IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB7512

Chapter 10

Assessing and Improving the Quality of Knowledge Discovery Data

Herna L. Viktor and Niek F. du Plooy
University of Pretoria, South Africa

Copyright Idea Group Inc.

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 chapter 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.

Copyright Idea Group Inc.

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.

Copyright Idea Group Inc.

Previously Published in *Managing Information Technology in a Global Economy*, edited by Mehdi Khosrow-Pour, Copyright © 2001, Idea Group Publishing.
This chapter appears in the book, Data Warehousing and Web Engineering by Shirley Becker. Copyright © 2002, Idea Group Publishing.

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 chapter 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 chapter is organized as follows. The next section introduces the KDD process and discusses the impact of data quality on the final results of KDD. The following section presents methods to improve the quality of the data through the use of data mining techniques. Finally, the last section concludes the paper.

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

Copyright Idea Group Inc.

6 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/chapter/assessing-improving-quality-knowledge-discovery/7868

Related Content

General Model for Data Warehouses

Michel Schneider (2005). *Encyclopedia of Data Warehousing and Mining (pp. 523-528).*

www.irma-international.org/chapter/general-model-data-warehouses/10653

Exploiting Captions for Web Data Mining

Neil C. Rowe (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 1461-1485).*

www.irma-international.org/chapter/exploiting-captions-web-data-mining/7710

Data Mining for Credit Scoring

Indranil Bose, Cheng Pui Kan, Chi King Tsz, Lau Wai Kiand Wong Cho Hung (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 2449-2463).*

www.irma-international.org/chapter/data-mining-credit-scoring/7774

Off-Line Signature Recognition

Indrani Chakravarty, Nilesh Mishra, Mayank Vatsa, Richa Singhand P. Gupta (2005). *Encyclopedia of Data Warehousing and Mining (pp. 870-875).*

www.irma-international.org/chapter/off-line-signature-recognition/10719

Justifying Data Warehousing Investments

Ram L. Kumar (2002). *Data Warehousing and Web Engineering (pp. 100-102).*

www.irma-international.org/chapter/justifying-data-warehousing-investments/7863