Data Quality-Oriented Data Integration in Peer-to-Peer System

1 Author(s)

Zhigang Zhao

View All Authors

Abstract:
In current applications, P2P system has become a robust environment for data integration. But the data sources of P2P system are autonomous and heterogeneous, which increase the difficulty of data integration. Current researches focus on how to manage information of data resources and ensure the data quality of data integration in P2P system. This paper presents a quality-oriented data integration model. In this model, we mainly deal with data integration of databases resources. We propose a structure to manage the schema information in P2P system. Then, we build a model for evaluating their data quality to select data sources the query of data integration, which is important for improving the quality of data integration. Finally, experimental results have shown that our quality-oriented data integration can provide higher quality of integration result and even lower time cost.

Published in: Hybrid Intelligent Systems, 2009. HIS '09. Ninth International Conference on

Date of Conference: 12-14 Aug. 2009

INSPEC Accession Number: 10891294

Date Added to IEEE *Xplore*: 22 September 2009

DOI: 10.1109/HIS.2009.199

ISBN Information:

Publisher: IEEE

SECTION I.
Introduction

P2P system can provide both efficient executing environment and mass shared data sources for data integration applications. In P2P environment, data sources are all autonomous and heterogeneous; this is the most obviously different between P2P system and federal database system. And, the data quality of result has become an important requirement in data integration. Data sources are often autonomic, heterogeneous, dynamic, which increase the complexity of data integration. So how to manage data sources and seamlessly integrate these distributed data with high quality has been a challenging issue in P2P system.

To improve the quality of integration, there are some problems must be discussed. Firstly, comparing with traditional data integration, the metadata of data source in P2P system is more complex, since the schema is heterogeneous. Secondly, the requirement of data integration is dynamic, so we need dynamically select data sources in P2P system for data integration. Moreover, a discovering mechanism for data sources is also required based on the structure of P2P system. Thirdly, query optimization techniques for traditional data integration can not nontrivially and directly be applied to P2P environment, because there are some fundamental differences between them. There are also other issues in data integration, such as security, but we focus on issues that are related with data quality in this paper.

To address these problems, we present a quality-oriented data integration model for the purpose of implementing the data integration on demand and improving the data quality of integrated result in this paper. So, we focus on resolving the following issues of data integration in a P2P system:

The structure of P2P system, and the mechanism of data sharing;

Build a model to evaluate the data quality of data sources;

Match the capabilities of data sources with the requirements to select data sources for data integration;

Guarantee the quality of the data integration.

We first propose a multiple layer structure for p2p system to register data sources and manage metadata information of data sources in global. Then we build the model of data quality evaluation based on schema of data source and statistical character of data.

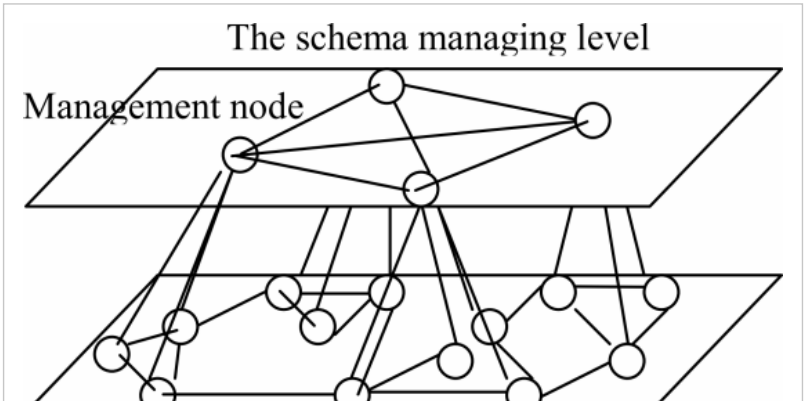
The rest of this paper is organized as follows. Section 2 presents related works. Section 3 describes the architecture of P2P system. Section 4 presents the evaluation model of data quality for data sources, while in Section 5, the algorithm for data source selection is proposed. Section 6 discusses experimental results on data quality-oriented data integration model. Finally, Section 7 summarized the conclusions.

