2005

Data Quality in Very Large, Multiple-Source, Secondary Datasets for Data Mining Applications

Marilyn G. Kletke
Oklahoma State University, mkletke@okstate.edu

Dursun Delen
Oklahoma State University, delen@okstate.edu

Data Quality in Very Large, Multiple-Source, Secondary Datasets for Data Mining Applications

Marilyn G. Kletke
Oklahoma State University
mkletke@okstate.edu

Dursun Delen
Oklahoma State University
delen@okstate.edu

ABSTRACT

The data mining research community is increasingly addressing data quality issues, including problems of dirty data. Hand, Blunt, Kelly and Adams (2000) have identified high-level and low-level quality issues in data mining. Kim, Choi, Hong, Kim and Lee (2003) have compiled a useful, complete taxonomy of dirty data that provides a starting point for research in effective techniques and fast algorithms for preprocessing data, and ways to approach the problems of dirty data. In this study we create a classification scheme for data errors by transforming their general taxonomy to apply to very large multiple-source secondary datasets. These types of datasets are increasingly being compiled by organizations for use in their data mining applications. We contribute this classification scheme to the body of research addressing quality issues in the very large multiple-source secondary datasets that are being built through today's global organizations' massive data collection from the Internet.

Keywords

Data quality, dirty data, preprocessing, data mining, very large dataset, multiple-source data

INTRODUCTION

Data mining is becoming increasingly important as organizations continue to collect terabyte-class and larger amounts of data from their Internet-related activities. One of the problems in working with very large datasets is the poor quality of data that over time, often from many sources, finds its way into the organizational database. Chaudhuri, Ganjam, Ganti and Motwani (2003) address this quality issue as the crucial task of detecting and correcting anomalies in data. They focus on data warehousing, and further mention that many data problems therein are caused by missing data and by inconsistent inputs from multiple sources. Mueller, Leser, and Freytag (2004) focus on the problems of integrating information from different data sources, and to do so, they have developed a model for identifying inconsistencies and eliminating the source of those problems.

According to Arts, de Kaizer, and Scheffer (2002), data quality is a combination of quality assurance procedures and quality control procedures. Quality assurance procedures are those activities carried out prior to the collection of data to ensure that the study design and collection activities are consistent and well-conceived. Quality control refers to those activities that take place during and after data collection and focus on identifying and correcting sources of data errors. Because of the nature of very large, multiple-source secondary datasets, quality assurance procedures are not likely to be as well put together or as well known as are those for single-source primary datasets. Likewise, information to identify and correct sources of data errors are likely to be less dependable and therefore effective than are those for single-source primary datasets. With today's very large datasets, errors in data can be magnified during the data mining process, thereby biasing results. Zheng, Padmanabhan, and Kimbrough (2003) demonstrate in a case study of web-usage mining that three different preprocessing techniques in that domain resulted in significantly different conclusions, thereby highlighting the message that a careless or unknowledgeable choice of preprocessing activities and techniques can lead to biased and inaccurate results. The quality of the input data is extremely important for successful data mining.

In order to achieve better data quality, the preprocessing of such data is frequently necessary to remove anomalies and to increase the homogeneity of the data; and to produce a consistent and accurate dataset. The challenge of preprocessing in data mining is to modify the dataset so that it demonstrates better analytical behavior without changing its essential information content. This challenge is magnified in the sharply increasing size and dimensionality of datasets being collected by today's e-commerce activities, because there are potentially many different data problems that must be corrected or otherwise treated. A variety of techniques and tools exist that may be appropriate for improving data quality, while

simultaneously preserving the information content embodied within the data. Preprocessing takes a disproportionate share of the time and effort involved in a data mining project; especially so where multiple sources of data are involved. It would be very useful (and possibly essential) to examine the causes and underlying structures for dirty data and accordingly develop methodologies to effectively and efficiently preprocess the data. Two relevant research questions are:

1. Is dirty data as identified for data in general meaningful in the same way for very large, multiple-source secondary datasets?
2. If not, how should we label and explain dirty data for very large, multiple-source secondary datasets?

