Frank, A. U., and Andreas Gruenbacher. "What Is Relevant in a Dataset?" Paper presented at the Agile 2002, Palma, Spain, 25.-27. April 2002.

What is relevamt in a dataset?

Andrew U. Frank and Andreas Grünbacher
Institute for Geoinformation, Technical University Vienna
{ frank, gruenbacher } @ geoinfo.tuwien.ac.at

Introduction

Not all the data in a dataset are relevant, but what does it mean that a piece of data is relevant? Relevance of data can be decided always only with respect to a decision. A test for relevance is to see if the decision has the same outcome if the dataset is improved or degraded: if the outcome of the decision is the same, then the improvement or degradation of the dataset has not affected relevant aspects of the dataset.

Relevance for a decision is a crucial point in the emerging market for geoinformation. Pricing of information means pricing for the relevant part of the information. Nobody is willing to pay for the irrelevant data delivered. Identification of relevant information is also necessary to produce datasets which are geared towards use in one specific decision situation and which are constructed to serve this and only this decision. Such specificity of a geoinformation product is necessary to allow setting of an appropriate price corresponding to the benefits a user draws from these data and to avoid that the same data are used for other decisions that yield higher benefits and thus, where a higher price would have been justified. In marketing jargon we speak of 'cannibalism' if a low price good geared to low benefit use is used in a high-benefit situation, where a high-cost good should have been used instead.

Example from airlines: It is well known that the market for airline tickets is segmented in a leisure and a business market. Business users draw higher benefits from transportation and are therefore willing to pay higher prices. Tourists draw smaller benefits, and their willingness to pay the same prices is low. The products are separated in 'leisure' class tickets requiring a stay over Saturday night, and in the higher-priced 'business' class tickets, which allow a return on the same day (or any other day). The business logic behind this marketing decision is that business people are not expected to stay during Saturday and return only on Sunday. Cannibalism happens, if business people buy 'leisure' class tickets because the ticket price difference is larger than the extra cost of staying longer.

The question of relevance is also related to the question of buying an update to revise a previous version of a dataset: A dataset is updated with some new information. Is this supplement of new information relevant for a decision? Should a customer buy the update? Again, we can compare the outcome of the decision with and without the additional supplement of new data: if the outcome is the same, then the supplement did not contain relevant information, if the outcome changes, then the supplement was relevant and its acquisition is recommended.

Formal framework

We consider the following formal framework for this discussion:

Given a dataset K_i and an additional dataset A, there is an operation to merge the two in a new dataset K_j. Consider a decision function d. When applied to K_i the decision function gives the outcome $d(K_i) = o_i$. When applied to K_j, the outcome is $d(K_j) = o_j$. The dataset A contains relevant information for the decision, if o_i is different from o_j.

$$\begin{array}{ccccc} K_i & * & A & = & K_j \\ \vdots & & & & \vdots \\ d(K_i) = o_i & & & & d(K_j) = o_j \end{array}$$

The same framework applies to the degradation of a dataset: instead of merging with additional data we merge with noise, which effectively degrades the dataset. The noise must not be relevant for the decision at hand, but should be relevant for all other decisions for which the data could be used.

Key in this discussion is the operation *, which merges the dataset with the additional data, which is either real data to improve the dataset K, or noise to degrade it.

We represent the dataset as a set of binary relations *'a* **R** *b'*, with no consistency constraints. We represent the supplement as a set of binary relations *'a* **R** *b'* and *'c* Ø **R** *d'*. When merging a dataset with a supplement, all

relations of the form '*a* **R** *b*' are added to the dataset, while for each relation of the form '*c* Ø**R** *d*' the corresponding relation '*c* **R** *d*' is removed from the original dataset.

In the presence of consistency constraints, which is the case for all realistic datasets, care must be taken that the dataset resulting from such a merge is again consistent, and so a more complicated merge operator may be needed, but the fundamentals are not changed.

This merge operation works as well for the contribution of noise: the data in the supplement dataset either change some of the values in the original data set, or they add or remove tuples.

Application and Example

The street network shown in Figure 1 (a) is represented by the tuples shown in Table 1. **C** stands for a connection between two nodes; **X** and **Y** stand for the x- and y-coordinates of a node, respectively. We call this data set K_i.

A **C** B	A **X** 1	A **Y** 2
A **C** D	B **X** 4	B **Y** 2
B **C** D	C **X** 5	C **Y** 3
C **C** D	D **X** 4	D **Y** 4
	E **X** 4	E **Y** 7

Table 1: Street network dataset

Table 2 shows a supplement to this street network that results in the network shown in Figure 1 (b) when merged using the merge rules described before. Let this supplement be dataset A. The relations 'B **C** C', 'B **X** 3' and 'D **C** E' are added, while the original relation 'B **X** 4' is removed. The pair of relations 'B Ø**X** 4' and 'B **X** 3' effectively updates the x-coordinate of node B. According to the formula $K_i * A = K_j$ we call the resulting data set K_j.

