# DATA QUALITY MINING
– Making a Virtue of Necessity –

Jochen Hipp
*DaimlerChrysler AG, Research & Technology, Ulm, Germany*
*Wilhelm-Schickard-Institute, University of Tübingen, Germany*
*Email: jochen.hipp@acm.org*

Ulrich Güntzer
*Wilhelm-Schickard-Institute, University of Tübingen, Germany*
*Email: guentzer@informatik.uni-tuebingen.de*

Udo Grimmer
*DaimlerChrysler AG, Research & Technology, Ulm, Germany*
*Email: udo.grimmer@daimlerchrysler.com*

**Abstract**

In this paper we introduce data quality mining (DQM) as a new and promising data mining approach from the academic and the business point of view. The goal of DQM is to employ data mining methods in order to detect, quantify, explain and correct data quality deficiencies in very large databases. Data quality is crucial for many applications of knowledge discovery in databases (KDD). So a typical application scenario for DQM is to support KDD projects, especially during the initial phases. Moreover, improving data quality is also a burning issue in many areas outside KDD. That is, DQM opens new and promising application fields for data mining methods outside the field of pure data analysis. To give a first impression of a concrete DQM approach, we describe how to employ association rules for the purpose of DQM.

## 1 MOTIVATION

Since the early nineties knowledge discovery in databases (KDD) has developed to a well established field of research. Over the years new methods together with scalable algorithms have been developed to efficiently analyze even very large datasets. However, KDD has not been broadly established outside academia. Although there are numerous success stories of practical applications today many of the people concerned with KDD seem to be somehow disillusioned. “Crossing the chasm” as Rakesh Agrawal formulates in (Agrawal, 1999) is overdue. Otherwise KDD might end like many promising technologies that were welcomed enthusiastically but finally missed to satisfy the expectations they generated.

The research community is aware that KDD implies much more than running a highly optimized algorithm on a large dataset. Nevertheless today we see a tendency to concentrate on quite special and unfortunately somehow “academic” questions instead of tackling the crucial problems. What KDD urgently needs for its breakthrough is more business driven research, e.g. by studying real world applications and a better understanding of the KDD process, e.g. (Fayyad et al., 1996; Brachman and Anand, 1996; Wirth and Hipp, 2000)

In this paper we want to draw the attention to data quality in the context of KDD. We believe the connection of both fields currently does not get the attention that is implied by its potentials. Basically, there are two aspects of data quality that make it interesting for successfully transferring KDD from academia to the business world:

(a) We experienced that in most cases KDD is employed in the presence of deficiencies of the underlying data. So dealing with data quality is crucial and in practice already part of many data mining projects. Appropriately reflecting this importance by KDD research and supplementing current ad hoc solutions by a systematic approach would help a lot to improve the results of many KDD projects.

(b) Often, the quality of the underlying data is significantly improved as a byproduct of a KDD project. So why not make a virtue of necessity? Deficiencies in data quality are a burning issue in many application areas. In other words, improving data quality by KDD methods opens a new and promising application field for KDD from the academic and the business point of view.

The main contribution of this paper is to give a first impression of how data mining techniques can be employed in order to improve data quality with regard to both improved KDD results and improved data quality as a result of its own. We illustrate our basic idea by giving a concrete example. That is, we describe a first approach to employ association rules for the purpose of data quality mining.

Our work is based on the experience collected at our research lab in various practical KDD applications over several years, e.g. (Wirth and Reinartz, 1996; Handley et al., 1998; Hotz et al., 1999; Kauderer et al., 1999; Hipp and Lindner, 1999; Gersten et al., 2000), and our research on KDD methods and algorithms, e.g. (Nakhaeizadeh et al., 1998; Hipp et al., 1998; Hipp et al., 2000b), and the KDD process, e.g. (Bartlmae and Riemenschneider, 2000; Wirth and Hipp, 2000).

## 2 DATA QUALITY MINING

According to Total Quality Management we define data quality as "consistently meeting customer's expectations", c.f. (English, 1999). There are several aspects of data quality, like integrity, validity, consistency, and accuracy but we do not want to go into details here.

As mentioned, poor data quality is nearly always a problem in practical applications of KDD. This might surprise, but the explanation is quite simple. Data normally does not originate from systems that were set up with the primary goal of mining this data. That is, we often have to deal with operative systems that produce data only as a byproduct. Even if during design and implementation of such systems data quality was considered an important aspect, experience shows that such "ideals" get worn out in the long run. The reason is that data quality pays off only from a strategic point of view. So after several years in business and a multitude of operative decisions to adapt the system to its changing environment, data quality typically suffers. We found that often this is even true for data warehouses, at least for subsets of data that are not frequently accessed.