The objective of this study is to address these two research questions associated with data quality. Because we believe that dirty data in very large, multiple-source secondary datasets differs in significant ways from dirty data in single-source primary datasets, we propose a classification scheme for the types of data errors that exist in these types of datasets. The classification scheme will allow the development of a set of preprocessing activities associated with the categories in the model. Ways of carrying out these preprocessing activities can then be suggested. This classification will serve as the blueprint for the future construction of a normative process/activity model for preprocessing data mining input from very large, multiple-source secondary datasets.

BACKGROUND

Considerable research is being done on quality issues with respect to data used in analysis (Davidson, Grover, Satyanarayana, and Tayi, 2004, Pierce, 2004). It is true that most data collected from multiple sources, including, for example, legacy databases, is, to some extent, dirty. Some metrics for assessing data quality have been proposed. Pierce (2004) and Davidson et al (2004), for example, are working with quality and control matrices to assess and improve data quality. Pipino, Lee and Wang (2002) identify dimensions of quality to try and bound the problem. They maintain that to date principles to guide the development of usable metrics for quality haven't been developed. Orr (1998) constructs some general data quality rules that can reduce the incidence of dirty data. Kim et al (2003) have developed an overarching taxonomy in which they define dirty data as that which generates the wrong results due to something inherent to the data. They, too, point out that the value of high quality data for use by applications is only recently being studied, and that there is lack of research that defines "quality" data or that provides a set of metrics for measuring this quality.

Reacting to Kim et al's (2003) taxonomy, Hand and Bolton (2004) suggest that there are high-level and low-level data quality issues. High-level issues affect entire population distributions; and low-level issues reference individual values in individual records that are missing or wrong. According to Hand and Bolton (2004), Kim et al's (2003) taxonomy does not categorize distributional distortions arising from entire records being missing, which qualifies as a high-level data issue. Quality issues become particularly meaningful when related to very large datasets because data from multiple sources is frequently aggregated into (conceptually) one dataset to form the input source for the data mining process (Kletke, Delen, and Kim, 2004). With single-source datasets, accompanying information about the data can likely be used to help resolve some data errors. When data is obtained from more than one source, there is an increased danger of heterogeneity in the way that the data is collected, structured, and coded; and this can lead to significant problems in the data and the analyses that rely on that data. Information about the data may not help in resolving data errors that stem from the way data was collected and/or coded. It is also possible that the original sources of individual data items can be lost in the aggregation process, and that as a result, information that could shed light on data errors can be obscured or lost. Because we are focusing on low-level data quality issues, Kim et al's (2003) general taxonomy provides us with a good starting point for classifying low-level dirty data for multiple-source secondary datasets.

RESEARCH APPROACH

Many data mining applications operate on very large datasets, and the data quality can significantly impact results. In this study, we consider the state of data after it has been gathered and aggregated from (possibly) multiple sources, but before any preprocessing has been done. This is commonly called "secondary data" because it has already been entered, coded and stored in a prior stage of data generation by someone else. Secondary data, before any preprocessing is done, can be problematic or dirty in ways that differ from data obtained from primary datasets, and especially single-source datasets. A robust classification scheme for data errors in these very large, multiple-source secondary datasets is needed to lead to understanding of quality and to help develop appropriate quality metrics. The purpose of such a scheme is to clearly delineate classes of errors such that each subclass can be treated with a specific and unique methodology for eliminating or repairing data errors. From this, "correct" and productive preprocessing activities can be designed and tailored for each type of error as defined in the classification scheme. The classification will provide a foundation for designing quality metrics for each of its subclasses. It is apparent from the derived classification that preprocessing activities for multi-source secondary datasets will

be different from those that can be constructed for single-source datasets where there is more consistency in the data collection and more information about the origin of the data is known.

To build such a scheme, we started with Kim et al (2003)'s general taxonomy, which applies generally to data whose origin is well known, and transformed it into a classification scheme specific to very large multiple-source secondary datasets. Kim et al's (2003) taxonomy is able to specify categories and subcategories that are associated with causal details for dirty data that are very specific because of the fact that the origin is known. For example, in their taxonomy, missing data results from two possible problems: "no Null-not-allowed constraints" or "Null-not-allowed constraints not enforced." In the secondary data that is compiled for data mining applications, we generally do not know why certain data is missing – only know that it is missing. Another example is that in secondary data, although the nature or type of data problem is often known, the specific reason for the data problem is generally not known. It is generally impossible, for instance, to tell whether there is a data problem in a particular value due to failure of transaction concurrency control in the original collection of the data, much less whether it was a lost update or a dirty read that led to the problem in the data.

