International Journal of Hybrid Information Technology
Vol.8, No.1 (2015), pp.427-434
http://dx.doi.org/10.14257/ijhit.2015.8.1.38

Analysis of the Problem of Semantic Heterogeneity in the Integration of Railway System

Haiming Jing
School of Economics and Management, Shijiazhuang Tiedao University, Shjiazhuang, Hebei, 050043
jimperlov@yeah.net

Abstract

Every information system has its own domain model for its environment and efficient task operation, which results in diverse heterogeneities, especially the semantic Heterogeneity.

In this paper, order to adapt to the data request model of railway system, and semantic conflict type existing in the process of data integration, Ontology technology is employed in heterogeneous information integration, and a heterogeneous information framework is given. By finding semantic conflict initiative and constructing semantic mapping relations. In the case of SCADA data in railway system, Ontology mapping discovery algorithm is verified.

Key words*:ontology; mediator; semantic heterogeneity; data integration; semantic mapping*

1. Introduction

Heterogeneity and field relativity is a major cause of information semantic heterogeneity, semantic heterogeneity has become the main bottleneck of information integration. For the railway system, the data is miscellaneous, the representation and storage are different. To ensure the safe and stable operation, the need for a variety of monitoring, analysis, decision-making power, control system simulation analysis application is urgent. SCADA (Supervisory Control And Data Acquisition) application in railway electrification on the motion of the system is early, it ensures the safety and reliability of power supply of electrified railway, plays a large role in improving the management level of railway transportation dispatching. Multiple semantic conflict will occur when system integrated, this cause great difficulties for the data integration. The traditional solution to solve semantic conflicts is determined by domain experts manually through semantic matching table when integrating, With the increasing of the amount of data source and the emergence and application of all kinds of new technology, taking the initiative to find semantic conflict, to dissolve and eliminate the semantic conflicts in data integration has become a hotspot and difficult problem.

In order to eliminate conflict, the establishment of consistent information understanding of inter system is very important and necessary. Ontology is a new kind of business data description specification appears in recent years, it can accurately describe the data semantics, and reason implicit semantic relation data [1] and the inner relation between concept, relationship implicated in concepts can be obtained through logical reasoning [2]. For the parameter estimation system, There may exist various types of conflict when integrating, a complete solution based on ontology and semantic technology is proposed, through the realization of active recognition semantic conflict to resolve semantic conflict when integrating. Semantic conflict is when describing the objects of the same real world, the inconsistency of two objects in the way of description, structure and content caused by different semantic.

ISSN: 1738-9968 IJHIT

2. The Data Integration Problem

The data used in date integration are related to some common semantics uniform, but from a set of heterogeneous, distributed and autonomous (not related) source data. This integration process is provide the user unified understanding for data. These data can be distributed in different host and connected through a network. The data sources are independent, *i.e.*, users and applications can access them through the local or federal system.

Semantic heterogeneity may cause semantic conflicts between different data sources. When the data seems to have the same meaning, hybrid conflicts will occur, but because of the difference of the actual situation the time background will be different. The usage of different reference systems to measure a value causes scaling conflicts, for example gallons versus litres.

2.1. Feature of Data Integration Framework based on Ontology

(1)Automatic generation of local ontology Face with heterogeneous data source with rich source, in order to make the analytical preprocessing more efficiently and establish local ontology, processing analytical interface design that can adapted to dynamic adaptation of unified heterogeneous data sources are designed, the strategy pattern is used to implement the interface. The unity of the heterogeneous data source analysis, pretreatment and data extraction are achieved, and the automatic establishment of local ontology is realized by the interface. A key part of the implementation of the interface is to extract semantic information to construct the local ontology from structured, semi-structured and non structured data file. Because the structured relational database schema and local ontology model are very similar, When the 3 types of is pre-processed, semi-structured data and unstructured data will first be transform into structured data, and then the local ontology will be constructed by constructed data uniformly. Through the implementation of automatic construction of local ontology, physical conflict exists in the data integration can be solved.

(2)Semantic Conflict Detection Mechanisms: Semantic conflict detection mechanism can realize the initiative find of semantic conflict in the process of data integration, and accomplish the construction of semantic mapping relation by discovered semantic conflict, complete the ontology mapping process. Ontology mapping refers to there exists concept association in semantic level two ontology. Through the semantic association, source ontology will be mapped to the target ontology, the most important process of mapping is the discovery of semantic conflict. Through the ontology mapping and the semantic conflict resolution, thus eliminating the semantic heterogeneity. In order to implement the initiative find of semantic conflict type, mapping finding strategy based on semantic tree is designed. Mapping discovery strategies respectively tailored finding strategy of property couple, concept couple and instance couple, thus the initiative finding problems of table, field and recording conflict have been solved from different levels respectively, while the mapping finding rules based on semantic tree is defined to improve efficiency and accuracy of finding.

