Proceedings of the ISCA
2nd International Conference

INFORMATION REUSE AND INTEGRATION

Honolulu, Hawaii U.S.A.
November 1 - 3, 2000

Editors: S. N. J. Murthy, S-C. Chen

A Publication of
The International Society for
Computers and Their Applications - ISCA

ISBN: 1-880843-36-6

Knowledge Management for Database Interoperability*

Naphtali D. Rishe[1], Rukshan I. Athauda[2,**], Jun Yuan[1], Shu-Ching Chen[1]

[1]High-Performance Database Research Center
School of Computer Science
Florida International University
Miami, FL 33199, USA

[2]Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399, USA

rishen@cs.fiu.edu, rukshana@microsoft.com, yuanj@cs.fiu.edu, chens@cs.fiu.edu

Abstract

The availability of multiple heterogeneous, autonomous, distributed data sources containing related information has created a need for integrated access to these information systems. Heterogeneous/multi-database systems address this issue when the component data sources are database systems. Resolution of heterogeneities for integrated access requires discovering and managing certain types of knowledge/facts. A generally accepted methodology or approach for managing this knowledge and information is lacking in research and industry. In this paper, we provide a framework for managing knowledge for interoperable access to heterogeneous database systems. The framework uses knowledge bases at the integration and component sites. Sample schemas of these knowledge bases are presented. A multi-database prototype system utilizing the techniques presented in this paper is being developed.

Keywords: multi-database, knowledge base

1. Introduction

The availability of multiple independently developed databases containing related information and networks that interconnect them, has created a need for integrated access to this data/information. Heterogeneous/multi-database research has focused on this issue resulting in many different frameworks for database integration. These different approaches can be classified into three main groups: (i.) *Global schema approach* ([1] and others) creates a global schema/view over the component database systems that capture the union of the information content of the component schemas; (ii.) *Federated database approach* ([4] and others) exports schemas of distributed database and integrates with the local schema to provide the necessary views for the local users; and (iii.) *Multidatabase language approach* ([3] and others) provides powerful multidatabase languages for querying a group of non-integrated schemas.

A general problem that is common to all of the above approaches is the resolution of heterogeneities caused by the autonomous, distributed, heterogeneous data sources. Heterogeneities occur at several levels: (i.) Semantic level and schema level heterogeneity: This occurs with the same real-world objects and concepts being represented in different databases using multitude of data models and user perspectives; (ii.) Database, platform and network level heterogeneity: This occurs due to the use of different DBMSs, networks and platforms at the distributed sites. The heterogeneities of platform and database tool level have been addressed in the industry using technologies such as CORBA, standardized query languages such as SQL and interfaces such as ODBC/JDBC. Semantic/schema level heterogeneity is usually resolved by developing a homogenizing layer over the heterogeneous distributed data sources. In this paper, we consider knowledge management techniques for resolving these types of conflicts.

A key issue for resolving semantic heterogeneity is the acquisition of appropriate metadata and discerning the semantic relationships between constructs of the different database schemas ([6]). The management of this knowledge in a modular and efficient way is crucial for building interoperable database systems. A multitude of approaches can be found in literature for this purpose. In [2], a knowledge base is used for storage and manipulation of meta-data. In [4], a semantic dictionary is proposed for this purpose. In [5] ontologies are utilized for knowledge reconciliation. In [10], a global thesaurus is discussed as a means for storing the meaning of terms and resolving semantic heterogeneity. In this paper, we describe knowledge management techniques used in MSemODB ([8]), a multidatabase prototype system being built by us.

This paper discusses techniques for knowledge management with an application for database interoperability. The contributions of this paper include: (i.) a framework for managing knowledge in a distributed, heterogeneous, autonomous database environment (which is discussed in Section 2); (ii.) knowledge management techniques at the component database sites (which is described in Section 3); and (iii.) knowledge management techniques at the global site for database interoperability (presented in Section 4). Finally, concluding remarks are provided in Section 5.

