A Hybrid Strategy for Integrating Sensor Information

Koly Guilavogui, Laila Kjiri and Mounia Fredj
ENSIAS, Mohammed V-Souissi University, Avenue Mohammed Ben Abdellah Regragui, Rabat, Morocco

Keywords: Integration Systems, Data Warehouses, Multidimensional Databases, Mediation, Sensor Networks.

Abstract: The combination of sensor networks with databases has led to a large amount of real-time data to be managed, and this trend will still increase in the next coming years. With this data explosion, current integration systems have to adapt. One of the main challenges is the integration of information coming from autonomously deployed sensor networks, with different geographical scales, but also with the combination of such information with other sources, such as legacy systems. Two main approaches for integrating sensor information are generally used: virtual and warehousing approaches. In the virtual approach, sensor devices are considered as data sources and data are managed locally. In contrast, in the warehousing approach, sensor data are stored in a central database and queries are performed on it. However, these solutions turn out to be difficult to exploit in the current technology landscape. This paper focuses on the issue of integrating multiple heterogeneous sensor information and puts forward a framework for decision making process.

1 INTRODUCTION

During the last years, sensor networks have been an area of intense investigation, from low-level protocols to application level databases. They may be wireless or not. Wireless Sensor Networks (WSNs) are composed of small devices with low processing power and data storage. In general, sensor devices present limited energy and communicate with each other by short-range radios (Akyildiz et al., 2002). They produce raw data continuously, thus providing data streams. The elements of a stream can be numeric (e.g. temperature), geospatial (e.g. GPS coordinates), or multimedia (images, text, videos, and sounds). We have then to deal with complex data.

The combination of WSNs with databases has led to a large amount of real-time data to be managed, and this trend will still increase in the next coming years. With this data explosion, current integration systems have to adapt. One of the challenges is the integration of information coming from autonomously deployed WSNs, with different geographical scales, but also with the combination of such information with other sources.

Nowadays, WSNs have been employed for monitoring and controlling real world data in several indoor and outdoor environments (Aggarwal et al., 2013). They are gaining more and more attention by both research communities and industries because of their use in many applications (Akyildiz et al., 2002) such as environmental monitoring (forest fire prevention, flood control, etc.), monitoring traffic (road, air, etc.), home automation, etc.

The great diversity of these sensors makes them excellent tools for distributed sensing of phenomena, processing and dissemination of information collected to one or more observers (Aggarwal et al., 2013).

This type of information is spatial, temporal, thematic, and is inherently dynamic. This complexity of information causes new needs in terms of acquisition, structuring, integration, analysis and visualization of this information for decision support systems (Kimball and Caserta, 2004), (Favre et al., 2013).

It is undeniable that the use of sensors allows a better monitoring of events that occur in the real world. However, data produced by these devices turn out to be difficult to exploit by conventional approaches in data warehousing. Two main approaches for integrating sensor information are generally used: virtual and warehousing approaches. In the virtual approach, sensor devices are considered as data sources and data are managed locally. In contrast, in the warehousing approach,

Guilavogui K., Kjiri L. and Fredj M..
A Hybrid Strategy for Integrating Sensor Information.
DOI: 10.5220/0004936302810286
In *Proceedings of the 16th International Conference on Enterprise Information Systems* (ICEIS-2014), pages 281-286
ISBN: 978-989-758-027-7

sensor data are stored in a central database and queries are performed on it. We note that the approaches mentioned above are not mutually exclusive, and can be combined in only one system (this refers to the hybrid approach).

The design of an integration system requires not only the choice of the approach (virtual, warehousing or hybrid) but also the model of integration. Basically, there are two models of integration for specifying the correspondence or the mapping between the data at the sources (local schemas) and those in the global schema: the Global-as-View (GaV) and the Local-as-View (LaV). The other variants (hybrid approaches) are just the combination of GaV and LaV (Halevy, 2001). It is exactly this mapping that will determine how the queries posed to the system are answered (Lenzerini, 2002).

