Semantic measures as information quality criteria for query routing processes

Crishane Freire*, Bruno F.F. Souza and Ana Carolina Salgado

Center for Informatics,
Federal University of Pernambuco,
Av. Jornalista Aníbal Fernandes, s/n, 50.740-560,
Recife, Pernambuco, Brazil
E-mail: caf4@cin.ufpe.br
E-mail: bffs@cin.ufpe.br
E-mail: acs@cin.ufpe.br
*Corresponding author

Damires Souza

Federal Institute of Education, Science and Technology of Paraíba,
Av. Primeiro de Maio, 720, Jaguaribe, 58.015-430,
João Pessoa, Paraíba, Brazil
E-mail: damires@ifpb.edu.br

Maria C.M. Batista

Federal Rural University of Pernambuco,
Rua Dom Manoel de Medeiros, s/n, Dois Irmãos, 52.171-900,
Recife, Pernambuco, Brazil
E-mail: ceca@deinfo@ufrpe.br

Abstract: Query answering has been addressed as a key issue in 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 results according to the submitted query. In this process, queries are reformulated and propagated through network peers using the semantic mappings between neighbour peers’ schemas. The successive processes of query reformulation may result in a semantic loss of the original query, i.e., concepts which belong to the original query may be lost when reformulated queries are produced. Thereby, this work proposes the use of semantic measures obtained from information quality (IQ) criteria aiming to avoid or minimise this semantic loss. Moreover, it combines semantic information and IQ, by presenting a model, which is instantiated to illustrate how this proposal produces the semantic measures and enhances query routing processes.

Keywords: peer data management systems; PDMS; query routing, information quality; IQ; query reformulation; semantic measures.
Reference to this paper should be made as follows: Freire, C., Souza, B.F.F., Salgado, A.C., Souza, D. and Batista, M.C.M. (2013) ‘Semantic measures as information quality criteria for query routing processes’, *Int. J. Business Intelligence and Data Mining*, Vol. 8, No. 2, pp.167–183.

Biographical notes: Crishane Freire is a PhD candidate at Federal University of Pernambuco. Her research interests are information quality, semantic information and query routing in distributed and dynamic environment. She is also an Assistant Professor in Computer Science at Federal Institute of Education, Science and Technology of Paraíba.

Bruno F.F. Souza is currently an MSc candidate at Federal University of Pernambuco in Computer Science. His main research interests are database, information quality, semantic information, query reformulation, semantic web and dynamic distributed environments.

Ana Carolina Salgado is currently an Associate Professor at Center for Informatics of Universidade Federal de Pernambuco – UFPE – (Brazil). She obtained her Doctorate in Computer Science from the University of Nice (France) in 1988, her Master (1983) and her Bachelor degree (1980) in Computer Science from UFPE. Her main research interests are in the areas of databases, information integration on the wand context-aware systems. She has published over a hundred technical articles in conference proceedings and journals. In her academic activities she has advised 40 MSc and 12 PhD theses. She is member of the ACM and the Brazilian Computer Society. She also was the head of Center for Informatics (2001–2005) and coordinator of the Computer Science undergraduate course (1992–1996) at UFPE.

Damires Souza is an Associate Professor in Computer Science at Federal Institute of Education, Science and Technology of Paraíba. She is the leader of the database research team at IFPB. She is also a member of the databases research team at Federal University of Pernambuco. She obtained her PhD in Computer Science from the Federal University of Pernambuco in 2009 and during her PhD studies she did a research internship at PRiSM laboratory of Versailles. Her main interests are concerned with data integration, semantics, context-aware systems, information quality and semantic web.

Maria C.M. Batista is currently working as an Assistant Professor at Federal Rural University of Pernambuco. Since 1988 she works in software development projects. She recently acted as a Project Manager and Software Factory Manager in the Court of the State of Pernambuco. She also worked in baking automation and business intelligence projects for companies such as Banorte and Pointer Software where she worked as a Project Manager. She has a PhD in Computer Science in the area of Information Quality in Data Integration systems, having published several papers in several areas of interest, including: quality of data and information, software engineering and databases.

