Semantic Heterogeneity Reduction for Big Data in Industrial Automation

Václav Jirkovský[a,b] and Marek Obitko[b]

[a]Czech Technical University in Prague, Prague, Czech Republic
[b]Rockwell Automation Research and Development Center,
Pekařská 695/10a Prague, Czech Republic
{vjirkovsky,mobitko}@ra.rockwell.com

Abstract. The large amounts of diverse data collected in industrial automation domain, such as sensor measurements together with information in MES/ERP[1] systems need special handling that was not possible in past. The Big Data technologies contribute a lot to the possibility of analyzing such amounts of data. However, we need to handle not only data volume, which is usually the major focus of Big Data research, but we also need to focus on variety of data. In this paper, we primarily focus on variety of industrial automation data and present and discuss a possible approach of handling the semantic heterogeneity of them. We show the process of heterogeneity reduction that exploits Semantic Web technologies. The steps include construction of upper ontology describing all data sources, transformation of data according to this ontology and finally the analysis with the help of Big Data paradigm. The proposed approach is demonstrated on data measured by sensors in a passive house.

Keywords: Big Data, Semantic heterogeneity, Data integration, Industrial automation

1 Introduction

The large amounts of diverse data can be collected in virtually any domain today, including industrial automation domain. A typical example is an assembly line — at the lowest level, there are sensors reading values important for low level control such as moving machines, while in the upper levels, there are quality checks and monitoring with a possible flow of customer orders etc. The data that can be captured are of huge quantities and variety. In addition, for the meaningful use of the results of their analysis, it is often necessary to react quickly. These characteristics overlap with the Big Data three Vs - volume, velocity and variety. However, let us emphasize from the beginning that the area of automation is much wider than just assembly lines — a control system can be used for any processes where it is needed to reduce human intervention for various reasons.

[1] MES - Manufacturing Execution System; ERP - Enterprise Resource Planning

Similarly to other domains that need to use the Big Data paradigms, in industrial automation data are already captured, processed and analyzed, however, the properties of these data were enabling only smaller scale processing. For example, sensor data are used for low-level control of a single machine. In addition, the data can be kept in historian software for further offline analysis — and even this analysis of single time series stored from measured data streams is hard with larger amounts of data for classical historian software packages.

The Big Data technologies were developed primarily by web companies to handle large amounts of user clickstreams and other information to be able to offer suitable products and ads in real time. Using these technologies it becomes possible and meaningful to analyze much larger amounts of data than when using classical data analysis packages. However, we must note that the focus of the existing technologies has been mainly on the *volume* of data. This is clear — handling the volume even for simple analysis was something that was needed most. As we discuss later in the paper, the *variety* of data is something that is important and not satisfactorily solved as well. In fact, one of the main outcomes of the survey [1] among the Fortune 500 leading companies is that "it's about variety, not volume" and that data integration will continue to be one of the greatest challenges faced.

Handling the semantic heterogeneity for Big Data application in industrial automation is the main topic we discuss in this paper. The rest of this paper is organized as follows: We describe Big Data features and existing technologies related to the types of data that are to be processed in industrial automation. Then we discuss the data heterogeneity problems we are facing and we propose how to handle the problem at least in our domain. We illustrate the approach on an example of data measured from a passive house, for which automation helps to maintain conditions suitable for humans. After that we provide a conclusion with the outlook for our future work.

2 Big Data

The term Big Data is used for data growing so that it becomes difficult to manage them using existing database management concepts and tools [17]. The difficulties may be related to data capture, storage, search, sharing, analytics etc. Such data noticeably exhibit the following properties: the amount of data is large and is quickly growing (volume), the data are generated and need to be processed with higher speed, usually in "real" time (velocity) and the data include less structured forms, including texts, images and videos (variety).

The main reason for the necessity of handling such data is better data driven decision making. The ability to handle these data presents an opportunity to create business advantage; however, it also requires new infrastructures and new ways of thinking different from the classical business intelligence approaches.