* This research was supported in part by NASA (under grants NAGW-4080, NAG5-5095, NAS5-97222, and NAG5-6830) and NSF(CDA-9711582, IRI-9409661, HRD-9707076, and ANI-9876409).

** This paper resulted from the research preformed while the author was employed at the High-Performance Database Research Center.

2. Knowledge Management Framework

A well-known approach for database interoperability is presented in figure 1. The schemas of component database systems are transformed into a canonical data

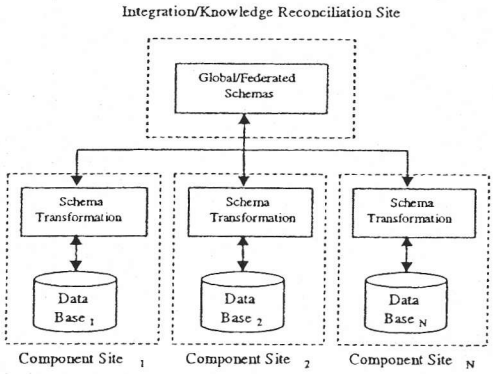

Figure 1. A well-known architecture for global and federated database approach

model and exported to the *Integration/Knowledge Reconciliation* site for the creation of global/federated schemas and views for querying. The schema transformation phase resolves conflicts caused by the different data models of component databases. Also, a canonical data model provides a uniform query facility for each component site. This allows for less complex query processing techniques at the *Integration/Knowledge Reconciliation* site.

Recently, wrappers have been developed ([9] and others) for performing schema transforming and query translation tasks at the component sites. At the *Integration/Knowledge Reconciliation* site, users are presented with integrated views/schemas for accessing multiple data sources in a uniform data model and query language. The resolution of conflicts at component and integration sites requires acquiring and managing knowledge and meta-data. The framework, discussed below, extends the architecture (presented in Figure 1) by introducing the use of knowledge bases at the different sites (i.e. integration and component sites). Figure 2(a) and 2(b) depict the high-level architecture at the component and integration sites respectively.

At the component site, schemas and relevant meta-data are imported and transformed into the canonical data model by the *Schema Loader & Transformer* component. These meta-data and schema mapping information are stored in the *Knowledge Base*. The *KDBTool/Semantic Enrichment* component interacts with the DBA for advanced knowledge acquisition and conflict resolution processing. The transformed schemas including the relevant meta-data are exported to the *Integration/Knowledge Reconciliation* site. *Query Translator* component translates queries posed on the transformed schema of the canonical data model into semantically equivalent queries of the component database schema. A wrapper developed using the architecture described above is SemWrap. The knowledge base schema of SemWrap is discussed in Section 3.

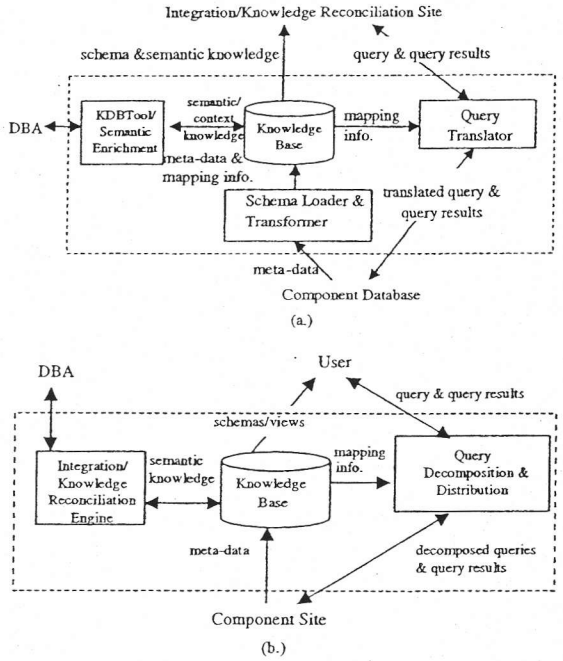

Figure 2. (a.) High-level architectural components for Component Site (b.) High-level architectural components for the Integration/Knowledge Reconciliation Site

At the *Integration/Knowledge Reconciliation* site, heterogeneities that occur due to a multitude of homogeneous database schemas are resolved. The knowledge required for resolving such heterogeneities are discussed in [6]. In Section 4, we present a knowledge base schema that captures this information content.