This paper is a revised and expanded version of a paper entitled ‘Towards an information quality approach to enhance query routing processes’ presented at the 14th International Conference on Information Integration and Web-based Applications & Services (iiWAS 2012), Bali, Indonesia, 3–5 December.
1 Introduction

Peer-to-peer (P2P) systems have become popular due mainly to scalability, self-configuration and autonomy features (Sassatelli et al., 2009; Sung et al., 2005). Peer data management systems (PDMS) came into the focus of research as a natural extension to data integration systems in the P2P setting (Souza et al., 2011; Mandreoli et al., 2009; Halevy et al., 2003). In this sense, PDMSs are P2P applications where each peer represents an autonomous data source which exports either its entire data schema or only a portion of it. Such schemas, named exported schemas, represent the data to be shared with the other peers. To enable data sharing, mappings (i.e., correspondences) between elements of the exported schemas are generated and maintained.

The environment in which a PDMS operates is very dynamic (e.g., peers may join and leave the system at any time) and the system usually works considering an open world assumption in which peers may have incomplete information (Halevy et al., 2004). As a result, answers to user requests, in most cases, are not supposed to be complete. Queries are submitted considering partial information provided by peer schemas and can be answered by other peers which are reached through the existing mappings. It means that, to query route among the peers in the network the query needs to be reformulated considering the neighbour peer schema concepts. During this reformulation process, in some cases, the original query semantics may be lost.

The idea of semantic loss during query reformulation has been addressed in a few works (Delveroudis and Lekeas, 2007), and in the context of information loss problem (Aberer et al., 2003; Roth and Naumann, 2007; Kantere et al., 2009; Mena et al., 2000). Usually, in a PDMS, a peer reformulates a query to its neighbours which, in most of the cases, have a different schema. In this sense, the query may lose some of its significance, i.e., occurring loss of information from the original query to the reformulated query (Delveroudis and Lekeas, 2007). As a result of this information loss, the query answer may be discrepant when related to the information requested by the user or may be an empty result.

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 to answer a given submitted query. This process is not easy due to some factors: 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’ relevancy or even the query semantic loss are important criteria that may be considered in the peer selection or the estimated routing paths.

The use of information quality (IQ) measurements may be a relevant discriminator in PDMSs environments. Nevertheless, few approaches consider IQ control mechanisms, for instance, the management of quality of the data that are obtained as query answers as well as the quality of the existing data sources (peers) (Green et al., 2010; Heese et al., 2005). The objective of embodying IQ analysis in a PDMS is to provide improvements over the system’s processes. In this work, we are particularly interested in using IQ in query routing processes (Freire et al., 2012). In this light, we propose the use of semantic measures as IQ criteria as a way to enrich the routing. We defined two IQ criteria to help the query routing process: query loss measure \((QuLM)\) and query enrichment measure \((QuEM)\). We argue that the use of such information may reduce the query search space by forwarding queries only to peers that may contribute with relevant answers, i.e.,
answers that match the semantics of the submitted query. Furthermore, 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. (2012). To illustrate, we present a running example
exemplifying the contributions of this work.

This paper is organised as follows: Section 2 introduces IQ and PDMS; Section 3
defines the quality criteria to be used in our approach; Section 4 presents the model to
represent IQ related concepts; Section 5 details the query routing problem with the
defined IQ criteria in practice; Section 6 discusses some related work. Finally, Section 7
points out some considerations and highlights important topics for further research.

2 IQ and PDMS

The notion of IQ has emerged during the past years and shows a steadily increasing
interest (Duchateau and Bellahsene, 2010; Keeton et al., 2009; Roth and Naumman,
2005; Batini and Scannapieco, 2006). One of the most known quality dimensions
classification is presented by Wang and Strong (1996). They have conceived one of the
first sets of structured and classified quality dimensions that have been a strong reference
for most of the studies in this area.

In that respect, IQ may be used as a relevant discriminator in PDMSs settings. IQ is
usually characterised via multiple dimensions or criteria, each of which captures a
high-level aspect of quality. The role of each one is to assess and measure a specific IQ
criterion (Batista, 2008). In general, quality metrics are defined to measure a particular
quality criterion.

PDMSs are different from traditional data integration systems since they do not
provide a mediator schema to which user queries are submitted. When a peer enters the
system, it exports its data schema that will be shared with the other peers. A peer knows
about its neighbour peers by mappings (i.e., associations or correspondences between
concepts from peer schemas), which help to translate queries during query reformulating
process and transform data into results during the integration process. Queries submitted
to one peer are answered by that peer and by peers reached along paths of mappings
through the P2P network. In this light, a peer can contain all or part of a global answer to
given query. In addition, other peers can provide the answers to the same query with
different levels of quality and response time.

