COMBINING SEMANTIC INFORMATION AND INFORMATION QUALITY ON THE ENRICHMENT OF WEB DATA INTEGRATION SYSTEMS

Damires Souza[1], Bernadette Farias Lóscio[2] and Ana Carolina Salgado[2]
[1]*Federal Institute of Education, Science and Technology of Paraiba (IFPB), João Pessoa, Paraíba, Brazil*
[2]*Center for Informatics, Federal University of Pernambuco (UFPE), Recife, Pernambuco, Brazil*

Keywords: Semantic Information, Information Quality, Web Data Integration Systems.

Abstract: The emergence of the Web and its permanent growth has caused a big impact on the database research community. Thereby, Database research areas have evolved in order to consider the new problems arising from the need of managing the huge volume of data available on the Web. One of such areas is Data Integration (DI), which is considered a pervasive challenge faced by applications that need to query across multiple autonomous and heterogeneous data sources. To help matters, we argue that semantic information like ontological and contextual information, combined with Information Quality (IQ) provided by IQ measures, may be employed together in order to enrich processes in DI (e.g., schema matching and query answering). In this paper, we present our ideas regarding what we mean by semantic information and IQ and why and how they may be combined in order to produce semantic knowledge to be used in Web Data Integration Systems. Furthermore, we propose a preliminary version of a metamodel, which presents a formal description of relationships between concepts associated with semantic information and IQ.

1 INTRODUCTION

The increasing use of the Web and the development of communication infrastructures have led to a demand for high-level integration of distributed, autonomous and heterogeneous data sources. In order to meet such demand, different kinds of Web data integration systems, including Mediation Systems (Halevy *et al.*, 2006), Peer Data Management Systems (Sung *et al.*, 2005), and Dataspaces (Hedeler *et al.*, 2009) have been proposed. In general, these solutions are characterized by an architecture mainly constituted by: i) a set of autonomous data sources ranging from traditional databases to semi-structured or non-structured data repositories, ii) an optional set of global schemas representing integrated views of the distributed data, and iii) a set of mappings, i.e. associations between data source elements as well as associations between data source elements and global schema elements. On the other hand, to offer a uniform view of heterogeneous and distributed data, a data integration system must provide solutions for several processes as, for example, query answering (Souza *et al.*, 2009; Stuckenschmidt *et al.*, 2005) and schema matching (Pires *et al.*, 2009; Giunchiglia *et al.*, 2004).

Despite a lot of research done in this area, Data Integration (DI) on the Web remains a challenging problem mainly due to the heterogeneous and autonomous nature of the data sources. Among the techniques employed to help to overcome these problems (Halevy *et al.*, 2006), the adoption of semantic knowledge has shown to be a helpful support to deal with this.

In this work, we consider that semantic knowledge may be produced by considering the combination of semantic information and information quality. In a general way, semantic information concerns the information that helps to assign meaning to elements (e.g., schema elements) or expressions (e.g., queries) that need to be interpreted in a given situation (Souza *et al.*, 2011; Mandreolli *et al.*, 2009). On the other hand, information quality (IQ) is a multidimensional aspect of information systems and it is based on a set of dimensions or criteria, which are used to assess and measure a specific IQ aspect (Batista and Salgado, 2007; Wang and Strong, 1996).

Souza D., Farias Lóscio B. and Salgado A..
COMBINING SEMANTIC INFORMATION AND INFORMATION QUALITY ON THE ENRICHMENT OF WEB DATA INTEGRATION SYSTEMS .
DOI: 10.5220/0003961602190224
In *Proceedings of the 8th International Conference on Web Information Systems and Technologies* (WEBIST-2012), pages 219-224
ISBN: 978-989-8565-08-2

Specifically, we are interested in semantic information provided by ontologies and context. Ontologies formally represent knowledge of a given domain through the definition of concepts, the relationships between them, axioms and individuals (Baader *et al*., 2003; Gruber, 1995). Context may be defined as a set of elements surrounding a domain entity of interest (e.g., user, query, or data source), which is relevant in a specific situation during some time interval (Bolchini *et al.*, 2009; Souza *et al.*, 2008; Dey 2001).

