Using Portfolio Theory for Automatically Processing Information about Data Quality in Data Warehouse Environments

Robert M. Bruckner, Josef Schiefer

Institute of Software Technology (E188)
Vienna University of Technology
Favoritenstr. 9-11 /188, A-1040 Vienna, Austria
{bruckner, js}@ifs.tuwien.ac.at

Abstract. Data warehouses are characterized in general by heterogeneous data sources providing information with different levels of quality. In such environments many data quality approaches address the importance of defining the term "data quality" by a set of dimensions and providing according metrics. The benefit is the additional quality information during the analytical processing of the data.
In this paper we present a data quality model for data warehouse environments, which is an adaptation of Markowitz's portfolio theory. This allows the introduction of a new kind of analytical processing using "uncertainty" about data quality as a steering factor in the analysis. We further enhance the model by integrating prognosis data within a conventional data warehouse to provide risk management for new predictions.

1 Introduction

In the past data quality has often been viewed as a static concept, which represents the write-once read-many characteristics of many data warehouses: The quality of a piece of information (fact) is evaluated only once (e.g. during the cleansing process) and stored in the data warehouse. From the data consumers' point of view, data quality is an additional property of a fact, which can be physically stored by an attribute or data quality dimensions. Depending on the context, the refinement (e.g. more details) of a fact can result in either of two possibilities 1) the creation of a new fact (with adapted data quality settings), or 2) updating the old one.

The trend toward multiple uses of data, exemplified by the popularity of data warehouses, has highlighted the need to concentrate on dynamic approaches, where the quality depends on the context in which data is used. Furthermore, data is viewed as a key organizational resource and should be managed accordingly [1].

In this paper we propose the adaptation of Markowitz's portfolio theory [3] to data warehouse environments. Our approach can be viewed as an additional layer above an existing representation of data quality in a data warehouse, where information about the "uncertainty" (of the quality evaluation) is used to select "optimal" portfolios, described by weight vectors for the base facts. The model considers the "fitness for use" [11], because data quality is an integral part during the analytical processing of the data. Furthermore the examined data quality can become the steering factor in each analysis.

The remainder of this paper is organized as follows. Section 2 contains a description of the goals of this approach. Section 3 gives a short overview of Markowitz's portfolio theory, which is the basis for the adaptation, described in section 4. In section 5 we apply the approach to "conventional" data warehouses. The integration of prognosis data is discussed in section 6 and followed by an investigation of the general limitations of this approach (section 7).

The paper concludes with section 8 (related works) and section 9, where we give a summary and conclusion of this work.

2 Goal of this Approach

The automatic consideration of data quality by using portfolio selection is suitable for "traditional" requirements of data warehouse users, such as ad-hoc queries and predefined reports, as well as analytical processing where the data quality of the involved facts is an important issue.

