Enriching Query Routing Processes in PDMS with Semantic Information and Information Quality

Crishane Freire[2], Damires Souza[1], Bernadette F. 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
damires@ifpb.edu.br, {caf4, bfl, acs}@cin.ufpe.br

Keywords: Semantic Information, Information Quality, Query Routing, Peer Data Management Systems

Abstract. Query answering has been addressed as a key issue in dynamic distributed environments such as Peer Data Management Systems (PDMS). An important step in this process regards query routing, i.e., how to find peers (data sources) that are most likely to provide matching results according to the semantics of a submitted query. To help matters, we argue that semantic information like contextual information, combined with Information Quality (IQ) provided by IQ measures, may be employed together to enrich query routing processes. In this work, we propose an instantiation of a metamodel which combines both concepts as a means to produce semantic knowledge to be used in query routing processes. We also present an example of such instantiation.

1 Introduction

Peer Data Management Systems (PDMS) are P2P applications, which provide data sharing and query answering capabilities considering that data sources are organized in a network of peers [12; 6]. In a PDMS, data sources (peers) are connected with each other through a set of semantic mappings in such a way that peers directly connected are called semantic neighbors. In this light, query answering in a PDMS means to provide capabilities of answering a query considering that such query is submitted over one of the peers and there is a set of mappings between the peer and their neighbors.

A key issue in query answering in PDMS regards query routing, i.e., the process of identifying the most relevant peers among the ones available in the network according to a given submitted query. This process is not easy due to the large number of peers, the dynamic setting and the heterogeneity of the sources that compose the system. During query routing, some conditions such as peers' unavailability or even a low-degree history of answers are important criteria that may be considered in the peer selection or the estimated routing paths.

Some works have been proposed to improve the query routing process by using semantic information in peers' clustering [8] and quality criteria (e.g., completeness of mappings) to select relevant peers [6; 10]. In our work, we propose the use of semantic information, combined with Information Quality (IQ) provided by IQ measures, in order to enrich query routing processes. We argue that the use of such semantic information may reduce the query search space by considering only peers that may contribute with relevant answers, i.e., answers that match the semantics of the submitted query as well as the user preferences. Specifically, we propose a model, which combines semantic information and IQ as a means to produce semantic knowledge to be used in PDMS query routing processes. Such model is based on a metamodel defined in Souza *et al.* [11].

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 [11; 8]. On the other hand, information quality (IQ) is a multidimensional aspect of information systems and it is based on a set of criteria, which are used to assess a specific IQ aspect [2, 7; 14]. In our approach, we are interested in semantic information provided by ontologies [1; 5] and context [13; 3].

This paper is organized as follows. Section 2 introduces the semantic knowledge metamodel. Section 3 proposes a model to enrich query routing processes. Section 4 points out some considerations and highlights important topics for further research.

2 Semantic Knowledge Metamodel

A metamodel can be viewed as a model of a modeling language [4] that defines the semantics for the main concepts that should be used to build other models. In order to provide means to build specific models that combine IQ with semantic information, we have built a metamodel, described in Souza *et al.* [11]. The metamodel has been developed as an ontology. The reasons underlying this choice are: (i) ontologies are formalized by means of Description Logics (DL) which provides expressiveness and reasoning mechanisms [1] and; (ii) ontologies may enable sharing and reusability [12].

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 concepts (*Semantic_Information* and *Information_Quality*) 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 *Measure* 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 *Se-*

mantic_Knowledge. To this end, rules and axioms are being developed as a way to allow inference and consistency conditions check.

In the next section, we instantiate this metamodel by building a model which aims to deal with PDMS query routing issues.

3 A Model to Enrich Query Routing Processes

According to the metamodel described in Section 2, we have been working in a model layer which aims to deal with semantic information and IQ as a means to enrich PDMS query routing processes. To this end, we have particularly considered contextual information (e.g., peer availability, user preferences) and IQ criteria related to peers (e.g., reputation, access frequency, reliability). The idea is to use such information and the produced knowledge to assist the task of selecting a subset of candidate peers to route a query to. As a result, a set of peers, called relevant peers, are defined as the ones to send the query. We present the main concepts of the model in Figure 1. Such model is indeed a conjunction of pertinent entities (in this case, *peers* and *queries*) and the contextual and IQ elements related to them.

