EMERGENT DATA QUALITY ANNOTATION AND VISUALIZATION

(Research-in-Progress)

Paul Führing
Hasso Plattner Institut, Germany
paul.fuehring@hpi.uni-potsdam.de

Felix Naumann
Hasso Plattner Institut, Germany
felix.naumann@hpi.uni-potsdam.de

Abstract: The systematic assessment, storage, and retrieval of data quality scores has proven to be an elusive problem, often tackled only with classifications, questionnaires, and models. We present a concrete solution for the graphical annotation of data with quality scores, to enable their efficient storage and retrieval, and ultimately their graphical display on top of the actual data. Our tool, VIQTOR, enables users to assign quality scores using simple point and click techniques in a natural data display environment, such as spreadsheets.

The particular challenges we tackle are the support of multiple users assigning different quality scores, the flexible assignment of quality scores to any subset of the data (rows and columns), the assignment and storage of scores in multiple quality criteria, and the graphical display of those scores aggregated both across users and across quality criteria. As multiple users assess scores, an overall quality assessment emerges.

Key Words: Data Quality, Information Quality, Assessment, Visualization, Quality Criteria

DATA QUALITY ASSESSMENT

While the necessity and usefulness of reasoning about information quality when making business decisions is apparent, obtaining quality scores to reason about remains a challenging task. Five obstacles face such a project.

1. **Multiple Criteria.** Information quality is made up of many facets, called quality criteria, such as completeness, accuracy, believability, etc. [9, 11]. To reason about them in a systematic fashion, each relevant score must be assigned some numerical value. Because each of these criteria has a different natural scale, to deal with them together one has to scale and normalize those numerical values.

2. **Subjectivity.** While some quality scores, such as completeness or accuracy, can be determined automatically, assuming the necessary metadata is at hand, many criteria are of subjective nature. Only the end-user can assess the reputation of a source or the understandability of its data.

3. **Multiple Users.** An important benefit of tediously collecting quality scores is that other consumers of the data benefit from the assessment. Quality scores should be collected from all consumers and collectively presented in an aggregated fashion, while still retaining the data lineage of each individual assessment step. A difficult problem is to decide upon an appropriate

aggregation function.

4. **Granularity.** Hardly ever can an entire data source be assigned a single numerical score for some quality criterion. Rather, the data obtained from a source can be assumed to have areas of higher quality and areas of lower quality [4, 6]. An area can be defined as a set of records and/or a set of attributes within those records. These areas can differ across criteria.

5. **Sporadic scores.** A further problem is the lack of time of users and the resulting lack of a representative number of ratings for any particular data item: An interpretation of the collected quality scores is only possible when enough ratings are available. Sporadic very high or very low ratings are then balanced with the average of all ratings.

Our research tackles all five obstacles. While our goal cannot be to automatically assess all quality scores, it is to enable users to efficiently perform that task right where the data is used. Using this human opinion as a source of information for quality ratings avoids limiting the ratings to those fields, which are amenable to automatic classifications. Although those classifications nowadays apply to multiple levels of granularity (from value, tuple, and single relation classifications to multiple relations and multiple data sources) the user's subjective impression, which is based on more overall knowledge or experience, cannot be transferred into a comprehensive and common rule [8].

We present with VIQTOR (*VIsual Quality evaluaTOR*) a tool that promises a solution to all five problems.

1. Multiple criteria: As VIQTOR supports several criteria within a single user interface for data input, the user can submit his/her opinion to all respective fields. While analyzing the collected ratings switching the context of the displayed criterion is easy, nevertheless all criteria can be aggregated to one screen.
2. Subjectivity: The user's subjectivity is channeled using pre-defined domains for each criterion. The user is thereby supported on how to rate the data. While rating data, the user can decide to already see the ratings of other users or to suppress their visualization to avoid foreign influence.
3. Multiple Users: We plan to allow a wide range of quality aggregation functions to reflect both the number of consumers who have assessed a particular data value, and to reflect the different quality scores assigned each time. Further, quality scores can be allowed to diminish over time.
4. Granularity: The actual processing of a user's rating takes place at the smallest level of granularity – single data values. However the user is not forced to enter a rating cell by cell but can select almost arbitrary areas within the data and give feedback about them.
5. Sporadic scores: To motivate users not to only consume quality scores but also to create them, VIQTOR offers instant gratification in a way that the opinion of other users who rated the same area as the current one is only shown after the ratings were submitted.

