Semantic Loss in Query Reformulation in Dynamic Distributed Environments

Bruno F. F. Souza [1], Maria C. M. Batista [2], Ana Carolina Salgado [1]

[1] Federal University of Pernambuco, Center for Informatics, Pernambuco, Brazil
{bffs, acs}@cin.ufpe.br
[2] Federal Rural University of Pernambuco, Informatics Department, Pernambuco, Brazil
ceca@deinfo.ufrpe.br

Abstract. Dynamic environments are descentralized systems that provide users with querying capabilities over a set of heterogeneous, distributed and autonomous data sources. Data Integration Systems, Peer Data Management Systems (PDMS) and Dataspaces are examples of such systems. They are composed by data sources (peers) that belong to a specific domain and are linked to each other by mappings (correspondences). Nonetheless, a challenge inherent to dynamic environments is to analyze the semantic loss during query reformulation. A semantic loss may occur when a query is reformulated from a peer to another in the system. To minimize the consequences of this problem, we propose the use of information quality criteria to help the semantic loss analysis. The semantic loss analysis is a step executed to verify the query routing possibilities.

Keywords: semantic loss, dynamic environments, PDMS, information quality, quality criteria

1 Introduction

Nowadays, there is a demand for high-level integration of autonomous and heterogeneous data sources through the development of distinct types of distributed environments, including Data Integration Systems [1], Peer Data Management Systems (PDMS) [2] and Dataspaces [3]. These dynamic environments (DEs) are composed by various autonomous data sources (e.g. site, files, database) referred here as *peers*, which maintain information about a certain domain and which are linked to other peers by mappings (i.e. associations between schemas) called hereafter as *correspondences*.

In a PDMS, when a user poses a query at a peer, the query is executed in that peer and then reformulated to its neighbors' peers in order to acquire more information. This reformulation process may lead to the query degradation, i.e., the query suffers some transformations in a way that the concepts used in the query will not be present in the target peer, in another words, the concepts will be left out during the reformulation process among peers.. The semantic loss is one aspect of query degradation.

DEs still suffer with inadequate control mechanisms to address, for instance, the quality of the query answers as well as the quality of the generated correspondences between peer schemas. Including Information Quality (IQ) analysis in a DE improves systems processes such as query evaluation and peer clustering. IQ is usually characterized via multiple criteria, each of which captures a high-level aspect of quality. The role of each one is to assess and measure a specific IQ dimension [4,5]. Thus, quality metrics are used to measure a particular quality criterion.

The goal of this work is to show the idea of using IQ for helping the semantic loss analysis after query reformulation in PDMSs. For this purpose, we propose two IQ criteria to analyze such a loss and provide this information to improve query routing process. This paper is organized as follows: Section 2 provides an example of query routing; Section 3 introduces the semantic loss problem; Section 4 considers IQ criteria for semantic loss analysis, Section 5 discusses the related work and finally, Section 6 points out some considerations.

2 Query Routing in Dynamic Environments

A DE has a set of autonomous peers that offer information and services to be shared. There is an issue that rises in this scenario. When a query is posed by the user how the system can choose the best possible peer or group of peers to send that query to in order to retrieve relevant information? This process is called query routing and has been addressed in [6]. As an example of query routing and query reformulation, let's analyze the hypothetical *research center* PDMS depicted in Figure 1. The arrows indicate schema correspondences between connected peers, which are used to reformulate the query (to transform a query based on one schema to another schema) over the peers' immediate neighbor, and so on. In this illustration, consider a user in *Brazil* that poses query Q_B based on his/her local schema. Q_B will first be reformulated to peer *Portugal*, according to the set of correspondences $C_{OB\text{-}P}$. Then, peer *Portugal* should decide to which peer reformulates the query in order to retrieve the best possible results. There are two possible paths to follow: *France* and *Germany*. The query routing mechanism is responsible for dealing with such issue and, in this case, if the query reformulation process generates semantic loss *semantic loss* it may compromising in a bad way the query result (imprecise answers for example).

