A Process Model for a Data Fusion Factory

Peter van der Putten[a] Martijn Ramaekers[b]
Marten den Uyl[c] Joost Kok[a]

[a] LIACS, P.O. Box 9512, 2300 RA Leiden
[b] Ordina Finance Business Solutions, Kastanjelaan 4, 3833 AN Leusden
[c] Sentient Machine Research, Singel 160, 1015 AH Amsterdam

Abstract

Data fusion is the process of enriching data sets by combining information from different sources, to provide a single data set to mine in. Data fusion projects are complex, and to structure these we have built a process model for data fusion, inspired by the CRISP process model for data mining. The end goal is to build a fusion factory, where fusion projects are automated much as possible.

1. Introduction

One may claim that the exponential growth of information provides great opportunities for data mining. In practice however, this information may not be directly accessible. It is fragmented over an even faster growing number of sources that only provide information on a small number of cases. This results in a barrier to more widespread application and successful exploitation of data mining.

Data fusion can provide a way out. It is the process of combining information from different sources into a single, enriched data set for further data mining. It facilitates the application of data mining by providing more data to mine in and allows the data miner to find valuable patterns that would otherwise go unnoticed [7].

Data fusion projects can get quite complex. Instead of a single data set, several heterogeneous data sources are involved in the procedure that need to be mapped onto each other. Source data sets with hundreds to thousands of variables in a wide range of logical and physical formats are not uncommon. The fusion process itself consists of many intertwined phases and steps, and a lot of choices have to be made. What the right choices are is predominantly determined by factors outside the fusion procedure itself, namely the business and data mining goals for which the enriched data set will be used.

Despite these challenges, we envision a streamlined fusion procedure where the core steps can be carried out in less than a working week instead of weeks or months. To standardize and structure fusion projects we decided to develop a data fusion process model, borrowing some key concepts from data mining process models like CRISP-DM [2,4,8]. The end goal of the fusion process model is to rationalize the process – and automate it where possible. The development of the process model took place in parallel

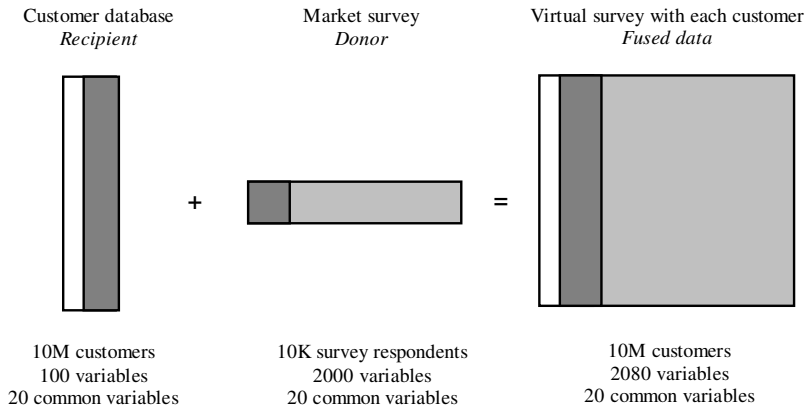

Figure 1: ***Data Fusion Example***

with three major commercial data fusion projects carried out by Sentient Machine Research, for a financial services company, a charity and a marketing data provider. Further input was provided by previous experimentation on a variety of data sets and some 25 data fusion cases from research literature.

The goal of this paper is to present a general overview of the data fusion process model. As may be expected, it takes more than just running a fusion algorithm on a given data set to carry out a successful real world data fusion project. It is our aim to give relevant examples of typical, expected or unexpected issues that can arise when one wants to apply data fusion in the real world instead of a lab environment, as an example of the difference between AI research and practice.

Note that a process model is primarily focused at identifying, grouping and prioritizing the right questions to be asked, rather than providing technical solutions itself, thus providing a framework for fusion research and development.

We will first illustrate the main idea of data fusion (section 2). In section 3 we describe the goals for the process model. Then an overview of the process model itself will be presented, illustrated with examples of real world cases (section 4), followed by the conclusion (section 5).

2. Data Fusion

Valuable work has been done on data fusion in areas other than data mining. From the 1970s through the 1990s, the subject was both quite popular and controversial, with a number of initial applications in economic statistics mainly in the US and later in the field of media research mainly in Europe and Australia [1,3,6,7]. It is also known as micro data set merging, statistical record linkage, multi-source imputation and ascription. The application focus of most of this work has been on merging surveys with surveys to reduce respondent fatigue or find relations between surveys from different domains.