Typically the owner of the data is not fully aware of data quality deficiencies. The system might have been doing a good job for years and the owner probably has its initial status in mind. Doubts concerning data quality may raise astonishment or even disaffection. We often have been facing exactly this situation. By trying to make the best of it we employed our skills – data mining techniques – as a patched-up solution to measure, explain, and improve data quality. In fact, this pragmatic behavior usually was astonishingly successful. A systematized approach, however, is inevitable in order to support analysts in this crucial situation. By introducing data quality mining (DQM) we hope to stimulate research to reflect the importance and potentials of this new application field.

> *DQM can be defined as the deliberate application of data mining techniques for the purpose of data quality measurement and improvement. The goal of DQM is to detect, quantify, explain, and correct data quality deficiencies in very large databases.*

There are many starting points to employ today's common data mining methods for the purposes of DQM. Namely, methods for deviation and outlier detection seem promising. But also clustering approaches and dependency analysis are straight forward to be employed for data quality purposes. In addition, if we are able to supply training data prepared by a human then also classifiers might do a good job. It is even conceivable that neural networks are trained to recognize data deficiencies.

Basically we see four important aspects:

(a) Employment of data mining methods to measure and explain data quality deficiencies

Research and practical experience are needed in order to understand what methods fit with which data and how KDD can actually be employed to measure and explain deficiencies in large databases.

(b) Employment of data mining methods to correct deficient data

Recollecting data is usually preferred to correction. Unfortunately, recollection is often impossible or too expensive. In such cases KDD can help to detect and exclude deficient datasets or even to guess missing or incorrect values. Of course fully automated correction should always be seen critically.

(c) Extension of KDD process models to reflect the potentials of DQM

Although current process models are aware of data quality aspects, we suggest that process models for KDD should have an explicit data quality phase. Furthermore, measuring and improving data quality is not only a central aspect of the initial phases of a KDD project. Also during the actual mining and deployment phases it is important to consider the insights from a data quality phase.

(d) Development of specialized process models for "pure DQM"

As mentioned, DQM needs not necessarily be embedded into a KDD process. That is, improving data quality is worth to be seen as a goal of its own, outside the context of data analysis. Then of course specialized process models for DQM need to be developed which reflect the change of scope from pure data analysis to data quality measurement and improvement.

# 3 A FIRST APPROACH: DQM WITH ASSOCIATION RULES

An association rule is an implication $X \rightarrow Y$ where $X$ and $Y$ are non empty and disjunct sets of items, c.f. (Agrawal and Srikant, 1994). Given a database of transactions – where each transaction is a set of items – an association rule $X \rightarrow Y$ expresses that whenever we find a transaction which contains all items $x \in X$ then this transaction is likely to also contain all items $y \in Y$ with probability $c$. This probability is called the rule confidence and is supplemented by further quality measures like support or interest, c.f. (Agrawal and Srikant, 1994; Brin et al., 1997a).

A transaction can be a single row of a relational table or might be collected from several tables or several rows of a table, c.f. (Hipp et al., 2001). The decision which entity corresponds to a transaction is up to the analyst. For example in address data taking each customer as a transaction probably makes sense, whereas in a database which stores information on vehicles, the vehicles themselves are potential transactions.

## 3.1 Basic Idea

To a given database $\mathcal{D}$ of transactions all rules that satisfy thresholds on the quality measures can be efficiently generated, e.g. (Hipp et al., 2000a) for an overview. Then, for each transaction from $\mathcal{D}$ we test its consistency with the generated rules. For example there might be a rule "Zip code: 80801 → City: Munich" that holds with confidence slightly below $100\%$. Obviously, each transaction contradicting this rule must be suspected of deficiencies.