Query answering has been recognised as the main service that a PDMS can provide.
Due to the lack of a global schema that represents all knowledge regarding the distributed
peers, each peer individually must decide to which existing neighbours the query should
be forwarded. This decision is usually made considering the available local knowledge at
that peer.

As in any data integration system built over autonomous and distributed sources,
PDMSs are supposed to have poor IQ in some aspects such as (Karnstedt et al., 2008;
Roth and Naumann, 2005; Herschel and Heese, 2005): peer (data source), peer schema,
mappings, data and query answers. In a PDMS, the IQ of query answers depends not only
on the data quality of a particular data source (peer), but also on the quality of the
mappings between neighbour peers (Yatskevich et al., 2007).

In fact, considering a PDMS, quality criteria can give trust to the system and enhance
its processes by defining metrics to assess the quality of data, query, peers and query
answers. One possible usage of IQ criteria in this environment should be the assistance in query routing process. In this process, every time that a peer receives a query it must decide, based on its local knowledge, to which of its semantic neighbours it should forward the query. In our approach, IQ criteria are used to decide the more suitable neighbours to forward a query. To better understand the query routing process and how the IQ metrics are obtained, we should consider two aspects: the generation of semantic correspondences between neighbour peer schemas and the query reformulation process itself. The first one is the initial step a peer undergoes when join the system and establishes semantic correspondences with semantic related neighbours peers. The latter aspect is related to how the query reformulation works when a query needs to be translated into another query according to a set of inter-schema correspondences. In the next sections, we describe the semantic matcher (Pires et al., 2009) and query reformulation process (Souza, 2009) used in our work.

3 Semantic matcher

The semantic matcher (SemMatcher) (Pires et al., 2009) considers domain ontologies (DO) as reliable references that are made available on the web. It uses them in order to bridge the conceptual differences or similarities between two peer schemas represented by ontologies. In this sense, first concepts and properties from the two schemas are mapped to equivalent concepts/properties in the DO and then their semantic correspondences are inferred based on the existing semantic relationship between the DO elements. To specify the correspondences, it take into account four aspects:

1. the semantic knowledge found in the DO
2. if the schema concepts share super-concepts in the DO
3. if these super-concepts are different from the root concept
4. the depth of concepts measured in number of nodes.

Considering that \( x \) and \( y \) are elements (concepts or properties) which belong to ontologies \( O_1 \) and \( O_2 \), the approach considers seven types of semantic correspondences between elements (Souza, 2009):

- **isEquivalentTo**, denoted as \( O_1:x \equiv \rightarrow O_2:y \)
- **isSubConceptOf**, denoted as \( O_1:x \sqsubseteq \rightarrow O_2:y \)
- **isSuperConceptOf**, denoted as \( O_1:x \sqsupset \rightarrow O_2:y \)
- **isPartOf**, denoted as \( O_1:x \sqsupset \rightarrow O_2:y \)
- **isWholeOf**, denoted as \( O_1:x \sqsubseteq \rightarrow O_2:y \)
- **isCloseTo**, denoted as \( O_1:x \approx \rightarrow O_2:y \)
- **isDisjointWith**, denoted as \( O_1:x \perp \rightarrow O_2:y \).
Figure 1 illustrates an example of how we use the DO for specifying semantic correspondences between ontologies $O_1$ and $O_2$. In this example, $O_1:x$ has an equivalence relationship with $DO:k$ and $O_2:y$ has an equivalence relationship with $DO:z$. As the element ‘$k$’ is a subconcept of ‘$z$’ in the DO, we can infer that the same relationship occurs between ‘$x$’ and ‘$y$’. Then, we can express that $O_1:x$ is subconcept of $O_2:y$.

Figure 1  A DO specifying semantic correspondences (see online version for colours)

The SemMatcher approach brings together a combination of already defined matching strategies (Euzenat and Shvaiko, 2007). A linguistic-structural matcher and a semantic matcher are executed in parallel. The former may be any existing matching tool including linguistic and/or structural matchers, e.g., H-Match (Castano et al., 2006). The latter uses the DO as background knowledge and identifies the seven kinds of semantic correspondences. The obtained similarity values of both matchers are combined through a weighted average. Each matcher receives a particular weight according to its importance for the matching process, as follows: $isEquivalentTo (1.0)$, $isSubConceptOf (0.8)$, $isSuperConceptOf (0.5)$, $isCloseTo (0.3)$, $isPartOf (0.1)$, $isWholeOf (0.1)$, and $isDisjointWith (0.0)$. The weights reflect the degree of closeness between the correspondence elements, from the strongest relationship (equivalence) to the weakest one (disjointness). We call this weights of similarity degree between two concepts $C_1$ and $C_2$ ($Sim(C_1, C_2)$).