2.1 Data Fusion Concepts

We assume that we start from two data sets. These can be seen as two tables in a database that refer to data sets that may be disjoint. The data set that is to be extended is called the recipient set *A* and the data set from which this extra information has to come is called the donor set *B*. We assume that the data sets share a number of variables. These variables are called the common variables *X*. The data fusion procedure will add a number of variables to the recipient set. These added variables are called the fusion variables *Z*. Unique variables are variables that only occur in one of the two sets: *Y* for *A* and *Z* for *B*. In general, we will learn a model for the fusion using the donor *B* with the common variables *X* as input and the fusion variables *Z* as output and then apply it to the recipient *A*. In most cases, statistical matching is used as the core fusion algorithm. The statistical matching approach can be compared to *k*-nearest neighbor prediction with the donor as a training or search set and the recipient as a test set.

2.2 A Marketing Example

Assume figure 1 describes the situation at an insurance company with 10 million customers (recipient set *A*). They have collected a lot of internal data by tracking customer interactions, such as policy ownership, claims data, response to direct marketing, etc.

What they lack however is all kinds of external information, such as ownership of competing products, media consumption, lifestyle etc. Fortunately there is a survey available, paid for by all major players in the insurance industry, for which 10.000 respondents were interviewed extensively (donor set B). The company can learn a lot by mining the survey.

Still there are some questions that cannot be answered easily. The company could be interested in the relation between unique variables *Y* and *Z*. Examples are the relation between survey questions and high value customers, customers at risk of switching to another company or owners of a specific product not in the survey. The company could also be interested in information on an individual level instead of an aggregate level, for instance for the purpose of selecting prospects for direct marketing campaigns and other forms of one-to-one interactions, either through manual selection or further predictive modeling [7].

Data fusion can provide a solution. For instance, assume that the customer database contains 100 core variables, the survey contains 2000 variables and 20 variables are common to both. By fusing the database with the survey, we carry out a virtual survey by predicting answers to all survey questions for all 10 million customers – for a very small fraction of the cost of interviewing all those customers, which would be infeasible anyway.[1]

3. Goals for the Process Model

Though the main idea behind data fusion is simple, carrying out real world data fusion projects from start to finish can be quite complex, as explained in the introduction. In

[1] Today, in marketing, data fusion is often limited to fusion on zip code alone.

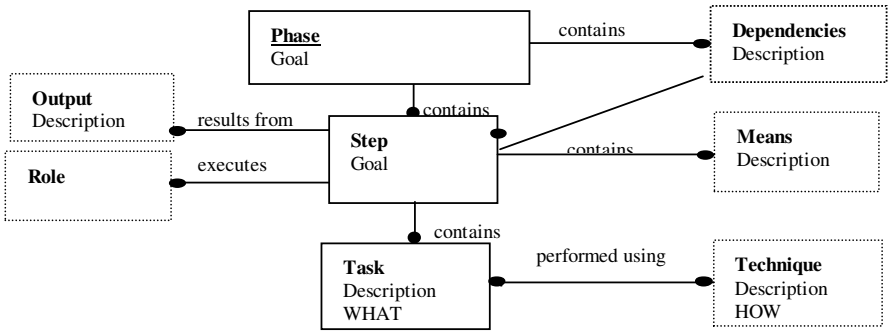

Figure 2: ***Process model elements***

addition to these challenges, we envision that the core of the fusion project should be carried out in five days. So it is no surprise that process automation is the main goal of the process model [8].

This can be detailed into a number of objectives. The process model should function as a basis for project management, reporting and planning. For project participants it should provide a generic approach for projects. Less experienced users may follow it as a strict guideline, more experienced users can use it as a checklist to verify whether everything has been covered. It should provide a knowledge management framework – interesting project experiences or results from literature can be generalized and stored in the process model. Given this experience about what the major pain points are during projects and what possible solutions could be, it provides a framework for continuous process improvement and automation.

4. Process Model Structure and Content

As a major (and obvious) solution direction we have chosen a hierarchical structure for the process model, just like in the CRISP_DM process for data mining [4]. This way we can satisfy both the requirement of a high level overview for setting priorities in planning and automation and the requirement for a concrete, low level of detail.

4.1 Detailed structure

We distinguish between phases, steps and tasks (figure 2). A phase (e.g., orientation phase) in the process generally results in a clear deliverable (e.g., project agreement). Phases differentiate in terms of the level of interaction with the client (extensive for the orientation phase). A phase is described in terms of its goal (e.g., determine fusion needs and agree), the steps it contains, the dependencies between the steps and the outputs (both internal and for the client). Steps contain a set of activities that are preferably performed out by a single person. Steps are generic enough to appear in most projects. Steps are described by the goal, people involved (e.g. client, account manager, project manager, analyst), outputs, means like tools and templates, tasks and techniques. Only techniques describe *how* a certain task may be achieved; phases, steps, tasks only define