THE DERIVATION OF THE MULTIPLE-SOURCE SECONDARY DATASET CLASSIFICATION SCHEME

To classify dirty data for multiple-source secondary datasets, we considered each of the categories in Kim et al's (2003) taxonomy. We evaluated whether the causal information specified in the subcategories could be identified apart from the parent category. If not, then the specific subcategory was collapsed into the more general parent category. For convenience in visualizing the derivation, Table 1 contains an outline of the higher-level categories in Kim et al's (2003) taxonomy.

Missing
Missing data where there is no Null-not-allowed constraint
Missing data where Null-not-allowed constraint should be enforced
Not Missing
Wrong
Non-enforcement of automatically enforceable integrity constraints
Non-enforceability of integrity constraints
Not wrong, but unusable
Different data for the same entity across multiple databases
Ambiguous data
Non-standard conforming data
Different representations due to non-compound data
Different representations of compound data

Table 1. The higher-level hierarchical categories of Kim et al's (2003) taxonomy

The taxonomy's top-level hierarchical category is "missing data" or "not-missing data." Missing data consists of data where there is no "Null-not-allowed" constraint and data where "Null-not-allowed constraint" should be enforced. In secondary data we don't necessarily know the reason for missing data, so we eliminate the two specific subcategories and have one general parent category called "missing data."

The "not-missing" data category contains two subcategories: "wrong data" and "not wrong, but unusable data." The parent category, "wrong data," contains two subcategories: "Non-enforcement of automatically enforceable integrity constraints," and "Non-enforceability of integrity constraints." It cannot be assumed that data collected for use in data mining comes from relational databases having enforced integrity constraints, although some may well do so. In fact, the data may not come from a relational database at all. As a result, we have collapsed Kim et al's (2003) 17 types of wrong data into three remaining categories: this data is either out of the domain; or it is in the domain, but keyed in wrong; or it contains extraneous data. We cannot count on having the information about the source(s) of the data that is needed to expand these three categories due to specific sources of dirty data.

The "not-wrong-but-unusable" data category contains 14 subcategories. We kept the three second-tier categories, but eliminated some indistinguishable subcategories, yielding 9 types of not-wrong-but-unusable data. Eliminated were the two subcategories of ambiguous data (we left the parent category called ambiguous data); and the subcategory of "Different

representation of compound data" named "Hierarchical data" along with its 3 subcategories, which we believe are incorporated into the different representations of the compound data category. We added a category to represent the data problems that may occur when datasets from significantly different time periods or spatial locations are compiled together. For example, if one is using salary data coming from the 1970 period, say, and from the 2000 period. These data must be normalized before working with, for example, univariate statistics to compensate for the effects on salary of time (e.g., inflation). Our classification scheme contains 15 problem data areas, down from 33 in Kim et al's (2003) taxonomy. Table 2 shows our classification scheme of errors for multiple-source secondary datasets.

1. Missing data
2. Not missing
   - 2.1 Wrong data
     - 2.1.1 Out of allowable domain (illegal domain value or category code)
     - 2.1.2 Wrong entry (in domain, but keyed in or generated wrongly at the source)
     - 2.1.3 Extraneous data
   - 2.2 Unusable data
     - 2.2.1 Different (conflicting) data from multiple sources (one says birth year of 1945 and another says birth year of 1955)
     - 2.2.2 Ambiguous data
     - 2.2.3 Data not conforming to same reporting standard
       - 2.2.3.1 Due to different representations of non-compound data
         - 2.2.3.1.1 Algorithmic transformation not possible
           - 2.2.3.1.1.1 Different abbreviations
           - 2.2.3.1.1.2 Aliases for a name/place/thing
         - 2.2.3.1.2 Algorithmic transformation possible
           - 2.2.3.1.2.1 Different encoding formats (classification code changes; change in the way something is reported)
           - 2.2.3.1.2.2 Different display formats (currency, dates, negative numbers, etc.)
           - 2.2.3.1.2.3 Different measurement units
       - 2.2.3.2 Due to different representations of compound data
         - 2.2.3.2.1 Abbreviated version (J. Kennedy instead of John Kennedy)
         - 2.2.3.2.2 Inconsistent use of special characters
         - 2.2.3.2.3 Different orderings of data (e.g. Smith, John vs. John Smith; or city county state vs. city state)
     - 2.2.4 Inconsistent temporal or spatial data (e.g., salary from 1970 merged with salary from 1990)

Table 2. Classification scheme for dirty data for very large, multiple-source, secondary datasets

Table 3 shows the numbers of categories in Kim et al (2003)'s taxonomy and the numbers of categories in our classification scheme, which is an adaptation for very large multiple-source secondary datasets.