In the GaV approach, the global schema is modeled as a set of views over the schemas of the sources. The major drawback of the GaV approach is that there is a necessity to redefine the view of the global schema every time a new source is integrated. Therefore, it is an optimal choice when the source schemas are not subject to frequent changes. The LaV approach is the dual of the latter, since the source schema is modeled as a set of views on the global schema.

The major contribution of this paper is the proposition of a hybrid approach which combines the benefits of both virtual and data warehousing approaches to efficiently and effectively integrate data collected from WSNs and other sources. Our framework follows the BGLaV (BYu Global-Local-as-View) approach for the mapping between the global schema and local schemas representing sources of data. The BGLaV (Xu and Embley, 2004) approach combines the best of LaV and GaV approaches that uses source-to-target mappings based on a predefined conceptual target schema. The latter is specified ontologically and independently of any of the sources.

Our objective is to provide an information integration system that uses not only any type of data sources (relational, semi-structured, ontologies, etc.) but also ensures the freshness of data and provides more flexibility with respect to adding or deleting a new information source. We also aim at building and composing multidimensional databases (data cubes) on-the-fly using our integration system. Existing approaches for integrating sensor information have some drawbacks and do not meet aforementioned requirements.

The remainder of this paper is organized as follows. Section 2 reviews the state of the art concerning the integration of heterogeneous sensor information with other sources. Section 3 puts forward the proposed architecture for the hybrid strategy. Section 4 concludes with a perspective of our on-going work.

2 RELATED WORK

As WSNs can be considered as autonomous, distributed and heterogeneous sources, data integration was needed in order to give the user an integrated view of information sources.

2.1 Data Integration Challenges

Data integration is vital in large organizations in order to deal with a research problem or environmental issues (e.g. forest fire, flooding, water quality, etc.). These organizations generally own a multitude of data sources, where data sets are being produced independently. In practice, sensor schemas typically refer to relational database schemas, semi-structured information or ontologies (Konstantinou, 2012).

The main challenge for an integration system is to solve different conflicts among information sources and to represent them in a single coherent schema (Ziegler and Dittrich, 2004). These conflicts may be encountered at schema and instance level. Beyond the issues of structural integration of heterogeneous information, the challenge is to identify conflicts among concepts in different sources that are semantically related, and then to propose a resolution of those conflicts. This data integration process is quite complex, since the integration system has also to take into account issues related to autonomy, distribution, dynamicity and volume of data.

In the next section, we review the virtual, warehousing and hybrid approaches for sensor information integration and we make a comparative analysis.

2.2 Virtual Approach

In this approach in Figure 1 proposed by (Wiederhold, 1992), the integration system is based on a mediator with a global schema and wrappers. Users' queries are expressed on the global schema. Two steps are required: the mediator accepts a query and determines the appropriate set of information sources to answer the query, and then generates the appropriate sub queries or commands for each information source. Wrappers perform translation, filtering and data collection from local sources. Then, the mediator obtains results from the information sources via wrappers, merges information and returns the final answer to the user or application.

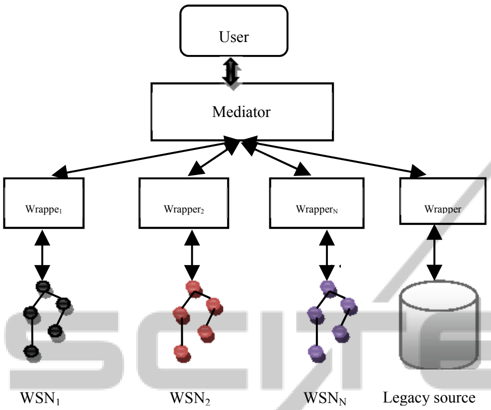

Figure 1: Virtual approach of information integration (Wiederhold, 1992).

