

Towards Intelligent Integration of Heterogeneous Information Sources*

Shamkant B. Navathe

Michael J. Donahoo

College of Computing
Georgia Institute of Technology
Atlanta, Georgia 30332-0280
{*sham, mjd*} @ cc.gatech.edu

Abstract

Current methodologies for information integration are inadequate for solving the problem of integration of large scale, distributed information sources (e.g. databases, free-form text, simulation etc). The existing approaches are either too restrictive and complicated as in the “federated” (global model) approach or do not provide the necessary functionality as in the “multidatabase” approach. We propose a hybrid approach combining the advantages of both the federated and multidatabase techniques which we believe provides the most feasible avenue for large scale integration. Under our architecture, the individual data site administrators provide an *augmented export schema* specifying knowledge about the sources of data (where data exists), their structure (underlying data model or file structure), their content (what data exists), and their relationships (how the data relates to other information in its domain). The augmented export schema from each information source provides an intelligent agent, called the “mediator,” knowledge which can be used to infer information on some of the existing inter-system relationships. This knowledge can then be used to generate a partially integrated, global view of the data.

1 Introduction

Much of the research in database interoperability has focused on two extremes: multidatabase and federated systems. Multidatabase [Lit90, Spe88] systems provide a uniform access language to a set of database systems. While this is a necessary first step in solving the problems of heterogeneity, it places most of the integration responsibility on the user which may be unacceptable. Federated systems [She90] propose to create a global view of the underlying systems making the heterogeneity completely transparent to the user. While this approach is enticing, the complexity of constructing a global schema for large scale integration makes this approach infeasible because it requires an administrator who understands the semantics of all underlying systems and can resolve all inter-system schematic conflicts [Bat86]. In addition, the maintenance of a global schema in the face of addition/deletion of systems is difficult.

A better approach to interoperability involves the combination of techniques of reasoning and learning with techniques of data modeling and access to provide a partially integrated, global view. To accomplish this, the administrator of each underlying system presents a semantic description (augmented export schema) of their information to the “mediator.” This augmented export schema may be as simple as the typical export schema or as detailed as a knowledge-based data description of the data, its relationships, and the system’s domain. A knowledge-base system, such as Loom [Bri94], provides the capability to represent knowledge about the underlying information repositories and to make inferences as to the relationships among the various autonomous systems and generalizations concerning the information in each system. We have previously demonstrated that classification hierarchies can be effectively used to carry out integration of schemas [Sav91]. In this paper, we

*To appear in Proceedings of 6th International Hong Kong Computer Society Database Workshop, Hong Kong, February 1995

review the goals and strategy of the project HIPED, Heterogeneous Information Processing for Engineering Design, which we are currently pursuing at the Georgia Institute of Technology.

2 Related Work

Earlier work in integration provides the motivation and framework for our efforts. Batini et al. [Bat86] detail the problems of schema integration and provide a methodology for comparison of proposed solutions. Unlike many earlier integration efforts, we do not limit ourselves strictly to integration of databases. Instead, we focus on the integration of *information sources* including databases, free-form text, hypertext, etc. One possible method of dealing with this wide variety of information is to use Stanford's Object Exchange Model (OEM)[Pap94] which allows information exchange via *self-described* objects[Mar85] between different types of information sources. We propose to adapt the mediator paradigm[Pap94, Wei92, Wei93, Are94] to perform integration of the augmented export schemas. Integration of heterogeneous information sources requires a semantically rich data model. Earlier work has shown that the CANDIDE[Bec89, Nav91] model provides unique integration capabilities not found in traditional models. One major feature of the CANDIDE model is its ability to compute class-subclass relationships even among classes from dissimilar systems by subsumption from class relationship information[Sav91, She93, Wha93, Bra85]. Work with classification in the object-oriented model has produced similar results[Nav95, Are]. A variety of such systems supporting description logics are surveyed in [Bor94].

3 Approach

Our main objective is to build and demonstrate an intelligent interface to a set of (possibly autonomous) information sources including structured databases, knowledge bases, and unstructured data. Figure 1 shows our proposed architecture. The parenthetical references are made to applications developed under the ARPA I3 Initiative. KQML (Knowledge Query and Manipulation Language)[Cha92] allows remote access to knowledge/data bases. LIM (Loom Interface Module)[Par93b] allows import of external database information into Loom data structures. IDI (Intelligent Database Interface)[Par93a] is a common access language to several commercial database systems.