3 Semantic Loss

As shown in the previous section a peer should reformulate a query to its neighbors which, in most of the cases, have a different schema. This process is done recursively for many peers in the network. In this sense, the query may loss some of its *significance* due to the reformulations over different peers' schemas. This problem is called semantic loss and is stated as follows [8]: Let suppose that a query Q is initially submitted over peer P_1 schema. Then the query is reformulated to Q' over peer P_2 sche-

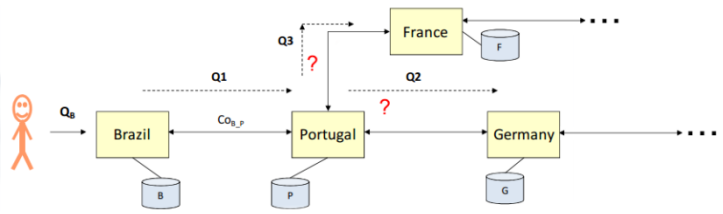

Fig. 1. Query Routing example in a PDMS, adapted from [8].

ma. Let also Q'' be a reformulation of Q' from P_2 to P_1. The difference $Q - Q''$ is called the semantic loss of the original query Q when submitted to P_2.

The semantic loss is detected by comparing two queries, the original one Q, and the reversed one Q''. There exist two ways of comparison: to compare the queries syntactically, and the syntactic differences lead to estimate the semantic loss. A second way to compare Q and Q'' is to verify their results. It is important to point out that the semantic loss does not occur every time. If a peer P_j has the same schema of *Pk*, there is no semantic loss in reformulating a query from P_j to P_k and vice versa.

In this work, we work on the semantic loss problem analysis as follows: how to use IQ criteria to help the detection and minimize the semantic loss in a PDMS network? We use IQ criteria analysis to track such loss and also use it to guide the query routing process. The idea is to calculate the semantic loss after reformulating a query and send the reformulated query only to peers in which the evaluated semantic loss is acceptable. The next section explains the IQ criteria used to evaluate the semantic loss.

4 Quality Criteria for Semantic Loss Analysis

We state our approach in the following: let us suppose that a peer P_1 will reformulate a query Q to neighbors P_2 or P_3. If P_3 has a higher degree of information completeness than P_2, the semantic loss in reformulating Q to P_3 is smaller than reformulating to P_2. The IQ criteria that compose the information completeness of a peer are: *data completeness* and *schema completeness*. In the following, we provide the definitions of these two criteria as well as how they can be evaluated.

Data Completeness: due to dynamicity, query answers in a PDMS may not be complete, considering its original definition (data completeness is typically understood as the ratio of answer set size to the total amount of known data [11]), which requires the knowledge of the total amount of data in the system and relies on the closed world assumption. Instead, peer schemas in the set of available peers have an open-world assumption [12], i.e., the data returned by querying these peer may be incomplete. In this light, data completeness in PDMSs may be defined as the ratio between received results and the existing suitable data belonging to the available peers at query answering time. Thus, we define that the completeness of a peer P_j for a query Q originally submitted from a peer P_i (P_j is a neighbor of P_i) is calculated by the formula:

$$\mathrm{DataComp}_{\mathrm{QP_iP_j}} = \frac{\mathrm{QP_iP_j tuples}}{\sum_{k=1}^{n} \mathrm{QP_iP_k tuples}} \qquad (1)$$

where QP_iP_j tuples is the number of tuples returned by peer P_j for query Q reformulated from P_i to P_j;
$QP_iP_k tuples$ is the number of tuples returned by peer P_k for query Q reformulated from P_i to P_k;
n is the number of P_i neighbors and;
the set of peers P_k ($1 \le k \le n$) are P_i neighbors.

Schema Completeness: is the degree to which entities and properties of the peer are not missing when related with to the entities and properties requested in a submitted query *Q*. In routing a query *Q* from peer P_i to peer P_j the completeness of P_j when related to query *Q* may be assessed by taking the ratio between the number of schema elements queried *Q* and number of elements held by P_j as in Formula 2:

$$\mathrm{SchComp}_{\mathrm{QP_iP_j}} = \frac{\mathrm{Qelements}}{\mathrm{P_j\ elements}} \qquad (2)$$

where *Qelements* is the number of schema elements present in *Q* and;
$\mathrm{P_j\ elements}$ is the number of P_j schema elements.

Our proposal is to use these IQ criteria to analyze semantic loss in query reformulation and identify routing possibilities. In a practical way, we intend to compute these criteria dynamically in order to know whether it is worth to send the query to peers based on their IQ score (value). For example, if a peer has low information completeness score it probably means the query will suffer a semantic loss, otherwise the query may be sent without any or less semantic loss. Moreover, the loss of semantic in a query may be used as a criterion to stop query routing process.

