TUTORIAL
on
"The BLOOM Model for Database Interoperability"

(Invited talk)

by Prof. Felix Saltor, Universitat Politecnica de Catalunya, Barcelona

(Extended abstract)

Our BLOOM data model has been developed as the canonical model in our methodology for schema integration in database interoperability. In order to understand why BLOOM is the way it is, let me first place it in context by introducing database interoperability and canonical data models.

1. Database interoperability

When several preexisting databases, designed independently and operating autonomously, are interconnected to form a federation of databases, so that a query to the federation produces a single, consolidated answer (also called "integrated access"), a number of problems arise. This federation (also called "multidatabase") is an heterogeneous and distributed system of autonomous databases, and the interoperation between the databases is a research topic in many centres worldwide (see [Bukhres95], [Sheth90], [Saltor93], [Klas94], [Hsiao92]).

The three main characteristics of a such an interoperable or federated system are:

1) Autonomy: each database was designed autonomously and keeps its freedom to change its design; is free to decide which data to share with whom, and how to execute queries coming from other database systems;

2) Heterogeneity: differences in hardware, operating systems, DBMS (systems heterogenity), including differences in data model and data languages (syntactic heterogeneity); and differences in how the real world is perceived, conceptualized and represented in the databases (semantic heterogeneity); and

3) distribution: this is not one database that is distributed, but a number of separate databases that happen to sit in different nodes of a distributed system.

Several architectures for Federated database systems exist or have been proposed. A reference architecture, based on 5 levels of schemas, has been proposed in [Sheth90]. We presented extensions to this architecture in the previous EDRG Workshop [Saltor94].

In most architectures, syntactical heterogeneity is solved by adopting a common, Canonical Data Model for the federation, so that schemas, queries and data are translated to and from this model. We have classified semantic heterogeneities in [Garcia95c].

2. Canonical Data Models

The role of the Canonical Data Model (CDM) of a Federated Database System is crucial, and therefore which model to select as the CDM is critical.

We have developed [Saltor91] a framework of suitable characteristics for a CDM, based on (1) expressiveness and (2) semantic relativism, and we have analyzed several models according to this framework. We have concluded that object oriented and functional models are best suited as CDMs. This is in line with current research (as reported in [Bukhres95]), which tends to use object oriented models as CDMs, in contrast with what happened in the 80s, when some authors favored Extended E-R models.

3. The BLOOM Model

Our research project at the Universitat Politecnica de Catalunya focuses on semantic issues in database interoperability. We have developed a data model, called BLOOM, that satisfies all suitable characteristics of our framework for CDMs, and that is the basis for our methodology for schema integration, that has three phases. A tool to help the schema integrator, by performing semiautomatically these phases, is under construction.

Phase 1) Semantic Enrichment. Each relational database schema is enriched with additional dependencies (extracted from the extension and made explicit) and then converted to BLOOM: [Castellanos94b].

Phase 2) Detection of interdatabase semantic relationships. A pair of schemas, enriched and converted to BLOOM, are analyzed, by comparing specializations of a class in one schema with the specializations of a corresponding class in the other. Which specializations and classes to compare are selected by an strategy, based on the specialization dimension and on the aggregation dimension of BLOOM. The criteria to decide upon their similarity is based on the aggregation dimension: [Garcia95a].

Phase 3) Resolution of semantic conflicts and construction of a federated schema. Classes and specializations found equivalent or similar in Phase 2 are integrated, and remaining semantic conflicts [Saltor92] are solved. Integration is not done by the standard generalization operator, but by our "discriminated generalization" operator. In this way, the integrated schema preserves all the information, allows "source tagging" of the data, and supports multiple semantics: [Garcia95b].

The Canonical Data Model that makes possible this methodology is BLOOM (BarceLona Object Oriented Model). BLOOM [Castellanos94a] is a semantic extension of an object oriented model, with a clear distinction between three dimensions, and with particular abstractions (constructs) along each one:

1) Classification/Instantiation dimension: Objects, Classes, Metaclasses and the Metametaclass.

2) Generalization/Specialization dimension: Specialization "criteria", and four "kinds" of specialization: Alternative, Disjoint, Complementary and General.

3) Aggregation/Decompostion dimension: Three kinds of aggregations: Simple, Collection, and Composition.

The main peculiarities of BLOOM lie along its generalization/specialization dimension [Saltor95]. A class does not specialize directly into its subclasses, but according to criteria: a class PERSONS may specialize by criterion GENDER into subclasses MEN and WOMEN, and according to criterion OCCUPATION into EMPLOYEES and STUDENTS. The "specialization" by GENDER is of the kind Alternative, while the specialization by OCCUPATION is General.

