Proceedings of the 6th AGILE
April 24th-26th, 2003 – Lyon, France

HOW TO SELL THE SAME DATA TO DIFFERENT USERS AT DIFFERENT PRICES

Andrew U. Frank, Marianne Jahn

Institute for Geoinformation, Technical University Vienna

1. INTRODUCTION

It is a known fact that the use and sale of geographic information lags behind it's potential and expected market growth [1]. Several projects tried to explore the dynamics of products and their pricing in GI Marketplaces. They started with a solution ready-to-use for a customer, then added further aspects to attract other types of users and to be more competitive. This is one way to enlarge the market for GI products.

Potential new customers try a test version, which is usually free. They use the same dataset as in a full version, but with constraints in functionality. In a pre-selection, the user himself chooses the desired functions, while only having a smaller number available compared to the full version. The user can always vary the selection after restarting the application. The complete functionality can be tested with high quality data, but with limited functionality during each session.

Good datasets are expensive to produce and preserving the data quality is costly too. The user must be convinced to spend money and appreciate the value of both data and the solution being offered. If value and the high quality of the solution is apparent to the user, he is more likely to buy a full version or an update.

Another approach is to offer a solution having different, smaller datasets but full functionality. The basic idea is to merge data [2] with different quality and, depending on the user, one gets different outcomes. To satisfy a test user, a presentation of all details but not the complete extend is one solution. Some data is redundant and not relevant for decision-making. Using cheaper and not so detailed data can reduce costs. Customers should know the real value of data and appreciate all the efforts behind the product, which can be achieved by showing them the relevant differences between the full and the preview-version.

Main crucial aspects in relevance of data:

- calculating relevance only via a change of outcome [2] is a time-consuming and therefore expensive way because you have to assess the needs of a user, which is not always easy,
- a simpler approach is to find out in advance which datasets are redundant and which should not be changed to maintain a certain level of quality,
- another relevant aspect is how to define quality, which aspects in data are important and which can be neglected and finally whether there are some dependencies between data and how to take them into consideration,
- exact investigations about users and their demands lead to a classification of data quality, in order to maximize use and usability for the customer.
-

2. PRODUCTION OF 3 TYPES WITH DEGRADED QUALITY

2.1 Merging Data with different Quality

The basic framework is from [2], where a dataset K_i and an additional dataset A are given. These two are merged in a new dataset K_j with an operation. Consider a decision function d. When applied to K_i the decision function gives the outcome d(Kj)=oj. The dataset A contains relevant information for the decision, if oi is different from oj.

The two datasets are of different quality and characteristics. Problems may occur during merging if data has been retrieved in different ways. An example would be merging data collected in different time. In Austria data from population statistics 2000 is available, but actual data from 2002 is not. Some private departments have collected samples to approximate the actual numbers, but these are not being published. On one hand it is difficult to gain access to them and on the other hand there can be huge problems in process during merging.

Depending on the decision function d it is necessary do determine precision and quality of data. As mentioned in the previous chapter, exactness is one of the fundamentals of data quality.

2.2 Merging Data with noise

The same framework as described above is used in merging data with noise, which effectively degrades the dataset. The important aspect here is the relevance of data. Dependencies must be preserved and the inserted 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 as well.

Consistency is the key for “correct” merging, where under certain constraints the result of a merge is again consistent. Another subject is the morphinism, which guaranties a model and any reproduction conserving structural retentivity. The constraints must be analyzed, whether there can be a general model developed. The used algorithm for testing preservation of dependencies in the application part of this paper is from [3].

2.3 Deterioration of data quality using a random generator

The third possibility is to use only one dataset and to generate noise in between with a random generator. The same aspects as mentioned above must be considered. The datasets must be tested about their relevance and the inserted noise must not have effects on the consistency.

3. APPLICATION AND EXAMPLE OF MINIMAL COVERS AND PRESERVING DEPENDENCIES

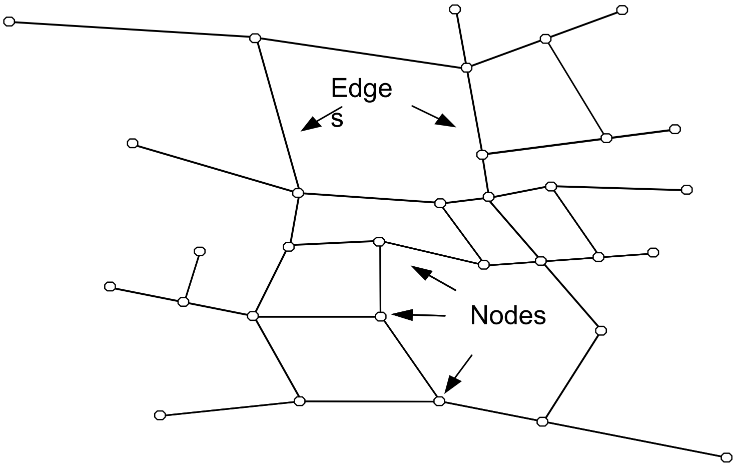

Fig. 1 Street network represented as a graph

The street network of Vienna is a system based on directed graphs, so called one-ways. Directions are defined on the street segments. Nodes are either an intersection point or an end point of a road and represent an intersection of two lines or a dead-end of a road. This classification is taken from [4].

