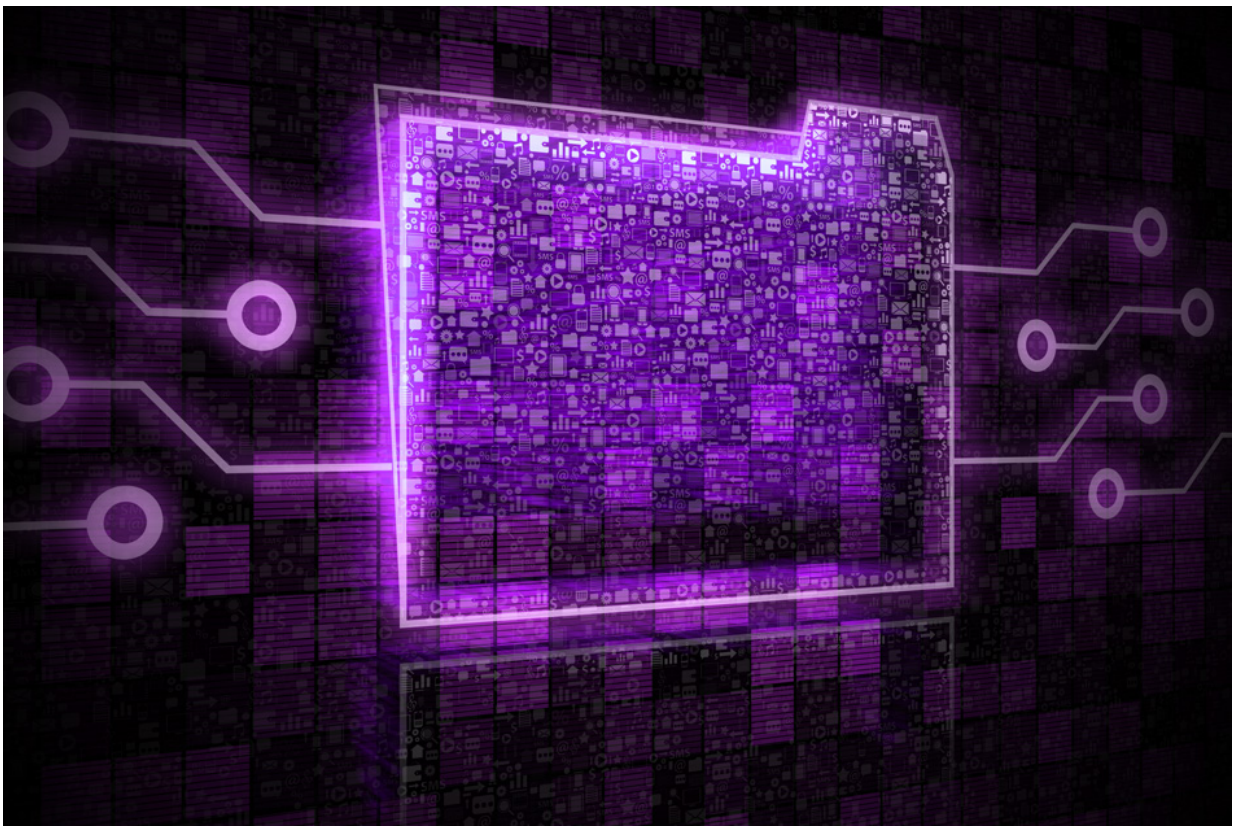
System finds and links related data scattered across digital files, for easy querying and filtering

January 19 2017, by Larry Hardesty



A new system called Data Civilizer automatically finds connections among many different data tables and allows users to perform database-style queries across all of them. The results of the queries can then be saved as new, orderly data sets that may draw information from dozens or even thousands of different tables. Credit: Massachusetts Institute of Technology

The age of big data has seen a host of new techniques for analyzing large data sets. But before any of those techniques can be applied, the target data has to be aggregated, organized, and cleaned up.

That turns out to be a shockingly time-consuming task. In a 2016 survey, 80 data scientists told the company CrowdFlower that, on average, they spent 80 percent of their time collecting and organizing data and only 20 percent analyzing it.

An international team of computer scientists hopes to change that, with a new system called Data Civilizer, which automatically finds connections among many different data tables and allows users to perform database-style queries across all of them. The results of the queries can then be saved as new, orderly data sets that may draw information from dozens or even thousands of different tables.

"Modern organizations have many thousands of data sets spread across files, spreadsheets, databases, data lakes, and other software systems," says Sam Madden, an MIT professor of electrical engineering and computer science and faculty director of MIT's bigdata@CSAIL initiative. "Civilizer helps analysts in these organizations quickly find data sets that contain information that is relevant to them and, more importantly, combine related data sets together to create new, unified data sets that consolidate data of interest for some analysis."

The researchers presented their system last week at the Conference on Innovative Data Systems Research. The lead authors on the paper are Dong Deng and Raul Castro Fernandez, both postdocs at MIT's Computer Science and Artificial Intelligence Laboratory; Madden is one of the senior authors. They're joined by six other researchers from Technical University of Berlin, Nanyang Technological University, the University of Waterloo, and the Qatar Computing Research Institute. Although he's not a co-author, MIT adjunct professor of electrical

engineering and computer science Michael Stonebraker, who in 2014 won the Turing Award—the highest honor in computer science—contributed to the work as well.

Pairs and permutations

Data Civilizer assumes that the data it's consolidating is arranged in tables. As Madden explains, in the database community, there's a sizable literature on automatically converting data to tabular form, so that wasn't the focus of the new research. Similarly, while the prototype of the system can extract tabular data from several different types of files, getting it to work with every conceivable spreadsheet or database program was not the researchers' immediate priority. "That part is engineering," Madden says.

The system begins by analyzing every column of every table at its disposal. First, it produces a statistical summary of the data in each column. For numerical data, that might include a distribution of the frequency with which different values occur; the range of values; and the "cardinality" of the values, or the number of different values the column contains. For textual data, a summary would include a list of the most frequently occurring words in the column and the number of different words. Data Civilizer also keeps a master index of every word occurring in every table and the tables that contain it.

Then the system compares all of the column summaries against each other, identifying pairs of columns that appear to have commonalities—similar data ranges, similar sets of words, and the like. It assigns every pair of columns a similarity score and, on that basis, produces a map, rather like a network diagram, that traces out the connections between individual columns and between the tables that contain them.

Tracing a path

A user can then compose a query and, on the fly, Data Civilizer will traverse the map to find related data. Suppose, for instance, a pharmaceutical company has hundreds of tables that refer to a drug by its brand name, hundreds that refer to its chemical compound, and a handful that use an in-house ID number. Now suppose that the ID number and the brand name never show up in the same table, but there's at least one table linking the ID number and the chemical compound, and one linking the chemical compound and the brand name. With Data Civilizer, a query on the brand name will also pull up data from tables that use just the ID number.

Some of the linkages identified by Data Civilizer may turn out to be spurious. But the user can discard data that don't fit a query while keeping the rest. Once the data have been pruned, the user can save the results as their own data file.

More information: Paper: "The Data Civilizer system"cidrdb.org/cidr2017/papers/p44-deng-cidr17.pdf

Provided by Massachusetts Institute of Technology

Citation: System finds and links related data scattered across digital files, for easy querying and filtering (2017, January 19) retrieved 20 March 2024 from https://phys.org/news/2017-01-links-digital-easy-querying-filtering.html