	Kim et al (2003)	Multiple-source Datasets
Missing Data	2	1
Not-missing but Wrong	17	3
Not-missing but Unusable	14	11
Total	33	15

Table 3. Comparison of number of categories in Kim et al's (2003) taxonomy with the classification scheme developed in this study (specific to very large, multiple-source, secondary datasets)

SUMMARY AND FUTURE RESEARCH

Very few researchers have attempted to address quality issues with respect to very large multiple-source secondary datasets, and appropriate metrics for measuring the quality of this sort of data have not been developed. This study advances quality assessment efforts by constructing a specific data error classification scheme that can be used to further the research of quality issues in very large multiple-source secondary datasets.

We plan to use this classification scheme as a foundation toward developing detailed preprocessing activities necessary for very large multiple-source secondary datasets; and to summarize and prioritize the techniques and tools that can be (or should be) used in those preprocessing activities. In order to outline the methodology, we will develop a normative process/activity model (using IDEF0 and IDEF3 methods) for preprocessing of data for data mining applications on very large, multiple-source secondary datasets. (The reader can learn about public domain IDEF modeling methods developed by the U.S. Air Force in the systems development literature, Whitman, Huff, and Presley 1997, and/or on the IDEF web site at http://www.idef.com).

REFERENCES

1. Arts, D., de Keizer, N., and Scheffer, G. (2002) Defining and improving data quality in medical registries: a literature review, case study, and generic framework, *Journal of the American Medical Informatics Association*, 9, 600-614.
2. Census (1990) - http://www.macalester.edu/econdata/United_States/pums.html
3. Chaudhuri, S., Ganjam, K., Ganti, V., and Motwani, R. (2003) Robust and efficient fuzzy match for online data cleaning, *Proceedings of SIGMOD 2003*, San Diego, CA.
4. Davidson, I., Grover, A., Satyanarayana A., and Tayi, G. (2004) A general approach to incorporate data quality matrices into data mining algorithms, *Proceedings of the 2004 ACM SIGKDD*, 794-798.
5. Hand, D., Blunt, G., Kelly, M., and Adams, N. (2000) Data mining for fun and profit, *Statistical Science*, 15, 111–131.
6. Hand, D., and Bolton, R. (2004) Pattern discovery and detection: a unified statistical methodology, *Journal of Applied Statistics*, 31, 8, 885-924.
7. Kim W., Choi, B., Hong, E., Kim, S. and Lee, D. (2003) A taxonomy of dirty data, *Data Mining and Knowledge Discovery*, 7, 81-99.
8. Kletke, M., Delen, D., and Kim, J. (2004) An approach to improve classification accuracy in very large datasets, *Proceedings of the AMCIS 2004*, New York.
9. Mueller, H., Leser, U., and Freytag, J. (2004) Mining for patterns in contradictory data, *Proceedings of the 2004 international workshop on Information quality in information systems*, Paris, France, 51-58.
10. Orr, K. (1998) Data quality and systems theory, *Communications of the ACM*, 41, 2, 66-72.
11. Pierce, E. (2004) Assessing data quality with control matrices, *Communications of the ACM*, 47, 2, 82-86.
12. Pipino, L., Lee, Y. and Wang, R. (2002) Data quality assessment, *Communications of the ACM*, 45, 4, 211-218.
13. Whitman, L., Huff, B., and Presley, A. (1997) *Structured models and dynamic systems analysis: the integration of the IDEF0/IDEF3 modeling methods and discrete event simulation*, Proceedings of the 1997 Winter Simulation Conference, 518-524.
14. Zheng, Z., Padmanabhan B., and Kimbrough, S. (2003) On the existence and significance of data preprocessing biases in web-usage mining, *INFORMS Journal on Computing*, 15, 2, 148-170.