In the setting of DI solutions, ontologies may be used, for example, as a standard model to represent data sources metadata or as background knowledge to help solving heterogeneity problems. In a similar way, context may help to deal with information that can be acquired only on the fly (e.g., the availability of data sources), and which is perceived at run time.

We also argue that IQ assessment is fundamental for the improvement of data integration processes (Keeton *et al.*, 2009; Roth and Nauman, 2005). IQ evaluation may contribute to minimize the query answering time as well as to enhance the quality of query answers, for example.

In this work, we are mainly interested in how semantic information and IQ may be combined in a properly way in order to bring substantial gains for the overall DI processes. Particularly, we present our ideas with regards to the following issues:

- What we mean by semantic information and information quality.
- How semantic information and information quality may be combined in order to produce semantic knowledge.

Regarding the second issue, we propose a preliminary version of a metamodel, which presents a formal description of relationships between concepts associated with semantic information and IQ. This metamodel has been developed as an ontology, being compliant to OWL standard (Baader *et al.*, 2003).

This paper is organized as follows: Section 2 introduces the main concepts underlying Semantic Information and IQ; Section 3 discusses a motivating scenario in the light of the presented concepts; Section 4 presents the preliminary metamodel. Finally, Section 5 points out some considerations and highlights important topics for further research.

2 BACKGROUND CONCEPTS

Semantic information has been increasingly used as a means to enhance DI processes by assisting them to deal with the heterogeneous and autonomous nature of distributed data sources. Meanwhile, IQ has become a critical aspect of Information Systems research (Ge and Helfert, 2007) and, consequently, of DI systems (Duchateau and Bellahsene, 2010; Wang 2010). Some works (Yasar *et al.*, 2011; Helfert and Foley, 2009; Molina and Olsina, 2008) have discussed the use of semantic information together with information quality, but they do not present how these concepts may be combined to properly enrich an information system. Particularly, we argue that in order to enhance processes in DI by using IQ criteria, it is necessary to take into account the data semantics as well as the context around the process at hand. In this section, we provide an overview of the semantic information and IQ concepts to provide a better understanding of our proposal.

2.1 Semantic Information

As mentioned earlier, semantic information concerns the information that helps to assign meaning to elements or expressions that need to be interpreted in a given situation. Specifically, we are interested in semantic information described through ontologies or provided by context.

An ontology is a representation of a shared understanding of concepts in a particular domain of interest as agreed by a community (Gruber, 1995). The knowledge captured in ontologies can be used, among other things, to annotate data, generalize or specialize concepts, and infer entirely new (implicit) information (Baader *et al.*, 2003).

There has been a growing interest in using ontologies for solving data heterogeneity problems. In the DI setting, for example, ontologies have been used for some purposes, including (Xiao, 2006): (i) *metadata representation*: each data source is represented by a local ontology; (ii) *global conceptualization*: an ontology, called global ontology, may be employed to provide a conceptual view over the schematically heterogeneous source schemas; and (iii) *support for high-level queries*: given a global ontology, users can formulate queries without specific knowledge of the different data sources.

In addition, ontologies may also be used as a way of providing a domain reference. Considering a given knowledge domain, an agreement on its terminology can occur through the definition of a domain ontology, which can be used as a semantic reference or background knowledge to enhance processes such

as ontology matching (Pires *et al*., 2009).

Another kind of semantic information increasingly used is context. Usually, context is concerned with some specific situation, most of the times perceived as a set of variables that may be of interest for an agent (Bolchini *et al*., 2009). Context may also be understood as the circumstantial elements that make a situation unique and comprehensible (Dey, 2001). More abstractly, Vieira *et al*. (2007) makes a distinction between contextual element (CE) and context. The former is any piece of data or information that enables to characterize an entity in a domain. The latter is the set of instantiated contextual elements that are necessary to support a task at hand.