This approach refers to virtual, mediation, on-demand or on-the-fly approach, since information is extracted from the sources only when queries are posed to the system. The Cougar system (Yao and Gehrke, 2002) at Cornell University is the first step toward sensor data management using the virtual approach. The work of (Ibrahim et al., 2005; Stocks et al., 2009) both also developed a mediation system using sensors. The work of Casola et al. (2009) proposed SeNsIM (Sensor Networks Integration and Management). SeNsIM is a framework that enables the integration of heterogeneous sensor systems using XML as modeling language.

2.3 Warehousing Approach

In this approach in Figure 2, a new physical database is created to integrate and import data from several sources, usually in one direction only (Widom, 1995; Inmon, 1996; Kimball and Caserta, 2004).

In the warehousing approach, the data model is multidimensional and queries are generally complex using sophisticated means of navigation in data across different dimensions with OLAP (On Line Analytical Processing) operators (Codd and Salley, 1993). In Figure 2, wrappers do the extraction and some transformation tasks while the integrator performs the final transformation and loading process. Queries are directly evaluated on the data warehouse without accessing the original information sources.

In this paper, we make a difference between the materialized and the warehousing approaches, although the two approaches are almost similar. The materialized approach generally refers to data warehousing in some studies (Zhou et al., 1995; Shokoh, 2010). In fact, data warehouses (DW) use materialize views, but the materialized approach does not all include the requirements for data warehousing. A materialized view is thus like a cache - a copy of data that can be accessed quickly (Gupta and Mumick, 1995). This approach is not intended to provide analytical processing. Opposite to DW, the historical data is not kept for long time storage and the new version of data replaces the last one.

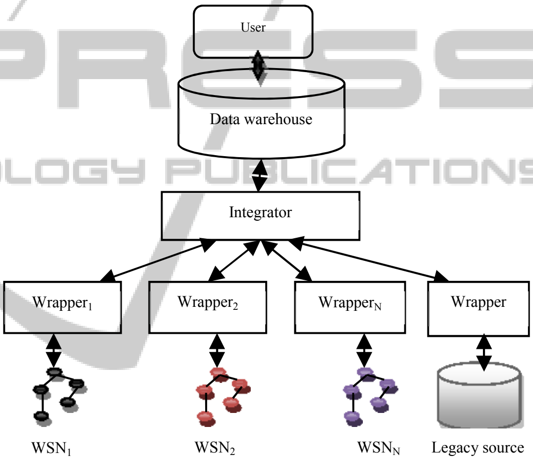

Figure 2: Warehousing approach of information integration (Widom, 1995).

Recent works (Shah et al., 2009; Ahmed et al., 2010; Li et al., 2013; Gökçe and Gökçe, 2014) used the warehousing approach for reducing the energy consumption in buildings equipped with sensors. (Da Costa and Cugnasca, 2010) also used a warehousing approach to manage sensor data for animal monitoring (pollinators). (Mathieu, 2011) proposed a web service based architecture based on OGC (Open Geospatial Consortium) standards for near real-time integration of sensor data streams into a geo-analytical data warehouse.

2.4 Hybrid Approach

This approach is usually used to overcome the drawbacks of the approaches mentioned above. In this work, we consider the hybrid approach to be divided in two: (a) the hybrid approach that combines virtual and warehousing approaches; and

(b) the hybrid approach that combines virtual and materialized approaches. The work of (Grosky et al., 2007; Huang and Javed, 2008; Roantree, 2009) followed the hybrid approach (virtual and materialized) for data integration. In the next section, we make a comparative analysis of the different approaches we consider in this work.

2.5 Comparative Analysis

As summarized in Table 1, there is no work that handles all the type of data sources that can be found in WSNs environment. However, all the authors used relational data sources in their respective work.

One can see from Table 1 that the virtual approach allows flexibility and freshness of data but avoids data replication. It is not intended to OLAP manipulation but for information retrieval. Generally, the virtual approach is preferable to the warehousing approach in the following cases (Convey et al., 2001): (a) the number of data sources in the integration system is very large, scalable, and sources are likely to be updated frequently, and (b) there is no way to predict the types of queries that the user will make.