3. Knowledge Management at Component Site

The knowledge base schema at the component site captures the following information: (i.) component database schema; (ii.) transformed database schema; and (iii.) mapping information between the component database schema and its transformed schema. This information is crucial for both schema transformation and query translation. Also, semantic enrichment of the transformed schemas (which includes incorporating context information) may be included into the knowledge base. However, this may differ according to the methodology used for semantic heterogeneity resolution and hence not included in our presentation of the schema.

In this section, we describe the schema of a knowledge base used in SemWrap. SemWrap is a wrapper over component relational databases providing a Semantic Binary Object-oriented Data Model (Sem-ODM) [7] interface. Hence, the component schemas are relational schemas and transformed schemas have a Semantic Binary Object-oriented Data Model. Sem-ODM is a powerful expressive data model capable of capturing advance complex modeling constructs and hence, we used

a Semantic Database Engine (Sem-ODB) [8] as the storage medium of the knowledge base.

Sem-ODM consists of *category*, which may be inherited and *relation*, which is a relationship between categories. Figure 3 (a.) presents the meta-schema of Sem-ODM. Graphically, the rectangles represent *categories*. The dashed-arrow represents *ISA* links (inheritance/super-category subcategory relationships). The dashed-arrows point from *sub-category* to *super-category*. The *attributes* of a particular category are placed in the respective category rectangle with ranges placed after the ":" (semi-colon). The thick (non-dashed) arrows represent *relations* between categories. The cardinalities and constraints of relations are represented inside brackets.

As shown in Figure 3 (a.), the primary constructs of Sem-ODM are *CATEGORY*s and *RELATION*s. A *CATEGORY* can be either *ABSTRACT* or *CONCRETE*. *ABSTRACT CATEGORY*s represent objects that are explicitly created representing real-world concepts, ideas or objects. *CONCRETE CATEGORY*s represents printable values. Subcategories of *CONCRETE CATEGORY* are not shown in this figure due to space limitations. A *RELATION* is a mapping between objects in the domain to objects in the range. A *RELATION* having a range of a *CONCRETE CATEGORY* is also termed an attribute of the domain.

The meta-schema of a relational database schema is shown in Figure 3 (b.). This sub-schema contains *TABLE*s, *FIELD*s which belong to tables and their respective *DATATYPE*s. Primary and foreign keys are represented by categories *PRIMARY KEY FIELD* and *FOREIGN KEY FIELD* respectively. The functional dependencies are represented by relation *refers-to*. This sub-schema is self-explanatory and will not be discussed further.

The subschema shown in Figure 3(c.) represents the mapping information among the transformed and component schemas. Categories *META OBJECT* and *COMPONENT META OBJECT* are the same categories represented in Figure 3(a.) and 3(b.). It is significant to note that category *META OBJECT* is not directly derived from *COMPONENT META OBJECT*, instead from category *VIEW META OBJECT*. *VIEW META OBJECT* is categorized to *COMPONENT META OBJECT* and *VIEW SPECIFICATION*, which is further categorized to categories *VIRTUAL CATEGORY*, *VIRTUAL RELATION* and *VIRTUAL ATTRIBUTE*. This is due to the fact that usually a transformed *META OBJECT* cannot be directly derived from *COMPONENT META OBJECT* due to the heterogeneities that may occur in the different representations. For instance, a *relation* in transformed schema is derived from a functional dependency in the component schema. This is represented by category *VIRTUAL RELATION*. Likewise, different types of heterogeneities are resolved with the addition of middle-level categories between transformed schema and component schema.