In works regarding data integration settings, context may be used to: (i) data tailoring, in order to define context-aware data views over large information systems (Bolchini *et al.*, 2009); (ii) schema reconciling, to identify in which context the elements occur and thus, to ease spell-check and schema-level sense disambiguation tasks (Belian *et al*., 2010) and (iii) query answering, where query results may be expanded with meaningful related answers according to the context acquired at query submission time (Souza *et al*., 2009).

2.2 Information Quality

The notion of Information Quality (IQ) has emerged during the past years and shows a steadily increasing interest (Duchateau and Bellahsene, 2010; Keeton *et al.*, 2009; Roth and Nauman, 2005). As mentioned earlier, IQ is a multidimensional aspect and it is based on a set of dimensions or criteria, which are used to assess and measure a specific IQ aspect. One of the best known quality dimensions classification is presented by Wang and Strong in (Wang and Strong, 1996). They have conceived one of the first sets of structured and classified quality dimensions that has been a strong reference for most of the studies in this area.

Regarding a DI system, there are some key points in which it is possible to consider IQ analysis, for instance: data source schema, global schema, data source selection, query processing, query routing, data integration and data materialization (Duchateau and Bellahsene, 2010; Wang, 2010; Keeton *et al.*, 2009; Batista and Salgado, 2007). Also, it is possible to enumerate several IQ criteria that can be associated with these DI components or processes (e.g., schema minimality, data source availability).

In the next section, we present a motivating example to better illustrate the use of semantic information and IQ on a DI scenario.

3 A MOTIVATING SCENARIO

Our motivating scenario regards a Web Data Integration System, which integrates data from a given domain (e.g., tourism, life science) distributed over a set of data sources related to each other by means of mappings. The data source schemas are represented by ontologies and the mappings between them are incrementally identified according to the queries posed to the system. Given this DI scenario, in the following, we discuss how semantic information and IQ may be used to enrich each one of the steps that compose the query answering process of the considered data integration system.

(1) *Query Submission*: whenever the user poses a query ***Q***, the current context of the user and the context of the data source where the query was posed must be acquired. For example, user preferences, user interface and data source's identification are relevant and should be gathered. Meanwhile, the query's context, by means of required information and operators, is also figured out.

(2) *Query Routing*: during this step, contextual information acquired at query submission time may be used to determine the most relevant data sources to answer ***Q***. Candidate data sources are also analysed according to their context (e.g., data model, location). At the same time, IQ metrics regarding the data sources reputation and access frequency may be taken into account. As a result, a set of data sources, called relevant data sources, are defined as the ones to send the query.

(3) *Schema Matching*: given the set of relevant data sources, an ontology matching is accomplished in order to identify mappings among the corresponding data source ontologies. To this end, a domain ontology is used as background knowledge to better identify the semantic mappings between the data source ontologies.

(4) *Query Reformulation*: during this step, the original query ***Q*** is reformulated into a set of queries to be submitted to the relevant data sources. This is done considering the mappings obtained during the *Schema Matching* step. IQ metrics such as the mappings confidence should be considered for the selection of the best mappings to be employed during the reformulation of ***Q***.

(4) *Query Execution and Answers Integration*: reformulated queries are submitted to the corresponding data sources. Next, the results are

returned to the data source where the query ***Q*** was originally posed and then they are integrated to obtain a single integrated answer.

(5) *Result Presentation*: during this last step, the integrated answer can be presented in various forms according to the user's preference, query interface and intended usage. At this moment, IQ metrics regarding accuracy and relevancy of the integrated answer may also be established.

In summary, this example aims to show some of the important usages of IQ and semantic information in order to enrich a DI query answering process:

(i) it enables the analysis of the user's query through its interpretation and identification of related entities and necessary operators on the fly;

(ii) it helps to identify the most relevant data sources that may contribute with answers to a given query, thus improving query answering results;

(iii) ontologies as background knowledge provide means to identify different kinds of mappings between pairs of data sources;

(iv) since the effects of collecting and integrating answers from various sources need to be handled, context may enrich the post-processing of the retrieved answers to adjust the final result representation according to the user preferences or intended level of detail and

(v) acquired contextual elements and IQ metrics may be stored in a knowledge base for later recovery, helping to identify trends in other future query answering processes.