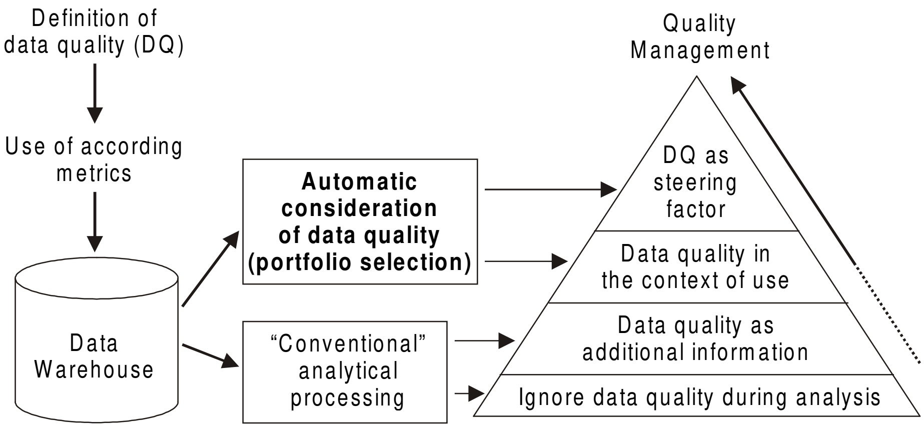

Figure 1: Achieving Data Quality in Decision Making

Besides *conventional* analytical processing, the approach described in detail in sections 4 to 6, enables the *automatic consideration* of data quality, as shown in Figure 1. In the first case the examined data quality is just an additional property of a fact. It is the user's responsibility to take advantage of the provided data quality information. Consequently only experienced data warehouse users (experts) are able to "exploit" this kind of knowledge. Similar to the far-reaching effects and consequences of poor data quality [7], the ignorance of that information can have diverse impacts, such as poorer decision making and difficulties to set and execute enterprise strategies.

The *automatic consideration* of data quality by using portfolio selections introduces a new layer between the data warehouse and its users by considering the context in which the data is used. Decision making often requires the combination of many aspects (represented by facts in a data warehouse) within an analysis. Therefore it is necessary to

- define data quality according to the "fitness for use" principle [11].
- consider and store "uncertainty" of the evaluated data quality.

This enables new perspectives on analytical processing of data for decision making:

- automatically calculate (optimal) weights that describe the impact of a specific fact on a decision according to its data quality and the involved uncertainty.
- consider the user preferences due to risk aversion, e.g. a venturesome user, who has high requirements on the data quality of the involved facts, nevertheless accepts high uncertainty of the assessed data quality of the facts.
- use the derived weights to combine the facts to a decision (with a resulting data quality and uncertainty assessment).
- introduce "*quality-driven*" *decisions*, where the user requirements due to the resulting data quality and uncertainty assessment of the decision are pre-specified.
- users need no further background knowledge to apply this approach for analytical processing. Even inexperienced users are able to take advantage of this approach.

Unfortunately it is not always possible to gain all the information from the data sources that are necessary to make good decisions. If there is a strong dependency between predictions about the future and those decisions (made to affect the future), it will be advantageous to integrate additional regular and prognosis data into a data warehouse, where all users have access to these valuable predictions represented in a consistent manner.

The integration of such data requires no alterations in the analytical processing through portfolio selection.

3 The Portfolio Theory

In 1952 Harry Markowitz published a landmark paper [3] that is generally viewed as the origin of the "modern portfolio theory" approach to investing. Markowitz's approach can be viewed as a *single-period*[1] approach, where money is invested only for a particular length of time (holding period). At the beginning the investor must make a decision regarding what particular securities to purchase and hold until the end of the period. Since a *portfolio* is a collection of securities (in the context of investing), this decision is equivalent to selecting an optimal portfolio from a set of possible portfolios and is thus often referred to as the "portfolio selection problem". Each security is characterized by the *expected return* and the *uncertainty* of this estimation (that is done by the investor).

Markowitz notes that the typical investor, in seeking both to maximize expected return and minimize uncertainty (= *risk*), has two conflicting objectives that must be balanced against each other when making the purchase decision at the beginning of the holding period. The investor should view the rate of return associated with any

[1] Markowitz shows in [5] that investing is generally a multi-period activity, where at the end of each period, part of the investor's wealth is consumed and part is reinvested. Nevertheless, this one-period approach can be shown to be optimal under a variety of reasonable circumstances (see also [9]).

one of the various considered securities to be non-deterministic. In statistics this is known as a *random variable* W_i, described by its moments: its expected value or *mean* r_i and the *standard deviation* σ_i (formula 1). Therefore a portfolio W_P is a collection of the weighted random variables.

$$W_i = (r_i, \sigma_i) \; . \tag{1}$$

$$W_P = a_1 W_1 + \ldots + a_n W_n \quad \text{with} \quad \sum_{i=1}^{n} a_i = 1 \; . \tag{2}$$

The calculated expected return r_P of a portfolio is the weighted average of the expected returns of its component securities. The calculated standard deviation σ_P (variance σ_P^2) of a portfolio is given by formula 4, where σ_{ij} denotes the covariance of the returns between security *i* and security *j* and ρ_{ij} indicates the correlation factor.

$$r_P = a_1 r_1 + \ldots + a_n r_n \; . \tag{3}$$

$$\sigma_P^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \, \sigma_{ij} \quad \text{with} \quad \sigma_{ij} = \rho_{ij} \sigma_i \sigma_j \; . \tag{4}$$

The *covariance* is a measure of the relationship between two random variables. A positive value for covariance indicates that the securities' returns tend to go together. For example, a better-than-expected return for one is likely to occur along with a better-than-expected return for the other.