Quality assessment or measurement is topic of several research activities, such as with the methodology AIMQ [5]. Neely provides a thorough overview and classification on such activities [7]. Methodology classifications and comparisons between the different strategies are described as well in [1] with a focus on the presented *Complete Data Quality Methodology* (CDQM). The latter approach is common to ours in that it proposes to interact constantly with the user of the data and to perform data quality assessment step by step. A practice-oriented tool on quality measurement is presented in [3]: Precisely defined rules, which define the domains and dependencies of data, allow conclusions about the quality of a certain data source. However, subjective criteria are not assessed and the improvement activities are tied to one specific business process.

FIELD OF APPLICATION

Wherever a large amount of relational data is displayed to the user who can interact with it, capturing the user's opinion about the data is useful. Examples for acquiring the quality ratings are:

- **Excel.** While analyzing sales figures in a large Excel sheet one often wonders about some numbers. Annotating these data fields instantly with certain attributes (e.g. «that sounds strange» or «this looks good») increases the usefulness of the data. Offering such kind of feedback possibilities also increases the acceptance of the tool as it respects the user as a contributor and source of information.
- **ETL process.** Within ETL tools or business process definitions a step for manual classification, assessment, or enrichment of incoming data is common. Beside the manual rating VIQTOR can also be used to apply automatic classification rules as described in [8].

When collecting feedback information about data, which ran through an ETL process, one needs pay special attention to the data lineage. Since the origin of the data is not identifiable at first sight the transformation process of the data needs to be examined. Cui and Widom provide in-depth descriptions of this topic [2].
By-and-by the collected ratings can lead to additional rules, which are extracted from the manual assessment. Feeding the pre-annotation with these newly defined rules a self regulating process can be established, which increases the acceptance of the data and of the data providing tool. Using VIQTOR the user can visualize the quality information and rate the data. Figure 1 provides a simple usage scenario. VIQTOR integrates into the frontend and has read-only access to the displayed data and stores collected quality ratings in its own database.

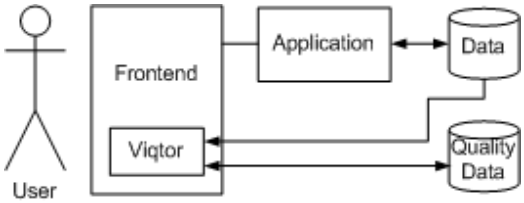

Figure 1: Schematic architecture chart of usage of Viqtor.

In general, analyzing the collected ratings leads to certain areas of interest in the (possibly huge amount of) data where the information quality is seen to be *worse* than in other areas. Thus searching for data with substandard information quality can be done much easier. Those areas (e.g. data from a specific data source or from a certain process) can be further investigated by the data or process owner or administrative personnel. As mentioned above the lineage of the data needs to be considered to retrieve the original source. The other way round areas with high ratings might be a candidate for further investigations as well, e.g., to share some best practices, which led to good results.

VIQTOR was not yet tested in real business scenarios. Oncoming activities include use cases within business processes and scalability tests. Initial performance tests have been conducted but further research is underway to find the best tradeoffs for storage alternatives and score calculations (see chapter «Challenges»).

SCORE CALCULATION FOR MULTIPLE CRITERIA

This section briefly describes our extensible concept of quality criteria, our flexible means of assigning score-ranges, and finally how we aggregate scores assigned by different users for different criteria.

Quality criteria and scores

To channel the user's rating several quality criteria are available. For each assignment the user can choose the criteria he/she likes to give feedback about and enter the score. The criteria are visualized independently of one another, however to summarize all criteria it is possible to chose a global criterion, which covers all criteria.

