IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4431

Assessing and Improving the Quality of Knowledge Discovery Data

Herna L Viktor and Niek F du Plooy
Department of Informatics, School of Information Technology, University of Pretoria, Pretoria, South Africa
Phone: +27 12 420 3376, Fax: +27 12 362 5287, Email: hlviktor@hakuna.up.ac.za, nduplooy@hakuna.up.ac.za

ABSTRACT

Data quality has a substantial impact on the quality of the results of a Knowledge Discovery from Data (KDD) effort. The poor quality of real-world data, as contained in many large data repositories, poses a serious threat to the future adoption of this new technology. Unfortunately, data quality assessment and improvement are often ignored in many KDD efforts, leading to disappointing results.

This paper discusses the use of data mining and data generation techniques, including feature selection, case selection and outlier detection, to assess and improve the quality of the data. In this approach, redundant low quality data are removed from the data repository and new high quality data patterns are dynamically added to the data set. We also point out that data capturing is part of the social practices of office work, and this fact must be taken into account in designing the data capturing processes.

1. INTRODUCTION

KDD is an exciting new technology that can be effectively used to obtain previously unknown patterns from large data repositories. However, experience shows that the quality of data in many real-world data repositories is unacceptably poor. According to Redman (1996), error rates of 1-5% are typical, with an estimated immediate cost of about 10% of revenue. These costs are amplified when poor data quality leads to the failure of KDD projects.

Poor data quality significantly impacts the application of the KDD process and the quality of the final results thereof. That is, large portions of data, which may contain important knowledge regarding the problem domain, may have to be discarded prior to data mining. The removal of substantial amounts of data may cause data mining tools to fail to find accurate and general concept descriptions. For example, our recent KDD efforts regarding the investigation of traffic accident reports, showed that the quality of the original data was so poor that the application of the discovery techniques could not be completed successfully without initiating new data capturing policies [Nel and Viktor, 1999]. The vast amount of available data could thus not alleviate the effect of poor data capturing and preprocessing.

Unfortunately, the importance of assuring high quality data is often understated [Weiss and Indurkhya, 1998]. Also, the implicit assumption that the data to be mined does in fact relate to the organization from which it was drawn and thus reflects the organizational processes, is often not tested [Pyla, 1999].

This paper proposes the use of data mining tools and data generation procedures to assess and improve the quality of organizational data. In this approach, data mining tools are used to identify low quality data. The resultant reduced data set is then used to generate new high quality data instances for subsequent data mining. In addition, we also emphasize the need for improved data capturing procedures.

The paper is organized as follows. Section 2 introduces the KDD process and discusses the impact of data quality on the final results of KDD. Section 3 presents methods to improve the quality of the data through the use of data mining techniques. Finally, Section 4 concludes the paper.

2. DATA QUALITY AND THE KDD PROCESS

The KDD process consists of three main stages, as shown in Figure 1. *Data preprocessing* involves the evaluation of the data to determine its appropriateness for the KDD project [Pyla, 1999]. Data preprocessing concerns the selection, evaluation, cleaning, enrichment and transformation of the data [Adriaans and Zantinge, 1997; Han and Kamber, 2000; Pyla, 1999]. The actual knowledge discovery stage in called *data mining*. Here, one or more techniques, such as decision trees or neural networks, are used to discover knowledge from the data. Finally, the *reporting* stage concerns the presentation of the results by means of a graphical user interface (GUI).

It can be argued that the results of the KDD process reflect the memory of the organization that is being investigated [Robey, *et al*, 1995]. That is, data are explored to discover knowledge about the organization, and ultimately, the world [Pyle, 1999]. Importantly, the KDD results can be viewed as a reflection of the quality of the data capturing and preprocessing processes. An understanding of the processes that are used to capture, generate, use and store the data are therefore essential to ensure data quality [Matheiu and Khalil, 1998] and to ensure the meaningfulness of the KDD process.

