93, 2001.

[5] Langner, R., Alpar, P. and Pfuhl, M., "Ein Vergleich ausgewählter Klassifikationsverfahren im Kontext von Finanzdienstleistungen", in „Wirtschaftsinformatik 2003", Vol. II, Eds. Uhr, W., Esswein, W., Schoop, E. pp495-517, 2003.

[6] Rapoport, R. N., "Three Dilemmas in Action Research", Human Relations, Vol. 23, No. 4, pp499-513, 1970.

[7] Wassermann, M., "Mining Data", Federal Reserve Bank of Boston, Vol. 10, No. 3, Quarter 3 2000.

[8] Wittmann, T., P. Kischka, M. Hunscher and J. Ruhland, "Data Mining - Entwicklung und Einsatz robuster Verfahren für betriebswirtschaftliche Anwendungen", Frankfurt am Main, 2000.