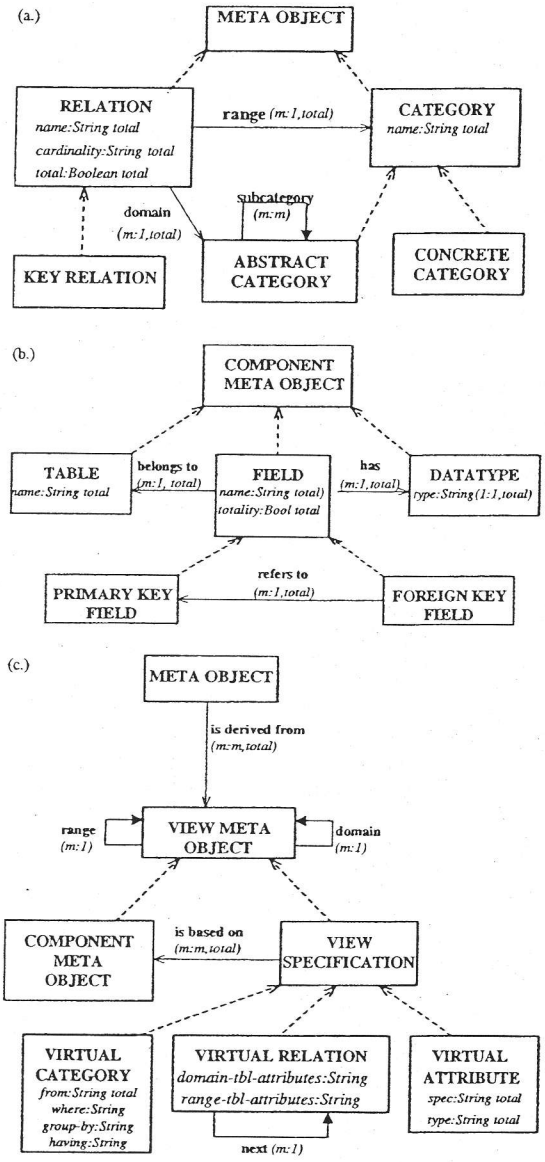

Figure 3. (a.) Meta-schema of Sem-ODM (b.) Meta-schema for a relational database (c.) Mapping from transformed schema (Sem-ODM schema) to component schema (relational schema).

4. Knowledge Management at Integration/ Knowledge Reconciliation Site

The heterogeneities between a set of Sem-ODM schemas are resolved at the *Integration/Knowledge Reconciliation* site. This process requires (i.) identification of semantic relations between constructs of component schemas; (ii.) acquiring means for determining object equivalences for related constructs; and (iii.) determining boundary conditions of related entities. The knowledge base at *Integration/Knowledge Reconciliation* site focuses on the storage of these types

of knowledge. The concepts mentioned-above are described in detail in [6] and hence not discussed in this paper.

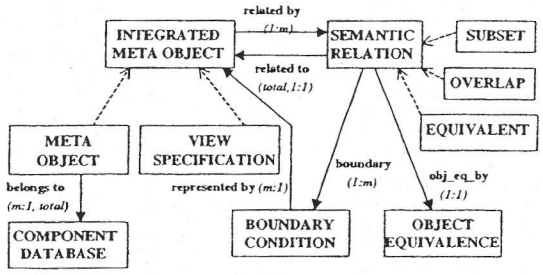

Figure 4. Schema of knowledge base at Integration/Knowledge Reconciliation site

A schema for the knowledge base of *Integration/Knowledge Reconciliation* site is presented in Figure 4. The categories *META OBJECT* and *VIEW SPECIFICATION* have similar definitions as in Figure 3. *Attributes* of each category are omitted to avoid complexity. Category *SEMANTIC RELATION* captures the different types of semantically related entities. Semantic relation "semantically disjoint" [6] is not represented and assumed by default. *Object equivalence* [6] (i.e. matching key attributes) is represented by category *OBJECT EQUIVALENCE*. The *boundary conditions* [6] are specified in category *BOUNDARY CONDITION* and are represented by an object of *VIEW SPECIFICATION* (using relation *represented _by*).