Because of our flexibility regarding scores and ranges, the set of IQ criteria can easily be defined and extended, covering areas such as reliability, completeness, or believability [9, 11]. Along with the criteria, the range of possible scores has to be defined, e.g., as a plain integer or using a set of distinct values. A plain integer range, such as [1 – 10], would also need the minimum and maximum descriptions, i.e., «low» to «high» or «incomplete» to «complete». Whereas the distinct values are predefined, e.g., «low», «medium», «high» or categories such as «<30%», «30-70%», «>70%». Of course when aggregating over multiple criteria and/or over multiple users, such categorical scores must in turn be interpreted numerically.

Score aggregation for multiple users

To make use of the collected ratings of multiple users it is suitable to calculate and export one aggregated quality score per data element (field in a table). The problem we face is similar to that of customer-rating in online-shops. Merely displaying the average quality score is not sufficient: The number of individual scores is also relevant. A data element rated as high-quality by many users should be treated differently from an element that was rated high by only a single user (i.e. one high rating vs. a hundred high ratings vs. a hundred low ratings etc.). Online-shops solve the problem by simply displaying both the average value and the count-value, i.e., the number of individual scores.

In analogy, to indicate the quality value we use a two dimensional approach with displaying two scores per cell (the average and the amount of ratings). This average calculation is the default algorithm – VIQTOR supports several others. The user can select an appropriate aggregation function, such as the sum of the ratings or the maximum rating or a logarithmic function.

The algorithm calculates one score per criterion, but as mentioned above, also all criteria can be displayed – again aggregating the quality score, not only across users but in this case also across criteria.

USER INTERFACE

VIQTOR has a spreadsheet-like user interface to show the relational or tabular based data source. For easier selection of the data elements to be rated, the user can re-shuffle the columns or sort the data. Because screen-size is limited users can filter the data by any column/value tuple or aggregate certain values to shrink the size of rows being displayed.

After the optional shuffling, filtering, and sorting, users select the cells to be rated by making a rectangular selection within the data. Future versions may include other forms of selection, which may be more intuitive to users, such as drawing a circle or ellipse. Annotating the data is then easy: the user chooses the criterion and enters the rating.

To display the calculated score for a cell, the background color of a cell is saturated according to the local score and the global maximum. Low saturation (almost white) signals low quality, high saturation (bright color) signals high quality. As mentioned above, the algorithm to calculate the score can be selected by the user. The entered ratings of a user are immediately analyzed and the coloring changes. To distinguish the different criteria, each of them uses a separate color space.

The screenshot in Figure 2 gives an impression of VIQTOR. At the top users can select a storage handler (for testing purposes only – in live applications the tool will take care of the best storage handler by its own) and the algorithm used to calculate a score (e.g. average of all ratings per cell). Beneath that, the user selects the criterion to be rated and enters the rating (currently only integer are allowable). The table in the middle is used for data selection, display, and quality score indication. The icon next to each data cell gives an impression how often the respective data element was rated. Thus the user has two information on one screen.
Mouse-over texts give information about the lineage of a score, i.e., the calculated score value, the individual scores, etc. To shrink the total number of rows the user can apply filters (beneath the table) and for information purposes a logging window shows the most recent activities.

