Intelligent Access to Heterogeneous Information Sources Report on the 4th Workshop on Knowledge Representation Meets Databases

Franz Baader
baader@informatik.rwth-aachen.de

Manfred A. Jeusfeld
jeusfeld@kub.nl

Werner Nutt
nutt@cs.huji.ac.il

1 Introduction

Integrated access to heterogeneous information is an increasingly important topic as more and more sources that developed independently from each other become accessible over networks. The structure of the information and the abilities of the sources to answer queries may vary widely. Therefore, systems are needed that are able to use knowledge about the contents and the capabilities of sources to break down a global query into portions that can be processed locally and to reassemble the answers. Problems of this kind arise when one wants to access information on the internet or on intranets, or when one wants to build a datawarehouse that integrates and consolidates data from different databases in a large institution. Some intelligence is required for these tasks, since systems must not only process data, but use information (or "knowledge") about data to determine in which way to organize the processing. So, certain database (DB) applications raise questions of a kind that are being dealt with in the area of knowledge representation and reasoning (KR).

The workshop series KRDB (= "Knowledge Representation Meets Databases") is a forum where the cross-fertilization of ideas from the two areas are discussed. This paper reports on the 1997 KRDB workshop [2] that took place in conjunction with VLDB'97 in Athens, Greece, and whose main topic was accessing heterogeneous information.

The KRDB workshop series started in 1994 as an offspring of the Esprit project on Computational Logic (CompuLog). CompuLog had the goal to establish logic programming on a broader basis for systems development by adding software engineering ideas (modules, program transformation), a better understanding of semantics, esp. negation, and knowledge representation techniques for knowledge base management. Roughly speaking, CompuLog tried to marry the static aspects of a system with the dynamic part. The static part, consisting of schema and integrity constraints, can be analyzed with knowledge representation techniques. Logic programming contributes expressive languages for querying. In CompuLog, expressive description logics (DL) have been developed in which rich schemata can be modeled. DLs are predicative languages having only unary predicates (set membership) and binary predicates (relationships between elements). The traditional purpose of a DL is to find out whether one DL expression is subsumed by another. Reasoning with DL's can be made fruitful for checking, e.g., the consistency of schemas or the containment of views [3]. The CompuLog experience showed that there is considerable potential of cross-fertilization between knowledge representation and databases.

Since then, KRDB took place every year. By assembling program committees with researchers from the two fields we wanted to highlight that KRDB is a forum for exchange and assure that contributions are of interest to each side. Now, after four workshops, it is time to present the topics to a larger audience in both communities. In the remainder, we first summarize the topics of previous KRDB workshops and then report in more detail the presentations and discussions at KRDB'97. We conclude with ideas on potential topics for future research in the cross-area of knowledge representation and databases[1].

2 Topics of previous workshops

KRDB started in 1994 in conjunction with the German AI Conference KI'94. It was initiated by Martin Buchheit and the authors of this article. We had the hypothesis that reasoning about schemas and queries would be a subject on which the two areas could exchange results [1]. When class descriptions of database schemas are expressed as formal concept definitions in a suitable description logic, then an AI tool can reason about them to

[1]Contact address: M.A. Jeusfeld, Tilburg University, Infolab, P.O. Box 9015, 5000 LE Tilburg, Netherlands

detect inconsistent descriptions and containment of classes. This reasoning is done independently of the specific content of a database. The instance at any given point in time is just one of the many possible models of the schema. At that time, object-oriented databases were seen as a major trend in the DB world that would create opportunities for advanced schema reasoning.

The second workshop, KRDB'95 in conjunction with KI'95, considered the connection of KR systems with relational databases. Topics were the implementation of KR systems on top of relational databases or the access to a database through a KR system. The lesson learned at KRDB'95 was that the assumption that DB schemas would be the place where KR could be employed was too simplistic: more knowledge about the application and its processes would be needed.

In 1996, KRDB joined the European AI Conference ECAI'96. Views in databases and multi-databases played a major role in the presentations and discussions. Still, description logics were proposed by many speakers as a unifying framework for describing and reasoning about schema elements—however now in a distributed environment. Other presentations studied these problems for conjunctive queries and attempted an amalgamation of conjunctive queries with DL's. New areas were presented where problems touching on both, KR and DB, arise, namely medical terminology bases and data mining.