(3)Semantic mapping types: the semantic mapping type defines the concept, attribute and instance between the ontology semantic relations, and provide the basis for resolving semantic conflicts between heterogeneous data sources. To resolve the semantic conflict problem found by conflict detection mechanism, four types of semantic mapping are defined the in railway data integration application.

2.2. Wrapper/Mediator

Wrapper/Mediator is a information integration technology method. Information integration system integrate the data from various source through the intermediary model,

and data is still stored in the local data source, through the wrapper of each data source (wrapper) to transform the data to conform to the intermediary model.

From a technical point of view, Wrapper/Mediator [3]which is based on the mode inheritance method theory is considered ideal solution to realize heterogeneous integration in recent years, this scheme does not need to change the original data storage and management mode, it coordinates various heterogeneous data sources system downward via middleware, provides a unified data model and common data access interface upward [4]. The typical system using this model are MIX, YAT, Nimble *etc.*, [5-7].

Wrapper/Mediator is a method in the information integration technology. Information integration system integrate the data of each data source through the intermediary model, and data is still stored in the local data source, through the wrapper of each data source transform the data to conform to the intermediary model.

We extends the construction of data integration system based on Wrapper/Mediator mode, Ontology as a solution to the semantic heterogeneous tools is introduced to the system. The semantic conflict detection mechanism is designed by Ontology's advantage in describing semantic, and by building a semantic mapping relation the problem of semantic heterogeneity in data integration can be well solved. The framework of the system is shown in Figure 1.

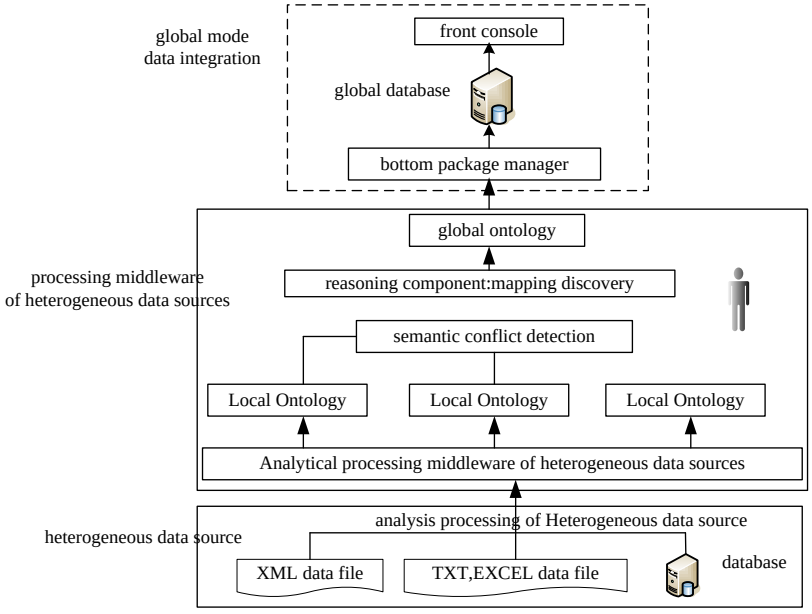

Figure 1. Heterogeneous Data Integration Frame Based on Ontology

2.3. Types of Semantic Conflict

To achieve the above data Integration, the traditional data integration framework of railway system mostly adopts Wrapper/Mediator method to build data integration system. It can well solve the system heterogeneity, data structure heterogeneity and syntax heterogeneity problem. But for the integrated data how to take the initiative to identify the semantic conflict problems, mainly the field conflict, table conflict and record conflict [8] is invalid. Semantic conflict is when describing the same real world objects, 2 objects in the way of description, structure and content caused by different semantic inconsistency.

To solve the problem from the use of semantic technology point of view, according to different levels, this paper on semantic conflict may meet the power data integration are summarized as follows:

(1) physical conflict: it is the conflict according to different storage format of data source. For example, SCADA data is unstructured TXT format data, and the grid model data is semi-structured XML format data.

(2) table conflict: it include naming conflict, structural conflict and relationship conflict. Naming conflicts including table name repetition, such as switch objects can be expressed by Breaker or Switch. Structural conflict refers to used different structure to express the same concept table. Relationship conflict refers to the inconsistent when table relationship integrated together. for example, in A system, the expression between the two table is father relationship, but in the B system, the expression of the similar two tables is the equivalence relation.