The warehousing approach allows data replication and it is well suited for OLAP of historical data, but the major drawbacks of this approach are freshness of data and flexibility when adding or deleting new information sources.

However, the hybrid approach is suited in case some data sources of underlying sources are frequently changing and other may change less frequently (Shokoh, 2010). Research using a hybrid approach is not widespread. The existing works that used this hybrid approach generally built their integration system for the purpose of information retrieval and not for OLAP. Our work is different, since our framework not only uses virtual and data warehousing approaches but also data sources could be from any types (relational, semi-structured, unstructured, geospatial or ontologies).

To the best of our knowledge, the hybrid approach (virtual and data warehousing) has not been used for managing sensor information combined with other sources. In the next section, our

Table 1: Comparative analysis of the three approaches.

Approaches Criteria		Virtual				Warehousing						Hybrid		
		(Yao and Gehrke, 2002)	(Ibrahim and al., 2005)	(Casola et al., 2009)	(Stocks et al., 2009)	(Shah et al., 2009)	(Ahmed et al., 2010)	(Da Costa and Cugnasca, 2010)	(Mathieu, 2011)	(Li et al., 2013)	(Gökçe and Gökçe, 2014)	(Grosky et al., 2007)	(Huang and Javed, 2008)	(Roantree, 2009)
Type of data sources	Relational (R) /Object-relational (O-R)	+	+	+	+	+	+	+	+	+	+	+	+	+
	Semi-structured (XML)	-	-	+	-	-	-	+	+	-	+	+	+	+
	Ontologies (RDF/OWL)	-	-	-	-	-	-	-	-	-	-	-	+	-
	Unstructured data (CSV, videos, etc.)	-	-	-	+	-	+	+	-	-	+	+	-	+
	Geospatial data type	-	-	-	+	-	-	-	+	-	+	+	-	+
Flexibility		+	+	+	+	-	-	-	-	-	-	+	+	+
Freshness		+	+	+	+	-	-	-	+	-	-	+	+	+
Data replication		-	-	-	-	+	+	+	+	+	+	+	+	+
Information retrieval		+	+	+	+	-	-	-	-	-	-	+	+	+
OLAP manipulation		-	-	-	-	+	+	+	+	+	+	-	-	-
Modeling language		ADT objects	R	XML	XML	OWL	R	R	R	R	R	-	RDF	O-R

ADT: Abstract Data Type; R: Relational model; OWL: Ontology Web Language; O-R: Object-Relational; RDF: Resource Description Framework; XML: eXtensible Markup Language proposed architecture is presented.

3 OUR PROPROSED ARCHITECTURE

We put forward a hybrid strategy named MEDWare (Mediation and Data Warehousing framework). We describe the main components of our framework which is divided into four layers (see Figure 3):

(a) Sensor network information sources: this layer consists of heterogeneous WSNs and other sources (e.g legacy systems).The information from the sensor nodes are gathered and can be accessed through a standard sensor gateway (a device uses for interfacing between the WSNs and a computer).

