Semint: A System Prototype for Semantic Integration in Heterogeneous Databases †

Wen-Syan Li
acura@eecs.nwu.edu

Chris Clifton
clifton@eecs.nwu.edu

Northwestern University/Dept. of EECS
Evanston, Illinois, 60208-3118

One important step in integrating heterogeneous databases is matching equivalent attributes: Determining which fields in two databases refer to the same data. In semantic integration, attributes are compared in a pairwise fashion to determine their equivalence. Automation is critical to integration as the volume of data or the number of databases to be integrated increase. Semint "discovers" how to match equivalent attributes from information that can be automatically extracted from databases; as opposed to requiring human knowledge to predefine what makes attributes equivalent.

System overview

Integration involves extracting semantics, expressing them as metadata, and matching semantically equivalent data elements. Semint (SEMantic INTegrator) is a system prototype for semantic integration being developed at Northwestern University. It utilizes both schema information and data contents to determine attribute equivalence. Semint supports access to a variety of database systems. Currently we have Oracle7 parser and are testing Ingres and "flat file" parsers; other parsers will be developed as resources allow. Figure 1 outlines this semantic integration process. In this process, DBMS specific parsers extract metadata (schema information and data content statistics) from databases and transform them into a single format. Then, a classifier is used to learn how to discriminate among attributes in a single database. The classifier output (the weights of cluster centers) is used to train a neural network to recognize categories; this network can then determine similar attributes between databases.

†This material is based upon work supported by the National Science Foundation under Grant No. CCR-9210704.

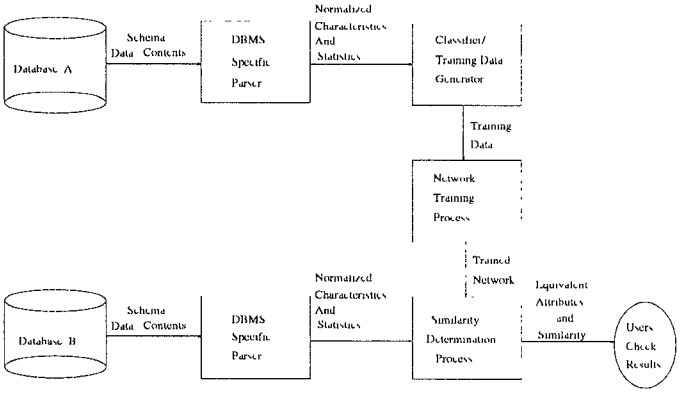

Figure 1: Semantic Integration in Semint

Demonstration

In this demonstration, we show how Semint extracts metadata from real databases, builds and trains neural networks to identify similar attributes and determines their similarity.

Status

Semint operates in a graphical interactive mode (allowing users to provide known information about the semantic integration problem at hand), or batch mode. It is implemented using C and Motif, and runs on IBM RS6000s under AIX and Sun workstations (Sun OS). Databases to be integrated are accessed directly using automatic "catalog parsers". DBMS specific parsers are implemented using SQL embedded C (e.g. Pro*C in Oracle7). The source code of the current system is also available through anonymous FTP from eecs.nwu.edu in /pub/semint. We are currently experimenting with Semint in collaboration with a Fortune 500 firm to gather practical results on large databases.

References

Wen-Syan Li and Chris Clifton, "Semantic Integration in Heterogeneous Databases Using Neural Networks", in *Proceedings of 20th International Conference on Very Large Databases*, Page 1-12, Santiago, Chile, September 12-15 1994.