The approach we have selected involves development of an Engineering Design Mediator (EDM) which utilizes meta-knowledge of the underlying information to aid a user in "browsing" the data for relevant information sources and to make informed decisions about a plan for retrieving the appropriate data. To demonstrate this technology, we intend to augment the capabilities of both an autonomous (KRITIK2) and an interactive (Canah-Chab[Goe93]) device design system by providing a mediated interface between the design system and a collection of data/knowledge based systems (D/KBS). The mediator will be responsible for processing queries from the device design systems by determining where relevant data is, sending the appropriate query to the information site, performing the appropriate translations on the data, and returning the data to the design system. The design of the mediator is predicated on the following design goals:

1. Autonomy of the remote systems. Additionally, the remote systems should not be required to perform any functions outside of those defined for the internetwork connecting the system to the mediator.
2. Meta-data query facilities which allow the design system to determine relevant information about component parameters, previous design specifications, device function descriptions, etc. The mediator may also take an active role in helping the design tool determine what information may be helpful (e.g. by use of a thesaurus, domain concept hierarchy, etc).
3. Separation of concerns of the device design system from the query system. This will facilitate reuse of the mediated query system for other intelligent tasks such as planning.

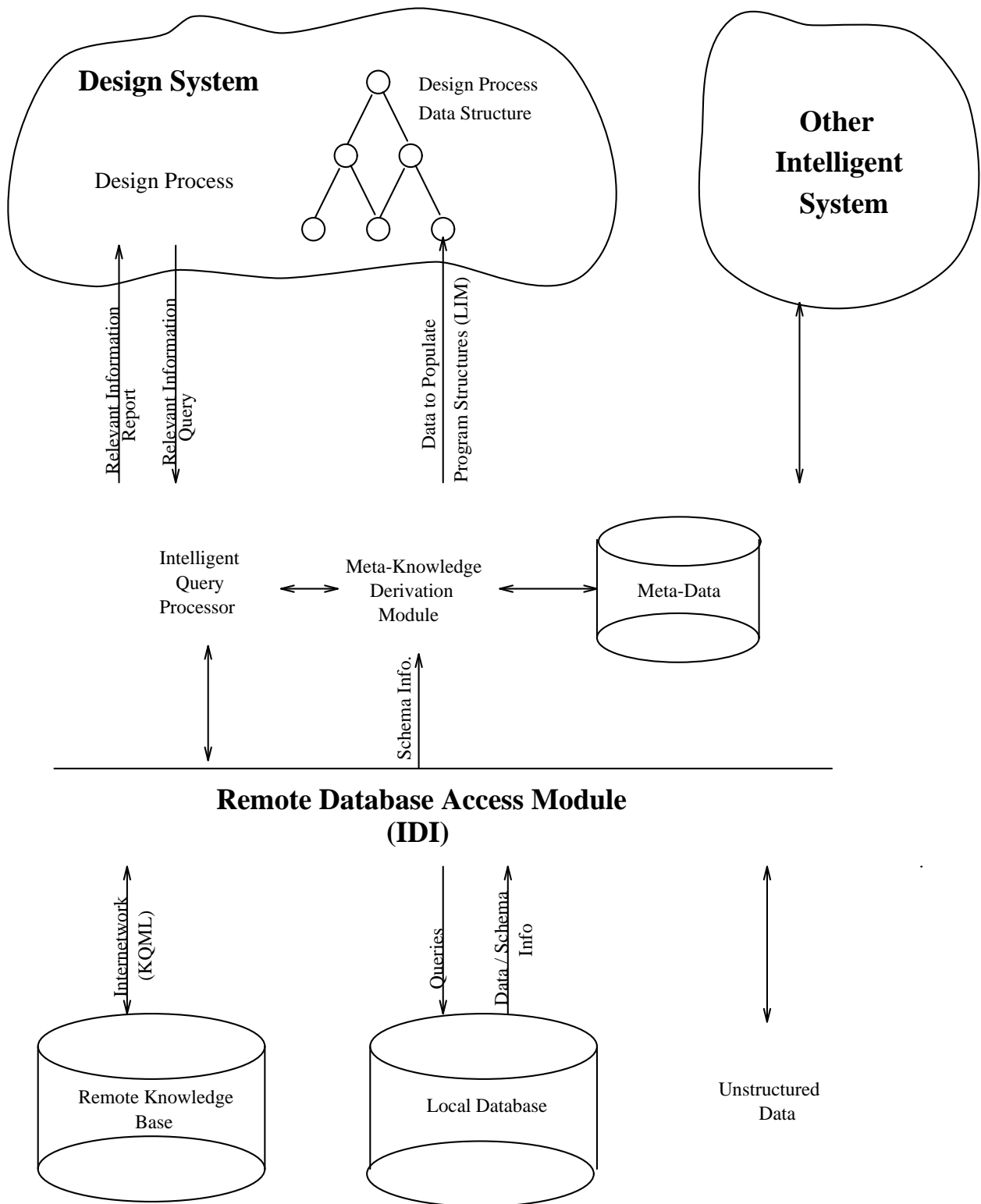


Fig. 1: Proposed Architecture for the Engineering Design Mediator (EDM)

4. Data location (remote vs. local) and data organization (relational, knowledge base, text, etc) transparency.
5. Easy import of external D/KBS information into existing design system data structures minimizing the required changes to the device design system.

These constraints are designed to facilitate reuse of the mediator and to make the use of the system as transparent to intelligent applications as possible. Figure 2 presents an example query processing scenario.

4 Ongoing Research

Research is currently under way in the following areas to facilitate construction of a prototype query system which can be integrated with the device design system:

- Selection and development of the appropriate export data model to represent the data stored at each information source.