3 Presentations on Intelligent Access to Heterogeneous Information

This year, KRDB moved to a major database conference in order to strengthen its connection to the DB field. Moreover, the message that there is the possibility of cross-fertilization should be made more popular in the DB community, which we felt largely tends to ignore developments in knowledge base management and KR. As main topic, access to heterogeneous information was chosen. A rough classification of the authors reveals that half of them come from the areas of KR and knowledge bases, and the other half from the DB side. Four papers on the DB side have their roots in deductive databases.

The call or papers had asked the contributors the following questions related to intelligent access to heterogeneous information:

1. What are adequate languages to describe a user's demand for information and the contents of information sources? Can more sophisticated schema languages support this task?

2. Which kind of reasoning is required to mediate between information demand and information supply? Can reasoning services from KR contribute to a solution?

3. What background knowledge about an application domain is needed to formulate and interpret queries over a set of heterogeneous information sources? Is there a role for ontologies?

4. Which is the view that an access tool for a data warehouse should provide to a user? Should the data warehouse appear as a relational database, a set of data cubes, a semantic network, or as a combination of all this?

5. What are adequate formalisms for representing and querying meta data? Should contradictions between different information sources be resolved or is it possible to give meaningful answers to queries in the presence of contradictions?

6. How does incompleteness of information affect system design and query processing?

7. What are suitable formalisms to represent data quality like accuracy, timeliness, and relevance?

Not surprisingly, the answers given are quite diverse. We organize them into six categories according to the approach taken.

3.1 Access by database views

Nick Roussopoulos started his presentation on "Materialized views and data warehouses" with the statement that (relational) views are perhaps the most important concept in the database domain. He argued that a view has multiple facets: a view is a program to generate data from other data, a view is a collection of derived data, a view is an index on other data. In the program facet, one can reason about the re-use of views when they are materialized. This is however not the only benefit. For example, one can derive valuable knowledge for the query optimizer from the value distribution in the view. Thus, the view itself is a premier source of metadata about the underlying database.

Elke Rundensteiner, Amy Lee and Anisoara Nica talked about "Preserving views in evolving environments." They noted that most research concentrated on updating or maintaining the content of views and neglected that the environment may force the view definition itself to change. They presented a taxonomy of view adaptation problems and showed a framework for view evolution for the case of SPJ (selection, project, join) views. Data sources publish their capabilities into a meta knowledge base that a view evolution tool consults whenever a re-definition is required. In their approach, a re-definition may well yield a different answer set. Problems like reuse of of materialized views thus appear in an even more difficult form than investigated in current research on containment and subsumption.

3.2 Information integration by KR reasoning

The representation of interdependencies among schemas was the topic of John Cardiff, Tiziana Catarci and Giuseppe Santucci in their paper "Exploitation of interschema knowledge in a multidatabase system." They start from a generic Entity Relationship data model which is able to represent data from heterogeneous data models. In a second step, a description logic is used to express intensional (schema-based) and extensional (data-based) statements about the inclusion and equality of data sets. This knowledge may then be used for global query optimization and schema integration. Moreover, the interschema assertions are interpreted as consistency constraints on the multidatabases to improve the coherence of the distributed data.

Tiziana Catarci with Luca Iocchi, Daniele Nardi and Giuseppe Santucci then showed that KR reasoning is also very promising for "Conceptual Views over the Web." The content of Web sources is described in a description logic framework. Data elements from the sources do not have a prescribed structure but they are classified into a given concept hierarchy using the CLASSIC reasoner. The result is stored in a database. The user then can use this database as an index to access the original Web sources.

Sonia Bergamaschi and Claudio Sartori presented "An approach for the extraction of information from heterogeneous sources of textual data." They see mediators just above the wrappers of data sources as the ideal place for schema reasoners based on description logics. The architecture follows ODMG's CORBA standard plus Stanford's OEM language for semi-structured data. Their dialect encodes the structural part of CORBA's object-oriented data model. If queries to the data sources are expressed in this language, then not only schema validation can be performed but also query optimization. A prototype called ODBTOOLS is available on the Web.

3.3 Information integration by logical techniques

Yangjun Chen and Wolfgang Benn presented a talk on "Building DD to support query processing in federated systems." The general architecture is similar to the one presented by Bergamaschi and Sartori, but their data sources are purely relational. Extended data dictionaries (DD) are attached to clients. They contain concept mappings (assertions about extensional equality and inclusion of schema concepts) and data mapping rules. The latter are represented as a dialect of Datalog rules to resolve conflicts in the structural representation of data. The meta data is used for query decomposition and query optimization.