(3) field conflict: it including naming conflict, type conflict, length conflict, precision conflict metering unit conflict, expression way conflict. Naming conflict refers to field homonyms, synonyms synonyms. type conflict refers to express the same features with different data types at different tables; length conflict refers to inconsistence when expressing the same characteristics of the different field data length; precision conflict refers to a field when expressing the same characteristics using different data accuracy in different table; measurement unit conflict refers to that data have different measurement unit when expressing the same characteristics of in different tables; expression conflict refers to inconsistency of data representation, such as data format differences, abbreviated difference.

(4) record conflict: it refers to the difference of numerical data records when describing the same data, caused by different units of measurement of conflict, such as the ratio and the stall of the conversion: 1.23 (ratio) =0.5 (gear)

3. The Key Technology of the System

Through the study, it is found that there has been a similarity calculation method only considering part of the information, other information only as auxiliary rules when found in the mapping relationship, which makes the final results are not accurate. Ontology mapping method that construct multi semantic similarity is a effective method. Semantic similarity is characterized by the word distance, it is a real number between 0 and infinity, the distance between a word with itsself is 0. The greater of the distance of two words, the lower of the similarity is; on the contrary, the smaller of the distance of two words, the greater the similarity is.

3.1. Multi Strategy Mapping Discovery based on Semantic Tree

Face with rich data types and mass data of railway system, efficient mapping found is important. Ontology mapping is to combine two or more different ontology as input, and then the process of establishing the corresponding semantic relationship according to the semantic relations for these ontology elements (concepts, attributes, relations), is the key to solve the problem of semantic heterogeneity.To fully consider the mapping relation between the concept, attribute and instance, multi-strategy mapping algorithm based on semantic tree should be taken account, and on the basis of it the rules of mapping discovery process are defined to improve the efficiency of mapping discovery.

3.1.1. Name Similarity Algorithm

The name similarity algorithm[9] is used to calculate the similarity of attribute couple in ontology. Name similarity algorithm is based on Wordnet semantic dictionary, Wordnet is a thesaurus dictionary, in which each node represents a word, synonymous words or phrases are saved in nodes, each word or phrase can be stored in a plurality of semantic nodes. The similarity word W_1 and word W_2 are definited below:

$$sim\ (w_1, w_2) = 1 - t\sqrt{(\alpha - 1)/\alpha \times \beta \times Dist\ (w_1, w_2)}$$

Where, $\beta = Dep\ (w_2)/(Dep\ (w_1) + Dep\ (w_2))$,T and alpha is a variable factor, it is the depth of C concept of a non-root node in the semantic tree, it is defined below:

$$Dep\ (c) = Dep\ (parent\ (c)) + 1$$

The basic idea of concept similarity is based thesaurus dictionary, the connection distance of 2 words by epistatic relationship (hypemym) is closer, the greater the similarity; conversely, the smaller. If they are on a node, *i.e.*, s_l=s_2, sim (W_1, W_2) =l, if they do not have a parent node in the upper level, sim (W_1, W_2) =0.

3.1.2. Concept Similarity Algorithm

Concept similarity algorithm is used to calculate the similarity of concept couple in ontology. Concept definition description information includes 2 aspects: synonym sets represented concept and feature sets of concept. Among them, the feature set can be divided into function, part and attributes. Synonym sets denotes a noun set of the same concept. The computing method of concept similarity is defined below:

In an ontology definition, concept attribute and relation has an important role to the concept description. Therefore, in the calculation these factors should be taken into account. Concept description similarity method refer to the calculation method put forward by [10] M.Andrea Rodrigue and MaxJ.Egenhofer. In this method, the definition of the of information of concept description includes two aspects, synonym sets represented concept and feature set described the concept. Synset is word set represented a name of concept, because of the presence of polysemy, synonym sets in the expression of word meaning is more accurate than a single word. The concept of feature set is divided into, function, part and atttribute.

$$sim(A,B) = \frac{\left|a \cap b\right|}{\left|a \cap b\right| + \alpha(A,B) + \left|a / b\right| + (1-\alpha(A,B))\left|b / a\right|}$$

A and B respectively represent description set of concepts A and B, including a synonym set and feature set; a∩b represents numbers of elements intersection of a and b, a/b represents numbers of elements that belong to set a but does not belong to the set b. elements. The proportion meet

$$\alpha(A,B) = \begin{cases} \dfrac{depth(A)}{depth(A)+depth(B)} & depth(A) \le depth(B) \\ 1 - \dfrac{depth(A)}{depth(A)+depth(B)} & depth(A) > depth(B) \end{cases}$$

Among them, depth (A) represents the shortest path distance from the concept A to the root.