The infrastructures, tools and frameworks developed to manage such data were originally developed by web companies for storing and analyzing data such as for searching, for analyzing clickstreams and user behavior etc. The required

outcome is typically offering relevant advertisements, products and services for individual users.

Let us mention an example from the automation domain [2]: a CPG (Consumer Packaged Goods) company generates 5000 data samples every 33 milliseconds leading to 4 trillion of samples per year. Such data need to be stored and processed in real time.

3 Related Work

There are many already developed solutions providing more or less capability how to deal with data heterogeneity problem within Big Data domain. Some of the interesting and promising solutions are listed below.

The Autonomy IDOL 10[2] developed by HP Company, where IDOL stands for Intelligent Data Operating Layer, is aimed at providing single processing layer for conceptual, contextual, and real-time understanding of data. The IDOL allows organizations to form a conceptual understanding of information based on a patented probabilistic algorithms to automatically recognize concepts and ideas expressed in all forms of information (documents, video, chat, phone calls, and emails).

The next available complex solution is IBM InfoSphere, especially Master Data Management [4] and Information Server [3]. These systems are cornerstones of the IIG — Information Integration and Governance. It is a unified set of capabilities that bring together data from diverse sources. The IIG delivers critical capabilities to Watson Foundations, the IBM big data and analytics platform.

The Optique [6] is an ontology-based data access solution for Big Data. The system aims at providing an end-to-end solution, where end-users will formulate queries based on ontology. The main components are the ontology and mappings that provide the relationships between the ontology and the underlying data. User queries over the heterogeneous data are transformed with the help of these mappings and ontology.

4 Data Heterogeneity

Information integration from heterogeneous data sources is significant problem in every complex IT system and every successful system has to be able deal to with this integration problem. In this section, the categorization of information integration and heterogeneity are presented.

General information integration consists of a set of local information sources potentially storing their data in different formats (RDF, XML ...) which have to be integrated to provide users with unified data access.

Schema integration [5] — it is the oldest scenario of information integration and represents situation when two (or more) local data sources integration

[2] http://www.autonomy.com/technology/idol-the-os/

is needed. The essential step of schema integration process is to identify correspondences between semantically identical entities of the schemas.

Catalogue integration [11] — In Business-to-Business (B2B) applications, trade partners store information about their products in electronic catalogues (product directories of electronic sales portal) — central warehouse of the marketplace. Finding correspondences among entries of the catalogues is referred to the catalogue matching problem. Having identified the correspondences between the entries, users of a marketplace can benefit from a unified access to the products which are on sale.

Data integration [10], also known as enterprise information integration — is an integration of information from multiple local sources without loading their data into a central warehouse. It ensures inter-operation of multiple local sources having access to the up-to-date data.

There are many different classifications of heterogeneity - e.g., in [9]. We consider the following types of heterogeneity according to [8]:

- **Syntactic heterogeneity** occurs when two data sources are not expressed in the same language, e.g., it can be caused when two ontologies are modelled by using different knowledge representation formalisms - OWL and F-logic.
- **Terminological heterogeneity** stands for variations in names when referring to the same entities in different data sources.
- **Conceptual heterogeneity**, also known as semantic heterogeneity or logical mismatch, denotes the differences in modeling the same domain of interest. it is caused due to the use of different axioms for defining concepts or due to the use of totally different concepts. We can distinguish a difference between the conceptualization mismatch (i.e., differences between modeled concepts) and the explication mismatch (i.e., concepts are expressed in different way).
  - *Difference in coverage* occurs when two data sources describe different regions of the world at the same level of detail and from a unique perspective.
  - *Difference in granularity* occurs when two data sources describe the same region of the world from the same perspective but at different levels of detail (e.g., geographic maps with different scales).
  - *Difference in perspective* (difference in scope) occurs when two data sources describe the same region of the world, at the same level of detail, but from different perspective (e.g., a political map vs. geological map).
- **Semiotic heterogeneity**, also known as pragmatic heterogeneity, stands for different interpretation of entities by people. This kind of heterogeneity is difficult for computer to detect and even more difficult to solve.