- Construction of an export knowledge model whereby information source administrators can express the relationships between their data and real world domain concepts. This in combination with the export data model will define the *augmented export schema*.
- Development of techniques for providing integration of the schemas of information sources into a partially integrated, global schema.
- Determination of optimization techniques for querying the remote information sources. Since the information sources may be interconnected with a WAN, a query processing bottleneck may arise with frequent remote data transmission.
- Provision of a query interface which aids the user in deriving the best answer to a query. Since no completely integrated schema exists and the user does not know what information is available, a query processor is required to guide users to the desired information.
- Capability of inferencing intersource knowledge from the augmented export schemas specifically concerning the relationships between information source entities.
- Ability to learn new, relevant knowledge about information sources based on user interaction.

5 Future Direction

Our initial focus is on providing access of integrated information to intelligent device design systems, but many other applications of this technology exist. With the advent of internetworks which connect thousands of computers all over the world, an explosion has resulted of the available data, both unstructured (text, graphical documents, audio, video, program sources) and structured (under DBMS control), accessible to hundreds of thousands of users. It would be difficult, if not impossible, to integrate all these sites with the current heterogeneous database techniques especially since most sites will not all be willing to provide services beyond those defined by the internetwork. Many query applications already exist for the Internet. WAIS servers provide keyword access to documents; however these documents must be under the control of a WAIS server. Gopher allows sites to setup directories of information that users can browse, but the information can only be accessed in the organization defined by the site manager. Archie provides a keyword query interface to find source code, but the keywords only work on the name of the source file (the user cannot ask for a program that performs some function, X; instead they must find the name of a program that performs X and search for it by name. World Wide Web (WWW) provides a nice interface

