PSemRef: Personalized Query Reformulation based on User Preferences

Thiago Arruda
Federal University of Pernambuco
50.740-540 Recife, PE, Brazil
+55 81 2126 8430
tan@cin.ufpe.br

Damires Souza
Federal Institute of Education, Science and Technology
58.015-430 João Pessoa, PB, Brazil
+55 83 3208 3062
damires@ifpb.edu.br

Ana Carolina Salgado
Federal University of Pernambuco
50.732-970 Recife, PE, Brazil
+55 81 2126 8430
acs@cin.ufpe.br

ABSTRACT
One key issue for query answering in dynamic distributed environments is the reformulation of a query posed at a peer into another one over a target peer. Making use of the semantics underlying a set of correspondences between peer schemas' elements, the *SemRef* approach has been developed as a means to enhance such process. Nevertheless, while such approach is able to provide users with a set of expanded answers, it lacks to tailor query results according to user's preferences on the different existing semantic correspondences options. In this paper, we address the issue of personalizing query results, in such a way that users may choose which types of semantic correspondences are important to their queries as well as the priority order in which these correspondences should be applied. More specifically, we address query personalization at reformulation time, producing a ranked set of answers according to user's preferences. We present the principles underlying our approach, examples illustrating how they work and some experimental results.

Keywords
Query Reformulation, Personalization, Ranking.

1. INTRODUCTION
Query answering has been addressed as a key issue in dynamic environments such as Peer Data Management Systems (PDMS) [10, 11]. An important step in this process is reformulating a query posed at a peer (data source) into a new query expressed in terms of a target peer – considering existing *correspondences* between peer schema elements. In previous work [7, 8], the *SemRef* approach has been developed as a means to explore semantic correspondences in order to improve query reformulation. The idea is to produce a resulting set of answers which expresses, as closely as possible, what the users define as important at query submission time, considering the dynamicity of the environment. However, a problem that still remains not dealt with is how we can rank such resulting set of answers in such a way that it actually reflects user's preferences.

Making use of a semantic underlying a set of correspondences between peer schemas' elements, the *SemRef* approach accomplishes query reformulation by means of query enrichment. To this end, besides equivalence, it uses other types of correspondences which go beyond the ones commonly found. The priority is to produce the best query reformulation through equivalence correspondence. However, if that is not possible, or if the user defines that it is relevant for him/her to receive semantically related answers, an enriched reformulation is also generated, considering the other types of correspondences.

In this work, we extend the *SemRef* approach by proposing a personalized query reformulation one – named *PSemRef*. In *PSemRef*, the user is enabled to choose what degree of approximation s/he is interested in as well as the priority order in which the set of semantic correspondences will be applied. By choosing that, *PSemRef* is able to produce different sets of query reformulations which are ordered according to users' priority preferences. We address our problem in a P2P network. We focus on reformulating a query posed at a source peer in terms of a target peer. In this paper, we present the principles underlying our approach. To clarify matters, we provide some examples illustrating how these principles work.

This paper is organized as follows: In Section 2 we provide background information on the *SemRef,* approach. Section 3 describes our method to generate the query results ranking and Section 4 illustrates such method by an example. Related work is discussed in Section 5. Finally, Section 6 draws our conclusions and points out some future works.

2. THE *SEMREF* APPROACH
SemRef approach has been instantiated in a PDMS, although it can be instantiated in any dynamic environment. In such systems, schema matching techniques are used to establish correspondences between schema elements which form the basis for query reformulation. Queries submitted at a peer are answered with data residing at that peer and with data that is reached on the basis of semantic correspondences over the network of peers.

In our approach, the peers are clustered according to the same knowledge domain (e.g., *Education*, *Health*), and an ontology describing the domain is available to be used as background knowledge.

The principle underlying the *SemRef* approach is to enhance query reformulation by using semantic correspondences between schema ontologies (which represent peer schemas) and contextual information. The idea is to provide users with a set of expanded answers, i.e., query answers provided by available peers, which are semantically related to the original submitted query and concern user's preferences defined at query submission time.

IIWAS2010, 8–10 November, 2010, Paris, France.

