National Conference on Recent Advances in Computer Science and IT (NCRACIT)
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 1 | ISSN : 2456-3307

Semantic Heterogeneity - An Overview

Waseem Jeelani Bakshi[1], Mujtaba Shafi[2]

[1]Department of Computer Sciences, University of Kashmir, Jammu & Kashmir, India.
[2]Bioinformatics Centre, University of Kashmir, Jammu & Kashmir, India.
waseembakshi@uok.edu.in[1],mujtabashafi@gmail.com[2]

ABSTRACT

Owing to the fast growing markets and the rapid evolution of the existing enterprises today different parties use different database schemas to develop their solutions for the same domain, which leads to what, is known as semantic heterogeneity. The importance of database systems in today's business world together with the fact that today business enterprises employ multiple coexisting information systems makes integration of these heterogeneous systems crucial for the growth and development of these enterprises. In this paper, we try to understand what semantic heterogeneity means and try to overview various classifications that have been proposed for classifying semantic heterogeneity.

Keywords: Heterogeneous Data, Schema, Semantic heterogeneity.

I. INTRODUCTION

In order to overcome the problem of semantic heterogeneity it is very important to first detect the contradictions, which in itself is a very difficult process due to the lack of sources of semantic knowledge.

According to Alon V. Halevy [1], there are many potential circumstances where semantic heterogeneity may arise, including:

- Enterprise information integration
- Querying and indexing the deep web
- Merchant catalog mapping
- Schema vs. data heterogeneity
- Schema heterogeneity and semi-structured data.

These along with many other sources in simple schema use and versioning create mismatches. Halevy further states that the possible drivers in semantic mismatches can occur from worldview, perspective, syntax, structure, versioning and timing:

- One schema may express a similar "world view" with different syntax, grammar or structure
- One schema may also be a new version of the other
- Multiple shames may be derived from the same source schema
- There may be many sources modelling the same aspects of the underlying domain ("horizontal resolution" such as for competing trade associations or standard bodies), or
- There may be many sources that cover different domains but overlap at the same ("vertical resolution" such as between pharmaceuticals and basic medicine)

II. TYPES OF SEMANTIC HETEROGENEITY

Resolution of semantic conflict is a key step in the integration of diverse information sources. For this, it becomes very important to identify and understand the different classes of semantic conflicts. The schematic mapping cannot be achieved without some comprehensive understanding of semantic conflicts.

Many classifications have been proposed for classifying semantic conflicts. However, it has been observed that there are similar semantic problems in many of the different types of conflicts listed. Further, the conflicts do not easily fall into discrete categories. Moreover, the classes of conflicts are overlapping and can be described with fewer dimensions.

Keeping in view the importance of classifying the semantic conflicts for the reconciliation process, Niaman and Ouksel [2] have proposed a list of functional requirements desirable features and properties of a classification of semantic conflicts. According to them:

- A classification must capture semantic conflicts, that is, it must capture the abstraction used to represent data in semantic modelling.

- The classification must allow the representation of alternative semantic conflicts. During dynamic reconciliation more than one interpretation of the conflicting schematic representations may be plausible. The classification must represent these plausible interpretations of the conflict.

- The classification must be sound, that is, if given a classification assertion, then there exists a semantic conflict which it represents.

- The classification must be minimal, that is, no classification with fewer dimensions or fewer values along the various dimensions can capture semantic conflicts. This results in disjoint classes, with any semantic conflict belonging to exactly one class. If the set of these classes would be reduced, then some semantic conflicts could not be represented. In this sense the classification must provide a minimal set of classes representing semantic conflicts.

- The classification must be complete, that is, if given a semantic conflict, there exists a classification assertion or a conjunction of assertions, which classifies this conflict.

- The classification must be shown in its application to dynamic reconciliation process. It is not enough that the classification provides criteria on which the semantic conflicts can be distinguished, but the classification must also be validated in its use in the dynamic reconciliation process.

Based on these requirements Niaman and Ouksel have proposed a classification of semantic conflicts which provides a formal representation of the knowledge necessary to map conflicting schematic representation between databases. They have organised this classification along the dimensions of naming, abstraction and level of heterogeneity and have shown that this classification is sound, minimal and complete.