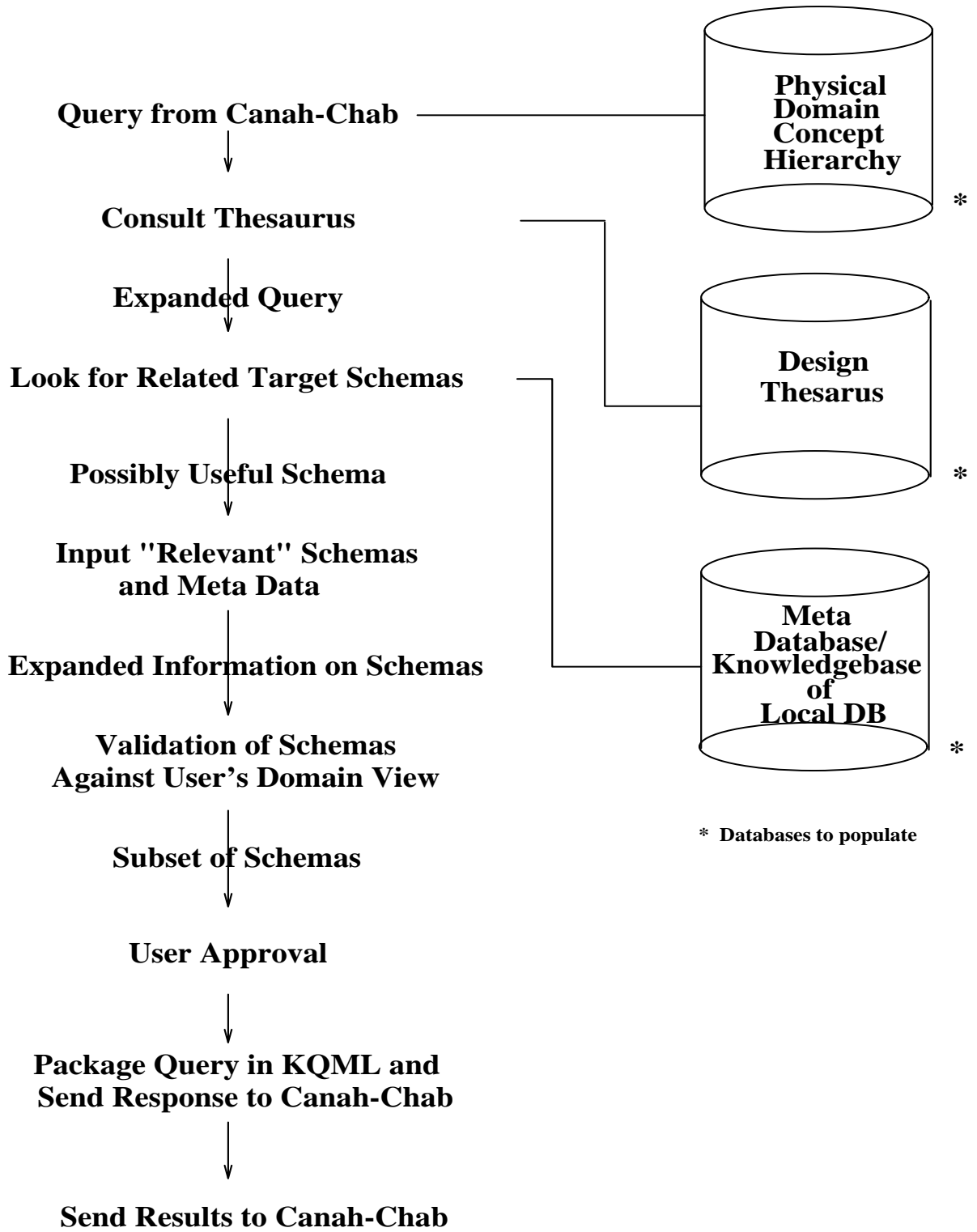


Fig. 2: Query Processing Scenario in the EDM

to information organized by site managers (similar to gopher), but users suffer from the “hypertext navigation problem” which creates difficulties in locating specific information and keeping track of where they are in the web of hypertext documents over time.

Several problems exist for the tools mentioned above. First, the tools access a particular type of data (e.g. Archie only finds source code). If a manual exists for a particular application whose source code is found by Archie, the user is not informed. Second, the tools lack relativism because the users must access the data in the manner dictated by the site manager (e.g. in WWW the data is explicitly organized by hyperlinks). Third, some of the applications require a particular site organization (e.g. Gopher requires a specific directory structure). If a site has information but no desire to organize it, a gopher search may not find the relevant information at that site. Fourth, the query processors provide little organization to the data (e.g. Archie does not organize its source code references by application type, instead all applications with a substring match on the query are returned). For these reasons, the Internet environment provides a true testbed for large scale, heterogeneous information source integration.