2.1 Using Domain Ontology to Define Semantic Correspondences

Domain Ontologies (DO) contain concepts and properties belonging to a particular knowledge domain and may be used as background knowledge in some tasks. In our PDMS, we consider DO as reliable references that are available on the Internet. Particularly, we use them in order to bridge the conceptual differences or similarities between two ontologies O_1 and O_2 representing the schemas of neighbor peers.

We say that $\{C\} = \{C_{ij}\}_{i<>j}$ refers to the set of correspondences between a source ontology (O_i) with a target ontology (O_j). Since terminological normalization is a pre-matching step in which the initial representation of two ontologies are transformed into a common format suitable for similarity computation, we consider that both ontologies O_i and O_j have been converted to a uniform representation format.

Figure 1 shows an overview of our approach for specifying the correspondences between peer ontologies. In this overview, O_1:x ≡ DO:k and O_2:y ≡ DO:z. Since k is subsumed by z in the DO, we infer that the same relationship occurs between x and y. Then, we conclude that O_1:x is subsumed by O_2:y, denoted by O_1:x $\sqsubseteq\!\!\rightarrow$ O_2:y.

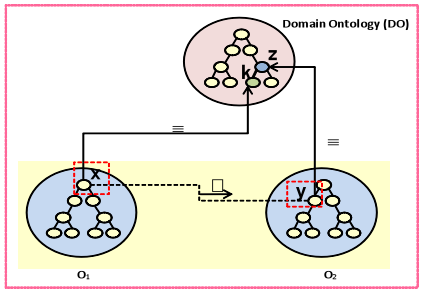

Figure 1. Semantic Correspondences between Peer Ontologies

We have defined seven types of semantic correspondences [7] which were formalized using a notation based on Distributed Description Logics (DDL) [3]. Considering two peer ontologies O_1 and O_2, the semantic correspondences we have defined may be of the following types [7]: *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 $\sqsupseteq\!\!\rightarrow$ O_2:y; *isPartOf,* denoted as O_1:x $\rhd\!\!\rightarrow$ O_2:y; *isWholeOf*, denoted as O_1:x $\lhd\!\!\rightarrow$ O_2:y; *isCloseTo,* denoted as O_1:x $\approx\!\!\rightarrow$ O_2:y; and *isDisjointWith*, denoted as O_1:x $\perp\!\!\rightarrow$ O_2:y.

To make definitions clear, we provide examples using a working scenario composed by two peers P_1 and P_2 which belong to the *Education* knowledge domain. In this scenario, peers have complementary data about academic people and their works (e.g., Research) from different institutions. Each peer is described by an ontology – O_1 (*Semiport.owl*) and O_2 (*UnivBench.owl*). We have considered as background knowledge a DO named *UnivCSCMO.owl*[1].

[1] The complete ontologies are available at our project's web site: http://www.cin.ufpe.br/~speed /SemMatch/index.htm

2.2 Formalizing the Query Reformulation Process

We use the Description Logics language ALC (Attribute Language with Complement) [2] to formalize ontologies as well as queries. In ALC, the constructors are: ¬C (negation), C ∗ D (conjunction), C + D (disjunction), ∀R.C (universal restriction) and ∃R.C (limited existential restriction) where C and D are concepts and R is a role.

In our work, we consider that a query Q is a formula consisting of a disjunction of queries which are themselves conjunctions of ALC concepts C_1, …, C_N where n ≥ 1, as follows:

Definition 1 – Query. A query Q expressed over P_1's ontology, has the following form: $Q = Q_1 + Q_2 + ... + Q_m$, where $Q_i = C_1 * C_2 * ... * C_n$, and where each C_j is an atomic concept, a negated atomic concept or a quantified atomic concept (C_j, ¬C_j, ∀R.C or ∃R.C_j).

Supposing a peer ontology concerning the domain of an academic research center, a query example is: Q_1 = [Student ∗ Worker] which asks for people who study and works.

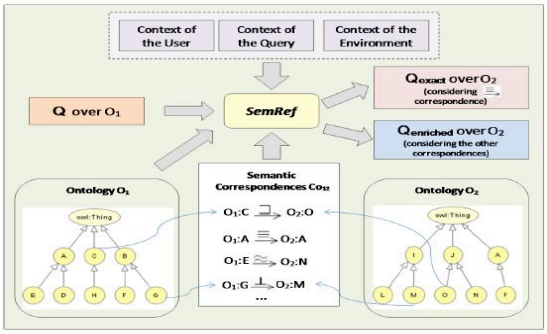

Figure 2. The *SemRef* Approach

Our approach is depicted in Figure 2. When a query Q is submitted in peer P_1, *SemRef* considers the semantic correspondences (Co_{12}) between the source and target ontologies (O_1 and O_2) along with the acquired context and produces two types of reformulations: Q_{exact} and $Q_{enriched}$.