In their study Sheth et al [3] define three categories of semantics-implicit, formal and powerful. Implicit semantics means what is either principally present or can be easily inferred: formal semantics involves the use of ontologies or other descriptive logics, and powerful semantics refers to fuzzy logic and is not confined to rigid set based assignments.

Working on Federated Database Systems, Saltor and Rodriguez [4] have classified semantic Heterogeneity into three groups:

- Heterogeneities between object classes

- Heterogeneities between class structures
- Heterogeneities between object instances

The first group, i.e., heterogeneities between object classes covers the corresponding classes, i.e., the classes in different component databases that represent the same concept in their respective context. This includes differences in extensions, differences in names, differences in attributes and methods, differences in constraints, etc.

The heterogeneities between class structures includes the inconsistencies due to generalization/specialization, inconsistencies due to the type and level of aggregation/decomposition used by the participating databases and inconsistencies due to the schematic discrepancies wherein some values or data in one component database are seen as part of the schema in another Component database.
Finally the heterogeneities between object instances include the value discrepancy (i.e., the differences in values for corresponding classes and corresponding attributes of the classes), the null/Non-null discrepancy (i.e., the presence or absence of null values of attributes) and the discrepancies in the number of values for multivalued attributes.

Firat [5] has classified Semantic Heterogeneity along three dimensions, namely, contextual, ontological and temporal. Contextual heterogeneity arises when different component databases represent the same concept differently. Ontological heterogeneity arises when different meanings are represented in different component databases by same terms. Temporal heterogeneity arises when both the contextual and ontological assumptions change over time.

According to Mustafa Jarrar (University of Birzeit) [6], there can be several heterogeneities between different database schemas. These include:

- Name Heterogeneities (differences in used vocabulary)
- Meaning Heterogeneities (different meaning for the same attribute in two schemas)
- Heterogeneity in structure and type
- Heterogeneity in rules and constraints
- Data Model Heterogeneity

Pluempitiwiriyawej and Hammer [7] classify heterogeneity into three broad classes:

Structural heterogeneity: different schemas in different sources represent similar or overlapping concepts. This includes generalization/aggregation conflicts, internal path discrepancies, missing items, elements ordering, constraint and type mismatch and naming conflicts between the element types and attribute names.

Domain heterogeneities: semantics of the participating data sources is different. This includes schematic discrepancy, scale or unit, precision and data representation conflicts.
Data heterogeneities: data values across multiple sources are different. This includes ID values, missing data, incorrect spellings, etc.

III. CONCLUSION

The study of Semantic Heterogeneity is very important for the understanding of the different challenges it poses so that efficient and appropriate solutions can be developed for their redressal. The diverse studies that have been conducted in this area lead us to realise that there are around 40 distinct potential sources of semantic heterogeneities, which can be broadly grouped into four major classes - Structural, Domain, Data and Language. A deep understanding of each class of Semantic Heterogeneity will help us to handle them appropriately and efficiently.

IV. REFERENCES

[1] A. V. Halevy, "Why Your Data Won't Mix", ACM Queue vol. 3, No. 8, October 2005.
[2] F. Channah, Niaman and Aris M. Ouksel, "A Classification of Semantic Conflicts in Heterogeneous Database Systems"

[3] A. Sheth, C. Rama Krishnan and C. Thomas, "Semantics for the Semantic Web: The Implicit, the Formal and the Powerful" International Journal on Semantic Web and Information Systems, 1(1), 1-18, Jan-March 2005.

[4] Rodríguez E., Oliva M., Saltor F., Campderrich B., On Schema and Functional Architectures for Multilevel Secure and Multiuser Model Federated DB Systems, in Conrad et al. (Eds), Proceedings of the Int. CAiSE'97 Workshop, Barcelona, Otto-von Guericke-Universität Magdeburg, June 1997, pp.

[5] A. Firat, 2003. "Information Integration Using Contextual Knowledge and Ontology Merging". Ph.D Thesis. Massachusetts Institute of Technology.

[6] M. Jarrar (University of Birzeit)

[7] C. Pluempitiwiriyawej and J. Hammer, "A Classification Schema for Semantic and Schematic Heterogeneities in XML Data Sources", Technical Report TR00-004, University of Florida, Gainesville, FL, 36pp, September 2000.