3.1.3. Finding Rules Mapping based on Semantic Tree

Ontology semantic model is a concept tree [11] instance or attribute, leaf nodes represent the concept of ontology in a concept tree, the other node represents the concept of Ontology (class). Therefore, the following rules based concept semantic tree is defined to improve mapping efficiency of discovery.

Rule 1 in the semantic tree, if there exists mapping relation between the parent node parent (A) parent (B) and sub node son (A) and son (B) respectively, then there maybe exists mapping relationship beween A and node.

Rule 2 in the semantic tree, if exists mapping relationship between brother node brother (A) and brother (B) has, then there may also exists mapping relationship A and B nodes.

Rule 3 In the semantic tree, if the A node and the B node is similar, the example nodes of A nodes and B nodes is similar.

Rule 4 in the semantic tree, if the A nodes and B nodes have the same attribute node, then A node and B node is similar.

3.2. Ontology Mapping Algorithm

The ontology mapping algorithm synthetically use the above similarity strategies to calculate the similarity of the ontology elements, take the initiative to find the semantic conflict between the concept, properties and examples in the two ontology, finally, ontology mapping types was output and then mapping discovery process have complete. The mapping finding algorithms is as below:

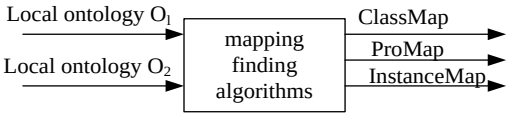

Figure 2. Mapping Finding Algorithms

The flow of attribute mapping algorithm is below:

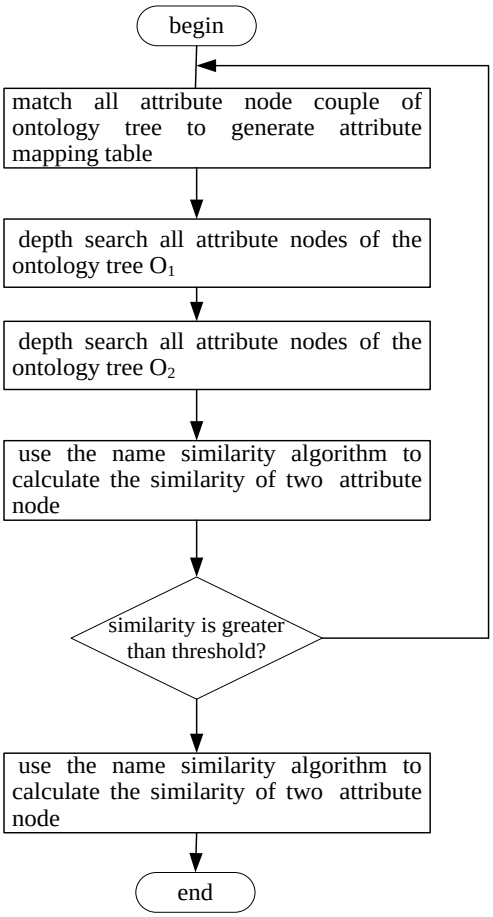

Figure 3. Flow of Attribute Mapping Algorithm is Below

The flow of concept mapping algorithm is below:

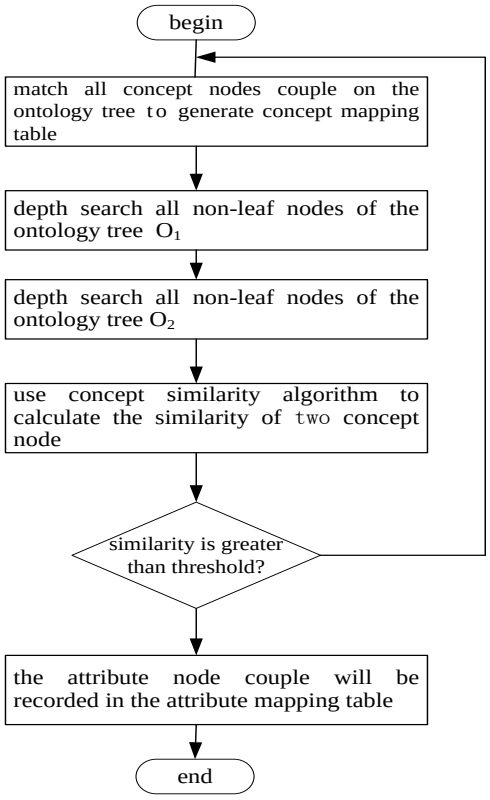

Figure 4. Flow of Concept Mapping Algorithm is Below

The test data include 16 concepts, 54 instances, 129examples, SCADA data contains 8 concept, 32 attributes, 156 instances.