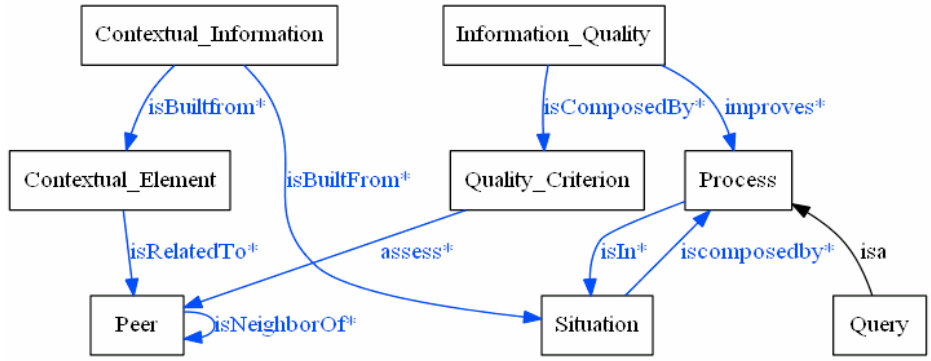

Fig. 1. Overview of the Model Main Concepts

As an illustration, consider a PDMS scenario composed by peers which integrates data from the Education knowledge domain. In this setting, we use ontologies as uniform representation of peer schemas. Since peers are grouped within the same knowledge domain, we use a domain ontology as background knowledge to identify the set of correspondences (i.e., mappings) between pairs of neighbor peers ontologies (schemas) [9]. Queries are submitted using SPARQL[1] language and are reformulated considering the set of such existing semantic correspondences between source and target neighbor peers.

Particularly, we consider three peers, named P1, P3 and P4, which are semantic neighbors, as depicted in Figure 2. Our goal is to route a query Q that has been submitted at P1 to each semantic neighbor (P3 and P4). Figure 2 also shows query Q1

[1] http://www.w3.org/TR/rdf-sparql-query/

submitted at P1 and query Q13 (reformulated from original query Q1) that is forwarded to P3. In this example, P4 is unavailable at query submission time, thus it is out of the relevant peers list. Since P3 is available and its reputation is under a defined threshold, it is chosen to receive the reformulated query. Therefore Q1 is submitted in P1 and routed to P3, according to the availability information (as an example of contextual element) and to the reputation measure (as an example of IQ criterion).

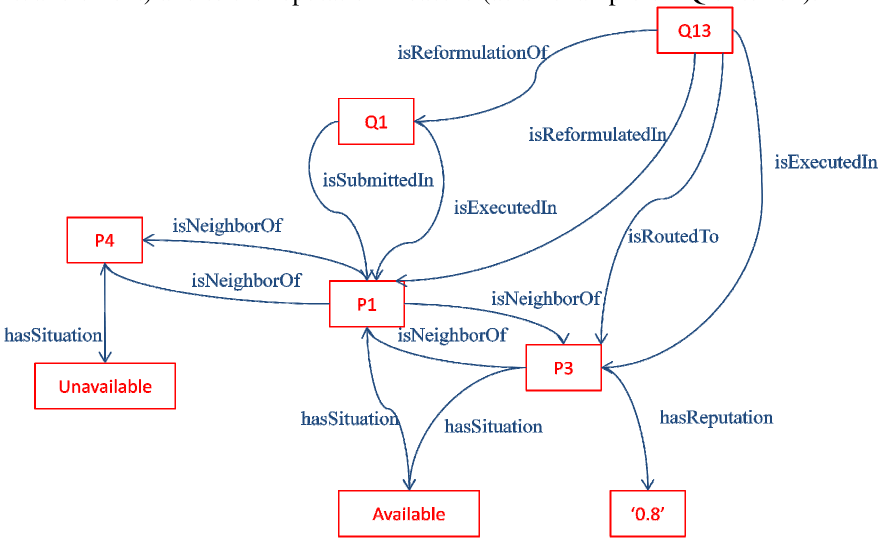

Fig. 2. Illustration of a Query Routing Process Using the Proposed Model

For the sake of space, this illustration is quite simple. However, it aims to show some of the important usages of IQ and semantic information in order to enrich a query routing process: (i) it helps to identify the most relevant data sources (peers) that may contribute with answers to a given query; (ii) it allows to minimize network traffic from reducing exchanged messages among peers; and (iii) acquired contextual elements and IQ metrics may be stored in a knowledge base for later access, helping to identify trends in other future query routing processes.