B **C** C
B Ø**X** 4
B **X** 3
D **C** E

Table 2: Supplement dataset

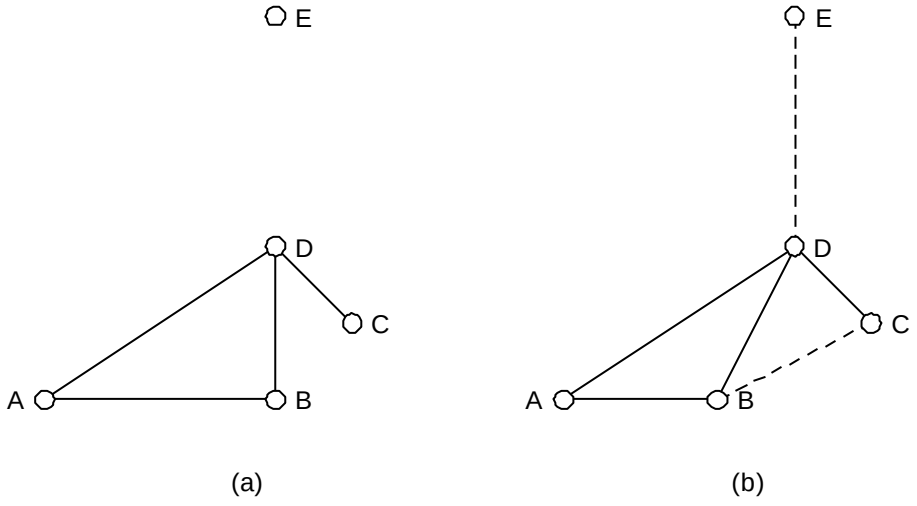

Figure 1: Street network

As a simplistic example consider the decision about the shortest path from nodes A to C. The outcome of the decision $d(K_i)$ is the path with D as an intermediary node, (A D C). If the same decision function is applied to dataset K_j, the outcome d(Kj) of the decision is (A B C). The update was clearly relevant for this decision.

Relevance of data in the dataset for a decision is restricted to the tuples the decision function d uses to compute the outcome of the decision. If the additional data does not affect these tuples, then the outcome will not change. This gives a way to see which additional data affect which decisions – both in the case of improving a dataset and in degrading it. For a shortest path decision, both connection information and coordinates are potentially relevant. The position of the node E does not affect the solution, merging with a supplement dataset that changes the position of E does not affect the decision about the shortest path from A to C; this data is not relevant. Likewise, degrading the coordinate values of all nodes by randomly adding or subtracting 0.5 to each coordinate value (e.g., Figure 2) does not affect the outcome for the shortest path.

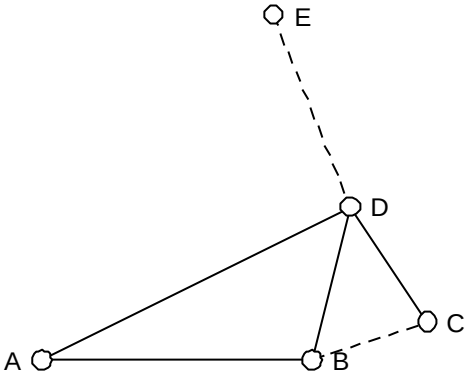

Figure 2: Randomly distorted network

Conclusion

We have developed a framework for the revision of datasets in order to produce better, updated datasets or to produce datasets that are degraded in some respect. Degradation of a dataset to have a determined quality is of high commercial interest to produce datasets specifically geared towards some decision and priced in accordance to the benefits it produces in this decision; limiting quality may make it uninteresting to use this dataset for a purpose where another (higher priced) dataset is expected to be used.

The framework demonstrates the basic logic of the merge operation that integrates the additional data with the existing data. Data merged is relevant, if the outcome of the decision for the original dataset and the updated dataset are different. Degradation of datasets can be achieved with non-relevant noise that affects aspects of the dataset that are not used by the decision function.

To simplify the cost model we have assumed that the actuality of the data set does not contribute to the value of a data set, i.e., we have assumed that users assign the same value to a data set and to a more recent, updated data set, and they put no value in knowing that the information used is “current”. A useful extension to the model presented would be to model the risk of false decisions due to the use of old datasets, which obviously also is a price determinant. (An update does not necessarily have no value to users just because their decisions shay the same; the increased reliability of a decision’s outcome may well be worth the update.)

Acknowledgements

This works is carried out under the REVIGIS project.