Ontology mapping discovery algorithm is implementated, by the mapping couple the concept mapping table, attribute concept mapping table, instance mapping table found examples of mapping, and the comparison outcome of actual number, as shown in Table 1.

Table 1. Number Comparison of the Actual Mapping and Experiment Mapping

number of mapping couple / mapping	actual existence of mapping	actual discovery mapping
Concept couple	10	8
Attribute couple	20	18
Instance couple	70	55

Precision of the information retrieval domain is used as the main index to evaluate mapping algorithm, the precision is defined as below

P= right mapping couple found/all of correct mapping couple

According to the formula of precision, the precision ratio of concept couple pobject=75%; precision ratio of property couple pproperty=85%; precision ratio of instance pinstance=78%. Precision ratio specify that mapping finding algorithm defined in this paper can take the initiative to find most of the semantic conflicts in data integration, but there are still a few conflicts have not identified by any method. These conflicts need be resolved by domain experts by developing semantic mapping rule.

4. Conclusion

Through the experimental analysis and the practical application results, the design of the mapping discovery algorithm can efficiently and accurately discover semantic conflict in integrated data, at the same time using the semantic mapping types defined in data integration, can solve most of the problem of semantic heterogeneity. This framework has been successfully applied in parameter estimation system without affecting the efficiency and correctness of data integration at the same time, it better ensure the consistency of the data integration process data.

The shortage of this article is that it does not design mapping finding strategy for the instance data, for the diversity and complexity of the data of railway system, so the general features of the instance data can not be extracted. The mapping algorithm should be studied in-depth, and making the mapping matching strategy for instance data, then automatic integration of heterogeneous data sources will be realized.

Acknowledgements

This work was supported by the follows: (1)The Young Foundation of Education Department of Hebei Province(No. 2011135) (2)The Young Foundation of Shijiazhuang Tiedao University. (No. 20133020)

References

[1] J. L. G. Dietz, "Enterprise ontology: Theory and methodology", Berlin: Springer, **(2010)**.
[2] B. Shi and L. Y. Fang, "Concepts Semantic Similarity Measure Based on Ontology", computer engineering, vol. 35, **(2009)**, pp. 83-85.
[3] G. Wiederhold, "Mediators in the Architecture of Future Information System", IEEE Computer Society, vol.25, **(1992)**, pp. 38-49.
[4] D. Kim, K. Jeong, H. Shin and S. Hwang, "An XML Schema-based Semantic DataIntegration.Grid and Cooperative Computing", Fifth International Conference of IEEE Computer Society, **(2006)** October, pp. 522-525.
[5] C. Baru, A. Gupta and B. Ludäscher, "XML based information mediation with MIX", ACM SIGMOD Record, vol. 28, **(1999)**, pp. 597-599.
[6] S. Chawathe, H. Garcia-Molina and J. Hammer, "Integration of heterogeneous information sources", Proceedings of the 10th Meeting of the Information Processing Society of Japan, Kyoto, Japan, **(1994)**, pp. 7-18.
[7] S. Cluet, C. Delobel and J. Simeon, "Your mediators need data conversion", Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, New York, **(1998)**.
[8] A. Doan and J. Madhavan, "Domingos Peta1. Leaming to Match Ontologics on the Semantic Web", The VLDB Journal, vol. 12, **(2003)**, pp. 303-319.
[9] W. Sadiq and M. E. Orlowska, "Analyzing Process Models Using Graph Reduciton Techniques", InformaitonSystems, vol. 25, **(2000)**, pp. 117-134.
[10] M. Andrew and M. J. Egenhofer, "Determing Semantic Similarity Amony Entity Classes from Different Ontologies", IEEE Transactions on Knowledge and Data Engineer, vol. 15, **(2003)**, pp. 442-456.
[11] Y. X. Li, "Research of technology of semantic heterogeneous elimination orin information integration", technology. Lanzhou: Northwest Normal University, **(2009)**.

Author

Haiming Jing, he received the B.S. degree in Industrial Automation from the Harbin University of Science and Technology for Nationalities of China, Heilongjiang in 1999, and the M.S. degrees inControl Theory and Contrl Engineer, Electrical Engineering College, Yanshan university for Nationalities of China, Hebei in 2002. His current research interests include digital convergence, Ontology Computing. He is a vice professor of the Informaion Science and Technology.

