

Software Increases Accuracy and Efficiency of Computer Searches



HOW DOES A COMPUTER program churn through a database of credit card transactions and identify potential identity thefts? How does a program analyzing a collection of census

records with tens of thousands of entries determine that three people named J. Smith, James Smith and James R. Smith are one and the same person—while 12 others with the same or similar names are unrelated?

Computer Science Assistant Professor Lise Getoor specializes in answering these difficult questions. To do so, she focuses on link mining and relational mining—these are statistically-based methods for extracting information about the relationships among different persons or things. “To increase the accuracy and efficiency of computer analyses, we’re adding statistical modeling to what computer science knows about structuring data and databases,” says Getoor.

Adding urgency to the effort is, as Getoor says, “the explosion of data” that requires analysis today. “Traditionally databases contained relatively limited and homogeneous sets of information, but that is no longer true. Now we confront more information than we are able to easily manage.”

Algorithms devised by Getoor and her doctoral students can analyze data collections and correctly identify relationships among the varied entities. The team has used their

methodology—in fact, their program encompasses a series of different methodologies—on bibliographic entries, email archives, public health datasets and other forms of data. A paper describing one of the algorithms developed by her student Indrajit Bhattacharya won the best paper award at the recent Society for Industrial and Applied Mathematics Data Mining Conference.

Getoor’s team has applied their algorithms in a variety of domains. For example, they tested their algorithms on epidemiological data related to the spread of tuberculosis. By looking at the characteristics of thousands of individuals with the disease, they were able to identify, “in a richer way than ever before who was at greater risk and the cases which should be followed up carefully to ensure they receive treatment,” notes Getoor.

The team’s success was documented last summer at a software competition organized by IBM. Contestants analyzed a large data set with millions of author references. “In all cases we were able to do a better job identifying matched sets in the database than traditional approaches based on attributes alone,” says Getoor. Now that their

methodology is recognized as superior, the team is making sure that its algorithms continue to perform well when data sets contain more diverse collections of entities.

Success in the experimental realm led Getoor and computer science doctoral students Louis

Licamele and Mustafa Bilgic to create a program that would use what they know about entity resolution to help ordinary computer users cleanse their databases. Together with Professor Ben Shneiderman from the college’s Human Computer Interaction Lab, they have developed an interactive environment for database deduplication. “Users can still make the final decisions. We point out possible or probable duplicates and other errors,” says Licamele.

Their tool, D-Dupe, employs a variety of different measures to analyze relationships among individuals. It can identify matched pairs even when names are misspelled or when they are parsed incorrectly. Users can easily access additional resources, depending on the domain. For example, when analyzing bibliographic databases, the tool can call to Google Scholar, while for a geographic location database calls to Google Earth are supported. Ordinary users may soon benefit from D-Dupe. “We are in the process of making this tool available open source to general users,” reports Getoor.

