Creating Informative Data Warehouses: Exploring Data and Information Quality through Data Mining

Herna L. Viktor and Wayne M. Motha
University of Pretoria, Pretoria, Republic of South Africa

hlviktor@hakuna.up.ac.za wmotha@hakuna.up.ac.za

Abstract

Increasingly, large organizations are engaging in data warehousing projects in order to achieve a competitive advantage through the exploration of the information as contained therein. It is therefore paramount to ensure that the data warehouse includes high quality data. However, practitioners agree that the improvement of the quality of data in an organization is a daunting task. This is especially evident in data warehousing projects, which are often initiated "after the fact". The slightest suspicion of poor quality data often hinders managers from reaching decisions, when they waste hours in discussions to determine what portion of the data should be trusted. Augmenting data warehousing with data mining methods offers a mechanism to explore these vast repositories, enabling decision makers to assess the quality of their data and to unlock a wealth of new knowledge. These methods can be effectively used with inconsistent, noisy and incomplete data that are commonplace in data warehouses.

Keywords: Data warehouse, Data mining, Data and Information Quality

Introduction

High performance organizations are able to produce high quality products and/or services with limited resources through the continuous improvement of productivity and quality. In order to achieve this objective, aspiring organizations are increasingly investing in data warehouses, constructed to obtain relevant information in order to ensure a competitive advantage in today's global market. It follows that poor quality data has a serious detrimental effect on the quality of the information and the resultant decisions made.

According to a survey conducted by Redman (Redman, 1996), a typical operational data repository contains 1% to 5% incorrect values. There is subsequently a temptation to dismiss the data quality problem as consisting only of a series of anecdotes. However, these anecdotes are just too numerous and involve every segment of the economy (Redman, 1996). In a data warehouse, which is a subject-oriented, time-variant data repository of integrated data, originating from various operational data sources, the negative effect of poor quality data is augmented. It follows that the poor quality of data may lead to incorrect decisions to be made, which in turn has a detrimental effect on the organizational performance. In addition, the implicit assumption that the data does in fact relate to the organization from which it was drawn, and thus reflects the organizational processes, is often not tested (Pyle, 1999).

Usually in an operational database, those attributes which are deemed necessary for the daily operations of the organization, are current and correct. For example, a client's billing address will usually be present, correct and up to date. The data requirements of data warehouses, which are mainly used for decision support, differ substantially from that of the operational repositories. Here, the "additional" data about clients, for example their job description and

number of dependents, may be of crucial importance for the decision maker.

Data mining and its related technologies provide a mechanism to assess and improve the quality of data warehousing data. Data mining tools, such as decision trees and rule induction programs, may be used to assess the quality of the data in a data warehouse and the information that is subsequently produced. These techniques determine not only the quality of the data, but also indicate the lack of attributes that needs to be captured. Data reduction through feature and case selection provides a mechanism to eliminate redundancies from data and to improve the quality thereof. The employment of outlier detection to highlight surprises in the data, which may indicate incorrect organizational practices and data capturing procedures, has shown to produce new insights, providing both data warehousing builders and organizational decision makers with indicators of how to improve the quality of the data warehouse.

This paper is organized as follows. Section 2 introduces data warehousing as a mechanism to ensure a competitive advantage in today's global market. This is followed, in Section 3, with a description of information quality issues. Section 4 introduces data mining and Section 5 discusses the use thereof to assess and improve the quality of data warehouse data. Finally, Section 6 concludes the paper.

Data Warehousing for a Competitive Advantage

A data warehouse is usually constructed from a number of organizational databases, together with external data repositories, typically including governmental Census databases and other repositories containing relevant market indicators. Data warehouses provide historical snapshots of the organization, its evolution and its data capturing processes. In this way, it provides organizational decision makers with an "Audit" of the relevance and quality of its operational data stores. Figure 1 shows the interplay between the organization and the data warehouse. The organization uses the data warehouse as an indicator to assess and to subsequently adapt the organizational processes, where appropriate. In particular, the information retrieved from data warehouses may subsequently be used as a yardstick to determine the quality of both the operational transaction data, as contained in the source databases, as well as the resulting data contained in the data warehouse.