At the Metaclass level, there is an SPECIALIZATION Metaclass, specialized according to criterion KIND into four sub(meta)classes: ALTERNATIVE, DISJOINT, COMPLEMENTARY and GENERAL. The specialization of PERSONS by GENDER is an instance of the ALTERNATIVE Metaclass, while the specialization by OCCUPATION is an instance of GENERAL. The behaviour of an specialization (for example, an instance of MEN cannot be made member of WOMEN) is specified only once as Metabehaviour of the corresponding Metaclass, and is automatically instantiated into behaviour of the specialization when this is created.

When a class has several specializations (according to several criteria), these should be orthogonal. BLOOM automatically generates the semilattice of all possible combinations of the criteria: in the example, subclasses FEMALE-EMPLOYEES, MALE-EMPLOYEES, FEMALE-STUDENTS and MALE-STUDENTS are generated.

In addition to its usage for database interoiperability, the BLOOM model is being implemented in an OODBMS, and is being used for application design.

References:

[Bukhres95] Bukhres & Elmagarmid (eds.) Object Oriented Multidatabases. Prentice-Hall, 1995.

[Castellanos94a] M. Castellanos, F. Saltor & M. Garcia-Solaco: "A Canonical Model for the Interoperability among Object-Oriented and Relational Databases". In: Ozsu, Dayal & Valduriez (eds) Distributed Object Management (Proc. Int. Workshop on Distributed Object Management, IWDOM, Edmonton, Canada, 1992). Morgan Kaufmann 1994, pp.309-314.

[Castellanos94b] M. Castellanos, F. Saltor & M. Garcia-Solaco: "Semantically Enriching Relational Databases into an Object Oriented Semantic Model". In: D. Karagiannis (ed.): Database and Expert Systems Applications (5th International Conference DEXA'94, Athens, 1994). Springer Verlag, LNCS 856, 1994, pp 125-134.

[Garcia95a] M. Garcia-Solaco, F. Saltor & M. Castellanos: "A Structure Based Schema Integration Methodology". In: Proc. 11th Int. Conference on Data Engineering (ICDE'95, Taipei). IEEE-CS Press, 1995, pp 505-512.

[Garcia95b] M. Garcia-Solaco, M. Castellanos & F. Saltor: "A Semantic-Discriminated Approach to Integration of Federated Databases". In: Proc. of the 3rd International Conference on Cooperative Information Systems (CoopIS'95, Vienna), Univ. of Toronto, 1995.

[Garcia95c] M. Garcia-Solaco, F. Saltor & M. Castellanos: "Semantic Heterogeneity in Multidatabases". Invited chapter in [Bukhres95].

[Hsiao92] Hsiao, Neuhold & Sacks-Davis (eds) Interoperable Database Systems (DS-5) (Proceedings, IFIP WG2.6 Database Semantics Conf. on Interoperable Database Systems (DS-5), Lorne, Victoria, Australia, 1992). North-Holland, 1993.

[Klas94] W. Klas: "Overview on Interoperable Database Systems". In: Deductive and Interoperable Databases, Report ERCIM-94-W005, ERCIM, Rocquencourt, 1994.

[Saltor91] F. Saltor, M. Castellanos & M. Garcia-Solaco: "Suitability of Data Models as Canonical Models for Federated DBs". ACM SIGMOD Record vol 20, #4, pp. 44-48 (special issue: A. Sheth (guest ed.):Semantic Issues in Multidatabase Systems, Dec. 1991).

[Saltor92] F. Saltor, M.G. Castellanos & M. Garcia-Solaco: "Overcoming Schematic Discrepancies in Interoperable Databases". In: [Hsiao92], pp.191-205.

[Saltor93] F. Saltor & M. Garcia-Solaco: "Diversity with Cooperation in Database Schemata: Semantic Relativism". Proceedings of the 14th International Conference on Information Systems (ICIS'93, Orlando 1993). Pp. 247-254.

[Saltor94] F. Saltor, B. Campderrich & M. Garcia-Solaco: "On architectures for federated DB systems". In: Deductive and Interoperable Databases (Proc. 6th EDRG Workshop, Barcelona, 1994). Report ERCIM-94-W005, ERCIM, Le Chesnay, 1994, pp 8-25.

[Saltor95] F. Saltor, M. Castellanos, M. Garcia-Solaco & Th. Kudrass: "Modelling Specialization as BLOOM Semilattices". In: H. Jaakkola (ed.) Information Modelling and Knowledge Bases VI (4th European-Japanese Seminar on Information Modelling and Knowledge Bases, Kista, June 1994). IOS Press, Amsterdam, 1995, pp 449-469.

[Sheth90] A. Sheth & J. Larson: "Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases". ACM Computing Surveys, 22:3 (Sept 1990).