But also rules at a much lower confidence level are worth considering. For example, a certain engine type may be typical for a vehicle model, resulting in the rule "Model: S-Class → Engine: Petrol" with confidence $90\%$. A S-Class vehicle equipped with a diesel engine contradicts this rule but of course this is not necessarily a sign of incorrectness. But if further rules also conflict then skepsis becomes more and more justified. For example a special air conditioning might also be typical for the luxury S-Class model but might not be installed in the vehicle under investigation. Thereby the rule "Model: S-Class → Equip: AirCondTypeC" holding with confidence $75\%$ is violated. Moreover, in contrast to the vehicle under investigation, $75\%$ of the S-Class vehicles might be equipped with an autonomous rain detection system triggering the windshield wiper. So the rule Model: S-Class → Equip: AutoWindshWiper is also violated etc. See Table 1 for hypothetical example rules.

| Association Rule | Confidence |
|---|---|
| Model: S-Class → Engine: Petrol | $90\%$ |
| Model: S-Class → Equip: AirCondTypeC | $75\%$ |
| Model: S-Class → Equip: AutoWindshWiper | $75\%$ |
| Model: S-Class → Equip: NavigSystemD | $75\%$ |
| ⋮ | ⋮ |

Table 1: Example association rules from a hypothetic vehicle database

It is important to note that basically we do not follow the typical test and evaluate approach. We characterize each transaction by rules that were derived from the set of transactions itself. Of course this assumes that deficiencies occur only exceptional and that not the whole database is cluttered with noise.

## 3.2 Suggested Realisation

First of all, an entity type that determines the transactions must be chosen. Next, association rules are generated, preferably with an algorithm implementation that can directly access tables stored in a relational database system. Depending on the underlying data it might be necessary to discretize data values. If additional information like a taxonomy is available, then these may also be taken into account during rule generation, c.f. (Srikant and Agrawal, 1995; Hipp et al., 1998).

After that we assign a score $s \in \mathbb{R}_0^+$ to each transaction that is computed on the basis of the generated rules. The score is a means to capture the consistency of a single transaction with the rule set as a whole. The idea is to assign high scores to transactions that are suspected of deficiencies.

Let $\mathcal{R}$ be a set of association rules and let $\mathcal{D}$ be a database of transactions, both containing only items from a universe $\mathcal{I}$. Let $r = X \rightarrow Y$ be an association rule with $\mathsf{body}(r) = X$ and $\mathsf{head}(r) = Y$. Let the mapping $\mathsf{violates}$ that determines whether a transaction $T \in \mathcal{D}$ violates a rule $r \in \mathcal{R}$ be defined as:

$$\mathsf{violates} : \mathcal{D} \times \mathcal{R} \rightarrow \{0, 1\} :$$

$$(T, r) \mapsto \begin{cases} 1 & \text{if } \mathsf{body}(r) \subseteq T \wedge \mathsf{head}(r) \not\subseteq T \\ 0 & \text{else} \end{cases}$$

We assign a score to each transaction by summing the confidence values of the rules it violates. Of course only rules should be taken into account that hold with a certain confidence. For that reason we restrict the rule set $\mathcal{R}$ to $\mathcal{R}_\gamma = \{r \in \mathcal{R} \mid \mathsf{confidence}(r) \geq \gamma\}$. Based on the definition from above we compute the scores as follows:

$$\mathsf{score}_{\mathcal{R}_\gamma} : \mathcal{D} \rightarrow \mathbb{R}_0^+ :$$

$$T \mapsto \sum_{r \in \mathcal{R}_\gamma} \mathsf{confidence}(r)^\tau \cdot \mathsf{violates}(T, r)$$

Some example vehicles together with score values for $\tau = 7$ and $\tau = 5$ can be found in Table 2. The scores are computed with regard to the rule set from Table 1. For example vehicle 1 violates all given rules leading to

$$\mathsf{Score}_{\mathcal{R}_\gamma} = 0.9^7 + 0.75^7 + 0.75^7 + 0.75^7 \approx 0.9$$

at $\tau = 7$.

We want to point out that we employ the score only as a means to order the transactions.

The rules from Table 1 and the vehicles from Table 2 are only introductory and simplified examples. Of course for real-world applications we expect upto several ten thousand or even hundred thousand rules and transactions containing items out of a very large set of possible items.

| ID | Vehicle / Transaction | Score at $\tau = 7$ | Score at $\tau = 5$ |
|---|---|---|---|
| 1 | Model: S-Class,<br>Engine: Diesel,<br>Equip: AirCondTypeB,<br>Equip: StdWindshWiper,<br>Equip: NavigSystemA,<br>. . . | $\approx 0.9$ | $\approx 1.3$ |
| 2 | Model: S-Class,<br>Engine: Diesel,<br>Equip:_AirCondTypeC,<br>Equip:_AutoWindshWiper,<br>Equip:_NavigSystemD,<br>. . . | $\approx 0.5$ | $\approx 0.6$ |
| 3 | Model: S-Class,<br>Engine: Petrol,<br>Equip: AirCondTypeB,<br>Equip: StdWindshWiper,<br>Equip: NavigSystemA,<br>. . . | $\approx 0.4$ | $\approx 0.7$ |
| 4 | Model: S-Class,<br>Engine: Petrol,<br>Equip:_AirCondTypeC,<br>Equip:_AutoWindshWiper,<br>Equip:_NavigSystemD,<br>. . . | $0$ | $0$ |