Data warehousing is often a daunting task, due to a number of factors, such as ensuring user buy-in, the merging of diverse data repositories and business requirements as well as the sheer complexity of the exercise. In particular, the extraction, transformation and transportation (ETT) of the data, and the associated data quality issues, are arguably the most time-consuming and tedious portion of any data warehousing effort. The ETT processes involve data format transformation, data consolidation and integration as well as metadata synchronization and management. Data warehousing practitioners agree that, when consolidating masses of data, the quality of the resulting data warehouse data is difficult to assess and guarantee. That is, good operational data may become poor quality data warehouse data, due to extractions and transformations that were not thought through carefully (Berson & Smith, 1997; Viktor, Pretorius, & Schoeman, & Blyth, 2001).

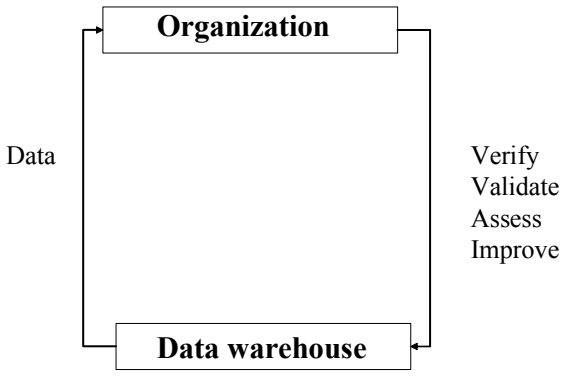

Figure 1: The interplay between the organization and the data warehouse

Numerous problems arise when integrating data from various data repositories, including the following (Berson & Smith, 1997; Viktor, Pretorius, & Schoeman, & Blyth, 2001):

- Source data anomalies, such as differences in naming and coding conventions, spelling mistakes, difference in meaning and missing values. A typical example of homonyms are where operational source A denoted the Gender field by values "M" and "F" where-as source B used values "1" and "2".
- Conflicting metadata and indexing, which need to be identified, resolved and transformed to a uniform format.
- Conflicting business rules and referential integrity issues, which are often not well documented. These rules reflect the business processes and implicit assumptions of the business unit that owns each operational database, which may be in direct conflict with the data warehouse's objectives.
- Different levels of summarization. For example, operational database A involves daily transactions, whilst the data as contained in an external source B contains monthly indicators.

The incorrect, unresolved extraction and transformation of the above-mentioned data conflicts will subsequently lead to a data warehouse containing unusable, poor quality data, as discussed next.

Information Quality Measures

According to Mallach (Mallach, 2000), data quality is measured by considering the degree to which the data produces information that contributes to a decision, i.e. the value thereof for decision-making. The information must be of sufficient quality for its purpose in terms of each of the following factors if it is to be of adequate decision support quality overall. The data should produce information that is relevant, correct, accurate, precise, complete, timely, usable, consistent, and conforms to expectations, as discussed below (Mallach, 2000).

- *Relevant:* Information is relevant if it applies to the task being performed – in our context, to making a decision. Its degree of relevance depends on how much of the information being supplied is related, and how much is not related, to that task. Knowing what information is relevant in a particular situation requires understanding the business. A data requirements study should be one of the first parts of any data-warehousing project even if the analysts are sure they know what is needed. However, it should be noted that data warehousing is especially subject to so-called "scope creep", where initially satisfied users continuously add additional information requirements. This important fact should be taken into account to ensure that the information is to be of the highest quality, and hence of the highest value, to decision makers and to the organization.
- *Correct, accurate, precise:* In addition to being relevant, high quality information must be correct. Incorrect information can lead to poor management decisions and eventually, dissatisfied customers. The correctness of a data element means that it is based on the right part of the "real world". If we were to look at the real world, what we would find would be consistent with what the data element led us to expect. Some correctness situations are absolute, where the data elements have no error tolerance. A telephone number is either correct or it is not. Other data elements are still correct, that is to say, they refer to the right item in the real world, if they are somewhat off from their actual value, e.g. stating that $\pi = 3.14159$. Approximate values are good enough for much numerical data. The question is often not whether numeric data are exact, because they cannot be, but whether they are close enough. As long as we take the possible error in our data into account when we use it, approximate data are not incorrect. Whether approximate data are of sufficiently high quality is the subject of the next quality factor, accuracy.