4 Query reformulation

In a PDMS, a query posed at a peer (data source) is reformulated into a new query expressed in terms of a target peer, considering existing correspondences between them. In this sense, it is important to verify if the reformulated query is still useful for users, i.e., if the new query is in conformance with users’ interest (i.e., with respect to the submitted query). Other aspect to be considered regards if concepts from the source peer have exact corresponding concepts in the target one, what may result in an empty reformulation and, possibly, no answer to the user.
The SemRef approach (Souza, 2009) uses semantics to enhance query reformulation through query enrichment and provides users with a set of expanded answers. The semantics is acquired from a set of correspondences that extend the ones commonly used (e.g., aggregation and closeness). Through this set of semantic correspondences, the SemRef may produce two kinds of query reformulations:

1. an exact one, considering only equivalence correspondences
2. an enriched one, resulting from the set of the other correspondences.

To define the query enriching mode the user sets four variables which specify what should be considered when a query \( Q \) is to be enriched. The variables are defined as follows: approximate: includes concepts that are close to the ones of \( Q \); specialise: includes concepts that are sub-concepts of some concepts of \( Q \); generalise: includes concepts that are super-concepts of some concepts of \( Q \); compose: includes concepts that are part-of or whole-of some concepts of \( Q \).

To better illustrate, we use a scenario composed by two peers \( P_1 \) and \( P_2 \). Each peer is described by the ontologies \( O_1 \) and \( O_2 \). In order to identify the semantic correspondences between \( O_1 \) and \( O_2 \), the rules of SemMatcher were applied. As a result, the set of semantic correspondences between \( O_1 \) and \( O_2 \) were identified. Since the correspondences are unidirectional, we present below a fragment of the set concerning the concept FullProfessor (from \( O_1 \)) with some related concepts in \( O_2 \):

- \( O_1: \text{FullProfessor} \equiv \rightarrow O_2: \text{FullProfessor} \) (isEquivalentTo)
- \( O_1: \text{FullProfessor} \rightarrow \rightarrow O_2: \text{Professor} \) (isSubConceptOf)
- \( O_1: \text{FullProfessor} \approx \rightarrow O_2: \text{VisitingProfessor} \) (isCloseTo)
- \( O_1: \text{FullProfessor} \rightarrow \rightarrow O_2: \text{Course} \) (isPartOf).

Suppose that a user submits a query \( Q = (\text{FullProfessor}) \) in \( P_1 \) and set all four enriching variables to TRUE. SemRef will produce two kinds of query reformulations: an exact reformulation, denoted by \( Q_{\text{exact}} = (\text{FullProfessor}) \), and another one, enriched, denoted by \( Q_{\text{enriched}} = (\text{VisitingProfessor} \sqsupset \text{Professor} \sqsupset \text{Course}) \). Thus, users may be provided with a set of expanded answers, according to their interest and the existing semantic similarity degree between queried concepts.

However, considering that a PDMS is composed by thousands of peers, the successive process of query reformulation may lead to a query semantic loss. In our PDMS, the query semantic loss may be produced in two ways: by the loss of query original concepts and by the new concepts acquired from the query enrichment process. In the next section we present our proposed IQ criteria for assessment of query reformulation.

5 Proposed IQ criteria

In a PDMS, the successive process of query reformulation may lead to some loss of information, i.e., loss of concepts from the original query to the reformulated ones.
Besides, in our approach, enriched queries may be also produced by the \textit{SemRef} algorithm, which works as a kind of query expansion between peers, resulting in new query results. These produced query results may be interesting to the user. In this light, suppose we are reformulating a query $Q$ from $P_i$ to $P_j$. We state that the more concepts queried by $Q$ present in $P_j$, the better is $P_j$ as a candidate peer for routing $Q$. We also argue that if the concepts queried by $Q$ in $P_i$ have a good similarity with the concepts present in $P_j$, $Q$ will obtain better results in being routed to $P_j$. Thus, we define quality criteria in order to obtain two measures: the semantic loss and the enrichment of a query. The semantic loss may occur when the reformulation is exact, i.e., only considers equivalent concepts between $P_i$ and $P_j$. The query enrichment may occur when the reformulation is enriched and the query gains some additional concepts due to the use of enrichment variables. The measures of loss and enrichment are used in query routing as choice parameters to decide if the query may be forwarded to another peer or not. These criteria are defined as follows:

