Semantic-based Query Routing for PDMS

Crishane Freire[1,2], Damires Souza[2], Ana Carolina Salgado[1]

[1]Center for Informatics, Federal University of Pernambuco (UFPE) - Recife, Pernambuco - Brazil

[2]Federal Institute of Education, Science and Technology of Paraiba (IFPB) - João Pessoa, Paraíba - Brazil

{crishane, damires}@ifpb.edu.br, {caf4, acs}@cin.ufpe.br

Nível: Doutorado
Ano de ingresso no programa: Agosto de 2009
Exame de qualificação: Abril de 2012
Época esperada de conclusão: Julho de 2013

Abstract. *Query routing is a key issue in dynamic distributed environments such as Peer Data Management Systems (PDMS). The dynamicity of the environment and the amount of heterogeneous and autonomous data sources available in the system have made hard the task of finding relevant results to user queries. We argue that 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. To achieve and deal with such knowledge, we combine two important aspects: semantic information and information quality. In this light, we present a semantic-based query routing approach for PDMS and highlight important issues related to this problem.*

Keywords
Query Routing, Semantic Information, Information Quality, Peer Data Management System.

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. This fact caused the appearance of diverse distributed environments such as Peer Data Management Systems (PDMS) [Souza *et al.* 2011; Kantere *et al.* 2009]. 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 each one of its neighbors.

A key issue in query answering in PDMS regards query routing. Query routing is defined as the process of identifying the most relevant peers among the ones available in the network that are most likely to provide matching results according to the semantics of a 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 poor history of answers are important criteria that may be considered in the peer selection or in the estimated routing paths.

In our work, we argue that the semantic knowledge produced by combining both semantic information and Information Quality (IQ) can be used in order to improve the query routing process. The idea is that the obtained semantic knowledge 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.

This paper is organized as follows. Section 2 introduces some important concepts related to this work; Section 3 presents our main contributions. The related works are discussed in Section 4, and Section 5 describes the current stage of the work.

2. Fundamental Concepts

This section introduces some concepts underlying this work, particularly PDMS and the query routing problem, semantic information and IQ.

2.1. PDMS and the Query Routing Problem

In PDMS, queries submitted at a peer are answered with data residing at that peer and with data that is reached through mappings that are propagated over the network of neighbor peers. Therefore, the query routing problem in PDMS occurs every time a peer receives a query and has to decide, based on its local knowledge, to which of its semantic neighbors it should forward the query. To avoid flooding, i.e., the query propagation among the entire network of peers, it is important to develop a process that may select relevant peers based on the circumstances that surround the process on the fly. The circumstances regard the entities around the activities that compose the process, for example, the submitted query and its semantics, the available peers, the user preferences, the existing mappings among the peer schemas, and all the factors that can be acquired at each step of the process.

2.2. Semantic Information

In general, semantics is the study of meanings of the message underlying the words or underlying certain elements that need to be interpreted in a given task or situation [Souza *et al.* 2011]. Due to the heterogeneity and to the dynamicity of a PDMS setting, the use of semantics in the form of ontologies and context has proven to be helpful in tasks such as query answering and peer clustering. In this work we are mainly interested in semantic information provided by context.

We define *Context* as a set of elements surrounding a domain entity of interest (e.g., user, query, and peer) which is considered relevant in a specific circumstance during some time interval [Souza *et al*. 2011; Vieira *et al.* 2010]. Vieira *et al*. (2011) makes a distinction between the concepts of 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 an activity at hand.

Contextual elements may improve the semantic interpretation of an entity by restricting or modifying the meaning of an element according to a circumstance [Souza *et al.* 2009]. Regarding query routing, the contextual information is related to any information that may influence the process activities such as query execution, peers selection, query reformulation, query forwarding and query results presentation.

2.3. Information Quality

The notion of IQ has emerged during the past years and shows a steadily increasing interest [Duchateau and Bellahsene 2010; Roth and Naumman 2007]. IQ is usually characterized by multiple dimensions or criteria, where each one captures a high-level aspect of quality. The role of each one is to assess and measure a specific IQ criterion [Wang and Strong 1996]. For our purposes, we will use the general definition of IQ – 'fitness for use' - which encompasses the aspects of quality.