Table 2: Example transactions with score values

The tuning parameter $\tau \in \mathbb{R}_0^+$ allows to assess the confidences depending on their value. For example vehicle 2 conflicts only with a single rule, $\mathsf{Model{:}\ S\text{-}Class} \rightarrow \mathsf{Engine{:}\ Petrol}$, but this rule has a relatively high confidence value of $90\%$. In contrast, vehicle 3 fulfills this rule but conflicts with all the three remaining rules from Table 1 which all have a confidence value of $75\%$.

At $\tau = 7$ vehicle 2 reaches a higher score than vehicle 3. That means violating a singe rule with confidence $90\%$ is rated as worse than violating three rules but each with confidence $75\%$. At $\tau = 5$ it is just the opposite and vehicle 3 is rated as more suspected of deficiencies than vehicle 2. In other words $\tau$ allows to appropriately calibrate the scoring function.

We suggest a minimal threshold of $\gamma = 75\%$ or higher for the confidence to restrict the rule set $\mathcal{R}$ in order to improve the results. In addition a threshold on the supplementary rule quality measure support is necessary for algorithmic reasons, c.f. (Agrawal and Srikant, 1994; Brin et al., 1997b; Han et al., 2000; Hipp et al., 2000a; Zaki, 2000).

The heart of the system will be the user interface.

The transactions will be presented as a list sorted according to the assigned scores. For each transaction the user will be able to retrieve the rules that are violated or hold. Based on this information together with her background knowledge the user will decide upon the trustworthiness of single transactions or groups of similar transactions and finally upon the quality of the whole data set.

With our approach we avoid a severe pitfall of association rule mining. Although thresholds on the rule quality measures significantly reduce the number of generated rules, in typical applications the resulting rule set can still easily overload the analyst. In our scenario the analyst is never confronted with the whole rule set. So even very large rule sets generated at very low thresholds for support can be employed for the purpose of DQM.

### 3.3 Open Issues and Future Work

The described framework is work in progress. Based on existing mining algorithm implementations we currently develop a first prototype that employs the scoring from above.

First of all we are going to evaluate the scoring function on real-world databases and learn more about appropriate values for the tuning parameter $\tau$. Based on the results we will enhance the scoring, e.g. by directly taking the supplementary quality measures like support, interest etc. into account. Moreover the question is open whether fulfilled rules should be allowed to counterbalance violated rules or not.

In addition the characteristics of the association mining problem in the context of DQM probably will differ from the typical scenario. For example we might not severely restrict the generated rule set by a threshold on support but probably by relatively strict thresholds on confidence, interest etc. The question is, whether current algorithms are suitable for this mining task or if better adapted algorithms need to be developed.

Another promising idea is to employ the described framework for the semiautomatic correction of data. Based on the generated rules the system can suggest changes that, when applied to a transaction, would lower its score. For example changing Model: S-Class to Model: A-Class in vehicle 1 from Table 2 will probably lower the assigned score drastically. Changing to Model: A-Class means no longer any violation of the rules from Table 1. At the same time a diesel engine, an air conditioning of type B, the standard windshield wiper and the navigation system of type A are not uncommon for an A-Class model. In other words not only the rules from Table 1 are no longer violated but also the rules typical for the A-Class model, e.g. Model: A-Class → Equip: StdWindshWiper, are likely to be fulfilled.

The above framework can easily be adopted to implement a quality monitoring system. First, during a training phase all rules are generated for a database. Then, in the application phase, these rules are employed to score new and unseen transactions that are intended to be added to the database. According to the score the system decides whether to accept or reject the data or at least whether to issue a warning. A typical application is data warehousing where data from operative systems is regularly loaded into the warehouse. Effective data quality monitoring allows to immediately detect deficiencies that otherwise might become a severe problem, e.g. if being discovered too late for easy correction or, even worse, if the deficient data has already been employed for decision support.

Finally, combining the described approach with other techniques, not necessarily from KDD but from the field of statistics, needs to be considered.