4 TOWARDS A SEMANTIC KNOWLEDGE METAMODEL

A metamodel can be viewed as a model of a modeling language (Fuchs *et al.*, 2005) that defines the semantics for the main concepts that should be used to build other models. Thereby, the task of putting together the concepts of semantic information and information quality may be more easily accomplished by using a metamodel layer. Particularly, such metamodel should specify the constructs related to semantic information and IQ, as well as their relationships, providing a conceptual infrastructure to support the building of specific models.

Our metamodel has been developed as an ontology. Since Description Logics (DL) provides the formal semantics for specifying ontologies (Baader *et al.*, 2003), we may gain benefits in terms of expressiveness and reasoning mechanisms. Also, ontologies have been considered an interesting approach because they enable sharing and reusability (Souza *et al.*, 2008; Wang *et al.*, 2004).

In order to figure out the metamodel constructs, we followed a participatory and incremental design methodology. The ontology has been developed during a series of face-to-face meetings between experts who are concerned with issues related to semantic information usage and IQ in DI Systems. The proposed metamodel with its main constructs is presented in Figure 1 and explained as follows.

The main concepts underlying the metamodel are *Semantic_Information* and *Information_Quality*. Both are subconcepts of *Information*. The former concerns information provided by ontologies, defined here as *Ontological_Information*, and context, defined as *Contextual_Information*. The latter concerns information obtained through IQ metrics. Both information are supposed to be identified and used when associated with a specific *Situation*, which is composed by a set of *Processes*. A *Domain_Entity* is defined as anything in the real world that is relevant to describe the domain we are dealing with. *Contextual_Elements* are used to characterize a given domain entity. Besides, a *Measurement* is defined as a score value characterizing a particular IQ criteria.

In this sense, combining both semantic information and IQ in a given situation may lead to relevant *Semantic_Knowledge*. To this end, rules and axioms are being developed as a way to allow inference and consistency conditions check.

5 CONCLUSIONS

Due to the ever increasing complexity of Web Data Integration systems, the usage of semantic information and IQ is becoming more and more a necessity, instead of an optional requirement. These systems are highly dynamic and the semantic knowledge around their processes is rather relevant to produce results which best meet the users' needs. In this sense, this work presented some ideas regarding the benefits of combining semantic information and IQ in order to enrich the common processes of a DI system. Furthermore, it was proposed a preliminary metamodel, which aims to bring together the relationships between both concepts in order to allow inferring knowledge more properly.

Developed as an ontology, such metamodel will be the basis for the development of other models. Furthermore, it will provide the ability of reasoning over the information.