Our reformulation algorithm is outlined in [7]. When posing a query, users must be aware that not only restricted answers, but also those that meet or complement their initial intention, can be relevant for them. Query reformulations are produced according to the following definitions:

Definition 2 - Exact Reformulation. A reformulation Q' of a query Q is said to be exact (denoted as *Qexact*) if each concept (or property) C' of Q' is related to a concept (or property) C of Q by a Co correspondence, where Co ∈ {$\equiv\!\!\rightarrow$} (equivalence).

Definition 3 - Enriched Reformulation. A reformulation Q' of a query Q is said to be enriched (*Qenriched*) if each concept (or property) C' of Q' is related to a concept (or property) C of Q by a Co correspondence, where Co ∈ {$\sqsubseteq\!\!\rightarrow$, $\sqsupseteq\!\!\rightarrow$, $\approx\!\!\rightarrow$, $\rhd\!\!\rightarrow$, $\lhd\!\!\rightarrow$, $\perp\!\!\rightarrow$}.

3. PERSONALIZATION IN *SEMREF*

SemRef mainly uses semantics underlying a set of correspondences between peer schemas to enhance query answering in dynamic distributed environments (in this case, a

PDMS). Nevertheless, the *SemRef* approach lacks to tailor query results according to the user's preferences on these semantic correspondence options. Our work – named *PSemRef* – is concerned with such task. It enables users to set the degree of relevance that underlying existing semantic correspondences have to their queries.

In *PSemRef*, the context of the users is effectively used in order to provide query personalization. As mentioned, users may state their preferences concerning the reformulation policy. These preferences are stated through the choice of four enriching variables that specify which types of semantic correspondences should be considered when a query Q is submitted.

The enriching variables are defined as follows: *Approximate* – which enables the use of *isCloseTo* correspondence; *Specialize* – which enables the use of *isSuperConceptOf* correspondence; *Generalize* – which enables the use of *isSubConceptOf* correspondence; and *Compose* – which enables the use of *isPartOf* and *isWholeOf* correspondences.

Users can set the priority order in which the chosen semantic correspondences are to be applied, resulting in a ranked set of expanded answers.

Query answers obtained by *PSemRef* as well as the priority ranking set are defined as follows.

Definition 4 – Set of Restricted Answers: Let Co be a semantic correspondence Co ∈ {$\equiv\!\!\Rightarrow$} between a peer schema ontology O_1 and a target peer schema ontology O_2, and let Q be a query submitted over O_1. The set of Restricted Answers to Q is the set of concepts c2 ∈ O_2 such that c2 is related to c1 of Q through the correspondence Co.

Definition 5 – Set of Expanded Answers: Let Co be a semantic correspondence Co ∈ {$\sqsubseteq\!\!\Rightarrow$, $\sqsupseteq\!\!\Rightarrow$, $\approx\!\!\Rightarrow$, $\rhd\!\!\Rightarrow$, $\lhd\!\!\Rightarrow$, $\perp\!\!\Rightarrow$} between a peer schema ontology O_1 and a target peer schema ontology O_2, and let Q be a query submitted over O_1. The set of Expanded Answers to Q is the set of concepts c2 ∈ O_2 such that c2 is related to c1 of Q through the correspondence Co.

In this sense, c2 is semantically related to c1 according to the existing correspondences between them. We are now concerned with finding the top-relevant ranked answers, according to the chosen semantic correspondence options and their underlying priority definition.

Definition 6 – Priority Ranking: A Priority Ranking PR is an ordered set of the enriching variables {R_1, … , R_n}, where n ≤ 4 which determines the generation of a ranking set of expanded answers.

This priority ranking is applied over answers from the target peer considering other types of correspondences rather than the *equivalence* one. Thus, it is defined according to the following list: {QR_1 from R_1, … , QR_n from R_n},where QR_n is the resulting set of *expanded* answers in conformance to the variable R_n, n ≤ 4. In this sense, we also define the top-relevant set of ranked answers for the users' queries as follows.

