Query Rewriting Based on Semantic Agreement in P2P Environment

I Wayan S. Wicaksana, Lily Wulandari,
Farah Virnawati, Tirta Paramitta, Wisnu Sukma M.
Universitas Gunadarma
{iwayan, lily}@staff.gunadarma.ac.id , {virtha_1408, eatha_020688,
wisnu_sm}@student.gunadarma.ac.id
www.gunadarma.ac.id

Abstract:
Collecting information from many sources brings some challenges. One of the main challenge 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 Semantic Agreement will be discussed.

Keyword : Data Integration, Ontology, Peer to Peer, Query Rewriting, Semantic Agreement

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 distributed data 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 closet neighborhood to send a query, a peer need to do many times before get to the right sources [3]. Query processing inside the data integration system, focused on LAV and GAV approach [1]. 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 appropriate domain which act as pivot point. More concretely, we study the problem of efficient answering of queries through a target schema, given a set of mappings / Semantic Agreement between the source schemas and the target schema by using the common ontology.

2 Approach

There are two approach 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, the method utilizes some variables as follow:

Q_U = Query which is submitted by user ;
M_U = The mapping between user and Common Ontology;
Q_O = Query which is rewrotten based on Mapping of user (M_U) to Common Ontology ($Q_O = Q_U + M_U$);

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 of sources (Semantic Agreement) M_N (Q' = $Q_O + M_N$)

Traditional Approach

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

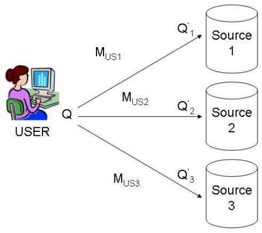

Figure 1: Traditional Approach

- $Q_1 = Q_U \rightarrow M_{US1} \rightarrow Q_1$'
- $Q_2 = Q_U \rightarrow M_{US2} \rightarrow Q_2$'
- $Q_3 = Q_U \rightarrow M_{US3} \rightarrow Q_3$' ; etc

P2P Approach

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

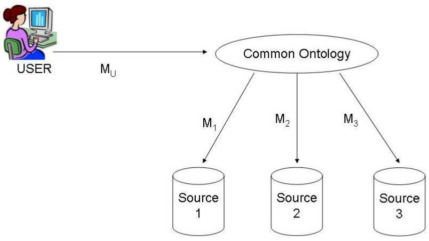

Figure 2: P2P Approach

- $Q_1 = Q_U \rightarrow M_U \rightarrow Q_O \rightarrow M_1 \rightarrow$ Q'
- $Q_2 = Q_U \rightarrow M_U \rightarrow Q_O \rightarrow M_2 \rightarrow$ Q'
- $Q_3 = Q_U \rightarrow M_U \rightarrow Q_O \rightarrow M_3 \rightarrow$ Q' ; 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 which called Semantic Agreement.

Result :
For Traditional Approach must create M_U as much as N Sources and Create Q' based on the M_U.
- Cost of Traditional Approach : X User + X.Y M_U + X.Y Q' ;
- Cost of P2P Approach : X User + X M_U + Y Q' ;
where X : Number of User , and Y : Number of Source ;

From description above we can say (i) traditional approach we must use more cost to rewrite query for many users to many sources. and (ii) however, P2P Approach we can reduce cost to rewrite query in significant number of process.

3 Running Example

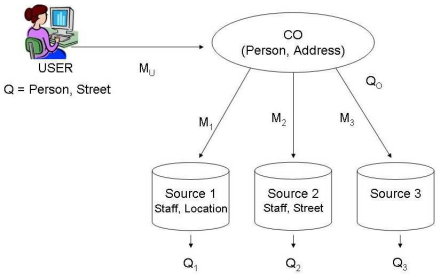

Figure 3: P2P Query Model

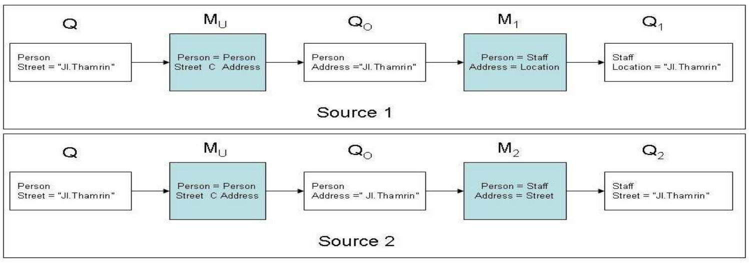

Figure 4: P2P Query Rewriting Process

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 = Person, Street = "Jl.Thamrin" ;

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

Then the query result which is got from Common Ontology is :
- Q_O = Person, Address = "Jl.Thamrin" ;

Assume Mapping between Common Ontology and Source :
- M_1 = Person ≈ Staff ; Address ≈ Location ;
- M_2 = Person ≈ Staff ; Address ≈ Street ; etc

Then the result of query :
- Q' = Staff, Location = "Jl.Thamrin" → Q_1
- Q' = Staff, Street = "Jl.Thamrin" → Q_2

So Query which will be got by user is Q' = $Q_1 + Q_2$. See 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 do not have to re-create all Mapping between Common Ontology to sources for every query that submitted by user. However, we just create Mapping between Common Ontology and user.

Even though the approach can reduce the cost in concept, but it need to be simulated and developed close to real condition. Further, we will develop query rewriting automatically based on Semantic Agreement which consider the changing of Common Ontology and Local Schema.

References

[1] Domenica Lembo Diego Calvenese and Maurizio Lenzerini. Survey on method for query rewriting and query answering using view. www.dis.uniroma1.it/ lembo/D2I/Prodotti/deliverable/D1.R5.ps, September 10th 2007.

[2] Christop Koch. Query rewriting with symmetric constraints. www.lfcs.inf.ed.ac.uk/research/database/publications/aicom September 10th 2007.

[3] Huiyong Xiao and Isabel F. Cruz. Ontology-based query rewriting in peer to peer network. www.cs.uic.edu/ advis/publications/dataint/ickeds06.pdf, September 10th 2007.

[4] Cong Yu and Lucian Polpa. Constraint based xml query rewriting for data integration. www.eecs.umich.edu/ congy/work/sigmod04.pdf, September 10th 2007.