- \textit{QuLM}: This criterion is used to extent the query semantic loss in terms of semantic concepts that are lost during exact query reformulation. When a query $Q$ is reformulated to another peer’s schema some of the concepts from query $Q$ could be lost due to the differences between peer’s schemas. Considering two peers $P_i$ and $P_j$, the \textit{QuLM} is stated as follows:

$$\text{QuLM}_{ij} = \begin{cases} 0 & \text{if } |C_{eq_i}| \geq |C_i|, \\ 1 - \frac{|C_{eq_i}|}{|C_i|} & \text{if } |C_{eq_i}| < |C_i| \end{cases}$$

where $|C_{eq_i}|$ is the number of concepts in $Q_j$ which have equivalents concepts in $Q_i$ and $|C_i|$ is the number of concepts in $Q_i$.

- \textit{QuEM}: This criterion is used to extent the semantic enrichment in terms of semantic extra concepts obtained during the query reformulation process. When a query $Q$ is posed at peer, the user can choose whether the query reformulation will consider more semantic extra concepts than equivalent concepts only. In this way, the query can be enriched with more concepts and, consequently, may retrieve more results. Considering two peers $P_i$ and $P_j$, the \textit{QuEM} is stated as follows:

$$\text{QuEM}_{ij} = \frac{|C_{sub_j}| x 0.8 + |C_{sup_j}| x 0.5 + |C_{clj}| x 0.3 + |C_{pw_j}| x 0.1}{|C_j|}$$

where

- $|C_{sub_j}|$ is the number of concepts in $Q_j$ which are subconcepts in $Q_i$
- $|C_{sup_j}|$ is the number of concepts in $Q_j$ which are superconcepts in $Q_i$
- $|C_{clj}|$ is the number of concepts in $Q_j$ which are close to concepts in $Q_i$
- $|C_{pw_j}|$ is the number of concepts in $Q_j$ which are part-of or whole-of concepts in $Q_i$
- $|C_j|$ is the number of concepts in $Q_j$. 
These measures will be dynamically calculated for every query submitted to the system. During a query routing process, a peer will decide based on QuLM and QuEM values if it is worthwhile to forward the query to its neighbours peer. In the next section, we present a model to represent IQ in a query routing scenario in the light of PDMSs.

6 A model to represent IQ in a query routing scenario

In order to allow IQ usage, it is crucial to define how IQ related concepts are represented and (possibly) persisted. A challenge to be faced is the fact that there is not a standard model for representing IQ concepts yet. In accordance with our analysis, some issues should be taken into account when evaluating techniques to represent IQ:

1. the model must be portable and lightweight
2. it is desirable to have validation tools which can be used for edition, type checking and for conversion
3. formality is welcome since it eases definition and reusability, and finally
4. it should allow inference mechanism in order to provide better assessment.

According to these issues, ontologies seem to be one of the best options for IQ and semantics related information representation. There are several advantages for developing ontology-based models (Calvanese et al., 2009; Souza et al., 2011), including: to provide knowledge sharing (services and agents can deal with the same set of concepts), to enable reuse of the information, to define semantics independently from data representation, and finally to enable the use of existing inference engines to reason about the represented information.

Therefore, our initial effort has been devoted to the definition of an ontology-based model which is based on the metamodel presented in Souza et al. (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 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.

According to the defined meta-concepts provided by the metamodel, we have developed an ontology-based model to represent IQ criteria and other IQ related concepts considering a query routing scenario in a PDMS. To help matters, we define our scenario, as follows. Consider a PDMS composed by peers which integrates data from some specific knowledge domain (e.g., education, geography). In this setting, we use ontologies as uniform representation of peer schemas. Such ontologies are named peer ontologies. Since peers are grouped within the same knowledge domain, we use a DO as background knowledge to help to identify the set of correspondences (i.e., mappings) between pairs of neighbour peers ontologies (Pires et al., 2009).