In dynamic distributed environments such as PDMS, IQ has received significant attention in the literature over the past decade. There are two major reasons for that (a) the phenomenal growth of information sources available for query, and (b) the highly accessible nature of this information by a diverse set of users [Arazy and Kopak 2011]. PDMSs are vulnerable to poor IQ in some aspects such as [Herschel and Heese 2005; Roth and Naumman 2007]: peer (data source), peer schema, mappings, data and query answers. In fact, considering a PDMS, quality criteria can give trust to the system and enhance its processes. In query routing process, for example, peer quality measures such as relevancy and reputation can be used as a choice parameter to select the best peers to forward a given query.

3. Contributions

In this section, we present the main contributions of this work regarding a query routing process and a model to represent information quality and contextual information in order to improve relevant peers' selection.

3.1. A Semantic-based Query Routing Process

The goal of the process is to select the best set of peers that are able to answer a submitted query. Thus, during the process, each peer that receives a query accomplishes the following activities:

- *Query Execution* - executes the query locally and stores the query answer in a result list maintained by itself;
- *Peer Selection* - identifies candidate peers from its semantic neighbors and selects relevant peers (i.e., target peers) from the set of candidate peers based on contextual information of domain entities (user, query and peer);
- *Query Reformulation* - reformulates the query to each relevant peer considering the target peer schema and the acquired contextual information [Souza *et al.* 2009];
- *Results Integration* – integrates the result list received from its neighbor peers;
- *Query Forwarding* - forwards the query to the chosen target peers preserving the contextual information acquired in the process.

A TTL (Time-To-Live) mechanism based on time unit and semantic information will be defined to limit the query routing process. When the TTL is reached, the result list maintained by the current peer is integrated and its result is routed back, following the reverse path of the received query.

In our approach, the semantic information and IQ are used in order to make the process decisions more specific. Thus, in each process activity, the circumstances that surround the system entities (e.g., peers) are analyzed in two perspectives: *context* and *IQ*. Three types of context are considered: (i) *the user context*, provided by his profile and defined preferences; (ii) *the query context*, acquired through its semantic analysis and (iii) *the peer context*, identified by peer availability and the associated quality criteria.

Regarding IQ, there are some criteria which have been considered and specified as follows: (i) *Reputation*: concerns the degree to which the information of a source is in high standing. In our approach the reputation criterion can be assessed by calculating the percentage of queries answered and queries not answered by a peer in a given time interval; (ii) *Relevance*: refers to the suitability of data to queries submitted by users. Usually, this criterion is subjective and user-dependent, since only the user can determine whether something is relevant or not. Mandreoli *et al.* (2009) considers relevance as a measure of semantic similarity between the query concepts and the concepts existing in target peer schema. In our work, we extend this definition, allowing the user to set weights (importance scores) to each concept that is being queried in a given query; (iii) *Query Degradation*: we define the concept of query degradation by extending the concept of semantic loss presented in Delveroudis and Lekeas (2007). The query degradation criterion is a metric obtained by the product of the percentage of concepts that may not be lost in a query reformulation process of query Q from peer P_i to peer P_j and the mean of similarity scores between the concepts queried in Q and the concepts present in a target peer P_j. In the routing process, for each candidate peer P_m to forward the query (i.e., neighbor of the peer that receives a query), a global quality score ($Global\_IQ(P_m)$) is calculated taking into account the defined quality criteria. A quality threshold will be defined to avoid query forwarding to peers that have a *Global_IQ* value below the threshold.

To help matters, Figure 1 depicts a query routing example scenario in a PDMS composed by six peers P1, P2, P3, P4, P5 and P6. In this scenario, suppose that a query Q has been submitted at peer P1 and that the system quality threshold value is 7.0. After receiving the query, P1 executes it and must select the relevant peers to send the query based on the circumstance that surrounds the system entities at runtime. Thus, the context (of user, query, peer) and IQ of neighbor peers are analyzed. Considering that P3 is unavailable and the relevance values of P2 and P4 are above the quality threshold, the query Q is reformulated (QR_{12}, QR_{14}) and forwarded only to peers P2 and P4. In the same way, after receiving the query, P4 executes the query and, based on the circumstance, decides to forward the reformulated query (QR_{46}) only to P6 because P5 has a relevance value below the quality threshold. At each peer that receives a query, the query answers (QA_{ij}) are integrated and routed back to the peer that sent the query. This activity will be repeated until reaching the peer that originated the query.