## 4 CONCLUSION

In this paper we introduced data quality mining (DQM) as a new and promising data mining approach. We defined DQM as the deliberate application of data mining techniques for the purpose of data quality measurement and improvement. The goal of DQM is to detect, quantify, explain, and correct data quality deficiencies in very large databases.

In practical applications KDD is typically employed in the presence of data deficiencies. So DQM supplements KDD and contributes to improve the results of KDD projects. Moreover, data quality deficiencies are not only crucial in the context of KDD. That is, improving data quality can be seen as a goal of its own and DQM opens many new and promising application areas for data mining techniques outside of KDD.

By introducing DQM we hope to stimulate research to consider the from our point of view high potentials and practical importance of the interaction between KDD and data quality.

## REFERENCES

Agrawal, R. (1999). Data mining: Crossing the chasm. Invited Talk at the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’99), San Diego, California.

Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Databases (VLDB ’94)*, Santiago, Chile.

Bartlmae, K. and Riemenschneider, M. (2000). Case based reasoning for knowledge management in kdd projects. In *Proceedings of the 3rd International Conference on Practical Aspects of Knowledge Management (PAKM 2000)*, Basel, Switzerland.

Brachman, R. J. and Anand, T. (1996). The process of knowledge discovery in databases. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, *Advances in Knowledge Discovery and Data Mining*, chapter 2, pages 37–57. AAAI/MIT Press.

Brin, S., Motwani, R., and Silverstein, C. (1997a). Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (ACM SIGMOD '97)*, pages 265–276.

Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997b). Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (ACM SIGMOD '97)*, pages 265–276.

English, L. P. (1999). *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing profits*. John Wiley & Sons, New York, USA.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.

Gersten, W., Wirth, R., and Arndt, D. (2000). Predictive modeling in automotive direct marketing: Tools, experiences and open issues. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '00)*, pages 398–406, Boston, MA USA.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM-SIGMOD International Confenerence on Management of Data*, Dallas, Texas, USA.

Handley, S., Langley, P., and Rauscher, F. A. (1998). Learning to predict the duration of an automobile trip. In *Proceedings of 1998 International Conference on KDD and Data Mining (KDD '98)*, pages 219–223, New York City, USA.

Hipp, J., Güntzer, U., and Grimmer, U. (2001). Integrating association rule mining algorithms with relational database systems. In *Proceedings of the International Conference on Enterprise Information Systems (ICEIS 2001)*, Setúbal, Portugal.

Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000a). Algorithms for association rule mining – a general survey and comparison. *SIGKDD Explorations*, 2(1):58–64.

Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000b). Mining association rules: Deriving a superior algorithm by analysing today's approaches. In *Proceedings of the 4th European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD '00)*, pages 159–168, Lyon, France.

Hipp, J. and Lindner, G. (1999). Analysing warranty claims of automobiles. an application description following the CRISP-DM data mining process. In *Proceedings of 5th International Computer Science Conference (ICSC '99)*, pages 31–40, Hong Kong, China.

Hipp, J., Myka, A., Wirth, R., and Güntzer, U. (1998). A new algorithm for faster mining of generalized association rules. In *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD '98)*, pages 74–82, Nantes, France.

Hotz, E., Nakhaeizadeh, G., Petzsche, B., and Spiegelberger, H. (1999). Waps, a data mining support environment for the planning of warranty and goodwill costs in the automobile industry. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD '99)*, pages 417–419, San Diego, California, USA.

Kauderer, H., Nakhaeizadeh, G., Artiles, F., and Jeromin, H. (1999). Optimization of collection efforts in automobile financing – a kdd supported environment. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining (KDD '99)*, pages 414–416, San Diego, California, USA.

Nakhaeizadeh, G., Taylor, C., and Lanquillon, C. (1998). Evaluating usefulness for dynamic classification. In *Proceedings of 1998 International Conference on KDD and Data Mining (KDD '98)*, pages 87–93, New York City, USA.

Srikant, R. and Agrawal, R. (1995). Mining generalized association rules. In *Proceedings of the 21st Conference on Very Large Databases (VLDB '95)*, Zürich, Switzerland.

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a standard process modell for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, pages 29–39, Manchester, UK.

Wirth, R. and Reinartz, T. (1996). Detecting early indicator cars in an automotive database: A multi-strategy approach. In *Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining (KDD '96)*, pages 76–81, Portland, Oregon, USA.

Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390.