Parke Godfrey and Jarek Gryz listed in their presentation "Semantic query caching for heterogeneous databases" multiple goals for storing answers to queries in caches: query optimization, security, fault tolerance, approximate answering and improved user interaction. Having caches, the problem is to find which part of a cache can be reused when a new query is posed to the system. They argue for a Datalog-based representation of cached queries. Theoretical intractability of reasoning on Datalog queries would be compensated by the the relative simplicity of queries in practice. Caches then should also be used for partial query answering (one part of answer is in cache, the rest is computed).

Paul Th. Kandzia and Christian Schlepphorst discussed in their presentation "DOOD and DL - do we need an integration" the advantages of a deductive language (query evaluation in minimized Herbrand interpretation) versus a description logic (reasoning on all interpretations). The latter is more general but also limited to simple assertions due to inherent intractability. As a case study, they took an example from computer linguistics on hypothesis generation that was originally represented in description logics. It turns out that a deductive representation (here F-Logic) is far more compact and that the query processing facilities of F-Logic were sufficient to do the required reasoning. The schema querying property of F-Logic proved to be very useful for this purpose[2].

3.4 Advanced query languages

Reasoning on database schema is just one way to support more intelligent data access. Kazumasa Yokoto, Yukuhaka Banjou, Takashi Kuroda and Takeo Kunishima presented a another approach: provide more information in the answer to a query. In their talk "Extensions of query processing facilities in mediator systems" they add two ideas. Firstly, a query may be augmented by conditional information. This is treated like data that is added to the database temporarily when processing the query. Secondly, the query processor itself generates hypothesis via an abduction process to form answers like "x is true when assumption A holds." Though the syntax resembles F-Logic, the semantics is rather different. The approach is applied to distributed

[2] This observation deserves a remark. When schema concepts are treated as data in a (meta) database and no negation occurs, then the closed-world semantics used in DB and the open-world semantics in DL are the same.

data sources which are wrapped by a logic-based language QUIK.

Vinay Chaudhri and Peter K. Karp tackled the query language issue from a different standpoint. In their talk "Querying schema information" the SQL-like object query language from ODMG is extended by expressions to extract information about the schema of a data source. Typically, such queries are about subclasses of a given class, on the attribute types, its cardinalities and so on. The subclass query should be answered by reasoning on the class definitions, e.g. in a description logic framework. The extended OQL language then not only supports query processing but also query formulation in a heterogeneous setting.

3.5 Architectures for information access

Previous talks already introduced distributed settings for information access. In the talk "Alamo - an architecture for integrating heterogeneous data sources," Daniel P. Miranker and Vasilis Samoladas addressed the client-side heterogeneity problem: a user may select object-oriented, active database, deductive database or data mining views on the heterogeneous data sources. Since all these front-ends have to deal with heterogeneous data sources, a middle layer called abstract search engine is proposed. It exports an abstract cursor to the client query tools. The cursor is a buffer for the retrieved data.

Martin Staudt, Jörg-Uwe Kietz and Ulrich Reimer presented "Adler - an environment for mining insurance data." The architecture is basically a central data warehouse plus a toolkit for data mining plus a meta data manager. The data mining toolkit constitutes a software bus where different tools can be plugged in without affecting the data warehouse itself. The meta data manager, using ConceptBase, maintains an enterprise model plus the schema of data sources. A data analyst can select appropriate data sources by browsing through the meta database. Data access and data transformation are expressed as concept definitions in the meta database. The authors then argue that the data homogenization and integration is the most urgent issue for today's data mining.

Bill Hills, Barry Florida-James and Nick Rossiter devoted their talk to "Semantic equivalence in engineering design databases." Of particular interest is the concept of object identity when a source data object is transformed to a view level in a federated schema. Three kinds of agents are used to manage access to objects. Resource agents are the wrappers for the data sources, behavioral agents reside in the middle and global agents at the client side. The behavioral agent is responsible to pass updates from sources to clients.

3.6 Interpretation of heterogeneous information

In their talk "Quality of service in knowledge collection and management," Eric Hughes with Daryl Morey and Arnon Rosenthal raised the issue of service quality when accessing different information sources. They first noted that different users need different quality factors like accuracy or timeliness. Secondly, existing sources come with different quality themselves. It is argued that a view definition must also contain such quality data. This becomes more and more relevant as more heterogeneous user groups want to access the same enterprise data.