SECTION II.
Related Work

Recently, data integration has become a popular research work in p2p area. In the work of Calvanese[1], [3], [7], they compare the commonly adopted approach of interpreting peer-to-peer systems using a first-order semantics, with an alternative approach based on epistemic logic. And Lenzerini[2] have proposed the Principles of peer-to-peer data integration and Quality-aware peer-to-peer data integration [4], in which they study the problem of data integration in peer-to-peer systems, with the aim of singling out the principles that should form the basis for the design of data integration systems in this architecture. Some research works[9] study the framework of data integration that aim at developing principles and techniques for peer-to-peer data integration on a Grid infrastructure. In p2p system, query is an important operation for integration, and Leopoldo Bertossi and Loreto Bravo [5] have studied the problem of answering queries posed to a peer who is a member of a peer-to-peer data exchange system. Moreover, an algorithm of robust data sharing and updates in p2p database networks is also proposed [6]. Now, more works concentrate on Semantic Data Integration in P2P Environment [8], [10]. In this paper, we study the Data Integration in P2P system on the data quality.

SECTION III.
The Architecture of Peer-To-Peer System for Data Integration

To support data quality-based data integration, we propose hierarchy architecture of p2p system with data sharing model. In our data integration, we mainly focus on dealing with database resources. The architecture is shown in Figure 1.



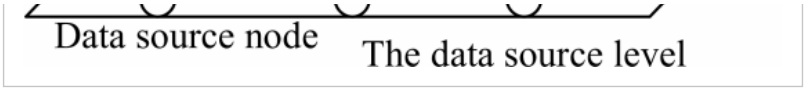


Figure 1.
The architecture of p2p system

In the architecture, we separate the nodes of p2p system into two levels: the schema managing level and the data source level. Nodes in schema managing level are used to manage the schema information and other metadata of data sources, so it is called management node. While nodes in data source level are providing data services and also submit data integration requirements, so it is called data source node. Each management node maintains a group of data source node by storing their schema and metadata. And each data source node is only belonging to one management node, and sent its schema and metadata. Management nodes communicate with each other for sharing schema and metadata of data source, while data source nodes communicate with each other to transfer data for data integration.

For data integration under this architecture, the requirement of data integration is firstly submitted to one of data source node. Then, the data source node transfers the requirement to its management node, which is called the start management node.

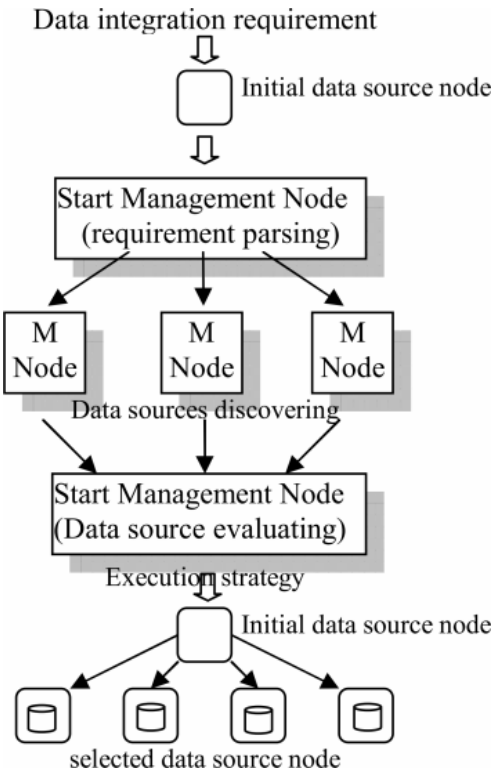


Figure 2.
The workflow of data integration

In the management node, the requirement of data integration is parsed into two information types. The first type is schema information, in which the schema of integrated data is defined with a set of sub schema, since there are multiple data sources needed in data integration. The second type is data quality requirement, which defines the Service Level Agreement (SLA) the result data need satisfied.

Then the management node sends the requirement to other management nodes, which search local data sources to discover data sources satisfied the requirement. After local data source discovering, management nodes submit information of matched data sources to the start management node. The start management node uses the evaluation model to select a set of data sources for executing the data integration and generate an execution strategy of data integration.

At last, the execution strategy of data integration is returned to the initial data source node to execute data integration. Finally, the initial data source node response the result to user.

SECTION IV

The Evaluation Model of Data Quality for Data Integration

In the process of data integration, the most important step is data source evaluating which determines the data quality of result. So the evaluation model is proposed in this section.

In our evaluation model, the most important part is the schema of data source, which provides integrality. And other metadata of data source describes the data quality of the data source. Therefore, we separate the evaluation model of data quality into two parts: the schema quality of data source and the service quality of data source.

The Schema Quality

Here, the schema quality, denoted as Qschema, is the matching degree between schema of data source and schema of data integration. The schema quality indicates the functional matching result, which means whether the data source is useful for the data integration.