Queries are submitted using SPARQL (http://www.w3.org/TR/rdf-sparql-query/) language and are reformulated considering the set of such existing semantic correspondences between source and target neighbour peers. Figure 2 depicts the model with its concepts and relationships between them. The concepts and relationships which are identified by the prefix meta belong to the metamodel and are reused in this model.
According to the defined scenario, in order to establish the appropriate IQ related concepts, at first, we have identified the domain entities that we needed to work with. A domain entity is anything in the real world that is relevant to describe the domain (e.g., data sources and users). Therefore, we determined three main domain entities around which we could consider IQ: peer, query and peer ontology. The concept Element is used to characterise a Domain_Entity. Concept concerns with elements of a peer ontology (represented by Peer_Ontology) and that are used in query formulation and reformulation. We also have defined the most important process in our setting. Thus, query routing has been defined as the root process to deal with. Activities such as peer selection, query forwarding, query execution, query reformulation and query results integration have been specified as activities of the query routing process.

The main concept underlying the model and reused from the metamodel is Information_Quality, a subconcept of Information. The concept Information_Quality is related with the information obtained through quality criteria. This is composed by the concept Quality_Criterion that has sub-concepts QuLM and QuEM criteria. These IQ criteria have been defined to produce quality measurements as discussed in Section 3.

Figure 2 Ontology-based model to represent IQ in PDMS query routing processes (see online version for colours)

7 Query routing with the defined IQ criteria in practice

To illustrate our work, we consider a motivating scenario in our running PDMS called SPEED (Souza, 2009), which integrates data from the Education domain distributed over a set of peers (data sources). The peers are linked to other peers by means of correspondences, as defined in Section 2. Each peer has an associated schema within the domain of interest and the schemas are represented by ontologies. In this setting, queries are submitted and reformulated and the query result is the union of the query answers (considering that data conversions which have already been done) returned by each accessed peer during the query routing process. Figure 3 depicts a set of five interconnected peers in SPEED environment.
As background knowledge, we have considered a public DO named *UnivCsCMO.owl*. The set of semantic correspondences between pairs of neighbour peers was identified and stored in OWL files (called alignment files) by the *SemMatcher*. A fragment of one of these files (concerning the concept *Book*) is shown in Figure 4.

Given this PDMS scenario, suppose that a user wants to know about scientific and technical documents. To illustrate the use of IQ criteria in query routing process our experiment considers exact and enriched reformulations. Thus, the user enters the system and connects with the peer P2178.

In the first example, the user wants to find only exact answers. Thus, the query $Q = \{\text{Book} \cup \text{BookArticle} \cup \text{Thesis} \cup \text{Manual}\}$ expressed in Description Logics language (Horrocks, 2005), is submitted by the user at peer 2178 and the enriching variables are not defined (see Figure 5).
The query $Q$ is processed locally without semantic enrichment and in order to acquire more results it is reformulated to its neighbour peer ($P_{2478}$). For each reformulated query, the QuLM is assessed and this process is repeated for each one of its neighbour peers. Thus, the query is reformulated from $P_{2178}$ to $P_{2478}$, then from $P_{2478}$ to $P_{2378}$ and, at last from $P_{2378}$ to $P_{2578}$. The Figure 6 presents the QuLM values obtained at each query reformulation step. It may be observed that, for all reformulation steps, the QuEM values are zero. This is due the choice of the user for only exact answers.

In a query routing process, to avoid the forwarding of queries with high degree of semantic loss, two semantics thresholds were defined: Loss_T and Enrich_T. At each step of the query routing, Loss_T and Enrich_T are used in a combined way to check if
the reformulation produces a suitable query that should be forwarded. In our example, Loss_T is 0.49 and Enrich_T is dynamically calculated considering the enriching variables defined in query submission by the user. For exact reformulations, the QuLM measure should be below 0.6 (Loss_T value), i.e., more than half of original concepts must be preserved in the reformulated query. For enriched reformulations, the QuEM measure should be above the Enrich_T, i.e., stating that the semantic enrichment is suitable to user interest. Thus, to decide about the query forwarding, the values of QuLM and QuEM should be considered.

In this example we can observe that the query reformulated between P2378 and P2578 generated QuLM 0.75. Thus, a high degree of semantic loss was produced and the query will not be forwarded to P2578.