Furthermore, it is common to face several types of heterogeneity together. One possible way, how to reduce (partially or in total) heterogeneity between data sources, is similarity matching.

4.1 Heterogeneous Data in Industry Automation

Data integration in industry automation domain can be expressed as an Enterprise Information Integration (EII) problem. The goal of EII is the ability to provide a uniform access to multiple data and information from an entire organization.

There are many different sources of information in an industry company — data from ERP system, MES, production line, and from outside of company. ERP system knows what customers want, MES systems know how to build it, and data from production line sensors shows how the production system works. However, oftentimes the enterprise data source systems are created by different vendors and these systems do not speak the same language.

Furthermore, it is no exception in the industry domain that a same kind of device (from different vendors) mounted at a production line has different interface, i.e., different output. This situation is caused from the fact that the vendors follow for example different standards.

5 Heterogeneity Reduction for Big Data

This section demonstrates main characteristics of data interoperability problem with focus on Big Data and proposes a solution for dealing with it. General process of the proposed heterogeneity reduction schema is depicted in Fig. 1.

The first type of heterogeneity we have to deal with is structure heterogeneity (or syntactic heterogeneity). Various data sources are essential to ensure quick and valuable decisions. In industry automation, it is also common to predict next steps for control from various sources as mentioned earlier — from low level systems as well as high level systems. It is necessary to take into consideration many different types of data sources for our application — text files, XML files, databases, etc.

The second type of heterogeneity is semantic heterogeneity. A possible solution for this problem is the creation of a shared ontology which ensures transformation of the data sources into the same "language".

We present a possible solution how to deal with previously mentioned types of heterogeneity and general heterogeneity reduction problem as well in the following subsection. The suggested approach has to take into account all the Big Data characteristics, not only *variety*, but also *volume* and *velocity*. It is confronted with processing huge data amount (e.g., data stored in databases or data streams from production lines) and a quickness of data processing, data loading, and data storing.

5.1 Ontology Construction

The goal is to deal with data heterogeneity and the first step in our approach is creation of a shared ontology which ensures knowledge sharing. The proper creation of the shared ontology is essential for subsequent processing.

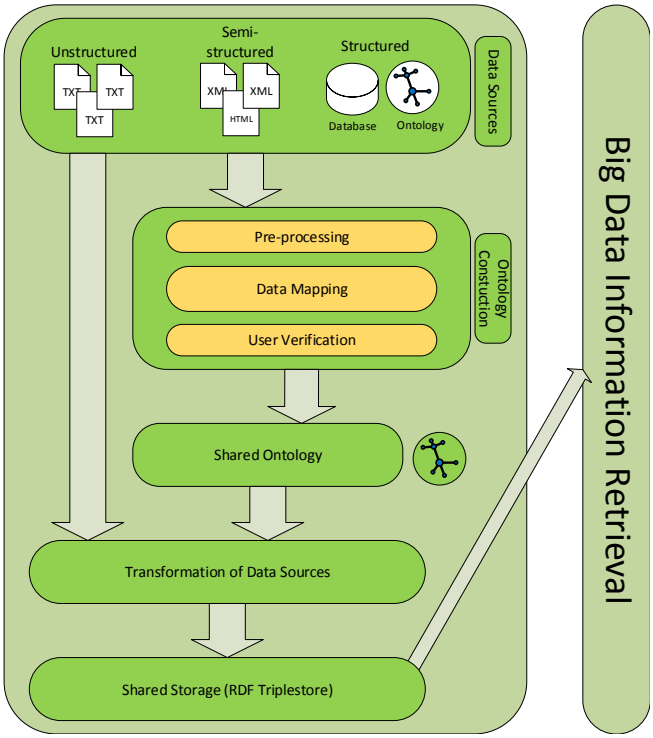

Fig. 1. Heterogeneity reduction process

First, we have to deal with structural diversity. Structural heterogeneity problem falls into the pre-processing step. Data source processing strategies differ depending on its category. Semi-structured documents may be transformed into a predefined relational structure. HTML documents can be indexed and reduced to free text. For processing textual data sources, the system must utilize language-specific natural language processing systems (e.g., GATE [7]).