Given a requirement of data integration, we denote Sr as the set of relation in requirement. For each Ri in Sr, we denote Ri as:

$$\mathrm{Ak} = \{\mathrm{Ak}, \mathrm{As}\}$$

View Source

Where Ak is the set of key attributes in integration, and As is the set of selective attributes.

We denote the schema of data source as Sd, and R'i as the relation in Sd. Then we compare each relation R'i in Sd with each relation Ri in Sr.

Definition 1.
The similarity $\mathrm{Sim}(\mathrm{Ri}, \mathrm{R'j})$ between two relations is defined as:

$$Sim(R_i, R'_j) = w_k \times M(A_k, A_d) + w_s \times M(A_s, A_d) \quad (1)$$

View Source

Where, Ad is the set of attribute in relation R'i, and M(A, Ad) is the recall of A in Ad. For example, there are five attributes in A, and only three of attributes in Ad matches with them. so the recall is 0.6. wk and ws is the weights of key attributes and selective attributes. We can adjust two values to satisfy different requirements.

Then we further define the formula to calculate the schema quality.

IV. Definition 2

The schema quality $\mathcal{Q}_{schema}$ is defined as:

$$Q_{schema} = \sum Max(Sim(R_i, R'_j))/count(S_r) \quad (2)$$

View Source

Where $Sim(R_i, R'_j)$ must be satisfied $Sim(R_i, R'_j) > \delta$ and δ is a threshold to determine the match relation between R'_i and R_i. For a given R_i, if there is no R'_j satisfied $Sim(R_i, R'_j) > \delta$ the value of $\max(Sim(R_i, R'_j))$ is 0.

The Service Quality The service quality, denoted as Qserv, is a general score for data source on parameters of data source. Here we just consider some important parameters for data quality.

The first parameter denoted as Pscale is the scale of relation in data source, which indicates how many potential records can be provides and impact the integrity integration result.

The parameter Pvalue is the value quality of relation, and can be calculate as the missing value rate in records of relation.

The parameter Pstat is the quality of relation on some statistic characters. For example, the number of different values in key attributes will influence the number of join result in data integration.

Based on above parameters, we define the formula of service quality.

IV. Definition 3

The schema quality Qschema is defined as:

$$Q_{serv}(R) = w_{sc}(P_{scale}/\max(P_{scale}) + w_v P_{value} + w_{st}(P_{stat}/\max(P_{stat})$$

View Source

Where max(Pscale) is the maximum scale among all candidate data source, and max(Pstat) is the maximum number of different values in the same key attribute. Wsc, wv and wst is three weight of these parameters and can be changed by different data integration applications.

So, we finally define the total data quality scoring of relation is denoted as:

Wsc Qschema +Wse Qserv,

Where, Wsc and Wse are the weight of schema quality and service quality in data source.

SECTION V. Algorithm for Data Source Selection

Based on the evaluation model, a great deal of data source are returned to the start management node, though is only transfer the requirement to part of management nodes.

So there are lots of execution strategies in solution space SS. In this section, we propose the selection algorithm to filter data sources for data integration. Generally, we consider the integrated result with high data quality should satisfy the following characters:

There are less reduplicate records and more records;

Integrity schema in result data;

No missing values of attributes in records.

To reach these results, we select data sources of integration following some heuristic rules, in which the most important two is:

Select data source with more records. This rule is not only helpful for ensuring the scale of result data, but reducing the reduplicate records, since it need less data source in data integration.

Select data source with more integrity schema. This rule ensures the integrity of result data and less data sources to be accessed.

Then we propose an algorithm to select integrating solutions. The main steps of algorithm work as follows.

SECTION Algorithm 1: Data Source Selecting Algoritm

Input: Candidate data sources $S\{S_1, S_2, \ldots, S_n\}$, SS=∅
Output: *ES(execution strategy)*
Begin:

Sort (S) by Q_{schema} and P_{scale};
Filter (S) by δ_{schema} and δ_{scale};
While $(SS<n)\{$
  While $(S_{ES} < S_r)\{$
    get Si from S and combine$(Si, S_{ES}) = S'_{ES}$;
    if $(S'_{ES} > S_{ES})$ $S_{ES} = S'_{ES}$;
  $\}$
Add S_{ES} into SS;
$\}$
Evaluate S_{ES} in SS by record scale;
Return $Max_{scale}(S_{ES})$
end

In algorithm, we first sort data sources by schema and record scale, then we composite solution to satisfy requirement with less data source. When we get a set of candidate solutions, we evaluate them by the scale of data to select an optimized one. Since finding the optimized solutions is a NP-hard problem, our algorithm is used to discover a better solution for data quality of integration.