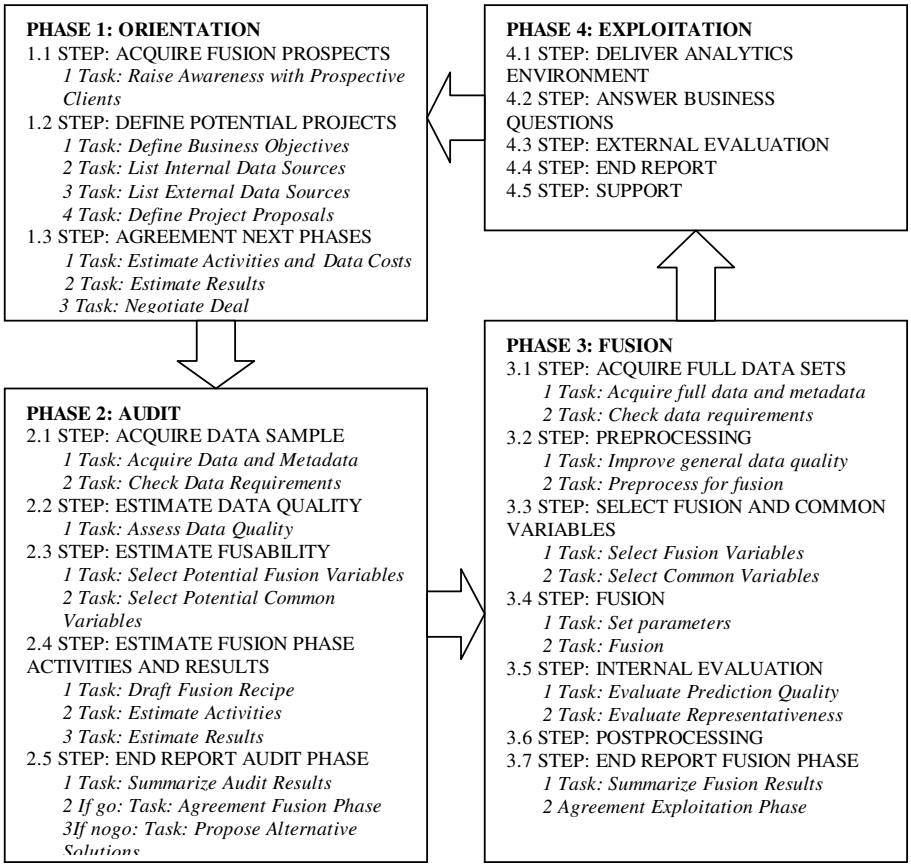

Figure 3: ***Process Model***

what needs to be achieved. In the following sections we will go through the various phases: orientation, audit, fusion and exploitation. We will not describe these phases in detail here, but rather single out some interesting experiences we had in the pilot projects, illustrating the practical steps needed before and after running the core fusion algorithm.

4.2 Orientation Phase

The goal of the orientation phase is to reach an agreement with a client for a well-defined data fusion project. The phase starts with acquiring a potential client and formulating potential projects. The possibility of projects mainly depends on the business questions and available data sets. The phase ends with formulating an agreement, containing at least the business questions to be answered, the datasets to be used and a high level agreement on the scope of the project, in terms of activities, costs and planning.

A key step involves choosing the donor and recipient set to be used (step 1.2). Before we started our pilot projects we assumed that the recipient would be an

internal customer database provided by the client and the donor would generally be externally acquired. In our first pilot project we immediately encountered an exception: the client wanted to enrich a commercially available address list with data that was generally only known for loyal, existing relations. Finding entirely new customers was a major issue for this organization, and this way they could analyze, profile and segment the external market using exactly the same data and methodology they used for existing customers.

An important guideline in this phase is to stay open minded: consider also other techniques than data fusion. Many questions can be answered just by mining on the donor only, or mining a small sample dataset of customers appearing both in the donor and the survey may be an alternative, if available.

This phase involves a lot of client interaction, so automation is less appropriate and obvious. Partial automation can be achieved by providing 'self service' for the client. For example a web site could be created with decision support for discovering interesting business opportunities that may be addressed with fusion, or search engines that provide intelligent search through available donor data.

4.3 Audit Phase

The goal of the audit phase is to minimize the risk that fusion benefits outweigh the costs. Core steps are analyzing the data quality and selecting the potential common and fusion variables. Together with the client we decide on the feasibility of the fusion phase, based on these findings. If fusion is expected to be infeasible, other donor and recipient sets are chosen or the project is stopped – with minimal loss of investments on both sides. The design of this phase is one more result of our experience in the pilot projects. Driven by the envisioned time goals we were first tempted to 'simply' start fusing, risking large investments in data preparation.

