Data Fusion and Semantic Web Mining: Meta-Models of Distributed Data and Decision Fusion

Vladimir Gorodetski, Oleg Karsaev, Vladimir Samoilov
St. Petersburg Institute for Informatics and Automation
{gor, ok, samovl}@mail.iias.spb.su

Abstract:

According to the Project funded by European Office of Aerospace Research and Development (AFRL/IF) we are developing mathematical model, multi-agent architecture and technology realized as a software tool supporting design and implementation of Data Fusion (DF) applications of broad spectrum. A multitude of tasks to be solved with regard to the development of DF software tool can practically be divided into two groups. The tasks whose solutions make use of methods, models and technologies of other adjoining scientific fields, for instance, data mining and knowledge discovery, multi-agent systems, object-oriented design, etc fall into the first group. The second group includes the tasks specific for DF systems and require development of specific methods, models and technologies. In fact, the most part of tasks of the last group fall into the interests of Semantic Web Mining. Although the tasks of both above groups are the subjects of the Project, below the DF specific tasks are only highlighted.

In its essence, DF task is one of making decisions (as a rule, classifications) on the basis of distributed data sources presented by distributed databases with access through Intra- or Internet. These sources contain data that can be represented by different data structures (temporal, sequential, transactional, relational), they can be of different physical nature (images, signals, truth values, etc) and measured in different scales (Boolean, categorical, real), be of different accuracy and reliability, they can be uncertain, contain missing values, etc. The objective of a DF system is to combine useful information from all of these sources to make decision, for instance, classification of an object, object state, situation, etc.

Within DF specific tasks two classes of them are of most significance. The first is development of *meta-model of distributed data* sources and the second is development of *meta-model of combining decisions* produced on the basis of particular sources. Note that the former task is purely Semantic Web related.

Three main issues of DF specific R&D of the Project are the subjects of presentation.

1. The *ontology-centric approach* which is considered as a basis of the development of meta-model of distributed data sources. In particular, ontology-based approach aims to answer the following questions associated with the development of meta-model of distributed data sources:
How to resolve the data non-congruency problem caused by *heterogeneity* and *distribution* of data sources? The particular questions are: How to provide for monosematic understanding of the terminology used in formal specification of distributed entities which, as a rule, are developed by distributed analysts? How to solve the entity identification problem, which arises due to the fact that the same entity specification is represented by its fragments in distributed databases? How to cope with the diversity of data physical natures, scales of attribute measurement, variety of data accuracy, duplication of the same attributes in different data sources? How to provide compatibility of ontology-based specifications of DF system notions and their interpretations represented in terms of a database language?

2. The second important issue of DF system design is *distributed learning* and meta-model of combining decisions. The questions to be answered here are as follows: What structures are used to combine particular decisions made on the basis of particular data sources to generate global decision? How to manage distributed data in order to provide correctness with regard to allocation of training and testing data used for learning particular classifiers and how to form and manage meta-data used for learning components responsible for combining decisions? What formal techniques are appropriate to combine decisions in accordance with the hierarchy of classifiers?

3. The third issue is of architectural kind. We use multi-agent architecture which to our opinion is the most appropriate for the implementation of DF systems and also for many other Semantic Web-based applications. This architecture includes specific components (intelligent agents) and protocols aiming at solving the questions formulated within both above issues.

In presentation we intend to highlight our results concerning to the above issues and their correlation with Semantic Web Mining-related tasks and problems. In addition, we intend to present an outline of a technology supported by a software tool used for design and implementation of DF software tool including design and implementation of its ontology component. In conclusion a brief outline of DF applications developed and being developed can be outlined.