We propose a query processing application which, using the native internet network capabilities, provides a single interface for accessing all types of data regardless of source or format. The following list proposes some of the necessary extensions to the EDM:

- The system should perform automated “net surfing” to create an intelligent index of each data store’s information. The intelligence of the index lies in the ability to discern between types of data (audio, text, source, etc), utilize an indexing methodology tailored to the particular data type (e.g. organize keywords of a text document by the document section), and facilitate determination of an object’s relevance for a query based on the knowledge of the user’s interests and technical expertise. This should require no a priori knowledge of the individual data site organization. Work is being done at the Georgia Institute of Technology in intelligent text document processing and work has been done at IBM Almaden Research Center in file classification[Vee95a]. Extensive work has been done on parsers for the various document types (e.g. html, LaTeX) on the Internet.
- The problem of data overload may result from this large scale integration. Our query processor should utilize user profiles so that only data of specific relevance and technical difficulty will be derived. Unfortunately, the user profile method of data overload reduction may eliminate relevant documents. To deal with this problem, the user needs feedback from the query processor in the form of a description of what information is/is not being considered and an explanation of why. Work in explanation is part of the Canah-Chab System[Goe93].
- Keyword searches should not be limited by the vocabulary of the query; instead, a thesaurus should be used to consider synonyms. This may result in synonym overload so user profiles should also be used in pruning the list of synonyms.
- The user is assumed to be “browsing” the available information; therefore, the query interface should provide reformulation capabilities. Reformulation techniques include iterative query alteration and positive/negative feedback from the user[Vee95b].
- The system should attempt automated knowledge acquisition to provide a better understanding of indexed objects and to find other available data stores. The following list orders levels of object knowledge in ascending complexity:

ID Knowledge - System only knows site assigned ID of object (e.g. filename)

Content Knowledge - System knows information about object content (e.g. keywords for text)

Description Knowledge - System knows content knowledge and an external specification of the object.

Interrelational Knowledge - System knows all of the above and interobject relationships (e.g. papers about cancer research grouped together).

- The system should be extensible with respect to “plugging-in” different types of data indexing components and user profiles. Additionally, the system should transparently handle adding/subtracting participating sites. Utilities already exist for component indexing including parsers for various document types, image recognition utilities, etc.
- Different server systems should be able to exchange information and knowledge. Work in KQML at the University of Maryland facilitates knowledge interchange even with differing ontologies[Cha92].
- Objects must be described in terms of a nested model. For example, a document may be composed of sections which are composed of text, subsections, and graphics. Stanford’s Object Exchange Model (OEM) provides “self-describing,” nested objects[Pap94].
- The distributed control of the system leads to problems of object identity. For example, identical application source code may reside in multiple locations; therefore, the system should attempt to provide object identity to facilitate replicated object identification. Additionally, object versioning will allow the system to keep track of more recent versions of a retrieved object. A primitive form of object identification is supported in Stanford’s OEM project [Pap94].
- External knowledge sources should be used to learn about objects in the system. For example, the query processor could inspect newsgroups or look at the manner in which objects are used in WWW to acquire knowledge about the objects and their relationships. Primitive forms of natural language understanding and concept derivation techniques may be used.
- Use of existing query systems should be considered (e.g. use WAIS server to augment search).
- Special consideration should be given to optimization including reuse of retrieved data[Don93].

6 Conclusion

We have presented a framework for research in the area of intelligent, large scale integration of information sources. Clearly, much more work needs to be done before any of the detailed functionality can be implemented. We believe that much of the research into the necessary technology has begun, and the main task lies in tailoring these technologies to the needs of large scale integration and applying them in a prototype environment. We intend to further study the concepts presented above in order to develop a flexible and extensible scheme for integrating information from heterogeneous sources. Although we wish to experiment by applying our research in the area of augmenting intelligent device design in engineering, the applicability of this technology obviously extends beyond the engineering domain.

References

- [Are] Yigal Arens, Chin Chee, Chun-Nan Hsu, and Craig A. Knoblock. Retrieving and integration data from multiple information sources. To appear in International Journal on Intelligent and Cooperative Information Systems.
- [Are94] Yigal Arens, Chin Chee, Chun-Nan Hsu, Hoh In, and Craig A. Knoblock. Query processing in an information mediator. ISI Technical Report, 1994.

