YEARBOOK DATA INTEGRATION BASED ON COMMON WAREHOUSE MODEL

Gaimei Miao and Juanqiong Gou
School of Economics and Management, Beijing Jiaotong University, Haidian District, 100044 Beijing, China

Keywords: Metadata, Integration, Yearbook.

Abstract: This paper summarizes the development of data integration and the present situation of the development in domestic and abroad. It also analyzes the yearbook data, find that different years, different regions of the data has very big differences in the structure, name, dimension, data type, and other aspects, and the current data integration technology cannot be effectively integrate yearbook data. However, metadata exactly can well solve the problems, using metadata not only can realize the effective integration about local yearbook data, but also can achieve different platforms, different subsystem, different software sharing yearbook data. So, metadata is helpful for effectively improving yearbook data utilization rate.

1 A SUMMARY OF DATA INTEGRATION

The worldwide Data Integration project index drawn by Patrick Ziegler, as shown in figure 1, describes the development of data integration.

The traditional data integration technology mainly includes many multiple database System and the federal database System. With the development of the distributed network technology, in order to increase the processing of Web data and semi-structured data and integrate data sources which have new forms, new technologies occurred , such as the data integration system based on agent , the integration technology based on ontology, etc. As web service technology developed, researchers launched a research based on the web service integration technology.

Compared with the overseas research, domestic research about the data integration technology started relatively late, however, it is developed very quickly. Relatively speaking, the researches that southeast university computer science and engineering researchers did was earlier that they developed versatile based on a distributed heterogeneous data sources integration system prototype — CORBA, aiming at integrating data from different data sources in the way of playing at the same time. Chinese people's university researchers paid more attention to doing research on the question of Web enquires and semi-structured data model.

To sum up, at present both at home and abroad, there is as yet a promising potential and development space in the research of the integration of heterogeneous data.

The mentioned representative research works as above, almost all didn't refer to the data in-consistency solution. And metadata can well solves the problem; therefore, here we did a research on how to achieve statistical yearbook data integration by metadata, in order to achieve better results.

2 ISOMORPHIC MANAGEMENT OF THE YEARBOOK DATA

The article fully analyzes the different years, different regions of the Yearbook, and thinks about metadata management strategy. Finally it definite the yearbook with four metadata (Yearbook, Special Topic, Report, Statistics Field), they include the relationship between layers. That is a yearbook includes a number of topics, a topic includes several reports, a report includes a lot of Statistics Field.

Meanwhile, it further refines the various metadata. it uses the metadata (Yearbook ID, Yearbook Name, Type, Start Date, End Date, Publisher, Date of Purchase, Note) to define Yearbook; uses the metadata (Special Topic ID,

Miao G. and Gou J..
YEARBOOK DATA INTEGRATION BASED ON COMMON WAREHOUSE MODEL.
DOI: 10.5220/0003586205690573
In *Proceedings of the 13th International Conference on Enterprise Information Systems* (PMSS-2011), pages 569-573
ISBN: 978-989-8425-56-0

Figure 1: Data Integration project index.

Belonged Yearbook ID, Chinese name, English name, notes) to define Special Topic; uses the metadata(Report ID, Belonged Special Topic ID, Report Chinese name, Store name, notes) to define Report; uses the metadata (SF_ID, Belonged Report ID,SF-Name, SF-S-name, SF-Unit, SF-Level, SF-Row-span, SF-Column-span, SF-Extend, notes) to define Statistics Field. Through these definitions, it can make the reports with same structure.

In order to realize isomorphic management of the yearbook data from different years, different regions, transforming the tables with different levels, different structure into the two-dimensional structure and storing it in Oracle database is critical. In the integration of the yearbook, it uses the rows of spanning, the columns of spanning, the rows of extensions to solve this problem, and has achieved good effect.

3 MAPPING OF THE YEARBOOK DATA

In order to complete mapping, it need to get out relevant field from mass data table, then store it in one or a few tables, as figure 2 shows.