The method that should be used in selecting the most desirable portfolio utilizes *indifference curves* [9]. These curves represent the investor's preferences for risk and return and thus can be drawn on a two-dimensional figure. In general two assumptions [3] hold for investors:

- *Nonsatiation*: It is assumed that investors, when given a choice between two otherwise identical portfolios, will select the one with the higher level of expected return.
- *Risk Aversion*: This means that the investors, when given a choice, will not want to take fair gambles, where a fair gamble is defined to be one that has an expected payoff of zero with an equal chance of winning or losing.

The *feasible set* simply represents the set of all portfolios that could be formed from a group of *n* securities. In general, this set will have an umbrella-type shape similar to the one shown by Figure 2. The *efficient set* can be located by applying the efficient set theorem (given by the assumptions above) to the feasible set. From this (infinite) set of efficient portfolios[2] the investor will find the *optimal* one by:

- Plotting indifference curves
  Plot the indifference curves on the same figure as the efficient set and then choose the portfolio that is farthest northwest along the indifference curve (Figure 2).
- Critical line method [4] for the identification of *corner portfolios*[3].
  Utilizing the information about the corner portfolios it becomes a simple matter for a computer to find the composition, and in turn the expected return and standard deviation, of the optimal portfolio by the approximation of the tangency

[2] All other portfolios are "inefficient" and can be safely ignored.

[3] A corner portfolio is an efficient portfolio where any combination of two adjacent corner portfolios will result in a portfolio that lies *on the efficient set* between these two portfolios.

point between the (calculated) efficient set and one of the investor's indifference curves.

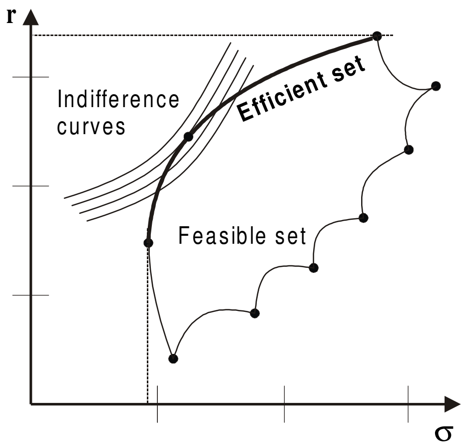

Figure 2: The Selection of the Optimal Portfolio

4 Adaptation of the Portfolio Theory for Data Warehouses

There are many approaches in the literature that can be applied to studying data quality (classified by a framework in [13]), but there is no commonly accepted model providing metrics to precisely assess data quality. One reason for this is the different context in which the data is produced (operational systems) and used (data warehouses). As a consequence every data quality assessment will be characterized by a kind of "*uncertainty*", which should be evaluated and stored as well.

The equivalences that can be identified between finance markets and data warehouse environments are presented in Table 1.

Table 1. Identified Equivalences: Portfolio Theory for Investments and Data Warehouses.

Data Warehouse Environment	Portfolio Theory for Investments
data warehouse user	investor
fact	security, random variable
data quality	expected return on security
uncertainty of the data quality assessment	standard deviation
covariance	covariance
analytical processing objectives: minimize uncertainty and maximize data quality of the involved facts	selection of an optimal portfolio

The adaptation of the portfolio theory (according to [5]) requires the *validation of the assumptions* and requirements for a data warehouse environment.

1. single-period decisions
   Analytical processing in data warehouses usually takes place when the update window is closed (all available facts are not changing within this period). New

information provided at the next update window can be used to rerun the analysis and compare the result with the old one.