5 Related Work

Semantic query reformulation has attracted significant attention. The work of Bonifati [13] provides a 'relevance' concept of a query wrt a mapping based on AF-IMF metric. This metric takes into account the semantic proximity between the query and the local and external mappings thus, creating only relevant mappings and minimizing the semantic loss. Delveroudis [9] shows an algorithm that estimates the semantic loss of rewritten queries based on the notion of containment queries. This information is used as the basis for extending the schema mappings and improving the quality of retrieved answers. A formal definition of semantics of query answering is presented in [14]. In this work, the authors show an algorithm that preserves semantics and reduces semantic loss among query reformulation. The authors in [15] highlight that the lack of IQ analysis may contribute to information loss as well as the completeness of query answers.

6 Conclusion and Future Work

In this paper we addressed the problem of semantic loss that affects the propagation of queries in DE, more specifically in PDMS. We illustrated the query routing process and how a query should be reformulated in a PDMS environment. We believe that the analysis of IQ criteria plays an important role in detection of query degradation by taking into account the IQ scores in query reformulation process. To this end, we showed two IQ criteria we consider relevant to that analysis and described how they may be assessed. This evaluation may improve the overall query reformulation process in terms of retrieving relevant information from peers as well as routing query only to peers that provide meaningful information. Currently, we are specifying the IQ criteria and preparing our environment to implement and test the results of our

approach in a PDMS called SPEED[1].We also plan to study the use of other criteria to enrich the semantic loss analysis.

7 References

1. Halevy A., Rajaraman, A., Ordille, J.: Data Integration: The Teenage Years. In: 32nd VLDB, Volume 32, p. 9 – 16, (2006).
2. Herschel, S., Heese, R.: Humboldt Discoverer: A Semantic P2P Index for PDMS. In: Proc. of the Internacional Workshop Data Integration and the Semantic Web, Portugal (2005).
3. Franklin, M., Halevy, A., Maier, D.: From Databases to Dataspace: A New Abstraction for Information Management. In: SIGMOD, Volume 34, p. 27 – 33, (2005).
4. Dustdar, S., Pichler, R., Savenkov, V., Truong, H.: Quality-aware Service-Oriented Data Integration: Requirements, State of the Art and Open Challenges. In: ACM SIGMOD Record, Volume 41, p.11 – 19, (2012).
5. Batista, M. C. M., Salgado, A. C.: Information Quality Measurement in Data Integration Schemas. In: 5th QDB, p.61 – 72, Viena (2007).
6. Ismail, A., Quafafou, M., Durand, N., Nachouki, G., Hajjar, M.: Queries Mining for Efficient Routing in P2P Communities. In: IJDMS, Vol.2, No.1, (2010).
7. Souza, D., Arruda, T., Salgado, A. C., Tedesco, P. C. A. R., Kedad, Z.: Using Semantics to Enhance Query Reformulation in Dynamic Environments. In: 13th ADBIS, p.78 – 92, Riga, Letônia (2009).
8. Delveroudis, Y., Lekeas, P. V.: Managing Semantic Loss during Query Reformulation in PDMS. In: SWOD IEEE, p.51-53. (2007).
9. Delveroudis, Y., Lekeas, P. V., Souliou, D.: On Estimating Semantic Loss in Peer Data Management Systems. In: AP2PS IEEE, p.51-53. NTUA, Greece (2009).
10. M. Karnstedt, K. Sattler, M. HaB, M. Hauswirth, B. Sapkota, R. Schmidt.: Approximating Query Completeness by Predicting the Number of Answers in DHT-based Web Applications. In: 10th ACM WIDM, p. 71-78, (2008).
11. F. Naumann, J.C. Freytag, U. Leser.: Completeness of Integrated Information Sources. In: Inform. Systems 29 (7), p. 583 – 615, (2004).
12. I. Tatarinov, A. Halevy.: Efficient Query Reformulation in Peer-Data Management Systems. In: ACM SIGMOD, p. 539-550, (2004).
13. A., Bonifati, G., Summa, E., Pacitti, F., Draidi.: Semantic Query Reformulation in Social PDMS. In: CoRR ABS, (2011).
14. A. Bonifati, E.Q. Chang, T. Ho, L.V.S. Lakshmanan, R. Pottinger, and Y. Chung. Schema mapping and query translation in heterogeneous p2p xml databases. VLDB J., 19(2):231–256, (2010).
15. K., Hose, A., Roth, A., Zeitz, K., Sattler, F., Naumann.: A Research Agenda for Query Processing in Large-Scale Peer Data Management Systems. In: Information Systems, (2008).

[1] The SPEED Project (http://www.cin.ufpe.br/~speed/)