Data Quality: Developments and Directions

Bhavani Thuraisingham* and Eric Hughes

The MITRE Corporation, Bedford MA, USA

** On Leave at the National Science Foundation, Arlington VA, USA*

Abstract: This paper first examines various issues on data quality and provides an overview of current research in the area. Then it focuses on research at the MITRE Corporation to use annotations to manage data quality. Next some of the emerging directions in data quality including managing quality for the semantic web and the relationships between data quality and data mining will be discussed. Finally some of the directions for data quality will be provided.

Key words: Data Quality, Annotations, Semantic Web, Data Mining, Security

1. INTRODUCTION

There has been a lot of research and development on data quality for the past two decades or so. Data quality is about understanding whether data is good enough for a given use. Quality parameters will include the accuracy of the data, the timelines of the data and the precision of the data. Data quality has received increased attention after the revolution of the Internet and E-Commerce. This is because organizations now have to share data coming from all kinds of sources, and intended for various uses, and therefore it is critical that organizations have some idea as to the quality of the data. Furthermore, heterogeneous databases have to be integrated within and across organizations. This also makes data quality more important. Another reason for the increased interest in data quality is warehousing. Data warehouses are being developed by many organizations and data is brought into the warehouse from multiple sources. The sources are often inconsistent, so the data in the warehouse could appear to be useless to users.

This paper describes various aspects of data quality. In Section 2 we discuss some of the developments in data quality. Section 3 discusses some of the research at the MITRE Corporation in data quality. We discuss some of the emerging directions in data quality in Section 4. In particular, data quality for the semantic web will be discussed. Data quality and data mining will be the subject of Section 5. Security, Integrity and Data quality will be discussed in Section 6. Finally the paper is concluded in Section 7.

2. DEVELOPMENTS IN DATA QUALITY

There has been a lot of work in data quality for the last several years. Notable among the efforts is the work carried out at Sloan School of the Massachusetts Institute of Technology (MIT) [TDQM]. Some of the early work focussed on incorporating data quality primitives for data models such as the entity relationship model. Data quality primitives would include information such as the timeliness of the data, the accuracy of the data and precision of the data.

As mentioned earlier, much interest in data quality is as a result of the web. There is now so much data on the web that we need some measure as to the accuracy of the data. The accuracy depends on the source of the data. Furthermore, data gets propagated from one source to another and as a result the quality could very well deteriorate. Therefore, we need some way to determine the quality as data is passed from one to another.

Data quality is also an important factor when integrating heterogeneous data sources. An organization often cannot be responsible for the quality of the data belonging to another organization. Therefore, when organizations share data, they need to have some idea of the quality of the data.

High quality data is critical for e-commerce transactions. These transactions may involve millions of dollars and therefore of the data is of poor quality, the consequences may be disastrous. Many of the businesses that have studied data quality have found that quality problems can be directly traced to significant loss of revenue.

There are web sites for data quality related information (see, e.g., [DQI]). These have information about data quality journals, products including data cleansing tools as well as information about the various organizations that do data quality work. These organizations include government organizations as

well as commercial organizations around the world. In addition, MIT also now has a series of workshops on information quality.

3. ANNOTATIONS FOR DATA QUALITY

Now that we have provided a brief overview, in this section we will discuss briefly the research that i.e. being carried out at the MITRE Corporation on data quality (see [HUGH01]).

In this effort, we have spent about two years learning about the data quality problems and efforts of a set of government organizations that share military intelligence information. These organizations share a database system, and connect it to many other databases and tools in a complex system of systems. We have begun advising these organizations on ways to address data quality problems with these systems, based on both our experience and the techniques that have been published by the data quality research community.

In summary, we have found much from the academic community that can be used by government organizations. Data quality is a significant problem, and one of the main barriers to addressing it is the need to include both technical changes to systems and managerial changes to the processes used in the organization. Many of the data quality annotations (precision, timeliness, accuracy, etc.) that have been defined for business information systems are equally relevant to government and military systems.

We also identified a need for data quality annotations, i.e., metadata that describe qualities of individual data records, rather than only using aggregate metrics (as is the focus of several prior efforts). For example, many organizations have devised ways of estimating the average accuracy of their data sets, improving that accuracy, and benefitting from the improvement. We showed that, in addition to such aggregate measures, our customers also needed to understand the quality of individual data records.