- [Bat86] C. Batini, M. Lenzenini, and S. B. Navathe. A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys*, 18(4):325–364, Dec. 1986.
- [Bec89] Howard W. Beck, Sunit K. Gala, and Shamkant B. Navathe. Classification as a query processing technique in the CANDIDE semantic data model. In *1989 IEEE Conference on Data Engineering*, pages 572–581. IEEE, 1989.
- [Bor94] Alexander Borgida. Description logics in data management. Technical report, Rutgers University, July 1994.
- [Bra85] R. Brachman and G. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9(2):171–216, 1985.
- [Bri94] David Brill. *Loom Reference Manual (Version 2.0)*. ISX Corp, October 1994.
- [Cha92] Hans Chalupsky, Tim Finin, Rich Fritzon, Don McKay, Stu Shapiro, and Gio Weiderhold. An overview of KQML: A knowledge query and manipulation language. Technical report, KQML Advisory Group, April 1992.
- [Don93] Michael J. Donahoo. *Integration of Information in Heterogeneous Library Information Systems*. Master’s thesis, Baylor University, May 1993.
- [Goe93] Ashok K. Goel, Andres Garza, Nathalie Grue, M. Recker, and T. Govindaraj. Beyond domain knowledge: Towards a computing environment for the learning of design strategies and skills. Technical report, College of Computing, Georgia Tech, 1993.
- [Lit90] Witold Litwin, Leo Mark, and Nick Roussopoulos. Interoperability of multiple autonomous databases. *ACM Computing Surveys*, 22(3):267–293, September 1990.
- [Mar85] Leo Mark. *Self-Describing Database Systems - Formalization and Realization*. PhD thesis, Computer Science Department, University of Maryland, 1985.
- [Nav91] Shamkant Navathe, Sunit K. Gala, and Seong Geum. Application of the CANDIDE semantic data model for federations of information bases. In *Invited paper, COMAD '91*, Bombay, India, December 1991.
- [Nav95] Shamkant B. Navathe and Ashoka N. Savasere. A practical schema integration facility using an object-oriented model. To be published in *Object Oriented Multidatabase Systems: A Solution for Advanced Applications* (O. Bukhres and A. Elmagarmid, eds), Prentice-Hall, January 1995.
- [Pap94] Yannis Papakonstantinou, Hector Garcia-Molina, and Jennifer Widom. Object exchange across heterogeneous information sources. Stanford University, Department of Computer Science, Technical Report, 1994.
- [Par93a] Paramax System Corporation. *Computer System Operator’s Manual for the Cache-Based Intelligent Data Interface of the Intelligent Database Interface*, revision 2.3 edition, Feb. 1993.
- [Par93b] Paramax Systems Corporation. *Software Design Document for the Loom Interface Module (LIM) of the Cache-Based Intelligent Database Interface*, revision 2.0 edition, Jan. 1993.
- [Sav91] Ashoka Savasere, Amit Sheth, Sunit Gala, Shamkant Navathe, and Howard Marcus. On applying classification to schema integration. In *First International Workshop on Interoperability in Multidatabase Systems*, pages 258–261. IEEE Computer Society, IEEE Computer Society Press, April 1991.

- [She90] Amit P. Sheth and James A. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22(3):183–236, September 1990.
- [She93] Amit P. Sheth, Sunit K. Gala, and Shamkant B. Navathe. On automatic reasoning for schema integration. *International Journal of Intelligent and Cooperative Information Systems*, 2(1):23–50, 1993.
- [Spe88] R. Speth, editor. *Global View Definition and Multidatabase Languages - Two Approaches to Database Integration*. Amsterdam: Holland, April 1988.
- [Vee95a] Aravindan Veerasamy, Scott Hudson, and Shamkant Navathe. Visual interface for textual information retrieval systems. To appear in Proceedings of IFIP 2.6 Third Working Conference on Visual Database Systems, Lausanne, Switzerland, Springer Verlag, March 1995.
- [Vee95b] Aravindan Veerasamy and Shamkant Navathe. Querying, navigating and visualizing an online library catalog. Submitted for Publication, January 1995.
- [Wei92] Gio Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, pages 38–49, March 1992.
- [Wei93] Gio Wiederhold. Intelligent integration of information. In Arie Segev, editor, *ACM SIGMOD International Conference*, volume 22, pages 434–437. ACM, ACM Press, June 1993.
- [Wha93] Whan-Kyu Whang, Sharma Chakravathy, and Shamkant B. Navathe. Heterogeneous databases: Toward merging and querying component schema. *Computing Systems*, 6(3), August 1993. (a Univ. of California Press publication).