The next step is to convert the network into a model using directed graphs. A user or agent wants to move from its current position on one node to the next node until reaching the final destination. The direction of the graphs picture the ways he can move. One-ways are shown as a single directed graph from A → B. The two way traffic roads are split into two graphs with contrarious directions in this simple graphical model from X → Y and Y → X. Finally the model is a list of functional dependencies.

A → BD, C → ABDG, D → A, E → F, F → EG, G → CDE, B → E.

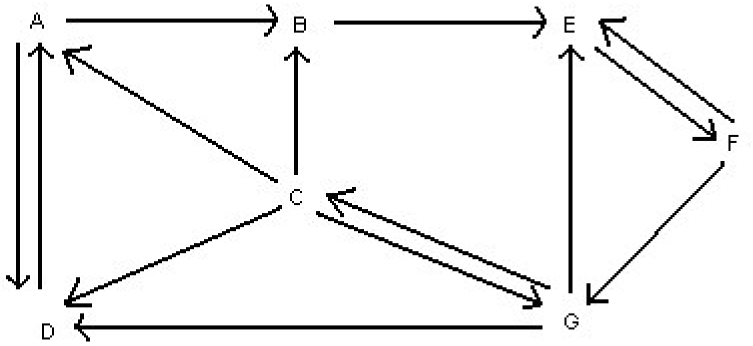

Fig. 2 Street network represented with directed graphs

This example shows that every point on the map is reachable from each starting point. The proof of this statement is given by the cover from [3] chapter 7 (Reasoning about functional dependencies). Point A has a relation to point B and D. The first step is to divide the assembled relations into single relations (A → B, A → D, C → A, C → B, C → D, C → G,….). The full list of relations is not shown in this paper but easy to reproduce. To compute the cover you just check this list in the following way to get the cover of A: A → A, A → B, A → D, B → E, E → F, F → G, G → C. The cover of A is formally: A^+ ={A,B,D,E,F,G,C} or in chronological order: A^+={A,B,C,D,E,F,G}. The remaining covers of B^+, C^+, D^+, E^+, F^+, G^+ can be figured out with this procedure and they evince that each point is reachable from every start.

Some of these relations unnecessary and can be eliminated, but the previous calculated covers must not change. It is desirable for a decomposition to have a lossless-join property, because it guarantees that any relation can be recovered from its projections. [3] For a given set of dependencies, an equivalent set with a number of useful properties can be found. A simple and important property is that the right sides of dependencies be split into single attributes. A set of dependencies F is minimal if:

- every right side of a dependency in F is a single attribute,
- for no X → Y in F is the set F- {X → Y} equivalent to F,
- for no X → Y in F and proper subset Z of X is F-{X → Y}∪{Z → Y} equivalent to F.

Intuitively, (2) guaranties that no dependency in F is redundant. Condition (3) guaranties that no attribute is redundant on the left side (3). As each right side has only one attribute by (1) (canonical), no attribute on the right is redundant. This algorithm is also called minimal covers algorithm.

The relation C → D is redundant. From start C, D is still reachable with C → A and A → D. There are several results of the this algorithm, which have no effects on reachability, but they differ in the chosen way and the length of the path.

4. CONCLUSIONS AND FURTHER WORK

One of the main problems is to know the decision function d. It is necessary to explore user demand and his expectations in the solution/application. Otherwise it is hard to predict the constraints under which a solution maintains a certain level of quality for the customer. With information about the user's needs it is possible to select data as floatable, meaning that data can be merged with noise without having considerable influence in the decision process. Some users don't select reachability as a sufficient condition of a tool, they also seek precision and completeness. So the different views of quality of each user group lead to different requirements of quality and functionality.

The minimal covers algorithm is one possible approach to find non-relevant data in a huge dataset. The remaining relevant data must not be changed or degraded by noise. As mentioned in the introduction, degradation of data to have a determined quality is of high commercial interest to enable price differentiation.

By knowing more about user preferences, it is possible to further differentiate user types and pricing within one group of users. The collection of information from the demand side enables the supply side to react on changes on the market. If a product is rejected from the selected user group, the outcome is not sufficient or not acceptable. Sometimes only small adaptations are necessary. It is critical to know whether differentiation creates value for buyers or not.

Another possible distinction lies within the different use of data and the application. Developers and companies spend a lot of money in branding and promotion of their offered goods. [5]

In the field of dominating the market, specialization is one possibility – in contrast to generalization of applications. Some customers are interested in the combination of different functionalities and sometimes they require to insert new algorithms as extension to enhance usability. The final goal of marketing research in GI Marketplaces is to analyze users and classify users in order to allow further investigations about user demands and to optimize the offered goods for usability.

5. ACKNOWLEDGEMENTS

This work is carried out under the REVIGIS project.

6. REFERENCES

[1] Brox, C. and Krek, A. «Products and pricing in GI marketplaces», *5th AGILE Conference*, Palma, 2002.

[2] Frank, A. U. and Grünbacher, A., «What is relevant in a dataset», *5th AGILE Conference*, Palma, 2002.

[3] Ullmann, J. D., *Principles of database and knowledge base systems*, Computer Science Press, 1988.

[4] Krek, A., *An agent-based model for quantifying the economic value of geographic information*, Institute for Geoinformation, Technical University Vienna, 2002.

[5] Varian, H. R., *Differential Pricing and Efficiency*, First Monday, 1996.