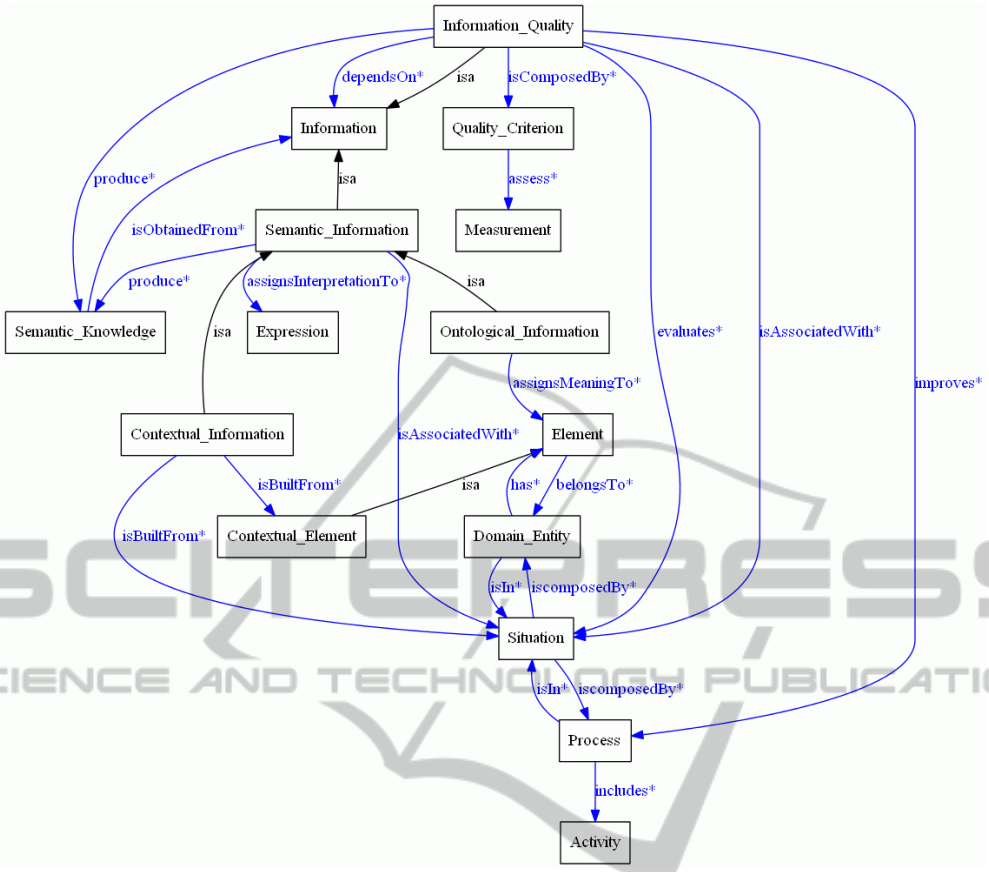

Figure 1: Semantic Knowledge Metamodel.

As further work, we will completely formalize the metamodel. We also plan to support developers with a framework that may provide the combined use of semantic information and IQ.

REFERENCES

Baader F., Calvanese D., McGuinness D., Nardi D., Patel-Schneider P. editors., 2003. The Description Logic Handbook: Theory, Implementation and Applications. *Cambridge University Press*.

Batista, M. C., Salgado, A.C., 2007. Data Integration Schema Analysis: An Approach with Information Quality. *In Proceedings of the 12th International Conference on Information Quality (ICIQ)*, MIT, Massachusetts, USA, October 2007.

Belian, R., Salgado, A. C., 2010. A Context-based Schema Integration Process Applied to Healthcare Data Sources. In *Proceedings of the International Conference On the move to meaningful internet systems*, Springer-Verlag.

Bolchini, C., Curino, C., Orsi, G., Quintarelli, E., Rossato, R., Schreiber, F., Tanca, L., 2009. And what can context do for data? In: *Communication of the ACM*, Volume 52 (11), pp. 136-140.

Dey, A., 2001. Understanding and Using Context. *Personal and Ubiquitous Computing Journal*, Volume 5, pp. 4-7. .

Duchateau, F., Bellahsene Z., 2010. Measuring the Quality of an Integrated Schema. In *Conceptual Modeling – ER* 2010, Lecture Notes in Computer Science.

Fuchs, F., Hochstatter, I., Krause, M., Berger, M., 2005. A Metamodel Approach to Context Information. In: *PerCom Workshops 2005*, 2005, pp. 8-14, Kauai Island, HI.

Ge, M., Helfert, M., 2007. A Review of Information Quality Research - Develop a Research Agenda. *In Proceedings of the 12*th *International Conference on Information Quality (ICIQ)*, MIT, Massachusetts, USA November 2007.

Giunchiglia, F., Shvaiko, P., Yatskevich, M., 2004. S-match: an algorithm and an implementation of semantic matching. In: *European Semantic Web Symposium (ESWC)*. pp. 61-75.

Gruber, T., 1995. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43:907-928.

Halevy, A., Rajaraman, A., Ordille, J., 2006. Data Integration: the Teenage Years, *In Proceedings. of the 25th International Conference on Very Large Data Bases (VLDB)*, pages 9-16, Seoul, Korea, September 2006.

Hedeler, C., Belhajjame, K., Fernandes, A.A.A., Embury, S.M., Paton, N.W., 2009. Dimensions of Databases, In *Proc. of 26th British National Conference on Databases*, Birmingham, UK, pages 55-66.

