Enriching and Answering Proteomic Queries Using Semantic Knowledges

Kunale Kudagba[1,⋆], Omar El Beqqali[1], and Hassan Badir[2]

[1] USMBA University, Computer Science Department,
P.O. Box. 1796, 30000 Fes, Morocco
{kkunale,oelbeqqali}@fsdmfes.ac.ma
[2] National School of Applied Sciences, Computer Science Department,
P.O. Box. 1818, 45000 Tangier, Morocco
hbadir@ensat.ac.ma

Abstract. Querying and sharing Web proteomics is a challenging topic in Pharmaceutical Drug Discovery and Computational Biology. Given that, several data sources can be used to answer the same sub-goals in the Global query, it is obvious that we can have many different candidates rewritings. The user-query is formulated using Concepts and Properties related to Proteomics research (Domain Ontology). Semantic mappings describe the contents of underlying sources in order to reflect their query capabilities.

In this work, we propose to enrich the user query using WordNet and we give a characterization of query rewriting problem using semantic mappings as an associated hypergraph. Hence, the generation of candidates rewrititngs can be formulated as the discovery of minimals Transversals associated with this hypergraph. We exploit and adapt algorithms available in Hypergraph Theory to find all candidates rewritings from a query answering problem. In this context, some relevant criteria could help to determine optimal and qualitative rewritings, according to user preferences, and sources technical performances.

Keywords: Proteomics, Ontology, WordNet, XML, Trees, Semantic Web, ψ-terms, Query Rewriting, minimal Transversals.

1 Problem Formalization

Given a Global Query Q and a couple of semantic knowledges $Sch'O$ made by of $O'_{proteomics}$ Ontology and a set M of all semantic mappings between proteomic sources and $O'_{proteomics}$, the Query Rewriting consists of computing two sub-queries $Q_{invalide}$ and Q_{valide} on the basis of mappings set, such as:

$$Q = Q_{valide} \vee Q_{invalide} \tag{1}$$

Explicitly, we shall calculate:

1. $Q'' = Q_{invalide}$. Sub-Query Q'' can not be answered by underlying sources, at the moment of the sending of the Global Query Q.

⋆ Corresponding author.

R. Meersman, P. Herrero, and T. Dillon (Eds.): OTM 2009 Workshops, LNCS 5872, pp. 11–12, 2009.

2. $Q' = Q_{valide}$ is the part of that will be rewritten using semantic mappings. Sub-query Q' can be answered by registered sources. Our final goal is to propose an intelligent subdivision of Q' into sub-queries Q'_1, Q'_2,..., Q'_m with $1 \leq m \leq n$,

So, we need to determine:

- all candidates rewritings expressed as:

$$Q'_{reecriture} = Q'_1 \wedge Q'_2 \wedge \ldots \wedge Q'_m \tag{2}$$

- and all partial queries Q'_j composing these rewritings and answered by a Source S_j as follows:

$$Q'_j = \bigwedge_{i=1}^{k} C_{ij} \tag{3}$$

with $k \leq m$ and k denotes number of atomic constraints C_i satisfying by the partial query Q'_j de Q' while m denotes number of atomic constraints in Q.

The algorithm receives as input a global query Q, a schema $Sch'O$ and generate as output a set of all candidates rewritings $r_i(Q)$.

2 Rewriting Algorithm

The algorithm runs like that:

1. From a rewriting query problem, we need to give a mathematical characterization, by defining an associated Hypergraph $H_{Q,M}(V, E)$:
   - For every mapping m_i , describing a local concept from M, as a logical function of $O'_{proteomics}$ global concepts, we associate a vertice V_{m_i} in the hypergraph $H_{Q,M}(V, E)$ and $V = \{V_{m_i},$ with $1 \leq i \leq n\}$.
   - For every constraint C_i of the Global query Q, we associate an hyperedge E_{C_i} in the hypergraph $H_{Q,M}(V, E)$. To simplify, we suppose that all these constraints are describing atomics goals. So, each hyperedge E_{C_i} is a set of mappings, calculated by considering those mappings which are relevant to answer these goals.
2. From this associated Hypergraph, we generate its minimal Transversals, correponding to all candidates rewritings.
3. Ranking of Candidate rewritings and Selection of best ones according to criteria specified by an online Biologist.

3 Conclusion

This paper shows briefly our current research Work that aims to provide a semantic-driven, user-centric and scalable framework integrate and query XML Proteomic Sources on the Web.

A Test realized according to a scenario of six sub-queries and Three semantic mappings allow us to find 36 quadruplets, 6 Transversals but only 2 are minimals.