SECTION VI.
Experiment Results

We design a series of experiments for measuring the efficiency of the data quality-based data integration model. Experiments were executed on a simulation environment. Data source in p2p system were using multiple databases on Oracle 9i that were deployed on several servers with different schema of data. We also insert different scale of data into these databases. Each of databases is composed of fifteen relations. The scale of relation is from 1,000 records to 300,000 records.

In the experiment, we set the data integration with different number of relations in its schema, and select one or two solution to execute integration. We compare the total number (T) of result records and the number of reduplicate records(R) between our data quality-base method (DQ) and the simple schema-based tnethod(S). The experimental result is shown in figure 3, where our data quality based method can receive more records and less reduplicate records in the data integration.

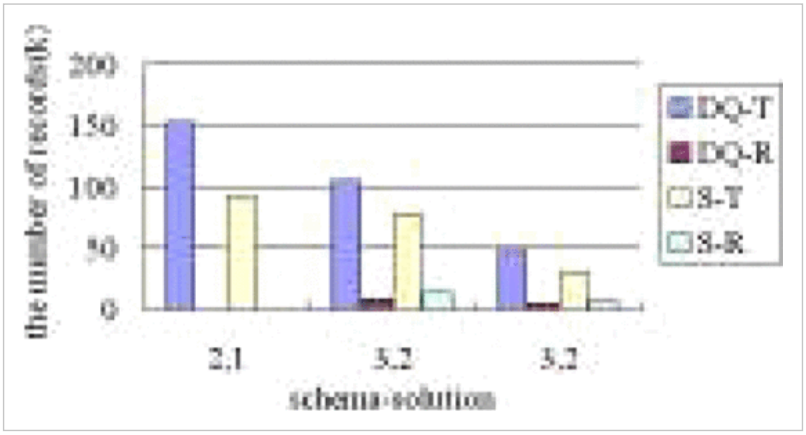


Figure 3.
The result of experiment

SECTION VII.
Conclusions

In this paper, we propose a data quality-based method for data integration in p2p system. We first present the architecture of p2p system with data sharing model is first proposed to support data integration. Then, the evaluation model of data quality for data sources is presented in detail. The

integration. Then, the evaluation model of data quality for data sources is presented in detail. The algorithm of data source selection is proposed based on evaluation model. Finally, the experiments have demonstrated our approach for the data integration worked more efficient on improving data quality in p2p environment. In the future of our work, we will study the optimization in data integration.

Keywords

IEEE Keywords
Peer to peer computing, Hybrid intelligent systems

INSPEC: Controlled Indexing
peer-to-peer computing, data handling, database management systems

INSPEC: Non-Controlled Indexing
databases resources, data quality-oriented data integration model, peer-to-peer system, P2P system, information management

Author Keywords
Algorithm, P2P, Quality-Oriented, Evaluation

Authors

Zhigang Zhao
Shenyang Normal Univ., Shenyang, China

Related Articles

Explicit construction of optimal exact regenerating codes for distributed storage
K. V. Rashmi; Nihar B. Shah; P. Vijay Kumar; Kannan Ramchandran

Privacy preserving social networking through decentralization
Leucio Antonio Cutillo; Refik Molva; Thorsten Strufe

2Fast : Collaborative Downloads in P2P Networks
P. Garbacki; A. Iosup; D. Epema; M. van Steen

A Data Reception Method to Reduce Interruption Time in P2P Streaming Environments
Suguru Sakashita; Tomoki Yoshihisa; Takahiro Hara; Shojiro Nishio

Bandwidth Trading in Unstructured P2P Content Distribution Networks
K. Eger; U. Killat

SeAl: managing accesses and data in peer-to-peer sharing networks
N. Ntarmos; P. Triantafillou

Comparison of Robust Cooperation Strategies for P2P Content Distribution Networks with Multiple Source Download
D. Schlosser; T. Hossfeld; K. Tutschku

LOADER: A Location-Aware Distributed Virtual Environment Architecture
Behnoosh Hariri; Shervin Shirmohammadi; Mohammad Reza Pakravan

Autonomous Data Replication Using Q-Learning for Unstructured P2P Networks
Sabu M. Thampi; K. Chandra Sekaran

Centralized P2P Streaming with MDC
Ivan Lee; Yifeng He; Ling Guan