2. risk aversion and nonsatiation
   As already mentioned decision making in every context (e.g. investments, analytical processing) is dependent on the quality of its input parameters. Therefore the two objectives (risk aversion and nonsatiation) apply to data warehouse environments as well.
3. methods for the approximation of all input parameters
   Many of these parameters can be derived automatically in data warehouse environments (e.g. data quality assessments), but in general the correlation factors (needed for the covariance matrix) between facts have to be approximated manually.

Based on these assumptions and identified equivalences the area of application of portfolio selection can be extended to decision support in data warehouse environments. Analytical processing of data in "conventional" data warehouses using this new approach is explained in section 5. The integration of prognosis data and automatic consideration through portfolio selection is discussed in section 6.

5 Portfolio Selection in "Conventional" Data Warehouses

"*Conventional*" data warehouses contain data from operational and legacy systems but do not include prognosis data as proposed in the next section. During the data cleansing process the data quality and its uncertainty is assessed (in general the quality of data depends on the applied context). The adapted portfolio selection method can be used two-fold in data warehouse environments:

1. Additional quality information for whole analysis results
   The result of the portfolio selection - the calculated weight vector (which describes the influence that each involved fact should have to achieve the optimal portfolio) is used to derive the data quality and uncertainty assessment for the whole analysis result. Therefore the components (a_i, $i=1,\ldots,n$) of the weight vector are inserted into formulas 3 and 4 mentioned in section 3. The derived quality statement for the whole analysis gives the user an idea of the quality of the involved facts.

2. *Risk-driven* decisions
   The indifference curves describing the user requirements on the quality properties of the analysis result are replaced by only one straight line that is equivalent to the user's requirement (e.g. resulting data quality should be high = 0.9 with minimized uncertainty in the quality assessment). We call this case a *risk-driven* decision because the data quality and the uncertainty assessments of the involved facts are the steering factor of such an analysis. If it is possible due to the quality of the involved facts, the user's quality requirement will always be met precisely by the approximation of the point of intersection between the (calculated) efficient set and the straight line (which represents the user's quality requirement).
   Furthermore the user's uncertainty of choosing the right weights for the influence of facts in a complex analysis will disappear, because Markowitz showed that the calculated weight vector is optimal (under the assumption of correct input parameters).

Example for using portfolio selection in the context of a risk-driven decision:

The analysis, which is the basis for the risk-driven decision considers three facts. They are given by their data quality and uncertainty assessments (see Table 2) and their correlation factors (ρ_{AB}=0.53, ρ_{AC}=0.706 and ρ_{BC}=0.209). All together there are 9 input parameters which are utilized to identify the *corner portfolios C(i)* by using the *critical line method* [4]. The result of this step is illustrated in Table 3.

Table 2. Characterization of the Three Facts Involved in the Analysis.

Fact	Data quality	Uncertainty
A	76.2%	12.08%
B	84.6%	29.22%
C	82.8%	17.00%

Table 3. Computed Corner Portfolios.

Corner portfolio	Assessments[4]		Weight vector of the corner portfolio		
	Data quality	Uncertainty	A	B	C
C(1)	84.60%	29.22%	0.00	1.00	0.00
C(2)	83.20%	15.90%	0.00	0.22	0.78
C(3)	77.26%	12.22%	0.84	0.00	0.16
C(4)	76.27%	12.08%	0.99	0.00	0.01

At this point in the analysis the user's data quality requirement is taken into consideration. In general a data quality requirement $r°$ leads to a linear equation (formula 5), where r^C and r^D denote the data quality assessments of two suitable adjacent corner portfolios. The factor *Y* in this equation describes the influence ratio of the corner portfolios to meet the user's data quality requirement.