Next step is the construction of a shared ontology from pre-processed data. A crucial step is the understanding of a content and identifying the correspondent entities across all data sources. Some ontology matching systems can be exploited for this task. We adopted our previously developed system MAPSOM presented in [12]. This system is semi-automatic matching system and is based on similarity measure aggregation with the help of Kohonen's self-organizing maps, and hierarchical clustering. The MAPSOM is focused on user involvement in ontology matching which is suitable for the domains as the industrial automation — i.e., for the application where the highest precision is preferred to the processing time.

The other possible supporting tool for shared ontology construction is formal concept analysis (FCA). The FCA is according to [15] a possible solution for supporting and simplifying ontology design process. It is a theory of data analysis

which identifies conceptual structures among data sets. This approach helps to design better ontologies that are more suitable for ontology sharing than pure taxonomies.

After shared ontology construction, a transformation of data is needed for a subsequent utilization. In the following paragraphs, two possible ways are described.

Data source transformation "on the fly" (on demand — i.e., when a user submits a query) is common in many solutions. This approach has one significant drawback and it is the additional computational and time requirement on every access. If a user submits a query then an initialization and transformation of data sources is needed. This is suitable solution only in the case where the storage capabilities are limited or when data access is not frequent.

The second way is a creation of a "snapshot" (a shared storage). This is more suitable in many application — it can speed up following analysis. On the other hand, it is important to take up-to-dateness of data sources into account — how often it is needed to re-transform source data into the new shared storage.

6 Testing Scenario

To demonstrate previously mentioned approach and to provide the wide usability of the methods, we have selected the following use case. It is based on our former project dealing with measuring physical behavior of passive houses and fine-tuning a dynamic simulation model approximating their operation. This use-case and the simulation model to be fine-tuned were presented in [14].

First of all, the starting point of the process is the pre-processing step. The data sources for this demonstrative example are the SSN Ontology[3] and measured sensors data stored in text files, i.e., temperature, carbon dioxide concentration, relative humidity, and air pressure. All sensor data are contained in text file divided day by day. Every text file consists of sensor headers and lists of sensors records. The SSN Ontology can be the other source for the information completion. The source ontology is a domain-independent and end-to-end model for sensing application by merging sensor-focused, observation-focused, and system-focused views. It is the backbone for a new shared ontology. In this case, the sensor data text file is parsed into a more useable structure (i.e., object representing properties and records of certain sensor) and SSN Ontology can stay in the original form. The pre-processing step is depicted in the Fig. 2.

The manual (semi-manual) ontology construction is possible in this simple example. It can be supported with the similarity mapping of the data source entities. Mapping step lies in finding similar concepts and properties with the help of similarity measures. It is not sufficient to use only one specific similarity measure, because one similarity measure is able to reflect only some specific heterogeneity aspect. Therefore the usage of similarity aggregation is suitable to capture the most of dissimilarity problems according to [12]. We used the