A survey conducted by Cykana *et al* (1996) lists the main causes of poor data quality according to four primary problem areas. These are process problems, systems problems, policy and procedure problems and data design problems. These problem areas are clearly also dependent on the social (work) practices that is followed in organizations and that are involved during data capturing in organizations. The way work is perceived (in the social sense) will determine the way information systems (and thus the associated capturing of data) is designed [Jones 1995; Käkölä, 1995; Pentland, 1995].

The fact that work (and thus data capturing) primarily is a social process is often ignored when information systems are designed. Systems are often approached mechanistically, that is, as a mere automation of current processes and procedures from the viewpoint of the systems designer. In this way the designed imposes his or her value system (eg. the way work should be done, and the order in which it should be done) on the data capturer [Du Plooy, 1998]. This may not be appropriate to the way the data capturer works or would like to work. It has been acknowledged

This paper appears in *Managing Information Technology in a Global Economy*, the proceedings of the Information Resources Management Association International Conference.

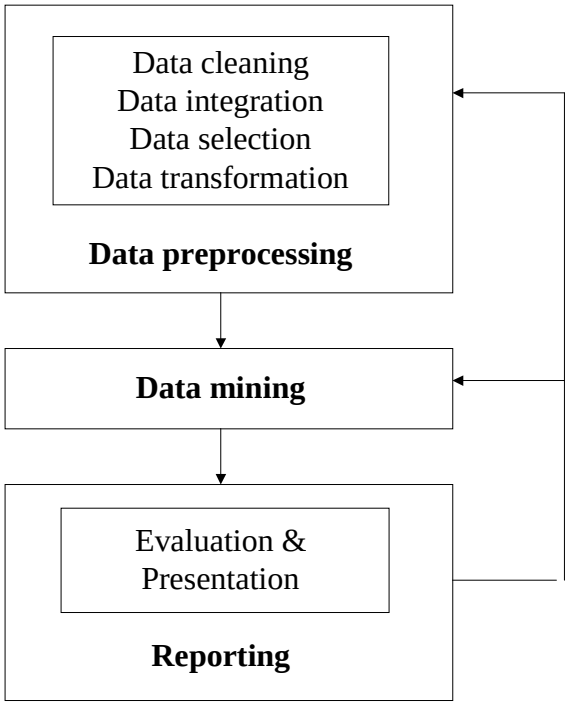

Figure 1: The KDD process (Adapted from [Han and Kamber, 2000])

that systems developers who are to design the data capturing techniques are not equipped (in the sense of the tools they use or the training they receive) to deal with the social processes intrinsic to information systems development [Hirschheim and Newman, 1991].

Rethinking the organization's approach to information system design may counter claims that office work in the information era often reflects Tayloristic work designs, focusing on the individual's task productivity to the exclusion of the social content of the work [Lyytinen and Ngwenyama, 1992; Fitzgerald and Murphy, 1994]. Mechanistic automation may result in increased control, but it also affects the autonomy of workers and depersonalises work. The effect of this may be that data capturing is done by rote, following the dictates of the system and thus resulting in erroneous and redundant data since the data capturers have no real interest in what they are doing.

The identification of specific problem areas, through the use of data mining tools, can improve the data capturing processes and thus facilitate organizational learning. The next section examines the use of data mining approaches that can be used to identify poor quality data and the problem areas that caused the creation thereof. The section also discusses how, after the removal of redundancies from the data repository, new high quality data can be generated.

1. DATA QUALITY THROUGH DATA MINING AND DATA GENERATION

Feature selection concerns the selection of those attributes that are deemed important to describe the data repository. That is, a subset of the features that are considered to be critical in order to adequately describe the data set is selected. Feature selection approaches include statistical analysis, sensitivity analysis and the use of data mining tools such as decision trees or rule induction algorithms to obtain the important features [Han and Kamber, 2000].

