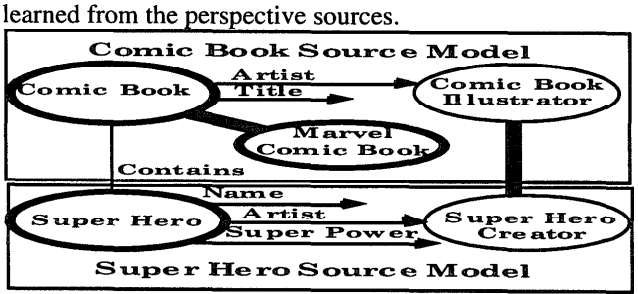
Learning Models for Multi-Source Integration

Sheila Tejada Craig A. Knoblock Steven Minton
University of Southern California/ISI
4676 Admiralty Way
Marina del Rey, California 90292
{tejada,knoblock,minton}@isi.edu

Because of the growing number of information sources available through the internet there are many cases in which information needed to solve a problem or answer a question is spread across several information sources. For example, when given two sources, one about comic books and the other about super heroes, you might want to ask the question "Is Spiderman a Marvel Super Hero?" This query accesses both sources; therefore, it is necessary to have information about the relationships of the data within each source and between sources to properly access and integrate the data retrieved. The SIMS information broker captures this type of information in the form of a model. All the information sources map into the model providing the user a single interface to multiple sources.

Presently, models are manually constructed by human experts who are familiar with the data stored in the sources. Automation of this task would improve efficiency and accuracy, especially for large information sources. We have conducted preliminary work in automating model construction. There has been related work conducted in this area (Perkowitz & Etzioni, 1995), which has focused on learning the attributes of sources. Their approach assumes that it has an initial model of the information to be learned from the perspective sources.



The diagram depicts a partial model of two sources. The ovals in the diagram represent classes of information and the lines describe the relationships between the classes. The two types of classes are basic classes and composite classes. Members of a basic class are single-valued elements, such as strings; while the members of a composite class are objects, which can have several attributes. **Comic Book Illustrator** and **Super Hero Creator** are basic classes; and **Comic Book**, **Marvel Comic Book**, and **Super Hero** are all represented as composite classes. Arrows represent the attributes of objects, and the thick lines represent the superclass/subclass relationships between classes.

Our approach to learning models examines all of the data to extract the relationships needed. Learning the model is an iterative process which involves user interaction. The user can make corrections or specify an information source to be added or deleted. The model proposed to the user is generated by heuristics we have developed to derive the necessary relationships from the data. These heuristics involve creating abstract descriptions of the data. We have assumed the data is stored as tables.

Properties of the data, such as type and range, are contained in an abstract description. For each column of data, an abstract description is computed. Each description corresponds to a basic class. For the basic class **Super Hero Creator** an example of an abstract description would be alphabetic strings with minimum length 8 and maximum 12. These descriptions are then used to help determine the superclass/subclass relationships that exist between the basic classes, like between **Comic Book Illustrator** and **Super Hero Creator**.

Potentially, all basic classes would need to be checked with every other basic class for a superclass/subclass relationship, but now only classes which have similar descriptions are tested. So, for example, the basic class **Super Hero Creator** would only be checked with other basic classes that contain alphabetic strings whose lengths are within the range specified in the abstract description. Our experimental results show that in every case using these restrictions reduced the running time and number of comparisions performed.

We are planning to apply statistical methods to assist in constructing the model, as well as integrating a natural language knowledge base into this process to help determine whether classes are semantically related. The relationships between the basic classes will be useful in the later steps of the modeling process which are to determine the relationships between composite class.

References

Ambite, J., Y. Arens, N. Ashish, C. Chee, C. Hsu, C. Knoblock, and S. Tejada. The SIMS Manual, Technical Report ISI/TM-95-428,1995.

Perkowitz, M. & Etzioni, O. Category Translation: Learning to Understand Information on the Internet. The International Joint Conference on Artificial Intelligence, Montreal, 1995.