Editing as a Data Repair Problem: a Logical Formalisation

Enrico Franconi[1]
Faculty of Computer Science, Free University of Bozen-Bolzano,
piazza Domenicani 3, I-39100 Bozen-Bolzano, Italy
http://www.inf.unibz.it/~franconi/

When the millions of compiled forms are returned to the national statistical agencies to be analysed, they may contain many errors, such as *missing* or *inconsistent data*. For example, it is possible that a person declared to be the spouse of the reference person, but at the same time he/she forgot to declare a marital status, or he/she declared to be single, or to be 6 years old. Before the data from the questionnaires is sent to the statisticians to be analysed, a cleaning phase is performed, in order to eliminate missing values or inconsistencies. It is very important that this step is done *without altering the (statistical) validity* of the collected data. For example, if the spouse declared to be 6 years old, and if there are other arguments to *enforce* that he/she is actually the spouse of the reference person (for example, he/she may have additionally declared to be married), then his/her age should be changed to make the data consistent: any age, say, greater than 16 years would be fine if only absolute validity of the data is taken into account. However, the age should be changed to an age in agreement with the average age for a spouse in his/her conditions, and not with a value which may alter the statistics. In this case, it would be more sensible to change the age to, say, 36 years, rather than to 96 years (there are very few people of that age) or to 16 years (there are very few people who are married at that age). Other corrections can be done *deterministically* – in the sense that there is only one valid change, which is also a statistically valid change – for example for a spouse who forgot to specify the marital status, and for which asserting that he/she is married does not introduce any other inconsistency.

1. Editing

For these reasons, a collection of *edit* rules is introduced by the statistical agencies for each questionnaire class, in order to understand whether a questionnaire is consistent, and to guide the repair of inconsistent questionnaires. This kind of restrictions can be expressed using denial integrity constraints (ICs), that prevent some attributes from taking certain values. Other restrictions may be expressed by means of aggregation ICs, e.g. the total number of very old people (older than 90 years) across all questionnaires can not be more than 5 percent of the total number of people; or the number of married men and married women must be the same.

The process of automatically producing a set of statistically consistent questionnaires from the raw ones is called *imputation*. Inconsistencies in numerical data are resolved by changing individual attribute values, while values in the key attributes are kept, e.g. without changing the household code, the number of children is decreased considering the admissible values. All the methodologies currently employed by the census agencies – including the most popular Fellegi and Holt methodology – are based on statistical principles, which do not have a clear semantics with respect to the complexity of the various errors that may happen, so that their behaviour while imputing the data is often unpredictable. Moreover, it is difficult to compare the various methodologies, since they are defined only on a procedural basis.

2. Data Repair

Imputation in the context of database system, called *data repair*, has been extensively studied as consistent query answering (CQA), i.e. the process of obtaining the answers to

[1]This work has been carried out with Leopoldo Bertossi, Loreto Bravo, Antonio Laureti Palma, Nicola Leone, Andrei Lopatenko, Simona Perri, Francesco Scarcello, and it has been partially funded by the Sewasie, KnowledgeWeb, and Interop European projects.

a query that are consistent wrt a given set of ICs. There, consistent data is characterised as invariant under all minimal restorations of consistency, i.e. as data that is present in all minimally repaired versions of the original instance (the *repairs*). In particular, an answer to a query is consistent when it can be obtained as a standard answer to the query from *every possible* repair. In our work we consider the problem of fixing numerical data according to certain constraints while (a) keeping the values associated to the keys of the relations in the database, and (b) minimising the quantitative global distance from the modified instance to the original instance. Since the problem may admit several global solutions, each of them involving possibly many individual changes, we are particularly interested in characterising and computing data and properties that remain invariant under any of these fixing processes.

3. A Logical Formalisation

The major contribution of our work is to define in a pure classical logic setting the semantics of the edit and imputation problems. The approach is very general and very expressive, since it makes use of classical First Order Logic. The advantage of this approach is the ability to specify in a declarative fashion the characteristics of the census questionnaire, the intra- and inter-questions constraints, and the additional statistical (i.e., aggregated) information required to guide the corrections. This is the first proposal in the literature making use of Logic at all the stages of the definition of the problem. We provide the semantic foundations for fixes that are based on changes on numerical attributes in the presence of key dependencies and wrt denial and aggregate integrity constraints. Fixing databases by changing numerical values while keeping the numerical distance to the original database to a minimum introduces interesting algorithmic and complexity theoretic issues. We study decidability and complexity of different decision and optimisation problems. We concentrate in particular on the "Database Fix Problem" (DFP), consisting in determining the existence of a fix at a distance not bigger than a given bound. We also consider the problems of construction and verification of such a fix. These problems are highly relevant for large inconsistent databases. For example, solving DFP can help us find the minimum distance from a fix to the original instance; information that can be used to prune impossible branches in the process of materialisation of a fix.

As far as computational complexity is concerned, we prove that DFP and CQA become undecidable in the presence of aggregation constraints. However, the DFP is NP-complete for linear denials, which are enough to capture census like applications. CQA belongs to Π_2^P and becomes *coNP*-hard, but a relevant class of denials is identified for which CQA becomes tractable. Considering approximation algorithms, we prove that DFP is *MAXSNP*-hard, but can be approximated within a constant factor. The complexity results refer to data complexity, i.e. wrt to the size of the database.

Another important contribution of our work is to provide a correct encoding of the edit and imputation problems in an executable specification using an extension of disjunctive logic programming. We provide the rationale of a *modular* encoding. For this purpose, disjunctive logic programming extended with two kinds of constraints has been used. The idea is that the preferred models of the logic program encoding the problem correspond to the repairs of a given questionnaire. It turns out that the solution of this challenging problem requires most of the expressive capabilities of disjunctive logic programming with constraints. Notably, this disjunctive logic programming language is supported by DLV, a system which is freely available on the web, and can therefore be used to obtain an immediate implementation.

REFERENCES

Franconi, E., Laureti Palma, A., Leone, N., Perri, S. and Scarcello, F. Census Data Repair: a Challenging Application of Disjunctive Logic Programming. In *Proc. Logic for Programming, Artificial Intelligence, and Reasoning (LPAR 01)*, Springer, 2001.

Bertossi, L., Bravo, L., Franconi, E., Lopatenko, A. Fixing Numerical Attributes Under Integrity Constraints. Technical Report, *Free University of Bozen-Bolzano, Italy*, 2004.