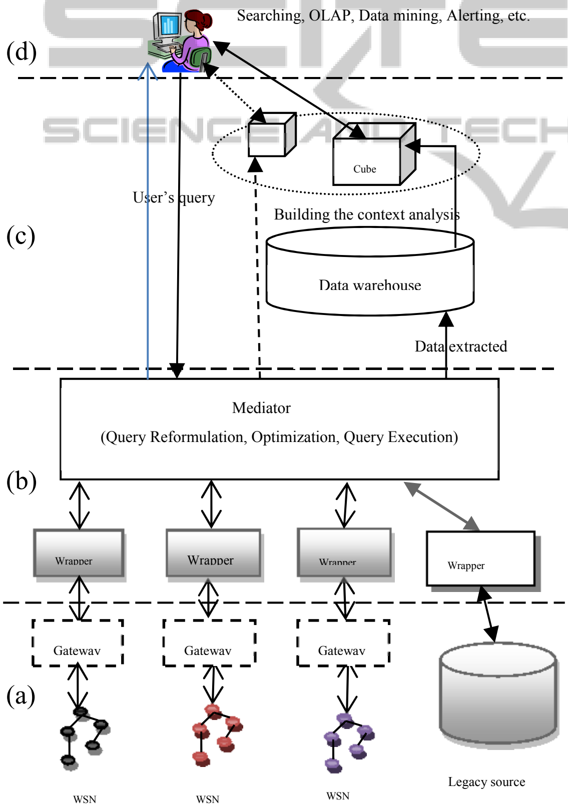

Figure 3: Our architecture based on a hybrid strategy.

(b) Mediation layer: consists of the mediator that maintains the global schema and mappings between the global and source schemas. The mediator also ensures query processing and solves semantic heterogeneity. Wrappers are also in that layer. They build a procedure to extract data from the sources and deal with syntactic heterogeneity.

(c) Data warehousing layer: this layer is composed of the data warehouse and multidimensional databases (OLAP data cubes) that can be built on-the-fly.

(d) Application layer: consists of different client applications used by the user to submit queries to the system for OLAP of sensor data, searching sensor information (e.g meta-data), or to receive notifications (e.g alerts).

The main advantages of this architecture are the following: (1) the development of a flexible decision support system (DSS) where new sensor networks or other sources can be easily added or removed from the system, (2) the freshness of data in the DSS, (3) the rapid building and composition of data cubes on-the-fly for better decision making, and (4) the ability to search information in heterogeneous sources.

The combination of the virtual and data warehousing approaches does not occur without problems. One can imagine the impact of such a combination on the performance analysis. Indeed, the proposed architecture generates many challenges to be addressed for future contributions such as:

- The query response time: when the number of supported users that interact with the system grows, it is necessary to deal with the issue of minimizing the query response time.
- Description of sensor data: there is a lack of adding semantics to sensor data in order to deal with semantic heterogeneity. Most of the time, the semantics of data is implicit.

4 CONCLUSIONS

In this paper, we proposed an architecture that integrates sensor information from multiple heterogeneous sensor networks and other sources for decision support systems.

Currently, we are working on formal description of the system, example scenarios and semantic data integration issue. Since then, many ontologies have been developed for describing sensor network information in the context of the semantic sensor web, sensor ontology integration is needed. The implementation is ongoing in order to validate the proposed architecture.

REFERENCES

Aggarwal, C. C., Ashish, N., Sheth, A. 2013. *The internet of things: a survey from the data-centric perspective*, in: Managing and Mining Sensor Data. Springer, pp. 383–428.

Ahmed, A., Ploennigs, J., Menzel, K., Cahill, B., 2010. Multi-dimensional building performance data management for continuous commissioning. *Advanced Engineering Informatics* 24(4), 466–475.

Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., and Cayirci, E. 2002. A survey on sensor networks. *Communications magazine*, *IEEE*, 40(8), pp. 102-114.

Casola, V., Gaglione, A., Mazzeo, A. 2009. *A reference architecture for sensor networks integration and management*, in: GeoSensor Networks. Springer, pp. 158–168.

Codd, E. F., Codd, S. B., and Salley, C. T., 1993. Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate. *Codd and Date*, 32.

Convey, C., Karpenko, O., and Tatbul, N., 2001. Data integration services. http://www.cs.brown.edu/people/atbul/cs227/chapter.pdf.

Da Costa, R., and Cugnasca, C. E. 2010. Use of data warehouse to manage data from wireless sensors networks that monitor pollinators. *In 11th International Conference on Mobile Data Management (MDM),* IEEE Computer Society, Missouri, USA. pp.402-406.