$$r° = Y * r^C + (1-Y) * r^D. \qquad (5)$$

In our example the user's data quality requirement is 0.8 which lies between the two adjacent corner portfolios: *C(2)* and *C(3)*. Solving the linear equation (formula 5) then yields *Y = 0.46*. The combination of this result with the weight vectors of the two adjacent corner portfolios, derives a new weight vector that describes the necessary influence of the considered facts to meet the user's data quality requirement exactly:

$$Y * C(2) + (1-Y) * C(3) = 0.46 * \begin{pmatrix} 0.00 \\ 0.22 \\ 0.78 \end{pmatrix} + 0.54 * \begin{pmatrix} 0.84 \\ 0.00 \\ 0.16 \end{pmatrix} = \begin{pmatrix} \mathbf{0.45} \\ \mathbf{0.10} \\ \mathbf{0.45} \end{pmatrix}$$

The combination of the three considered facts with the weights A=0.45, B=0.10 and C=0.45 will achieve an assessed data quality of 0.8 and *minimized uncertainty*.

We call this a *risk-driven decision* because the uncertainty about data quality is the steering factor in the analysis.

[4] The data quality and uncertainty assessments are computed using formulas 3 and 4 (mentioned in section 3) with the according weight vectors of the corner portfolios.

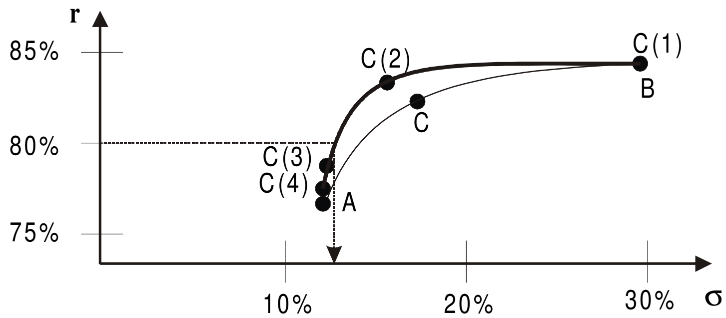

Figure 3: Analysis under Consideration of a Data Quality Requirement of 0.8 and Minimized Uncertainty.

6 Integration of Prognosis Data

As already mentioned in section 2 it will be advantageous to integrate additional regular and prognosis data into a data warehouse, if there is a strong dependency between predictions about the future and decisions made to affect the future.

The characterization of regular data through data quality and uncertainty assessments (section 4) is necessary to apply the presented approach and will be used to integrate prognosis data as well:

The *data quality* of a prediction is equivalent to the trustworthiness from the data warehouse users' viewpoint. This assessment is dependent on criterions similar to that of regular data [10], for instance completeness, reliability, consistency and believability of the provided predictions.

The *uncertainty* of data quality assessments can be provided by external sources, where each reasonable prognosis contains information about the (statistical) risk of deviations due to the assumptions (e.g. survey results on a specified target group). If such additional information is not available, the uncertainty has to be evaluated in the context of the data warehouse environment manually.

Table 4. Regular vs. Prognosis Data (in general)

	Regular Data	Prognosis Data
Data quality assessment	high	low - high
Uncertainty assessment	low	medium - high

- **Exclusive usage** of prognosis data in analytical processing:
  The adapted portfolio selection approach can be applied to regular and prognosis data as well without any alterations, because both are assessed by their data quality and uncertainty.
- The integration of **regular and prognosis** data in the same analysis to generate new predictions requires further investigation to achieve reasonable results:
  In general prognosis data is distinguishable from regular data by their mean quality and uncertainty assessments, as described in Table 4. If we apply the presented portfolio selection approach to an analysis involving both kinds of information at the same time without any alteration, we will achieve an

unintended result: Regular data will be privileged, because in general the mean data quality assessment will be higher than the one of the predicted data.
A solution for this situation is the introduction of a *displacement* that is added temporarily to the data quality assessments of prognosis data (only during the portfolio selection). This causes an increase in the data quality assessment, but the uncertainty assessment remains unchanged. Small displacements will favor regular data (suitable for short-term predictions).