In the following, we show another example where the same query is submitted by enabling the enrichment mode (Figure 7). The variables compose, generalise and specialise are selected. To define the Enrich_T value at each step of the routing, the semantic measure is computed based on correspondence weights related to each enriching variable and the number of correspondences produced in the reformulated query.

As in the first example, Q is executed locally in P2178. Then, in order to acquire more results Q is reformulated to its neighbour peer P2478 taking account the enriching variables. In the same way, at each peer which receives the query, QuLM and QuEM
measures are calculated. Figure 8 presents all the measures related to each query reformulation.

In enriched reformulation mode, approximated concepts will provide the user with a set of expanded answers. The reformulated queries presented in Figure 8 were produced taking into account the semantic correspondences existing between the peers’ ontologies.

In the query reformulation between P2178 and P2478 there is a semantic loss of 0.25. Three original query concepts were preserved. As the user defined enriching variables, the query forwarded to P2478 is enriched with the concepts Publication, ScientificPublications, UnofficialPublication, Article and DoctoralThesis. Thus, the reformulated query obtained a semantic gain of 0.3.

When the query reaches the peer P2478 its semantic neighbours are identified (P2278 and P2378). As the same way, the QuLM and QuEM measures are assessed to each neighbour peer in order to be used as a choice parameter in the routing process. From the peer P2478 to P2378, the query is reformulated without some original concepts such as Manual and BookArticle. Despite of semantic loss of 0.5, the query semantic is preserved and the QuEM measure states a semantic gain of 0.22. The measures obtained in each reformulation are associated to the received query by the peer. All measures are used as choice criteria to query routing. Figure 9 presents a partial model view during peer selection activity at peer P2478.

8 Related work

In general, the goal of query routing is to reduce the response time and to return the more suitable query results. For this purpose, the most of the routing strategies (Mandreoli et al., 2009; Ismail et al., 2011; Li and Vuong, 2007) reduce the number of peers to forward
the query by clustering peers with semantic similarity or using distributed index to improve the query performance.

Despite the fact that there is much research showing the importance of IQ for the improvement of query answering in PDMS (Green et al., 2010; Heese et al., 2005), few works discusses the use of IQ aspects specifically in query routing. Zhuge et al. (Zhuge 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. The objective is to forward query considering peers and mappings that promise large partial 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 several 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 in order to create a search space to route queries to relevant peers (Löser et al., 2003; Montanelli et al., 2010).

Different from the referred works, our approach defines two quality criteria (QuLM and QuEM) to be used in query routing processes and presents a model to represent and use these criteria in a combined way. The idea is to use this model to produce knowledge. In PDMS, queries need to reach the largest number of peers in order to obtain the best set of results. Otherwise, the need of scalability requires that the generated number of messages should be as lower as possible. So, it is important to reduce the search space of query selecting peers that may contribute with relevant results. In this sense, semantic criteria may be used to dynamically identify and select peers to forward a query to.

9 Conclusions and further work

Due to the ever increasing complexity of PDMS processes, such as query routing, the usage of semantic information and IQ is becoming very significant, instead of an optional requirement. PDMSs are highly dynamic and the semantic knowledge produced by IQ is rather important to assist the selection of the most relevant peers to send a query, according to its required semantics. Meanwhile, it helps to produce query results which best meet the users’ needs. IQ evaluation may contribute not only to minimise the query routing time, but also to reduce the search space by considering only peers that can indeed contribute with relevant answers.

To help matters, we have developed an ontology-based model which deals with IQ aspects in PDMS query routing processes. This model has been represented in OWL (http://www.w3.org/TR/owl-features/). It defines IQ related concepts in a PDMS scenario as well as IQ criteria metrics. Furthermore, we have specified more properly two important IQ criteria in the light of query routing: QuLM and QuEM. These criteria are represented in the proposed model that has been initially used in some preliminary tests of query routing. The initial results have shown benefits in our approach, since we can identify and select the most relevant peers to route a query by instantiating the suitable criteria, calculating their metrics and applying them on the fly. The preliminary results provide evidence that the application of IQ criteria by means of an IQ representation model has the potential to yield considerable improvements to query routing processes.
As further work, we will develop additional scenarios which may allow us to work with other instances, constraints, queries and IQ criteria as well as with larger and different datasets. We plan to extend our proposal to deal with other quality criteria, and to refine the model by defining rules/axioms for the inference of semantic knowledge. We are also integrating this work with an IQ manager which is being developed within our research group.

References