While other authors were much concerned with removing ambiguities, Felix Saltor and Elena Rodriguez took a different standpoint in their presentation "On intelligent access to heterogeneous information." They claim that interpretation of information is always relative to a persons conceptualization of the world. The latter is usually not explicit in an information system. The classified the arising semantic heterogeneities into class extensions, class structures and object instances. Whether the conflicts are resolved should be dependent on the user context. It may well turn out that it is more meaningful not to resolve a heterogeneity because the database designers of the heterogeneous data sources had in fact different concepts in their mind.

Geert-Jan Houben and Frank Dignum talked on "Integrating information for organized work." Like in the previous talk, they claim that knowledge about the pure structure of data is not sufficient for intelligent access to data. Their approach is to view data access as a communication act where a client proposes a request to a server, this requested is negotiated and then agreed and finally the answer is generated and returned to the client. The negotiation and reformulation of requests is based on quality goals of the client. In essence, the old query paradigm is replaced by an agent paradigm, where non-functional quality goals are explicitly passed from the client agent to the server agent.

4 Outlook

So, did the presentations provide answers to the seven questions in the call for papers? We leave the answer to the reader. Instead, we observe some trends in the intelligent access to information. Datalog-based and description logic-based techniques are competing when applied to the core reasoning questions like query subsumption. Datalog is apparently more suitable when query processing and data transformation is dominating the approaches. Description logics seem stronger when a conceptual model of the distributed information is desired. Recent work has come up with advanced reasoning techniques for conjunctive queries and classes of Datalog queries. Research is under way to combine the two paradigms.

CORBA-based environments as well as communicating agents were the two competing architectural paradigms. Approaches relying on CORBA try to augment the definitions in that standard to allow some more intelligent functions like access to schema information and then reasoning on them. Communicating agents appear to be much more flexible and also more fuzzy in terms of implementation.

A problem repeatedly faced by the authors was that of missing information on how to interpret data and schema. Some argue to augment a query at run time by additional information like hypothetical conditions or quality goals. Others vote for adding such information at compile time, i.e., when forming a federated schema. The compile time solution allows for some formal reasoning, the run time approach tends to be much more pragmatic.

Views were a central concept basically all presentations. It seems that they offer much more research opportunities than just the old "query subsumed by view" problem. Future research should not just take the structural aspects into account but also all kinds of meta data about the view like quality requirements and the conceptualization of the view in the user's mind. Unfortunately, we do not have a good way to represent such meta data up to now. Such meta data is not explicit in the database schema – it should probably be in the future.

Quite a lot of system implementations were reported at KRDB'97. The ODBTOOLS employ a DL reasoner in a CORBA-based architecture. The ADLER toolkit for data mining uses a meta data repository. Borgida's CLASSIC system is used for organizing semi-structured data. F-Logic and QUIK are systems for advanced querying and so on. Thus, there is already a record of experience of amalgamating KR and DB.

The summary should have shown that it is worth to do more research and to have more discussion on how knowledge representation and databases would meet. We are facing the appearance of a global information network where more and more humans have access to a vast amount of data. Tools helping humans to access data in a meaningful, knowledgeable way are desperately needed.

The topic of intelligent access to heterogeneous information will continue to be in the focus of KRDB. The next workshop is being organized in the first week of June 1998 adjoint to SIGMOD/PODS'98 in Seattle, USA. We invite researchers and practitioners from the KR and DB areas to join the workshop by contributing a position paper. More information on KRDB'98 can be obtained from http://www.ai.sri.com/krdb98/.

Interested readers can access all papers of KRDB'97 as well of its predecessor workshops via the WWW address given below [2].

Acknowledgements. We would like to thanks the organizers of VLDB'97, esp. Timos Sellis, for being excellent hosts for KRDB'97. Many thanks go to Spyros Ligoudistianos for printing the proceedings in the last minute.

References

[1] F. Baader, M. Buchheit, M.A. Jeusfeld, W. Nutt: "Reasoning about structured object - knowledge representation meets databases," *Knowledge Engineering Review*, 10, 1, 1996, pp. 73-76.

[2] F. Baader, M.A. Jeusfeld, W. Nutt: *Intelligent Access to Heterogeneous Information.* Proceedings of the 4th Workshop KRDB-97, Athens, Greece, August 30, 1997, ONLINE http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-8/.

[3] M. Buchheit, M.A. Jeusfeld, W. Nutt, M. Staudt: "Subsumption between queries to object-oriented databases", *Information Systems*, 19, 1, 1994, pp. 33-54.