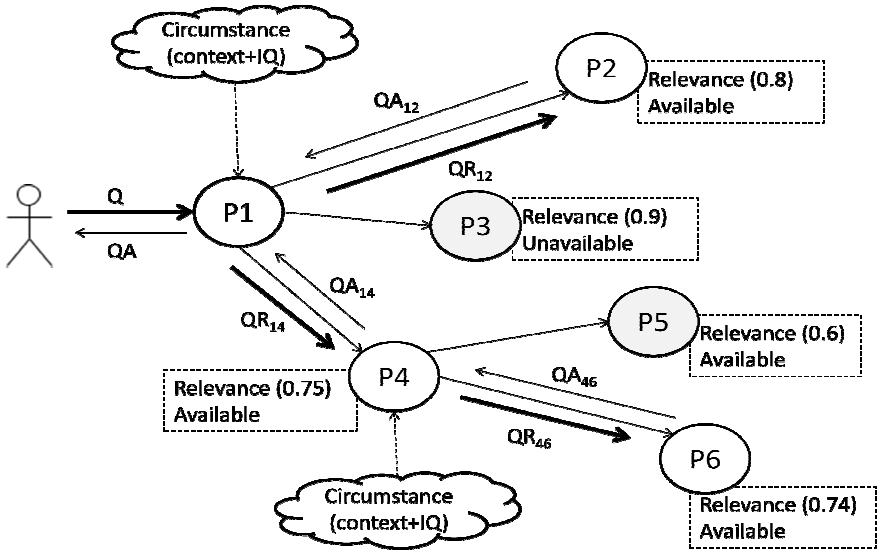

Figure 1: An Example Scenario for Query Routing

3.2. A Model to Represent Semantic Information and Information Quality

In order to allow semantic information and IQ usage, it is important to define how these related concepts are represented and (possibly) persisted. For this purpose, we have used the metamodel presented in Souza *et al.* (2012). Such metamodel has been developed as a way to provide constructors that can combine semantic information (e.g., ontological and contextual information), and Information Quality (IQ) provided by IQ measurements. By combining such concepts, it aims to produce semantic knowledge to be used in data integration settings. The defined meta-constructors (i.e., meta-concepts) can be reused in other models for specific purposes. In our proposal, we have generated a model for the query routing process based on such metamodel constructors.

Figure 2 depicts the model with its main concepts and the relationships among them. The concepts and relationships which are identified by the prefix *meta* belong to the metamodel and are being reused in this model.

