Hybrid Knowledge- and Databases

Merwyn Taylor mtaylor@cs.umd.edu
Department of Computer Science
University of Maryland
College Park, MD 20742
January 3, 1996[1]

In the modern era, databases have been created spanning many domains. Hospitals have databases containing patient information, schools have databases containing student records, and manufacturing companies maintain data about product defects just to name a few. However, these databases do not contain general knowledge about their respective domains. For example, whereas a medical database could contain an entry for a patient with some medical disorder, it would not normally contain taxonomic information about medical disorders, known causal agents, symptoms, etc. Collections of this sort of general information are usually called knowledge bases and powerful tools have been developed for querying these collections in complex and flexible ways. The research described in this abstract aims to develop methodologies for merging existing databases with knowledge bases, so that the power and flexibility of knowledge base technology can be applied to existing collections of data.

In traditional DB's, queries are limited to those that do not reference general information about the domains in which the DB's are created. To support queries that reference such information, one can merge a DB created in some domain with a knowledge base of the domain to create a hybrid knowledge base / database. Merging a knowledge base with a database extends the range of queries that can be issued. In some cases, ad hoc queries can be simplified when applied to the hybrid KB/DB. Besides expanding the range of queries that can be issued, a hybrid KB/DB also lends itself to research in attribute oriented data mining techniques.

Merging a KB with a DB requires a common representation scheme. In this research, a frame based semantic network, Parka, developed by the PLUS group at the University of Maryland was used to facilitate the merger. A DB is first converted to a semantic network. After the DB is converted, all

[1]This research was supported by grants from ONR (N00014-J-91-1451), AFOSR (F49620-93-1-0065), the ARPA/Rome Laboratory Planning Initiative (F30602-93-C-0039, DABT-63-95-C-0037)), the ARPA I3 Initiative (N00014-94-10907).

interesting terms from the DB are enumerated and a knowledge base is built on the basis of this list of terms.

As an example, we are working with a medical database of about 20,000 OB/GYN patients. The DB is a single table with 70+ columns. Using the original DB, a user could query the DB to list all patients with a particular type of infection. But if the user wanted to query the DB to list all patients with infections known to be caused by any form of bacteria, the query would have to be expressed as a disjunction over all the infections that the user knows to be caused by bacteria. Since the DB does not contain a general concept of *bacteria,* the query would have to be constructed without any references to that term. Rather it would look something like:

select *patient*
from *ptable*
where *infection = 'Peritonitis'* **OR** *infection = 'Pyelonephritis'* **OR ...**

This requirement limits the number of users that can issue interesting queries to those that have an extensive background in OB/GYN. With a hybrid KB/DB, a broader spectrum of users can issue interesting queries in a form much simpler than the one mentioned above. We have developed a hybrid OB/GYN KB/DB from the DB. (It contains 60,000 frames and 1.3 million assertions!) Figure 1 illustrates how the previous query can be expressed in a simpler form using Parka's Graphical Query By Example System.

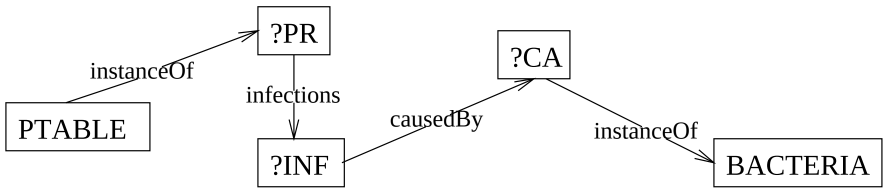

Figure 1: Sample Parka Query

In my current research I am developing general methodologies for merging DB's and KB's that allow this technique to be more widely applied. I am investigating techniques that automatically convert relational DB's to semantic networks by analyzing the schemas of relational DB's. I am also investigating techniques to automatically extract and integrate parts of existing KB's, such as NIH's medical term ontology, UMLS.