Definition 7 – Top-relevant Set of Ranked Answers: Given a query Q submitted on peer schema ontology O_1 of P_1, and reformulated on peer schema ontology O_2 of P_2, the Top-relevant set of ranked answers regards the ordered set of expanded answers in P_2 obtained according to the priority ranking PR.

Consider our running scenario of the Education domain. Suppose the following query Q_1 = Worker submitted in O_1 which asks for all people who works belonging to a university. This query is executed in *restricted mode*, i.e., it only produces an *exact* reformulation: Q_2 = Worker in O_2.

Suppose concept Worker is related to some concepts in O_2 according to the following semantic correspondences: (i) *isCloseTo* ($\approx\!\!\Rightarrow$) Student, (ii) *isPartOf* ($\rhd\!\!\Rightarrow$) ResearchProject, and (iii) *isSubConceptOf* ($\sqsubseteq\!\!\Rightarrow$) Worker. If a user sets a priority ranking PR_u as {$\sqsubseteq\!\!\Rightarrow$, $\approx\!\!\Rightarrow$}, then the *expanded* answers are presented to the user in the following order: {QR_1 from Worker, QR_2 from Student, QR_3 from ResearchProject}, where answers QR_1 and QR_2 are the top-relevant ranked answers. As QR_3 is an *isPartOf* correspondence, it is not present in the priority definition.

To clarify matters, in next section we provide some other query examples regarding *PSemRef* in practice.

4. EXPERIMENTS AND RESULTS

We have developed the *PSemRef* approach within a query submission module (implemented in Java) for our PDMS. Figure 3 shows a screenshot of the module's main window that is split into three parts: **(i)** the peer ontology area, **(ii)** the query formulation area and **(iii)** the query results area. Queries can be formulated using Sparql[2] or ALC-DL.

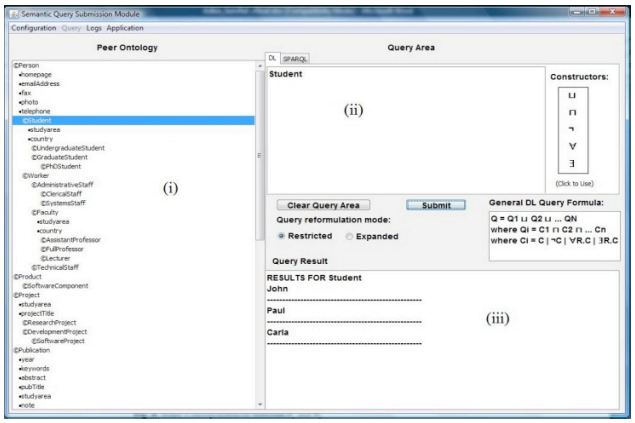

Figure 3. *PSemRef* Interface

4.1 PSemRef in Practice

We have identified a set of semantic correspondences between O_1 and O_2. Since the correspondences are unidirectional, we present examples of this set concerning the concepts Student and FullProfessor (from O_1) with some related concepts in O_2: O_1:Student$\sqsupseteq\!\!\Rightarrow$ O_2:GraduateStudent, O_1:Student$\sqsupseteq\!\!\Rightarrow$ O_2:UndergraduateStudent, O_1:Student$\approx\!\!\Rightarrow$ O_2:Worker, O_1:FullProfessor$\approx\!\!\Rightarrow$ O_2:VisitingProfessor, O_1:FullProfessor $\rhd\!\!\Rightarrow$ O_2:Course.

From this illustrative set, we have run a few query examples with Student and FullProfessor concepts from O_1 to O_2.

When we submit query Q_1 = Student without choosing any variables (i.e., an exact reformulation), only a few Student

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

concepts' identifiers are returned from O_1 and O_2. Then, choosing *approximate* and *specialize* enriching variables originates a set of expanded answers – QR_1 which is composed by some target concepts associated to GraduateStudent and Worker, still without any ranking order.

Supposing we are interested in Students again, but mainly in ones who don't work, then we set the priority ranking PR_1 as $\{\sqsupset, \approx\}$. In this way, QR_1 returns the same set of concepts originally got in *expanded mode*, but now presented in a ranking order {QR_1 from GraduateStudent, QR_1 from Worker}. Such ranking conforms to what has been established through preference variables.