7 General Limitations

The presented method for identifying an optimal portfolio for analytical processing has the following limitations:

- Complexity
  The number of input parameters that are needed to perform portfolio selection according to Markowitz increases in quadratic order with the number of the involved facts: $O(n^2)$. In general a portfolio selection with n facts involved requires
  $$\frac{n(n+1)}{2} + 2n = n\left(\frac{n-1}{2} + 2\right) \text{ input parameters.} \quad \textbf{(6)}$$
- Metrics for measuring data quality
  After more than a decade of intensive research in the area of data quality [13] there is still no consensus about a "fully featured" data quality model providing precise metrics. However data quality assessments are an important input parameter for the portfolio selection.
- Determination of the correlation between facts
  Data warehouses are characterized by a huge amount of facts that is continuously increasing. Whereas there are several methods [11] to examine data quality aspects, there is no one to automatically determine the correlation between facts.
- Application of the results
  In general weight vectors are only applicable to continuous value domains (e.g. numbers) but not to discrete domains, like strings or booleans.

8 Related Work

This work is related to [2], where a first draft of the approach is described and is further extended in this paper.

In the area of data warehouse research the management of data quality is an emerging topic, where approaches from other research fields (as we did in this paper) are investigated [12], [6] to provide better information for data consumers.

In the context of investments there are some other models that require fewer input parameters (e.g. Index Model [8]: n securities → $3n+2$ parameters), but they are not directly applicable to data warehouse environments. Furthermore they require additional information, which can be provided by finance markets but not by data warehouses (at this point in time, because of the lack of appropriate methods).

9 Conclusion

Based on the portfolio theory we identified equivalences in decision making in the context of finance markets and data warehouses and adapted Markowitz's approach.

The introduction of this new kind of analytical processing provides two benefits: 1) a derived quality statement for the whole analysis, that gives an idea of the properties of the involved facts and 2) risk-driven decisions, where uncertainty about data quality is a steering factor in the analysis. The approach can be applied to regular and prognosis data as well.

The presented approach for identifying an optimal portfolio that minimizes the assessed uncertainty and maximizes the data quality of the involved facts during analytical processing requires the further estimation of all covariances between the facts, which is in general difficult to automate.

Our future work will concentrate on discovering efficient solutions to reduce the high number of required input parameters to apply this approach.

References

1. Ballou, Donald P., Tayi Giri K.: *Enhancing Data Quality in Data Warehouse Environments.* Communications of the ACM, January 1999, Vol. 42(1), 73-78
2. Bruckner, Robert M.: *Datenqualitäts- und Evaluierungsaspekte von Data Warehouses (in german).* Master Thesis, Vienna University of Technology, Institute of Software Technology (E188), 1999
3. Markowitz, Harry M.: *Portfolio Selection.* Journal of Finance, Vol. 7(1), March 1952, 77-91
4. Markowitz, Harry M.: *The Optimization of a Quadratic Function Subject to Linear Constraints.* Naval Research Logistics Quarterly 3, nos. 1-2 (March-June 1956), 111-133
5. Markowitz, Harry M.: *Portfolio Selection.* Yale University Press, New Haven, Connecticut, 1959
6. Redman, Thomas C.: *Data Quality for the Information Age.* Artech House Publishers, Boston, 1996
7. Redman, Thomas C.: *The Impact of Poor-Data Quality on the Typical Enterprise.* Communications of the ACM, February 1998, Vol. 41(2), 79-82
8. Sharpe, William F.: *Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk.* The Journal of Finance, Vol. 19(3), September 1964, 425-442
9. Sharpe, William F., Alexander, Gordon J., Bailey, Jeffery V.: *Investments.* 6th Edition; Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1998
10. Strong, Diane M., Lee, Yang, Wang, Richard Y.: *Data Quality in Context.* Communications of the ACM, May 1997, Vol. 40(5), 103-110
11. Tayi Giri K., Ballou, Donald P.: *Examining Data Quality.* Communications of the ACM, February 1998, Vol. 41(2), 54-57
12. Wang, Richard Y.: *A Product Perspective on Total Data Quality Management.* Communications of the ACM, February 1998, Vol. 41(2), 58-65
13. Wang, Richard Y., Storey Veda C., Firth Christopher P.: *A Framework for Analysis of Data Quality Research.* IEEE Transactions on Knowledge and Data Engineering, August 1995, Vol. 7(4), 623-639