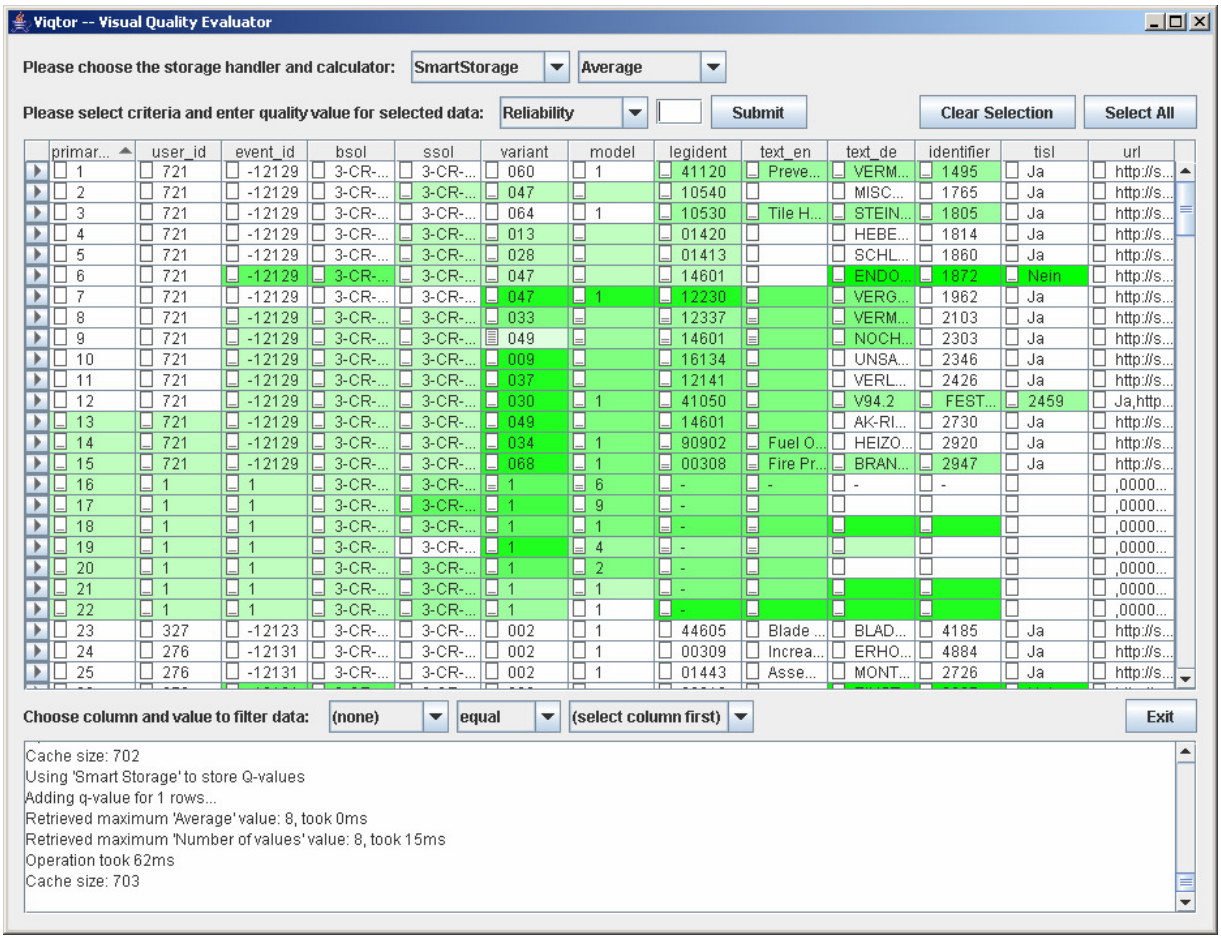

Figure 2: Several ratings for the criterion 'Reliability' of an example database.

Because only standard components, such as dropdowns, tables, and buttons are used, the UI can easily be adopted to a web interface or integrated into office tools, such as Microsoft Excel. An Excel-plugin will

collect the ratings for selections within the program and store them into a database using internal procedures or send the selection and rating to a server side application or service. This concept was already suggested in [10] and could be extended regarding quality information exchange.

CHALLENGES

Several research challenges arise from the described functionalities. As mentioned in the first chapter, we face several obstacles. In the following they are described from a concrete implementation point of view. However we address important research fields, such as lineage of data together with its quality and standardized interfaces to make the quality information usable by other application areas.

- **Selection Predicates.** Because the user can re-order the columns and sort the data as desired, the tool needs to transform the selected area into a common model, which can be restored at a later date. This transformation is necessary, because the user's selection only consists of relative information regarding the beginning and end of the selection, which needs to be converted into absolute values.
  Example: The user sorts the sales figures by the 'city' and shifts the columns 'shop owner' and 'shop name' aside. The selection that is afterwards drawn on the screen is not reproduceable using absolute coordinates (upper right and lower left) such as (2/3) to (10/7), because the row and column index change according to the column shift and sort order.

  In interaction with the user, VIQTOR can also recognize certain kinds of selections and store a more generic model of the selection as a cell by cell or row by row representation. Examples for such enriched models are selections where the selected values of a certain column occur only within the selection (see example below). Thereby the information, which rows where selected, can be stored easier. However the selection can always be coincidental, therefore the tool interacts closely with the user asking about the intended selection scope. Another example is a select-all rating, which again does not need to be stored row based.