With this knowledge specified, in a global/federated schema approach, we can derive global/federated schemas/views from the *INTEGRATED META OBJECT* category in a similar fashion as done in Figure 3(c.) (i.e. *GLOBAL SCHEMA META OBJECT* can be derived from *INTEGRATED META OBJECT* similar to Figure 3(c.)). In a multidatabase language approach, a user's query based on a set of *INTEGRATED META OBJECT*s can be directed translated into a set of queries based on the related *INTEGRATED META OBJECT*s and transmitted to the component databases to obtain complete answers.

5. Conclusion

This paper discusses the issues related to knowledge management for interoperability of databases. A framework for integrating and query processing a set of heterogeneous, autonomous database systems is presented. Schema designs for knowledge bases at component sites and integration sites are illustrated.

In future, we will consider exploiting knowledge in the knowledge bases for optimizing query optimization in a heterogeneous database environment. Extending knowledge bases to automate the process of discovering and managing semantic knowledge are also fruitful areas for future investigation.

6. References

[1] C. Batini, M. Lenzerini and S.B. Navathe, "A Comparative Analysis of Methodologies for Database Schema Integration," In ACM Computing Surveys, Vol.18, No.4, 1986, pp. 323-364.

[2] C. Collet, M.N. Huhns and W.M. Shen, "Resource Integration Using a Large Knowledge Base in Carnot," In IEEE Computer, Vol. 24, No. 12, 1991, pp. 55-62.

[3] L.V. S. Lakshmanan, F. Sadri and I. N. Subramanian, "SchemaSQL – A Language for Interoperability in Relational Multi-database Systems," In Proceedings of the International Conference on Very Large Data Bases, 1996, pp. 239- 250.

[4] D. McLoed and A. Si, "The Design and Experimental Evaluation of an Information Discovery Mechanism for Networks of Autonomous Database Systems," In Proceedings of the IEEE International Conference in Data Engineering, 1995, pp. 15-24.

[5] E. Mena, V. Kashyap, A. Illarramendi and A.P. Sheth, "Managing Multiple Information Sources through Ontologies: Relationship between Vocabulary Heterogeneity and Loss of Information," In Proceedings of the 3rd Workshop on Knowledge Representation Meets Databases, 1996, pp. 1-3.

[6] N. Rishe, R.I. Athauda, J. Yuan and S.C. Chen, "Semantic Relations: The Key to Integrating and Query Processing in Heterogeneous Databases," In Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics, Vol. 7, Computer Science and Engineering: Part I, 2000, pp.717-722.

[7] N. Rishe, *Database Design: The Semantic Modeling Approach*. McGraw-Hill, 1992.

[8] N. Rishe, J. Yuan, R. Athauda, X. Lu, X. Ma, A. Vaschillo, A. Shaposhnikov, D. Vasilevsky and S.C. Chen, "SemanticAccess: Semantic Interface for Querying Databases," To appear in the International Conference on Very Large Databases, September , 2000.

[9] M.T. Roth and P. Schwarz, "Don't Scrap It, Wrap It!: A Wrapper Architecture for Legacy Data Sources," In Proceedings of the International Conference of Very Large Data Bases, 1997, pp. 266-275.

[10] D. Weishar and L. Kerschberg, "Data/Knowledge Packets as a means of Supporting Semantic Heterogeneity in Multidatabase Systems," In ACM SIGMOD Record, Vol. 20, No. 4, 1991, pp.69-73.