  The correctness of information depends on the correctness of the inputs and their processing. The major reason for incorrect input data is incorrect human input – as in an operational database. A well-designed data warehouse incorporates data validation procedures to minimize the amount of incorrect data introduced into storage. These validation procedures can be divided into several categories. All types can be enforced automatically in some fashion, though different features of the database man-

agement system, the data-entry process, and the application program may be required in each case. The designer of any application that incorporates a data-entry process should include checks from each category, or verify conclusively that the category does not apply.

The accuracy of a data element is a measure of how close it is to its real world value. This measure of information quality applies to items that have some error tolerance. These include most numerical quantities, unless they are used as identifiers such as a student ID number. It also includes many text strings that are long enough for people to derive meaning despite errors, and many pictures.

When assessing the accuracy of information, the following two principles should be kept in mind:

- The accuracy of a numeric information item is a function of the accuracy of the data elements that made the most important contributions to it. When data is retrieved from many data sources and used in a summary in a data warehouse, the accuracy of the data in the data warehouse is dependent on the accuracy of the data elements in the original sources.
- The difference between two data elements may be more accurate than either one is by itself. This happens when both are calculated in the same way from the same assumptions.

In these cases sensitivity testing can be valuable. The inputs can be varied a reasonable range to see what happens to the difference between two figures or images.

The precision of a data element is the maximum accuracy that can be represented by the way a data element is stored in a computer or presented to its users. Precision can be visualized as a grid overlaid on the world, either internally to an information storage system, or externally when data are presented to their users. Excess precision raises two problems, i.e. it may not reflect the underlying data properly, and it may make the data hard to use.

The challenge with the issues of correctness, accuracy and precision of data is to be consistent as to where data, especially numeric data is to be rounded-off. As stated earlier, preciseness based on "inaccurate input", is useless.

- *Complete:* High quality information must be complete. Completeness means that it includes all the necessary elements for the decision that is to be made, and that each element is based on all relevant factors. Omitting the sales figures of an important region from a data warehouse could mean incompleteness to the user and therefore it is perceived as of lower quality.
- *Timely:* Although a data warehouse is a historic snapshot, based on summarized data, the timeliness of an information element refers to relationships among three items:
  - The time that information is needed.
  - The time that information is made available.
  - The time that its underlying data was obtained.

  Information timeliness has two characteristics:
  - Information must be available in time for its intended use. Information value often deteriorates with time. Many business decisions are worth more if made earlier.
  - Information must reflect up-to-date data. The information should thus not be outdated.

  These requirements can only be met if the designers of the data warehouse know and understand the user requirements. This asks for continuous user-involvement in the data warehousing design and development.
- *Usable:* The usability of information is how quickly and easily its intended users can figure out what they need to know from what they see. Information presented so that they can complete tasks without

extra effort is usable. High quality information is presented in a usable format. The best format for presenting information depends on the people who will use the information and how they will use it. The date warehouse designer should understand the user's needs and provide information in the format that best meet those needs.

- *Consistent:* Consistency means that all data elements that contribute to an information item, or to a set of related items, are based on the same assumptions, definitions, and time period. Consistency is an attribute of information, or processed data, that does not apply to individual data elements. Users of data warehouses must often bring together many data elements from a variety of sources or consolidate thousands of records from a database.

  Not only should the information retrieved from a data warehouse be consistent, but also the interfaces. The interface of a data warehouse must reflect the needs of the people from various fields who will use it, not those who will develop it.

- *Conform*: Usually, information, which conforms to the expectations of the users, is considered to be of a high quality. It should be noted that data warehousing and the associated technologies of OLAP and data mining, provide users with new insights into the data. That is, new information may be obtained which challenges the assumptions being made. The detection of surprises in the data should be carefully investigated, in order to distinguish between nonsense and novel, new knowledge.