  Example: If the aforementioned selection contains all values of e.g. «Detroit» in the column 'city' the tool recognizes this and instead of storing distinct values for each row the common characteristic «Detroit» is used to determine the selected rows.

- **Storage.** The information to be stored for each individual rating consists of the selected area, the criterion, and the score itself. Additionally, other metadata, such as the current user and the time stamp, are taken into consideration. However, storing the user's ratings in a cost efficient manner is surprisingly complex.

  To make the quality ratings available to other applications as well, VIQTOR does not store the ratings as a proprietary serialized object but as raw data into the database.

  Nevertheless several alternatives of storing the user's ratings are possible. To choose the best alternative VIQTOR analyzes the data to be stored (e.g. what size has the selection, are more columns than rows selected or vice versa etc.). Using an internal cost model the most appropriate storage behavior is identified on the fly and applied. Immediately after a vote the ratings are analyzed, stored, and the visualization is updated.

- **Score Calculation.** As described above finding an algorithm to calculate a meaningful score

needs to reflect both the user's opinions and the amount of users who rated a data item. Also metadata can influence the rating, e.g., decreasing a quality rating with its age, or use the information who rated, which area, e.g., intensifying the opinion of certain users.

- **Import and Export.** To make use of already existing quality ratings the application needs to import them or understand different languages of quality data descriptions. In turn, exporting the aggregated information leads to a more accepted usage of the tool as the results are not bound to the application. This export interface as an access layer to the collected quality ratings can allow other tools to select only those data that meet a certain level of quality.

  To import and export a common and widely accepted or standardized model is appreciated. An approach for XML data, which adds quality information to the data itself into the same document, is described with D²Q in [1].

- **Annotation Lineage.** Together with that import and export the lineage of the displayed data and of the ratings becomes important. Knowing when data was changed is significant as the rating and the rated data needs to be consistent. Particularly when data was added the existing ratings can seldom be used for the new data items.
  This applies as well for data, which was somehow modified within an ETL process. Tracking back the collected data quality to its data root can be an expensive and complicated step [2].

With VIQTOR we present a tool that already faces all of the described problems in a fundamental way. During future research and accompanying implementations we will continue to evaluate usage scenarios and integration possibilities for the tool and the collected quality information.

REFERENCES

[1] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methods and Techniques.* Springer Verlag, Heidelberg, 2006.

[2] Y. Cui and J. Widom. Lineage tracing for general data warehouse transformations. *VLDB Journal*, 12(1): 41-58, 2003.

[3] M. Gebauer and P. Caspers. Reproducible measurement of data field quality. In *Proceedings of the International Conference on Information Quality (IQ)*, 2005.

[4] I. A. Gelman. A theory of complementarity for extracting accurate data from inaccurate sources through integration. In *Proceedings of the International Conference on Information Quality (IQ)*, pages 143-157, 2005.

[5] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang. Aimq: a methodology for information quality assessment. *Information and Management*, 40(2):133-146, 2002.

[6] A. Motro and I. Rakov. Estimating the quality of databases. In *Proceedings of the International Conference on Flexible Query Answering Systems (FQAS)*, pages 298-307, Roskilde, Denmark, May 1998. Springer Verlag.

[7] M. P. Neely. The product approach to data quality and fitness for use: A framework for analysis. In *Proceedings of the International Conference on Information Quality (IQ)*, 2005.

[8] P. Oliveira, F. Rodrigues, and P. Henriques. A formal definition of data quality problems. In *Proceedings of the International Conference on Information Quality (IQ)*, 2005.

[9] S. Pradhan. Believability as an information quality dimension. In *Proceedings of the International Conference on Information Quality (IQ)*, 2005.

[10] A. H. von Thile and I. Melzer. Smart files: Combining the advantages of dbms and wfms with the simplicity and exibility of spreadsheets. In *Proceedings of the Conference Datenbanksysteme in Business, Technologie*

und Web Technik (BTW), 2005.

[11] R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal on Management of Information Systems*, 12(4):5-34, 1996.