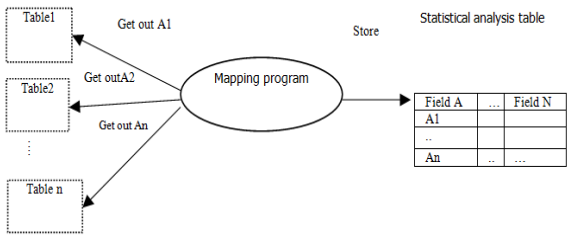

Figure 2.

It is too fussy to put forward extract field from many tables one by one, therefore it needs to be dealt with it in a structural processing, when unstructural fields switch into structural fields, we need to deal with the following conflicts:

(1) Naming conflicts

Description: if two elements E1 and E2 express the same entity, but different names, it is happened the conflicts of naming synonym. When the elements E1 and E2 have different names but means different entity or concept， it will occur the conflict of homonymy. For example, using permanent population at the end of years in the table T1 means the number of permanent population; at the table T2, using the total population expresses the number of permanent population, then it will occurs the conflict of class name synonymous between long-term populations at the end of the years. Absolute (number ／all workers) means the absolute number of the workers' salary in the table A1; Absolute (number/ all workers) means the absolute average wages in the table A2. The conflict of class name homonymy word is happened.

The forms of relational model conflicts: Using class in the relation bag represents entity or concept, so the name conflicts in the model performance class names with the same form and synonymous conflicts.

Solving strategy: the Abnormity synonyms conflicts tip the synonyms checked by HUB according to input the synonyms by users; Homonymy conflicts find the conflict according to matching class-name users choose the method to solve the conflicts

(2) Date type conflicts

Description; suppose that A1 and A2 describe the character of the same entity but have different date model A1 and A2 have date model conflicts. For example, the date model of agricultural output is plastic in the table 1. The date model of agricultural output is character type in the table2.

The form of the conflicts of relation model; relevance of column and date type expresses the date type of a field. So the conflict is happen, when data type of relevant field is different.

Solving strategy: Matching means the same character of column, which is relevant for the date type, users can choose date type.

(3) The conflict of Data dimension

Description: Suppose A1and A2 describe the same features of the same entity, but date dimension is different between A1and A2, they exist the conflict of date dimension. Such as using meter expresses height in the table 1; using inch expresses height in table 2.It cannot know the date dimension in the date dictionary. That is to say, database metadata does not provide dimension of semantics.

The form of conflict model; there is no the express way of the date dimension, when model integration, it does not check the date dimension. Suppose that users understand the conflicts of the date dimension in the two tables, user solve it by herself.

(4) Numerical range and Precision conflict

Description: related objects equivalent data elements have different range and accuracy settings.

For instance, in the table T1, Agricultural output value of the unit price is six-figure, the two-figure behind point, such as 1000.82;in the table 2,the unit price of the total agriculture production is five- figure, the behind of the point is one- figure, such as 1200.5.

The form of the relation model conflict; attribute setting in the relation bag express the scope and precision which is list in the date base, the conflicts of scope and precision in the model express the inconformity of the attribute setting of column.

Solving strategy: users make sure the scope and precision according to the need of statistical analysis, and get rid of noise data.

(5) The description of constraints conflict

Description: related objects equivalent data elements have different examples constraint. Such as the age of the adult in the tableT1 must be over 18 years old, and it is above 20 years old in the table T2.

The form of the relation model conflict; the relevant for the element in the relation bag and constraint is based on the constraints of the element. Such as the relevance of the column and constraint, the attribute of constraint .body is above 18 years old .which means examples must be above 18.In the model, the constraints conflict is the same element of the expression conflict which is based on the constraint.

Solving strategy: matching the same element, which is relevant for the constraint, if the relevant expression is different, users decide whether it has constraints conflict and conflict resolution or the conflict solved by user is only the constraints conflict which is possible to occur. The specific expression meaning is solved by users.

(6) The primary key conflicts

Description; Established in related objects of different only marks. Such as, the primary key is numbers which is in the table T1 (numbers, years, trade …),the primary key is trade in the table T2

(numbers, years, trade …).

The form of relation model conflict; the relevance among tables, columns, primary keys is used in the relation bag, which marks a column as primary key. Primary key conflict in the model is built up the relation among the same table, different columns, and primary keys.

Solving strategy: we should check column which is related with primary key, if column is different, users choose the way of solving conflict.