The challenge with the above-mentioned information quality issues is that, once users suspect these factors are not present in a data warehouse, the data repositories might become "useless data morgues". According to Redman (Redman, 1996), the slightest suspicion of poor quality data often hinders managers from reaching decisions, when executives spend "half of their decision making time just arguing whose data is better".

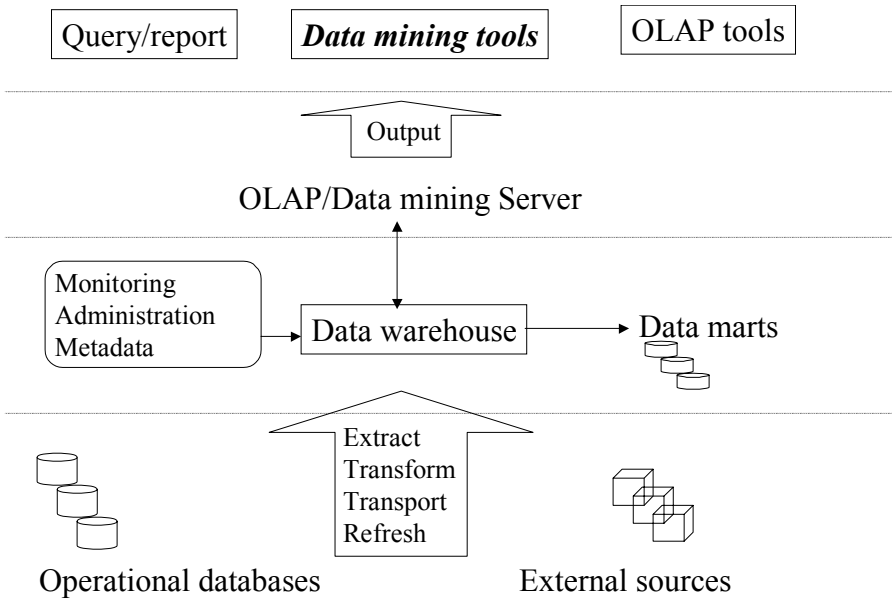

Figure 2: Data warehousing and data mining framework
(Adapted from [3])

It is therefore crucial that organizations eliminate data anomalies from the operational systems prior to migrating the data to the data warehouse. However, experience shows that anomalies are bound to exist also in the resulting data warehouse, leading to incorrect information and reporting. This has a serious detrimental effect on the users' perceptions of the information and reporting and also the frequency of use

thereof. Data mining provides a mechanism to assess the data warehouse and to improve its quality through data reduction techniques, as discussed next.

Data Mining from Data Warehouses

Data mining, which involves the automated extraction of so-called "knowledge chunks" from large data repositories, is increasingly being used as a tool for data warehousing access, as depicted in Figure 2. Data mining employs techniques from statistics, mathematics, database technology, artificial intelligence, economic theory and decision theory. Data mining tools are able to detect underlying structure and relationships in data warehouses that are difficult to find with traditional managed queries and OLAP tools. This new technology has been successfully used in many domains, including marketing, financing, medicine and engineering. Data mining can be used for classifying data, finding associations in data, the segmentation of data, the clustering of similar data into groups and outlier detection (Han & Kamber, 2000).

Data mining provides a mechanism to explore the data, as contained in the data warehouse, in order to find new information. It provides a tool for the decision makers within the organization to obtain his/her own new insights into the data without the explicit intervention of, or interaction with, the IT departments. That is, the decision maker is provided, especially through data visualization, with the possibility to "humanize" the mass of data (Berson & Smith, 1997).

In addition, the decision maker can assess the quality of the data through the application of data mining tools, such as decision trees, rule induction programs and association rule mining. In particular, visual data mining, which combines data visualization and data mining, is a highly attractive tool for the comprehension of data distributions, clusters, etc. Here, the visual display of the data and the results of data mining characterize the data within the data repository. In this way, the decision maker is provided with a clear impression and overview of the data characteristics (Han & Kamber, 2000). The decision maker can determine whether the data does in fact reflect his/her perception of the organization and its processes.