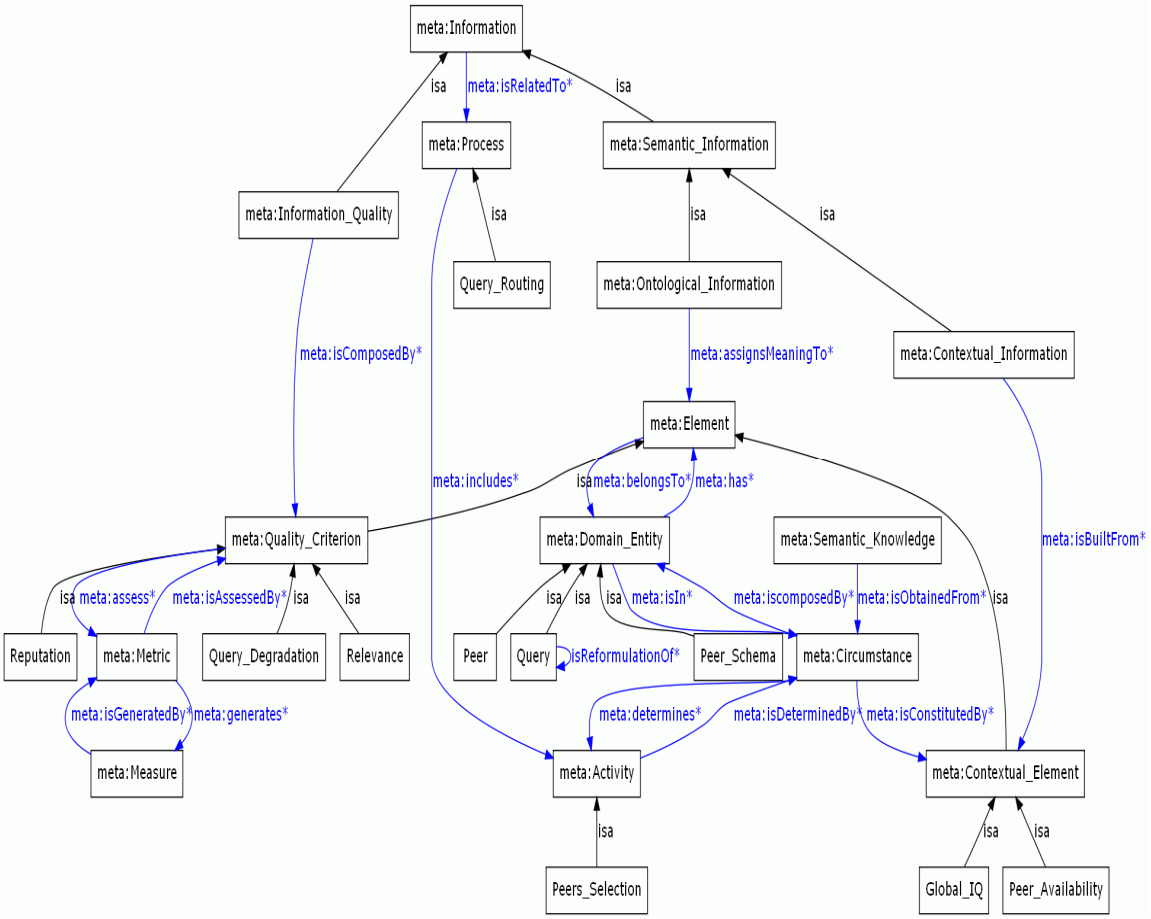

Figure 2. Model to Query Routing Process

The main concepts underlying the model that are reused from the metamodel are *Semantic_Information* and *Information_Quality*, subconcepts of *Information*. In this work, the former concerns information provided by *Contextual_Information*. The latter concerns information obtained through IQ metrics. *Information_Quality* is composed by a set of criteria (*Quality_Criterion)* such as *Relevance*, *Query_Degradation* and *Reputation*. The concept *Measure* concerns the values provided by quality criteria metrics (*Metric)*. A *domain entity* is anything in the real world that is relevant to describe the domain of interest. In our process, we determined three domain entities: *peer*, *query* and *user*. Also, we defined which *Process* would be the most important in our setting. The *query routing* process has been defined as the one of our interest. Query execution, peer selection, query reformulation, results integration and query forwarding are activities belonging to the query routing process. For the sake of simplicity, only the *peer selection* activity is represented in the model. *Element* is used to characterize a *Domain_Entity*. Quality measures and contextual elements are types of element. Elements under a given circumstance which are considered as relevant become contextual elements. *Peer_Availability* and *Global_IQ* represent contextual elements that are acquired at runtime. *Peer_Availability* indicates if the peer is available to receive a query and *Global_IQ* represents the final IQ score obtained from peer quality criteria assessment. *Semantic_Knowledge* concerns the knowledge obtained from the contextual elements and domain entities that compose the circumstance of an instantiated activity at hand.

4. Related Work

Despite the fact that there are a lot of researches showing the importance of IQ for the improvement of query answering in PDMS [Green *et al.* 2010, Heese *et al.* 2005], only few works discusses the use of IQ aspects and contextual information specifically in query routing.

Zhuge *et al.* (2005) deals with data inconsistency proposing a Quality of Peers (QoP) method. They also propose methods based on the notion of routing graphs for estimating query completeness. System P [Roth and Naumann 2007] provides a completeness-driven query planning. Its objective is to forward queries by considering peers and mappings that promise large result sets and mappings with low information loss. Herschel and Heese (2005) use a PDMS architecture [Heese *et al.* 2005], which extends the classical PDMS along three dimensions (Quality, Web and Semantics), enabling more efficient lookup of information sources and improving query routing. Other works offer a dynamic approach for clustering peers in semantic groups by using IQ [Löser *et al.* 2003] and IQ and contextual information [Montanelli *et al.* 2010] in order to create a search space for query.

