Query Rewriting Based on Semantic Agreement in P2P Environment

I Wayan S. Wicaksana, Lily Wulandari, Farah Virnawati, Tirta Paramitta, Wisnu Sukma M.

Abstract
Gathering information from many sources brings some challenges. One of the main challenges is how to write an appropriate query to the sources. Query Rewriting is an interesting research since traditional database. In this paper, query rewriting approach in P2P environment which has Common Ontology, Local Schema and Mapping will be discussed.

1. Introduction

Recently, the focus of research on integrated information systems has shifted to the definition of methodologies, architectures and tools to effectively manage and share data in heterogeneous distributed environments. Large volume of data various formats increasingly accessible in the web, including web pages, semi-structured documents (XML, RDF, etc.) and spatially referenced data. The need for sharing data stems from (1) the explosive growth of the web and the ability to interconnect a growing number of information sources, (2) the increasing availability of autonomous data sets, and (3) rising acquisition costs of complex non-traditional data.

Based on the architecture, there are two different kinds of systems: central data integration systems and peer-to-peer (P2P) data integration systems. A central data integration system usually has a global schema, which provides the user with a uniform interface to access information stored in the data sources. In contrast, in a peer-to-peer data integration system, there are no global points of control on the data sources or peers. Instead, any peers can accept user queries from the system, but it does not guaranty to provide appropriate answer.

Data sources can be heterogeneous in syntax, schema, or semantics, thus making data interoperability is a difficult task. Syntactic heterogeneity is caused by the use of different models or languages. Schematic heterogeneity results from structural differences. Semantic heterogeneity is caused by different meanings or interpretations of data in various contexts. To achieve data interoperability, the issues posed by data heterogeneity need to be eliminated.

The Ontology based on the approach toward the problem in data interoperability, especially focused on the problem of query process in setting the heterogeneity P2P. Pure P2P approach will use closest

---

1 Center for Information System Study, Gunadarma University, iwayan@sta®.gunadarma.ac.id, lily@sta®.gunadarma.ac.id, virtha1408@student.gunadarma.ac.id, eatha 020688@student.gunadarma.ac.id, wisnu sm@student.gunadarma.ac.id
neighborhood, so to send a query, a peer need to do many times before get to the right sources [7]. Query processing inside the data integration system, focused on LAV and GAV approach [5]. The selection is about to use LAV or GAV method, so the process need more computation cost.

Our work is focused on query rewriting in hybrid P2P environment. In hybrid model, we do not need global schema but utilize common ontology in an appropriate domain which use as pivot point. We study the query rewriting to sources by considering pivot concept of request Local Schema of Sources and Ontology of a domain. Set of mappings between the source schemas and the target schema by using the common ontology can utilize for query rewriting.

2. Approach

There are two approaches to view-based query-processing, called query rewriting and query answering, respectively. In the former approach, we are given a query and a set of view definitions, and the goal is to reformulate the query into an expression, the rewriting, that refers only to the views and provides the answer to the query. Typically, the rewriting is formulated in the same language used for the query and the views but in the alphabet of the view names, rather than the alphabet of the database. In query rewriting, query processing is divided in two steps, where first re-express the query in terms of a given query language over the alphabet of the view names, and the second evaluates the rewriting over the view extensions.

To do the query rewriting, there are several variables to figure out the query process. In this case, which utilize variables as follow:

\( Q_U \) = Query which is submitted by user/request;
\( M_U \) = The mapping between user and Common Ontology;
\( Q_0 \) = Query which is resulted from Mapping user \((M_U)\) to Common Ontology \((Q_0 = Q_U + M_U)\),
\( M_N \) = The set of mapping between Common Ontology and the source \(N\) \((M_N = M_1 + M_2 + ... + M_n)\),
\( Q' \) = Query Rewriting based on Mapping \(M_N\) \((Q' = Q_0 + M_N)\)

**Traditional Approach**

In sharing data by traditional method as Figure 1, we use many \( M_U \) for all sources. Because the system must map as many times of \(N\) number of sources to submit query to sources.

![Figure 1: Traditional Approach](image-url)
\begin{align*}
Q_1 &= Q_U \rightarrow M_{US1} \rightarrow Q_1' \\
Q_2 &= Q_U \rightarrow M_{US2} \rightarrow Q_1' \\
Q_3 &= Q_U \rightarrow M_{US3} \rightarrow Q_1' ; \text{ etc}
\end{align*}

**P2P Approach**

In sharing data by P2P approach we just use once MU for all sources as Figure 2. Meanwhile, by using P2P, we will use:

![Figure 2: P2P Approach](image)

- \( Q_1 = Q_U \rightarrow M_U \rightarrow Q_0 \rightarrow M_1 \rightarrow Q' \)
- \( Q_2 = Q_U \rightarrow M_U \rightarrow Q_0 \rightarrow M_2 \rightarrow Q' \)
- \( Q_3 = Q_U \rightarrow M_U \rightarrow Q_0 \rightarrow M_3 \rightarrow Q' ; \text{ etc} \)

From the explanation above, query writing to many sources is conducted just one mapping query from user to the Common Ontology and from Common Ontology view to sources, the query rewriting will utilize the mappings to Common Ontology which have previously created by every source.

From description above we can see, that with traditional P2P approach we must use more cost to rewrite query for many users to many sources.

**Mapping need** : Number of Request x Number of Sources.

However, for our Approach we can reduce cost to develop mapping. **Mapping need ; Number of Request + Number of Sources**

The other advantage, the approach based in Common Ontology of a community that can reduce miss understanding of concept during query rewriting process.
3. Running Example

Assume that an user want to find information about the name of people who live in Jl.Thamrin, then the query which is submitted by an user is Person with Street = "Jl.Thamrin", and represented as:
- \( Q_U = \text{Person, Street} = "\text{Jl.Thamrin}" \);

Assume the mapping between User and Common Ontology (see Figure 3) as the following:
- \( M_U = \text{Person} \approx \text{Person; Street} \approx \text{Address} \)

Then the query result which is got from Common Ontology is:
- \( Q_O = \text{Person, Address} = "\text{Jl.Thamrin}" \);

Assume Mapping between Common Ontology and Source:
- \( M_I = \text{Person} \approx \text{Staff; Address} \approx \text{Location} \);
- \( M_2 = \text{Person} \approx \text{Staff; Address} \approx \text{Street} \); etc

Then the result of query:
- \( Q' = \text{Staff, Location} = "\text{Jl.Thamrin}" \rightarrow Q_I \)
- \( Q' = \text{Staff, Street} = "\text{Jl.Thamrin}" \rightarrow Q_2 \)

So Query which will be got by user is \( Q' = Q_I + Q_2 \). As Figure 4.
4. Conclusion

In a peer-to-peer data integration system, there are no global points of control on the data sources (or peers). Each peer can accept user queries for the information distributed in the whole system with different source. In query rewriting of traditional approach, a query and a set of view definitions over the global schema are provided, and the goal is to reformulate the query into an expression, the rewriting, that refers only to the views and supplies the answer to the query.

With P2P Approach we can reduce cost as much as possible to do query rewriting to many Sources. It because P2P has Common Ontology that can map Query to the sources. With this P2P approach, we will also easily changing or adding in Source, because we don't have to change all Mapping every query that submitted by user. But we just change or adding Mapping between Common Ontology and Sources.

Even though the approach which we use it can reduce the cost which is used to do query rewriting, but it is still to be simulated and developed more. Further, we will try to develop query rewriting automatically and consider to changing of Common Ontology and Local Schema.

References