Data Mining for Information Quality Assessment

As stated above, data warehouses are, due to their huge size, susceptible to noisy, missing and inconsistent data. Data mining tools can help to identify and correct problems and are used to eliminate poor quality data.

Data mining can aid the decision maker to assess and improve the information quality as follows.

- *Relevant.* A large number of real world data repositories contain features (fields) or instances (rows) that are redundant and irrelevant to the organizational processes. These fields are usually obtained from legacy systems, where the outdated redundancies are included in the new systems. In addition, the daily organizational processes may capture data, which contains numerous irrelevant fields. The unnecessary use of organizational resources to capture these fields is detrimental to the organizational performance, since the redundant data merely lead to the creation of "data morgues". The inclusion of irrelevant fields indicates a data design problem, or hint at ineffective procedures or policies. For example, consider a traffic accident report that consists of 10 pages and includes 300 questions, many of which are redundant. Due to the time consuming nature of the report, many redundant fields are not completed or contain incorrect default values, which are thus useless for decision making (Viktor & du Plooy, 2001).

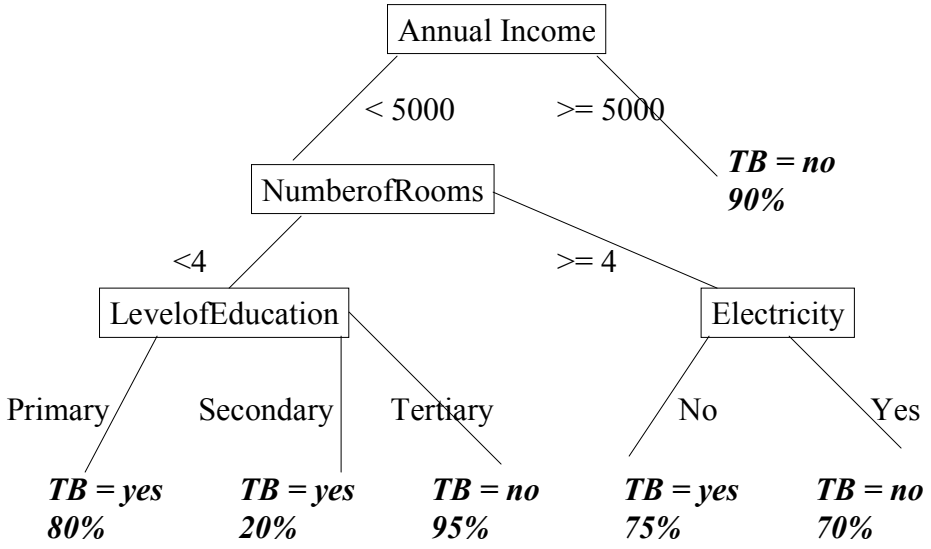

Figure 3: A (simplified) decision tree used to investigate the interplay between quality of life and TB occurrences

Feature selection involves the selection of those fields that are deemed important to describe the data repository. It can be used to identify irrelevant fields through the use of approaches such as sensitivity analysis, statistical analysis and the use of data mining tools such as decision trees and rule induction programs (Gray & Watson, 1996; Motha & Viktor, 2001). For example, consider a decision tree created from a data repository regarding interplay between quality of life and the occurrence of tuberculosis (TB) in South Africa (Viktor & du Plooy, 2001), as depicted in Figure 3. The figure implicitly indicates that the annual income, number of rooms, level of education and access to electricity are the important features to consider when attempting to eliminate the occurrence of TB within a particular region.

Case or instances selection involves the elimination of redundant instances (rows) from a data repository. Consider, for example, a data repository that captures an organization's daily operations. The existence of a large number of redundant or duplicate rows in a data warehouse not only shows poor ETT processes. It may, in addition, indicate that the organization captures unnecessary data and that the organizational databases need to be streamlined. For example, consider a Human Resource data warehouse constructed from a number of data sources. The duplication of a person's information indicates wasteful use of the organization's resources and shows that the procedures used to extract the data have been incorrectly designed (Viktor & du Plooy, 2001).

The removal of the appropriate poor quality instances from the repository not only improves the quality of the data warehouses, but also may subsequently be used when questioning the appropriateness of the data sources and ETT processes.