One case where these *per item* quality annotations are important is when quality varies significantly from one record to the next. In this case, if users cannot understand the quality of presented data, they may not differentiate the good from the bad. If the data is used to make decisions (as is usually the case), then either bad data can lead to bad decisions, or data of uncertain quality can lead to less well-informed decisions. As a result, users may not trust the system as a whole, even though it contains valuable data.

We defined a methodology for adding such annotations to a system, and using them to improve data quality. We are now working with two organizations to apply this methodology, and measure its effects.

4. SEMANTIC WEB AND DATA QUALITY

In the previous sections we provided an overview of the developments in data quality as well as one approach to manage data quality. In this section we discuss data quality for the semantic web.

The semantic web is essentially about machine understandable web pages. In the article on semantic web by Berners Lee et al, [LEE01], semantic web is described to be a web that can understand and interpret web pages and manage activities for people. These activities could be maintaining appointments, giving advice, and essentially making the life of the human as easy as possible.

If the semantic web is to be effective, then we need to ensure that the data and information on the web is timely, accurate, and precise. Note that with bad data one cannot make good decisions. Therefore, the techniques being used to understand data need to be widened to understand data quality. These technologies include XML, (eXtensible Markup Language), RDF (Resource Description Framework), and agents. There is little work reported on data quality for the semantic web. In fact, data quality is an essential ingredient in understanding the semantics of data, and the semantic web approaches should take advantage of the work that has been done to define data quality annotations and use them in database systems.

5. DATA MINING AND DATA QUALITY

As we have stated, having good data is essential for making effective decisions. Data mining is about posing queries and extracting information previously unknown from large quantities of data using pattern matching and statistics (see [THUR98]). One of the challenges in data mining is to get the data ready for mining. One has to determine where the data is, assemble the data and clean the data. This would also include removing redundant data as well as inconsistent data. Data also has to be complete. Various tools are being developed to ensure that the data is ready for mining. However, much remains to be done.

Another relationship between data mining and data quality is to use data mining to improve the quality of the data. That is, while data quality is important for mining, we could use mining to improve the quality of the data. Data mining techniques could be used to determine the bad data as well as data that is incomplete. That is, in the same way we use mining to detect unusual occurrences and patterns, our patterns here are bad data. We need to define what bad data means and perhaps train tools like neural networks to detect bad data. This is an area that is getting much attention recently.

6. SECURITY AND DATA QUALITY

There have been many discussions on the tradeoffs between security, and integrity. For example, if we are to enforce various access control rules, then the transactions may miss the timing constraints due to the time it takes to make all the checks. As a result, the data may not be current. As a result, the quality of the data may be compromised. Another issue is that to enforce high levels of security, one may need to give out cover stories. This will also compromise data quality, as the data given out may not be accurate. In other words, there are tradeoffs between security and data quality.

There have been various studies carried out between security and integrity as well as between security and real-time processing. We need all aspects. That is, security, integrity, real-time processing and as well as data quality are all important. The question is how can we have all of them enforced in a system. This is where quality of service comes in. We need to develop flexible policies that can deal with security, integrity, real-time processing and data quality essentially on a case by case basis. There will be cases where security will be absolutely critical and there will be cases where we cannot live with bad data. Some issues were discussed in [THUR99].

7. SUMMARY AND DIRECTIONS

This paper has provided a brief overview of the developments and directions for data quality. We discussed the need for data quality and then showed an approach to manage data quality. This approach is to use annotations. That is, quality information is defined and treated as part of the data. Next we discussed some of the emerging trends such as data quality for the semantic web and the relationship between data mining and data quality.

Because of the developments of the web and the importance of technologies such as data warehousing and data mining, the need to have good data cannot be overemphasized. It is critical that for many applications we need high data quality. There is already a lot of work that has been done on data quality. However much remains to be done.

8. REFERENCES

1. [DQ] www.dataquality.com
2. [HUGH01] Rosenthal, A., et al, Data Quality in the Small: Providing Consumer Information, Information Quality Workshop (IQ 2001), November 2001.
3. [LEE01] Berners Lee, T., et. al., The Semantic Web, Scientific American, May 2001.
4. [TDQM] MIT Sloan School, MIT Total Data Quality Management Program, <http://web.mit.edu/tdqm/www/>.
5. [THUR98] Thuraisingham, B., Data Mining: Technologies, Techniques, Tools and Trends, CRC Press, 1998.
6. [THUR99] Thuraisingham, B., and J. Maurer, Information Survivability for Evolvable Interoperable Command and Control Systems, IEEE Transactions on Knowledge and Data Engineering, January 1999.