[3] http://www.w3.org/2005/Incubator/ssn

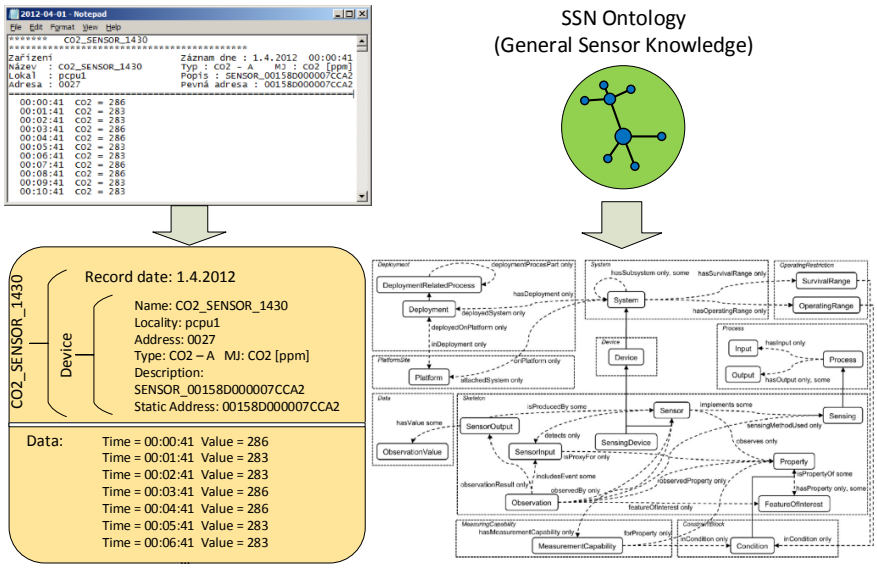

Fig. 2. Pre-processing step of heterogeneity reduction

string similarity measures (*prefix*, *suffix*, and *n-gram* similarity measure) and language-based techniques, i.e., extrinsic WordNet similarity measure (*Resnik*). The following corresponding entities are results of this step:

- Device — Device
- Sensor — CO2_SENSOR_1430
- ObservationValue — Value

The shared ontology construction is important step and has to be provided by skilled domain expert. The created ontology (Fig. 3) is the base for ontology-driven data transformation of the data sources. The shared ontology includes concepts and relationships for sensor properties, measurements, and taxonomies.

Finally, the last step is storage of transformed data from data sources. Naturally, it is possible to use original data sources with on-line ontology-driven data transformation, but this approach is inappropriate for the case of big amount of sensor data that require fast data access. Therefore, the storage of transformed data is required for faster data access. The Hadoop RDF triplestore, i.e., H2RDF [16], was selected due to the data characteristics and ontology-driven transformation as well.

Using the described approach, we can take the advantage of integrated data sources for knowledge discovery with the help of Map&Reduce programming as known from the Big Data technologies, directly using RDF processing. The implementation of this approach, which is in progress and extends data processing without prior integration [13], will show the advantages of direct usage of integrated data sources.

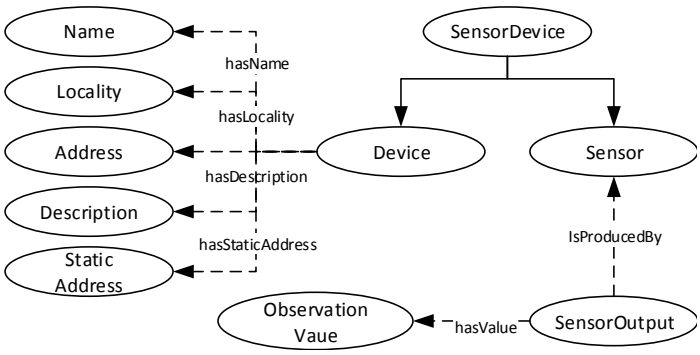

Fig. 3. Generated shared ontology

6.1 Summary

The cornerstones of the proposed heterogeneity reduction approach are transformation of the data sources into a simple common format, translation to the "same language", and shared ontology construction. All of the mentioned steps are complex tasks and we have demonstrated one of the possible approaches of how to deal with these challenges.

Our approach offers possibility to integrate data sources for batch processing (utilization of the joint RDF storage), real-time processing of sensor measurements (shared ontology serves for data source interface), and combination of the both mentioned processing.