Many real-world data sets contain a number of features that are unimportant. When considering a tuberculosis data set containing 345 features, only seven of them were of importance [Viktor, *et al*, 1997]. In another data mining effort, approximately ten of the 4000 features considered when describing small business loans were actually important [Weiss and Indurkhya, 1998]. The presence of a large amount of redundant data, as identified during feature selection, shows that the data capturing processes are not adequately focused. Rather, the data repository acts as a "data morgue" in which all features, relevant or not, are placed. Storing data in a haphazard and unfocused manner may therefore also have a serious detrimental affect of the usability of this part of the organization's memory. This implies that the organization should rethink the validity of the data that are captured, thus streamlining their operational processes. For example, a traffic accident report should not blindly be streamlined by removing redundant features that are normally discarded by the traffic officers. Rather, the traffic accident and road condition expertise of, for example, construction engineers, rally drivers and medical personal should be used to update the report. This 'rethinking' could well start by examining the manner in which information systems design is approached. If systems are designed in a mechanistic manner, data will also be captured mechanistically, thus robbing the 'data capturer' of the opportunity to remove the redundancy and enhancing the validity.

Case or instance selection is used to eliminate redundant instances from the data repository, mainly to limit the size requirements to store data in memory [Brodley and Friedl, 1996; Cherkauer and Shavlik; 1996]. This approach has also been used to reduce decision tree size, thus improving human comprehensibility [Sebban and Nock, 2000]. The existence of a large number of redundant and irrelevant instances again indicates that the data repository has not been fine-tuned for data mining and hints to possible organizational problems that may result in poor quality data. The removal of the appropriate poor quality instances from the repository not only improve the quality of the data set, but may be subsequently used when questioning the appropriateness of the data sources and data capturing processes.

Outlier detection highlights surprises in the data, that is data instances that do not comply with the general behavior or model of the data [Han and Kamber 2000]. 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 [Pyla, 1999]. Outliers are detected using statistical tests or data mining tools such as the nearest neighbor algorithms.

Most data mining methods and practitioners discard outliers as noise or exceptions [Pyla, 1999]. However, from a data quality perspective, the rare event can be more interesting that the regularly occurring patterns. That is, outlier detection is especially useful to assess the quality of the data, since it may indicate that the organizational processes and subsequent assumption may be wrongfully made. The detection of many diverse outliers often indicates the presence of problems within one or more of the four areas identified in Section 2, leading to poor data capturing processes.

Data generation utilizes the results of feature selection, case selection and outlier detection to produce new data. Here, the data selection approaches produces a reduced data set that contains high quality data that has been shown to be of importance to adequately describe the problem domain. This data set may thus substantially reduce the large data repository to a small set of usable data to be used for data mining. However, experience has shown that many data mining tools have difficulty to generalize well if the number of instances is few. This is especially evident in

difficult to learn domains. Data generation addresses this problem through the automating generation of new instances that are based on the high quality data, as contained in the reduced data set identified earlier [Viktor, 2000]. The application of this approach to a real-world repository concerning a Human Resources data warehouse, showed that data generation can be used to generate sufficient data for effective data mining, even in domains where the initial data quality was poor [Viktor and Arndt, 2000].

Note that the above-mentioned data mining and data generation processes should be executed with the active participation of the members of the organization and the subsequent adaptation of the organizational processes, where appropriate. In this way, the problem areas can be identified and be rectified as far as possible. If this is not done, KDD may fall into the selfsame trap as information systems design in general, namely, approach the discovery of data mechanistically in the firm belief that the technology alone can take care of things [Postman, 1993].

2. CONCLUSION

According to Matheiu and Khalil (1998), the improvement of the quality of data in an organization is often a daunting task. This is especially evident in KDD projects, which are often initiated "after the fact". However, it is the opinion of the authors that the assessment and improvement of data quality during KDD can positively influence the organizational processes, highlight problem areas and facilitate organizational learning. This however, is dependent on heading the injunction that information systems design, and thus the design of data capturing processes, is approached less mechanistically and by taking the social context of the work of data capturing into account.

This paper concerned the use of data mining techniques, namely feature and case selection and outlier detection, to assess and improve the quality of the data. Also, the use of automatic data generation, in order to improve the quality of the data, was discussed. In addition, it was pointed out how ignoring the social side of data capturing may influence the quality of the data. These methods can be effectively used to with inconsistent, noisy and incomplete data that are commonplace in large, real-world data repositories.