(7) Structural conflicts

Description; the same entity or concept use different representation methods, one represents entity, the other one represents features. Such as table T1 (numbers, years, animal husbandry and fishery output…) represents animal husbandry and fishery output; it is the attribute of table T2 (numbers, years, absolute number / (animal husbandry and fishery index) (last year is equal to 100/), animal husbandry and fishery).

The form of relation model conflict; table in the relation bag represents the table which is in the date base, column represents line which is in the date base, the structural conflict in the model express table which is in the model of address, in a model address is column.

Solving strategy: matching table class name and column class name which is under the different table ,we can find out the possible structure conflict, but the complexity of the class name across the namespace matching time is big, and there exists a lot of similar name which is under the different namespace, HUB submit mass of the possible conflicts, which is estimate by users .however, the structural conflict is seldom, so the benefit it brings is little, therefore we do not check the structural conflict. Suppose that users realize the structural conflicts of two tables, they solve structural conflict, otherwise there are too many date in the integrated table.

4 FACING THE HETEROGENEOUS PLATFORMS YEARBOOK DATA SHARING

Unified metadata standard is the most need of Yearbook data of cross-platform sharing, and the CWM (Common Warehouse Model), which the international object management group proposed to, is just such kind of metadata exchange standards in the fields of data warehouse and business analysis, which realizes the exchange of metadata in different software tools, as figure 3 shows.

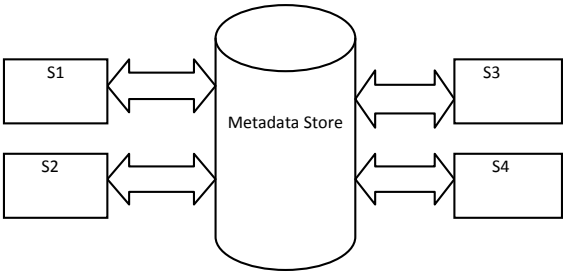

Figure 3.

CWM is in fact a kind of exchange technology, aiming at promoting the metadata exchange activities between different software tools. In the data integration based on CWM, I use a metadata store tool to store the metadata used in the exchange, and by accessing metadata store components directly have the access to the exchange metadata needed, adopting metadata store tool can lessen the workload involved in the exchange of metadata. At the same time if every kind of tool use CWM that is the general format to input and output metadata, all software can understand each other and exchange metadata, but also can exchange metadata with all other tools.

Create XML exchanging documents according to the document type definition in the metadata integration, which can directly convert to the object related to the metadata tools, realizing the exchange of metadata in the condition of CWM as sharing intermediary.

The essence of CWM metadata exchange, namely the exchange of class and associated examples, is the ability of exchanging in any middle format which can make them mutual agreement aiming at CWM. Once started the CWM metadata transforming mechanism, it would send the metadata of public format to any a kind of tool, not need to specify the concrete tools to create this exchange. Therefore, we can use a metadata store as the medium of CWM metadata exchange.

Due to the CWM metadata store is a base relational database, however the meta store extracted out based on the CWM modeling is object-oriented, therefore the first problem needed to solve is to ensure meta store of CWM keep object-oriented logic structure and make them map to the relation table.

5 SUMMARY

This paper makes full use of the characteristics of the metadata, firstly it makes the data of yearbook isomorphic management locally; Secondly by mapping and extracting the field needed, we can make evaluation and analysis using the yearbook data; Finally, using the CWM exchange technology, we complete the data sharing among different platforms, different subsystem and software. That is, it improves yearbook data usage rate effectively.

This paper was supported by “the Fundamental Research Funds for the Central Universities (2009BAG12A10-2)”.

REFERENCES

John Poole, Dan Chang, Douglas Tolbert, David Mellor et al. Common Warehouse Model Developers Guide Wiley Publishing, 2003.

Dr. Daniel, T. Chang. Common Warehouse Meta-model (CWM), UML and XML. Meta Data Conference, 2000, 6(4): 1 9-23.

William Rub, Enterprise Application Integration. John Wiley &Sons，2002.

Daniel T. Chang, Common Warehouse Meta-model (CWM), UML and XML Meta Data Conference, March 19·23, 2000.