With respect to the integration of query results obtained from the diverse peers and some quality problems resulting from the processes of reconciling and merging data are out of the scope of this work. They are subject of other researches that are being developed in our group.

4 Conclusions

Due to the complexity of PDMS query routing, 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 this process is rather important to select the most relevant peers to send a query and produce results, which best meet the users' needs. In this sense, this work has presented an instantiation of the metamodel proposed by Souza *et al.* [11] considering query routing processes. We have also presented an example to show the benefits of combining semantic information and IQ measures.

As further work, we plan to refine the model as well as to define rules/axioms for the inference of semantic knowledge. This model will provide the combined use of semantic information and IQ and will be used in query routing process of a PDMS.

5 References

1. Baader F., Calvanese D., McGuinness D., Nardi D., Patel-Schneider P. editors.: The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Pressb (2003).
2. Batista, M. C., Salgado, A.C.,: Data Integration Schema Analysis: An Approach with Information Quality. In: 12th International Conference on Information Quality (ICIQ), MIT, Massachusetts, USA (2007).
3. Dey, A.,: Understanding and Using Context. Personal and Ubiquitous Computing Journal, vol. 5, pp. 4-7 (2001).
4. Fuchs, F., Hochstatter, I., Krause, M., Berger, M.,: A Metamodel Approach to Context Information. In: PerCom Workshops 2005, pp. 8-14, Kauai Island, HI (2005).
5. Gruber, T.:. Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, 43,907-928 (1995).
6. Kantere V., Tsoumakos D., Sellis T., Roussopoulos N.: GrouPeer: Dynamic Clustering of P2P Databases. Information Systems Journal, vol. 34, n. 1, pp. 62–86 (2009).
7. Keeton, K., Mehra, P., Wilkes, J.,: Do You Know Your IQ?: A Research Agenda for Information Quality in Systems. ACM SIGMETRICS Performance Evaluation Review, vol. 37, Issue 3, (2010).
8. Montanelli S., Bianchini D., Aiello C., Baldoni R., Bolchini C., Bonomi S., Castano S., Catarci T., Antonellis V., Ferrara A., Melchiori M., Quintarelli E., Scannapieco M., Schreiber F., Tanca L.: The ESTEEM platform: enabling P2P semantic collaboration through emerging collective knowledge, Journal of Intelligent Information Systems. vol. 36 (2) (2010).
9. Pires, C. E. S., Souza, D., Pachêco, T., and Salgado, A. C.: A Semantic-based Ontology Matching Process for PDMS. To appear in 2nd International Conference on Data Management in Grid and P2P Systems (Globe'09), Linz, Austria (2009).
10. Roth, A., Naumann, F.,: 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).
11. Souza D, Lóscio B. F., Salgado A. C.: Combining Semantic Information and Information Quality on the Enrichment of web Data. In: 8th International Conference on Web Information System and Technologies. (WEBIST), Porto, Portugal (2012).
12. Souza D., Pires, C. E., Kedad, Z., Tedesco, P., Salgado, A.C.: A Semantic-based Approach for Data Management in a P2P System. In LNCS Transactions on Large-Scale Data- and Knowledge-Centered Systems, (2011).
13. Tanca L., Bolchini C., Quintarelli E., Schreiber F. A., Milano P., Orsi, G.: Problems and Opportunities in Context-Based Personalization. In Proc. of the VLDB Endowment, vol 4, n 11, pp.10-13, (2011).
14. Wang, R., Strong, D.,: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, vol. 12, N. 4, pp. 5-33, (1996).