Favre, C., Bentayeb, F., Boussaid, O., Darmont, J., Gavin, G., Harbi, N., ... and Loudcher, S., 2013. Les entrepôts de données pour les nuls... ou pas!. *2ème Atelier Aide à la Décision à tous les Etages (EGC/AIDE 13),* Toulouse, France.

Gökçe, H. U., Gökçe, K. U., 2014. Multi-dimensional energy monitoring, analysis and optimization system for energy efficient building operations. *Sustainable Cities and Society*, 10, pp. 161–173.

Grosky, W. I., Kansal, A., Nath, S., Liu, J., Zhao, F., 2007. Senseweb: An infrastructure for shared sensing. *Multimedia, IEEE,* 14, pp. 8–13.

Gupta, A., and Mumick, I. S., 1995. Maintenance of materialized views: Problems, techniques, and applications. *IEEE Data Eng. Bull*., 18(2), pp. 3-18.

Halevy, A.Y. 2001. Answering queries using views: A survey. *The VLDB Journal*, 10(4), pp. 270-294.

Huang, V., Javed, M.K., 2008. Semantic sensor information description and processing, in: *Sensor Technologies and Applications*. SENSORCOMM'08. pp. 456–461.

Ibrahim, I. K., Kronsteiner, R., and Kotsis, G., 2005. *A semantic solution for data integration in mixed sensor networks*. Computer Communications, 28(13), pp. 1564-1574.

Inmon, W. H. 1996. *Building the data warehouse*. Wiley Publishing, Inc.

Kimball, R., and Caserta, J. 2004. *The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data*. Wiley Publishing, Inc.

Konstantinou, N., 2012. Converting raw sensor data to semantic web triples: a survey of implementation options. *Journal of Sensors, Wireless Communications and Control* 2.1: 44-52.

Lenzerini, M., 2002. Data integration: a theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems,* ACM, pp. 233-246.

Li, Y., Wang, L., Ji, L., Liao, C., 2013. A data warehouse architecture supporting energy management of intelligent electricity system, in: *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering***.** Atlantis Press, Paris, France.

Mathieu, J. 2011. *Intégration de données temps-réel issues de capteurs dans un entrepôt de données géo-décisionnels.* Thesis (MSc), Université de Laval. Canada.

Roantree, M., 2009. A hybrid storage model for web information systems, *in: Proceedings of the 6th International Workshop on Web Information Systems Modeling.* Amsterdam, The Netherlands.

Shah, N., Tsai, C. F., Marinov, M., Cooper, J., Vitliemov, P., and Chao, K. M., 2009. Ontological on-line analytical processing for integrating energy sensor data. *IETE Technical Review*, *26*(5), pp. 375.

Shokoh, K. 2010. *A hybrid approach to data integration.* Thesis (PhD). Université Joseph Fourier-Grenoble I. France.

Stocks, K. I., Condit, C., Qian, X., Brewin, P. E., and Gupta, A., 2009. Bringing together an ocean of information: an extensible data integration framework for biological oceanography. *Deep Sea Research Part II: Topical Studies in Oceanography*, *56*(19), pp. 1804-1811.

Widom, J. 1995. Research problems in data warehousing. *In Proceedings of the fourth international conference on Information and knowledge management*, USA, ACM, pp. 25-30.

Wiederhold, G. 1992. Mediators in the architecture of future information systems, *IEEE Computer*, 25(3), pp. 38–49.

Xu, L., and Embley, D. W., 2004. Combining the best of Global-as-View and Local-as-View for data integration. *In ISTA,* 48, pp. 123-136.

Yao, Y., and Gehrke, J., 2002. The cougar approach to in-network query processing in sensor networks. *ACM Sigmod Record*, *31*(3), pp. 9-18.

Zhou, G., Hull, R., King, R., and Franchitti, J. C. 1995. Data integration and warehousing using H2O. *IEEE Data Eng. Bull*., 18(2), pp. 29-40.

Ziegler, K., and Dittrich, R. 2004. *Three decades of Data Integration – All Problems solved?* Building the Information Society, Springer.