Helfert, M., Foley, O., 2009. A Context Aware Information Quality Framework. In *Proceedings of the 4th International Conference on Cooperation and Promotion of Information Resources in Science and Technology (COINFO'09)*, November 21-23, Beijing, IEEE Computer Society Press, pp.: 187-193.

Keeton, K., Mehra, P., Wilkes, J., 2009. Do You Know Your IQ? : A Research Agenda for Information Quality in Systems. *ACM SIGMETRICS Performance Evaluation Review*, Vol. 37, Issue 3, December 2009.

Mandreoli, F., Martoglia, R., Villani, G., Penzo, W., 2009. Flexible query answering on graph-modeled data. In: *12th International Conference on Extending Database Technology (EDBT'09)*, Saint-Petersburg, Russia, pp. 216-227.

Molina, H., Olsina, L., 2008. Assessing Web Applications Consistently: A Context Information Approach. In *Proceedings of ICWE'2008*. pp.224-230.

Pires, C. E., Souza, D., Pachêco, T., Salgado, A. C., 2009. A Semantic-based Ontology Matching Process for PDMS. In: *2nd International Conference on Data Management in Grid and P2P Systems (Globe'09*), Linz, Austria, pp. 124-135.

Roth, A., Naumann, F., 2005. Benefit and Cost of Query Answering in PDMS. In *Proceedings of the Int. Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P)*, 2005.

Souza D., Arruda T., Salgado A. C., Tedesco P., Kedad, Z., 2009. Using Semantics to Enhance Query Reformulation in Dynamic Environments. In: *Proceedings of the 13th East European Conference on Advances in Databases and Information Systems (ADBIS'09)*, Riga, Latvia, pp. 78-92.

Souza D., Pires, C. E., Kedad, Z., Tedesco, P., Salgado, A.C., 2011. A Semantic-based Approach for Data Management in a P2P System. In *LNCS Transactions on Large-Scale Data- and Knowledge-Centered Systems*, 2011.

Souza, D., Belian, R., Salgado, A. C., Tedesco, P., 2008. Towards a Context Ontology to Enhance Data Integration Processes. In: *4th Workshop on Ontologies-based Techniques for DataBases in Information Systems and Knowledge Systems (ODBIS)*, Auckland, New Zealand, pp.49-56.

Stuckenschmidt, H., Giunchiglia, F., van Harmelen, F., 2005. Query processing in ontology-based peer-to-peer systems. In *V. Tamma, S. Craneeld, T. Finin, and S. Willmott*, editors, Ontologies for Agents: Theory and Experiences. Birkhuser.

Sung, L., Ahmed, N., Blanco, R., Li, H, Soliman, M. A., Hadaller, D., 2005. A Survey of Data Management in Peer-to-Peer Systems. *School of Computer Science*, University of Waterloo, 2005.

Vieira, V., Tedesco, P., Salgado, A.C., Brézillon P., 2007. Investigating the Specifics of Contextual Elements Management: The CEManTIKA Approach. *The Sixth International and Interdisciplinary Conference on Modeling and Using Context*. B. Kokinov et al. (Eds.): LNAI 4635, Springer-Verlag, pp. 493–506.

Wang, J.A., 2010. Quality Framework for Data Integration. In *Proceedings of the 27th British National Conference on Databases (BNCOD)*.

Wang, R., Strong, D., 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, Vol. 12, N. 4, pages 5-33, 1996.

Wang, X., H., Gu, T., Zhang, D. Q., Pung, H. K., 2004. Ontology based context modelling and reasoning using OWL. In: *Proceedings of the 1st Workshop on Context Modeling and Reasoning*, 2004, Orlando, Florida.

Xiao, H., 2006. Query processing for heterogeneous data integration using ontologies. *PhD Thesis in Computer Science*. University of Illinois at Chicago.

Yasar, A., Paridel, K., Preuveneers, D.,Berbers, Y., 2011. When efficiency matters: Towards quality of context-aware peers for adaptive communication in VANETs. *2011 IEEE Intelligent Vehicles Symposium (IV)* (June 2011), pg. 1006-1012.