7 Conclusions

We have shown how the heterogeneity of data sources can be reduced by means of shared ontology. This solution is proposed to enable dealing with integration of diverse data collected in industrial automation domain, such as sensor measurements or data from MES/ERP systems. The proposed solution was demonstrated on the passive house testing use case.

The capability of handling various data sources stored in databases or files as well as various streams of data is the significant advantage not only for the industrial automation domain. This capability can precisely capture various relationships among data sources and therefore can improve the process of decision making. As have demonstrated in the state of the art review, the variety aspect of Big Data is very important one, but also very hard to be solved.

Our future work includes further development of the proposed process for heterogeneity reduction in Big Data with the focus on industrial automation domain. This work involves construction of complex software solution which will be able to deal with huge amount of data from ERP/MES system, streams of data generated by sensors from a production line, and also integrating data from external data sources, such as weather conditions.

To conclude, the possibility of handling large volumes of data at high speed and especially the integration of various data sources is essential for gaining competitiveness in every enterprise. The heterogeneity of data was always an important issue to handle, but it is much more visible and important today when the trend of increasing processed amounts of data continues.

8 Acknowledgements

This research has been supported by Rockwell Automation Laboratory for Distributed Intelligent Control (RA-DIC) and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS12/188/OHK3/3T/13.

References

1. Big Data executive survey, Consolidated summary report. Consolidated summary report, NewVantage Partners (2012)
2. The rise of industrial Big Data. Whitepaper, GE Intelligent Platforms (2012)
3. Integrating and governing big data. Whitepaper, IBM Corporation (2014)
4. The MDM advantage: Creating insight from big data. Whitepaper, IBM Corporation (2014)
5. Batini, C., Lenzerini, M., Navathe, S.B.: A comparative analysis of methodologies for database schema integration. ACM computing surveys 18(4), 323–364 (1986)
6. Calvanese, D., Giese, M., Haase, P., Horrocks, I., et al.: Optique: OBDA Solution for Big Data, pp. 293–295. The Semantic Web: ESWC 2013 Satellite Events, Springer (2013)
7. Cunningham, H.: Gate, a general architecture for text engineering. Computers and the Humanities 36(2), 223–254 (2002)
8. Euzenat, J., Shvaiko, P.: Ontology matching, vol. 18. Springer (2007)
9. Goh, C.H.: Representing and reasoning about semantic conflicts in heterogeneous information systems. Ph.D. thesis, Massachusetts Institute of Technology (1996)
10. Halevy, A.Y., Ashish, N., Bitton, D., Carey, M., Draper, D., Pollock, J., Rosenthal, A., Sikka, V.: Enterprise information integration: successes, challenges and controversies. In: Proceedings of the 2005 ACM SIGMOD international conference on Management of data. pp. 778–787. ACM (2005)
11. Ichise, R., Hamasaki, M., Takeda, H.: Discovering relationships among catalogs. In: Discovery Science. pp. 371–379. Springer (2004)
12. Jirkovský, V., Ichise, R.: Mapsom: User involvement in ontology matching. In: Proceedings of the 3rd JIST Conference. Springer (2013)
13. Jirkovský, V., Obitko, M., Novák, P., Kadera, P.: Big Data analysis for sensor time-series in automation. In: Proc. of 19th ETFA Conference (2014)
14. Novák, P., Šindelář, R.: Design and verification of simulation models of passive houses. In: Proc. of 17th ETFA Conference (2012)
15. Obitko, M., Snášel, V., Šmíd, J.: Ontology design with formal concept analysis. In: CLA. vol. 110 (2004)
16. Papailiou, N., Konstantinou, I., Tsoumakos, D., Koziris, N.: H2rdf: adaptive query processing on rdf data in the cloud. In: Proceedings of the 21st international conference companion on World Wide Web. pp. 397–400. ACM (2012)
17. Singh, S., Singh, N.: Big Data analytics. In: 2012 International Conference on Communication, Information & Computing Technology (ICCICT). pp. 1–4 (2012)