REFERENCES

CE Brodley and MA Friedl, 1996. Identifying and Eliminating Mislabeled Training Instances, 13th National Conference on Artificial Intelligence, California: USA.

KJ Cherkauer and JW Shavlik, 1998. Growing Simple Decision Trees to facilitate Knowledge Discovery, Second International Conference on Knowledge Discovery and Data mining.

P Cykana, 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.

J Debenham, 2000. Knowledge Decay in a Normalized Knowledge Base, Data and Expert Systems Applications, Lecture Notes in Computer Science, Vol 1873, pp.417-436.

NF Du Plooy, 1998. An Analysis of the Human Environment for the Adoption and Use of Information Technology, (Unpublished) DCom dissertation, University of Pretoria, South Africa.

B Fitzgerald and C Murphy, C. 1994. Introducing Executive Information Systems into Organizations: Separating Fact from Fallacy, Journal of Information Technology, 9, pp.288-296.

J Han and M Kamberen, 2000. Data Mining: Concepts and Techniques, Morgan Kaufman, California: USA.

RA Hirschheim and M Newman, 1991. Symbolism and Information System Development: Myth, Metaphor and Magic, Information Systems Research, 2(1), pp 29-62.

M Jones, 1995. Organizational Learning: Collective Mind or Cognitivist Metaphor? Accounting, Management & Information Technology, 5(1), pp 61-77.

TM Käkölä, 1995. Increasing the Interpretive Flexibility of Information Systems through Embedded Application Systems, Accounting, Management & Information Technology, 5, 1, pp 79-102.

M Liskin, 1990. Can you trust your Database? Personal Computing, June 29, pp.129-134.

K Lyytinen and OK Ngwenyama, 1992. What does Computer Support for Cooperative Work Mean? A Structural Analysis of Computer Supported Cooperative Work, Accounting, Management & Information Technology, 2 (1), pp 19-37.

RG Mathieu and O Khalil, 1998. Data Quality in the Database Systems Course, Data Quality Journal, 4 (1), September.

C Nel and HL Viktor, 1999. Data Mining of Traffic Accident Reports, Progress Report, Department of Informatics, University of Pretoria, Pretoria, South Africa.

BT Pentland, 1995. Information Systems and Organizational Learning: The Social Epistemology of Organizational Knowledge Systems. Accounting, Management & Information Technology, 5, 1, pp.1-21.

N Postman, 1992. Technopoly: the Surrender of Culture to Technology, Vintage Books: New York.

D Pyle, 1999. Data Preparation for Data Mining, Morgan Kaufman, California: USA.

JR Quinlan, 1994. C4.5: Programs for Machine Learning, Morgan Kaufman, California: USA.

TC Redman, 1996. Data Quality for the Information Age, Norwood, MA: Artech House.

D Robey, NA Wishart and AG Rodriguez-Diaz, A.G. 1995. Merging the Metaphors of Organizational Improvement: Business Process Reengineering as a Component of Organizational Learning, Accounting, Management & Information Technology, 5. 1. pp.23-39.

M Sebban and R Nock, 2000. Contribution of Dataset Reduction Techniques to Tree-simplification and Knowledge Discovery, *to appear in* International Journal of Computers, Systems and Signals, December.

HL Viktor, I Cloete and N Beyers, 1997. Rules for Tuberculosis Diagnosis, Methods for Informatics in Medicine, 36 (2), pp.160-168

HL Viktor, 2000. Generating New Patterns for Information Gain and Improved Neural Network Learning, The International Joint Conference on Neural Networks (IJCNN-00), Como, Italy, 24 - 27 July.

HL Viktor and H Arndt, 2000. Data Mining in Practice: From Data to Knowledge using a Hybrid Mining Approach, *to appear in* the International Journal of Computers, Systems and Signals, December.

G Weiss and N Indurkhya, 1998. Predictive Data Mining: A Practical Guide, Morgan Kaufman, California: USA.