- *Correct, Accurate and Precise.* Recall that data warehousing is, due to its integrated nature, susceptible to errors. Outlier detection concerns the detection of surprises in the data, i.e. where data instances do not comply with the general behavior or model of the data (Gray & Watson, 1998). An outlier is a single, or very low frequency, occurrence of a value of a variable that is far away from the majority of the values of the variable (Pyle, 1999). For example, data mining may indicate that two patient records do not comply with the "normal" profile of TB patients. For example, a patient with an annual income of $12 000.00 was diagnosed as a confirmed TB case. Outliers are detected using statistical tests or data mining tools such as the nearest neighbor algorithms.

Many times, outliers are discarded as noise or exceptions (Pyle, 1999). However, from a data quality perspective, the rare event can be more interesting than the regularly occurring patterns. That is, outlier detection is especially useful to assess the quality of the data, since it may indicate faulty organizational processes and show that subsequent assumptions may have been wrongfully made. This is especially useful for auditing purposes, since outliers may, for example, indicate fraudulent behavior. From a data quality perspective, outlier detection may again indicate the presence of irrelevant data and hint at inadequate ETT processes. For example, the occurrence of a value of "N" in a Gender field (valid values "M" and "F") may indicate that a business entity was incorrectly included in a repository concerning individuals (Viktor & du Plooy, 2001).

- *Complete.* Data mining can fruitfully be used to point out those data items that are currently missing from the data warehouse. That is, the end results of a data mining effort may indicate that, due to incomplete data, little information could be found. For example, our recent KDD efforts regarding the investigation of traffic accident reports, showed that the amount of missing values in the original data was so high that the application of the discovery techniques could not be completed successfully without initiating new data capturing policies (Viktor & du Plooy, 2001). This is especially evident in data warehousing efforts that are initiated “after the fact”.

- *Timely.* It is crucial that decision makers obtain timely information to base their decision on. For example, recent data mining efforts concerned a data repository concerning a South African National Research and Technology (NRT) Audit conducted in 1995 (Viktor & Arndt, 2000). This study investigated, amongst others, the applicability of the technologies used within key industries, such as mining, rubber and plastics and civil construction, etc. Data mining results clearly indicated that the technologies used by the footwear and textiles industries were outdated, leading to the inability of these industries to compete internationally, as discussed next. The textile and footwear industry did not identify any key technology driver. However, the organisations that participated in the Audit indicated that exports should increase by the year 2000 due to the application of key technologies. The organisations did not indicate how this would be accomplished without the appropriate key technology drivers. It is interesting to note that the South African textile and footwear industry is currently experiencing serious difficulties and that a number of factories had to close down. This industry has difficulty competing internationally, since the production costs when using manual labour are higher than their international competitors’.
- *Usable.* In another data warehousing and data mining effort, the Census data regarding the Pretoria region of South Africa was analyzed (Viktor, Pretorius, Schoeman, & Blyth, 2001). The Census data includes information regarding the quality of life of the population, where quality of life is measured by considering a number of criteria, including the access to running water and electricity, health facilities, annual income, cooking facilities, number of rooms in house, household size and the level of education. Our initial investigation of the data showed that the usability thereof is very limited, due to a number of anomalies and the spatial nature thereof. For example, many individuals did not list their income or education level. Data mining efforts thus showed that decision makers who use this data for future prediction should proceed with caution.
- *Consistent*. Recall that, due to the integrated nature of data warehouses, they are susceptible to inconsistent data caused by homonyms and synonyms, amongst others. In many domains, the use of data mining, and especially outlier detection, to point out these inconsistencies has proved to be worthwhile. When considering the survey of human resources in science and technology, which formed part of the NRT Audit, many inconsistencies were discovered. For example, specific individuals’ gender was recorded as both Female and Male. Further investigation revealed that the data transformation process was incorrectly implemented. That is, all the values of data source A, transformed Female to

"1" whilst the Females from source B was transformed to "0". This human error was thus identified and could subsequently be rectified (Viktor, Pretorius, Schoeman, & Blyth, 2001).