Choosing the right common and fusion variables is key in this process (step 2.2). The fusion variables must provide added value with respect to the business goals, e.g. facilitating better selection of prospects. Too be feasible, there must be a strong relation between common and fusion variables, and this relation must be representative between donor and recipient. If the donor set is a subset of the recipient set the conditions are generally good – all recipient variables can then be used as potential commons.

This step poses interesting questions for the data mining and statistical communities. How can we estimate the quality of a data set, and how can we estimate the fusability of two sets, without performing extensive data preparation and analysis? In our pilot projects we found some rules of thumb for these questions, but more research has to be done, including the theoretical underpinning of these rules.

4.4 Fusion Phase

The goal of the fusion phase is simply to perform the fusion, following the requirements defined earlier. First all data sets are acquired, loaded and preprocessed. Then the fusion and common variables are chosen. The fusion parameters are set and the fusion is performed, evaluated and repeated if necessary. The resulting fused dataset is post processed to the desired format and the results are reported to the client.

Because in this phase very large datasets are used, performance becomes a much more serious issue than in the audit phase. This makes automation a must, for data preparation, choosing common and fusion variables and the core fusion step [1,5,6].

We discovered that the other steps in this phase were at least equally important, critical and painful. In one of the pilot projects the recipient contained hundreds of raw variables for millions of bank customers. We needed to cluster many variables in order to deal with sparseness in the data set. We had to transform common variables to the same format, remove autocorrelations, perform population weighting etc. Most importantly, we had to filter out commons and fusion variables that were bad predictors, hard to predict or irrelevant for the business objectives. In contrast to public domain data sets, where most of the variables have some relation the behavior to predict, we often encountered situations where up to 90% of the raw variables were not relevant.

Extraction, transformation and load (ETL) tools generally provide powerful facilities to implement data transformations – however, the main questions is: how can we determine, at least semi-automatically, what kind of transformations are needed given the business, data mining and data fusion objectives?

4.5 Exploitation Phase

The goal of the exploitation phase is to satisfy the business objectives, using the enriched data set. This is the bottom line, external evaluation of the added value of the fusion variables.

If the focus has been put on the wrong fusion variables, a fusion with a reasonable classification accuracy for 90 of the 100 variables can likely be worse than a fusion where 10 out of 100 were predicted well: the quality measured in the fusion phase is no guarantee for good quality in the exploitation phase. Another example is a situation where a customer wants to make cross-tabulations on the enriched data instead of using it as input for selection or predictive modeling. In this case, it is important that relations between variables are preserved, even for values that are not frequent. This can be achieved for instance by constraining the statistical matching procedure through penalizing donors that are used too often in matching [1,5,6].

So generally, it will be very important to know what the enriched data set will be used for, to make the right choices in the audit and fusion steps. However, for competitive reasons some clients may be hesitant to share what they want to do with the fusion result, as we encountered in the pilot for the financial services company.

5. Conclusion

We have presented a high level overview of a process model for data fusion. The model provides a framework for project planning, execution, reporting, process improvement and automation. It has been a very helpful tool for us to guide the fusion pilots and development of the data fusion factory. The model also provides some typical examples of problems that can arise in large scale, real world applications of data fusion and data mining.

References

[1] K. Baker, P. Harris and J. O'Brien (1989) Data Fusion: An Appraisal and Experimental Evaluation. Journal of the Market Research Society, 31 (2), 152-212.

[2] R.J. Brachman and T. Anand. *The Process of Knowledge Discovery in Databases.* In U.M Fayyad, G. Piatetsky-Shapiro, Padhraic Smyth and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, AAAI / MIT Press, Cambridge, MA, 1996.

[3] E.C. Budd. The creation of a microdata file for estimating the size distribution of income. Review of Income and Wealth (December) 17, (1971) pp 317-333

[4] P. Chapman., J. Clinton, T. Khabaza, T. Reinartz., R. Wirth. (1999). The CRISP-DM Process Model. Draft Discussion paper, Crisp Consortium, March 1999. http://www.crisp-dm.org/.

[5] X. van Pelt, *The Fusion Factory: A Constrained Data Fusion Approach.* MSc. Thesis, Leiden Institute of Advanced Computer Science (2001).

[6] D.B. Radner, A. Rich, M.E. Gonzalez, T.B.. Jabine and H.J. Muller (1980). Report on Exact and Statistical Matching Techniques. Statistical Working Paper 5, Office of Federal Statistical Policy and Standards, US DoC. See http://www.fcsm.gov/working-papers/wp5.html

[7] P. van der Putten. *Why the Information Explosion Can Be Bad for Data Mining and How Data Fusion Provides a Way Out*. Second SIAM International Conference on Data Mining, Arlington, April 11-13, 2002

[8] M. Ramaekers, *Procesmodel The Fusion Factory.* Sentient Machine Research, Amsterdam, (2000).