Experiments guided with users showed that the use of enriching variables and priority ranking definition were very useful for their queries. Returned results for their queries were also useful, mainly because of the semantic enrichment provided by *PSemRef* (complete results) as well as because of the personalization applied to the queries (ranked answers).

5. RELATED WORK

Query ranking techniques have been tackled in some environments. The work of Koutrika and Ioannidis [5] developed a personalization framework for database systems based on the users' profiles that are created with the allocation of preferences, which determine the ranking order of the query results. Besides users' profiles, the work of Stefanidis [9] also considers contextual information, according to the location of users at query time, which influences the query ranking, depending on the user characteristics.

Query reformulation techniques have been also studied and proposed in some works [4] [1]. Necib [6] has presented an approach for query reformulation within single relational databases using ontology knowledge, to transform a user query into another query that may provide a more meaningful answer to the user.

Comparing these works with ours, in our approach we apply personalization of queries in a P2P environment, which is a very highly dynamic one. Furthermore, we take into account users' preferences at query reformulation time, providing users with a ranked set of answers related to the degree of relevance they are interested in.

6. CONCLUSIONS AND FURTHER WORK

In a PDMS, largeness and heterogeneity are common features which characterize the datasets. These peculiarities make it impractical to exactly query the data. As a result, usually massive data are provided to users. Considering that, users should be enabled to include varying degrees of relevance in their submitted queries, so that they could better specify their own needs and preferences. Furthermore, it is of fundamental importance to provide users with answers made up of related data in a significant way, and, still better, presented in a ranked order.

In this sense, this work has presented the *PSemRef* approach, which uses personalization at reformulation time, exploring the existing semantic correspondence options present in *SemRef*. We have addressed the issue of personalizing query results in such a way that users may choose which level of approximation is important to their queries. Also, the priority order in which these variables should be applied is defined.

As future work we will work to improve the graphical user interface of *PSemRef*. Furthermore, we will enrich our personalization approach taking into account users-specific context.

Acknowledgments: The work was partially supported by the National Institute of Science and Technology for Software Engineering (INES), funded by CNPQ and FACEPE, grants 573964/2008-4 and APQ-1037-1.03/08.

7. REFERENCES

[1] Adjiman P., Goasdoue F., Rousset M.-C.: SomeRDFS in the Semantic Web. Journal of Data Semantics, 8:158–181. LNCS (2007).

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

[3] Borgida A. and Serafini L.: Distributed description logics: Assimilating information from peer sources. Journal of Data Semantics, 1:153–184. LNCS 2800, Springer Verlag. (2003).

[4] Kostadinov D.: Data Personalization: an Approach for Profile Management and Query Reformulation. PhD. Thesis, Université de Versailles Saint-Quentin en Yvelines (2007).

[5] Koutrika G., Ioannidis Y.: Personalization of Queries in Database Systems. In: Proceedings of the 20th International Conference on Data Engineering (ICDE'04), pp. 597-608, Boston, Massachusetts, USA (2004).

[6] Necib B.: Ontology-based Semantic Query Processing in Database Systems. Berlin, Humboldt Universitat, PhD. Thesis (2007).

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

[8] Souza D., Arruda, T., Salgado, A. C., and Tedesco, P.: SemRef: A Semantic-based Query Reformulation Tool for Dynamic Environments. In: 24th Brazilian Symposium on Data Bases (SBBD'09), Demo Session, pp.7-12, Fortaleza, Brazil. (2009).

[9] Stefanidis K., Pitoura E., Vassiliadis P.: A Context-Aware Preference Database System. International Journal of Pervasive Computing and Communications, vol. 3, n. 4, pp. 439-460 (2007).

[10] Stuckenschmidt H., Giunchiglia F., and van Harmelen F.: Query processing in ontology-based peer-to-peer systems. In V. Tamma, S. Crane¯eld, T. Finin, and S. Willmott, editors, Ontologies for Agents: Theory and Experiences. Birkhuser. (2005).

[11] Tatarinov, I., Halevy, A.: Efficient Query Reformulation in Peer Data Management Systems. In: Proc. of the SIGMOD International Conference Management of Data, pp. 539-550, Maison de la Chimie, Paris, France. (2004).