- *Conform.* Data mining involves the discovery of previously unknown, non-trivial knowledge from data. For example, the data mining results may indicate that some of the decision makers' assumptions are wrongfully made. When considering the survey of human resources in science and technology, as introduced above, the decision makers were convinced that females in the applied biological sciences aren't employed in the private sector. The perception was based on the fact that the government sector dominates biotechnology within South Africa. However, the trends as discovered during learning contradicted the human experts and indicated that the food processing industry is a major employer of females trained in this discipline (Viktor & Arndt, 2000).

  Also, the trend that employees with a Masters or Doctoral degree in applied physical sciences were strongly absorbed by the private sector was previously unknown. This is due to the fact that the scarce applied physical science discipline was dominated by chemistry. The South African chemical industry is recognised as an international competitive industry. Due to its sophistication, this industry employs human resources similarly distributed to corresponding organisation types in Germany and the USA (Viktor & Arndt).

Conclusion

Providing organizational decision makers with clean, consistent and relevant data is an important, but expensive and time-consuming goal. In a high performance organization, where the employees are constantly aiming at improving the quality thereof, the data repository should provide an undisputed "single version of the truth" (Cykana, Paul, & Stern, 1996). High quality data enables the employees to make good decisions fast, whilst lower quality data lead to poor decisions, which in turn have a detrimental effect on organizational performance (Mallach, 2000).

Experience shows that the use of data mining and its associated data reduction techniques to assess and improve the quality of organizational data substantially increase the employee's trust in the data warehouse and its related technologies. The resultant "buy in" from organizational employees in order to constantly improve the organizational processes and systems, is crucial for developing a data quality policy ensuring that high quality data are captured and transported to the data warehouse (Motha & Viktor, 2001).

It should also be noted, however, that the data reflects and influences the organizational performance, and stressed that this fact should be taken into account when designing policies and procedures to assess and improve the quality of the data in the data warehouse. In this way, organizations are aided to become high performance organizations, i.e. organizations displaying a constant increase in productivity and quality.

References

Berson, A. and SJ Smith, 1997. Data warehousing, data mining and OLAP, McGraw Hill, New York: USA.

Cykana, P. A Paul, and M Stern, 1996. US Department of Defense guidelines on data quality management. Proceedings of the 1996 Conference on Information quality, Cambridge, MA: USA, pp.154-171.

Gray, P. and HJ Watson, 1998. Decision support and the data warehouse. The Data warehousing institute, Prentice-Hall, New Jersey: USA.

Han, J. and M Kamber, 2000. Data mining: Concepts and techniques, Morgan Kaufman, California: USA.

Kennedy, R. et al., 1997. Solving data mining problems through pattern recognition. Prentice Hall, New Jersey: USA.

Motha, W. M. and HL Viktor, 2000. Expanding Organizational Excellence: The Interplay between Data Quality and Organizational Performance, International Conference on Systems, Cybernetics and Informatics (SCI'2001), Orlando: USA, July 22-25, Volume XI, pp.60-65.

Mallach, E. F. 2000. Decision support and the data warehouse system, McGraw-Hill, New Jersey: USA.

Pyle, D. 1999. Data preparation for data mining, Morgan Kaufman, California: USA.

Redman, T. C. 1996. Data quality for the Information age, Norwood, MA: Artech House.

Viktor, H. L. and H Arndt, Combining data mining and human expertise for making decisions, sense and policies, Journal of Systems and Information Technology, Perth: Australia, 4(2), pp.33-56, Dec 2000.

Viktor, H. L. and NF du Plooy, 2001. Assessing and improving the quality of knowledge discovery data, Managing IT in a Global Environment: 2001 Information Resources Management Association International Conference, Idea Group Publishers, Hershey, PA: USA, pp.511-513.

Viktor, H. L. 2002. Creating High Quality Data Warehouses for E-Business, the International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet (SSGRR'2002), L'Aquila: Italy, January 21-27.

Viktor, H. L., J Pretorius, J Schoeman and C Blyth, 2001. Mining Spatial Data: Lessons Learned, Working paper, Department of Informatics, University of Pretoria.