Different from the referred works, our approach defines a set of quality criteria (reputation, relevance, query degradation) and contextual information (e.g., user preferences, peer availability) to be used in a combined way to enhance query routing processes. It also presents a model to represent these concepts that will be used in the approach development.

5. Current Stage of the Work

At this moment, we are working on the formalization of the specified query routing process. A quality threshold will be defined to be used as a reference during the selection of relevant neighbor peers. The model defined to represent IQ and semantic information will be refined by specifying some rules/axioms for the inference of the semantic knowledge. To evaluate our proposal some experiments will be accomplished in a semantic-based PDMS, named SPEED [Pires *et al.* 2009].

References

Arazy, O. and Kopak, R. (2011) “On the Measurability of Information Quality”. In Journal of the American Society for Information Science, v. 62, n.1, p. 89-99.

Delveroudis, Y. and Lekeas, P. V. (2007) “Managing Semantic Loss during Query Reformulation in PDMS”. In SWOD IEEE, p.51-53.

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

Green, J., Ives, Z.G. and Tannen, V. (2010) “Provenance in ORCHESTRA”. In IEEE Data Eng. Bull, v.33, n.3, p. 9-16.

Heese, R., Herschel, S., Naumann, F. and Roth, A. (2005) “Self-extending Peer Data Management”. In Proceedings of the Datenbanksysteme in Business, Technologie und Web, v.65 of LNI.

Herschel, S. and Heese R. (2005) “Humboldt Discoverer: A Semantic P2P index for PDMS”. In Proceedings of the International Workshop Data Integration and the Semantic Web. Porto, Portugal.

Kantere, V., Tsoumakos D., Sellis T. and Roussopoulos N. (2009) “GrouPeer: Dynamic Clustering of P2P Databases”. Information Systems Journal, v. 34, n. 1, p. 62–86.

Löser, A., Wolpers, M., Siberski, W. and Nejdl, W. (2003) “Semantic Overlay Clusters within Super-Peer Networks”. In Proceedings of P2PDBIS. Berlin, Germany.

Mandreoli, F., Martoglia, R., Penzo, W. and Sassatelli, S. (2009) “Data-sharing P2P Networks with Semantic Approximation Capabilities”. In IEEE Internet Computing, p. 60-70.

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. and Tanca, L. (2010) “The ESTEEM platform: enabling P2P semantic collaboration through emerging collective knowledge”. In Journal of Intelligent Information Systems. v. 36, n. 2.

Pires, C. E., Souza, D., Kedad, Z., Bouzeghoub, M., and Salgado, A. C. (2009) “Using Semantics in Peer Data Management Systems”. In Colloquium of Computation: Brazil/INRIA, Cooperations, Advances and Challenges (Colibri’09), Bento Gonçalves, Brazil.

Roth, A. and Naumann, F. (2007) “System P: Completeness-driven Query Answering in Peer Data Management Systems”. In BTW, p. 1-4, Aachen, Germany.

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

Souza, D., Pires, C. E., Kedad, Z., Tedesco, P. and 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.

Souza, D, Lóscio, B. F. and Salgado, A. C. (2012) “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 .

Tanca, L., Bolchini, C., Quintarelli, E., Schreiber, F. A., Milano, P. and Orsi, G. (2011) “Problems and Opportunities in Context-Based Personalization”. In Proc. of the VLDB Endowment, v. 4, n. 11, p.10-13.

Vieira, V., Tedesco, P. A. and Salgado, A. C. (2011) “Designing context-sensitive systems: An integrated approach. In Journal Expert Systems with Applications, v.38, n.2, p. 1119-1138

Wang, R. and Strong, D., (1996) “Beyond Accuracy: What Data Quality Means to Data Consumers”. Journal of Management Information Systems, v. 12, n. 4, p. 5-33.

Zhuge, H., Liu, J., Feng, L., Sun, X., and He, C. (2005) “Query Routing in a Peer-To-Peer Semantic Link Network”. In Computational Intelligence, v. 21, n.2, p. 197-216.