Data 'hashing' improves estimate of the number of victims in databases

June 5 2018, by David Ruth

Destroyed tanks in front of a mosque in Azaz, Syria, in 2012. Credit: Christiaan Triebert via Wikimedia Commons

Researchers from Rice University and Duke University are using the tools of statistics and data science in collaboration with Human Rights

Data Analysis Group (HRDAG) to accurately and efficiently estimate the number of identified victims killed in the Syrian civil war.

In a paper available online and due for publication in the June issue of the *Annals of Applied Statistics*, the scientists report on a four-year effort to combine a data-indexing method called "hashing with statistical estimation." The new method produces real-time estimates of documented, identified victims with a far lower margin of error than existing statistical methods for finding duplicate records in databases.

"Throwing out duplicate records is easy if all the data are clean—names are complete, spellings are correct, dates are exact, etc.," said study co-author Beidi Chen, a Rice graduate student in computer science. "The war casualty data isn't like that. People use nicknames. Dates are sometimes included in one database but missing from another. It's a classic example of what we refer to as a 'noisy' dataset. The challenge is finding a way to accurately estimate the number of unique records in spite of this noise."

Using records from four databases of people killed in the Syrian war, Chen, Duke statistician and machine learning expert Rebecca Steorts and Rice computer scientist Anshumali Shrivastava estimated there were 191,874 unique individuals documented from March 2011 to April 2014. That's very close to the estimate of 191,369 compiled in 2014 by HRDAG, a nonprofit that helps build scientifically defensible, evidence-based arguments of human rights violations.

But while HRDAG's estimate relied on the painstaking efforts of human workers to carefully weed out potential duplicate records, hashing with statistical estimation proved to be faster, easier and less expensive. The researchers said hashing also had the important advantage of a sharp confidence interval: The range of error is plus or minus 1,772, or less than 1 percent of the total number of victims.

"The big win from this method is that we can quickly calculate the probable number of unique elements in a dataset with many duplicates," said Patrick Ball, HRDAG's director of research. "We can do a lot with this estimate."

Shrivastava said the sharpness of the hashing estimate is due to the technique used to index the casualty records. Hashing involves converting a complete data record—a name, date, place of death and gender in the case of each Syrian war casualty—into one number called a hash. Hashes are produced by an algorithm that considers the alphanumeric information in a record, and they are stored in a hash table that works much like the index in a book. The more textual similarity there is between two records, the closer together their hashes are in the table.

"Our method—unique entity estimation—could prove to be useful beyond just the Syrian conflict," said Steorts, assistant professor of statistical science at Duke.

She said the algorithm and methodology could be used for medical records, official statistics and industry applications.

"As we collect more and more data, duplication is becoming a more timely and socially important problem," Steorts said. "Entity resolution problems need to scale to millions and billions of records. Of course, the most accurate way to find duplicate records is having an expert check every record. But this is impossible for large data sets, since the number of pairs that needs to be compared grows dramatically as the number of records increase."

For example, a record-by-record analysis of all four Syrian war databases would entail some 63 billion paired comparisons, she said.

Shrivastava, assistant professor of computer science at Rice, said, "If you make assumptions, like dates that are close might be duplicates, you can reduce the number of comparisons that are needed, but every assumption comes with a bias, and ultimately you want an unbiased estimate. One statistical approach that avoids bias is random sampling. So perhaps choose 1 million random pairs out of the 63 billion, see how many are duplicates and then apply that rate to the entire dataset. This produces an unbiased estimate, which is good, but the likelihood of finding duplicates purely by random is quite low, and that gives a high variance.

"In this case, for example, random sampling could also estimate the documented counts at around 191,000," he said. "But it couldn't tell us with any certainty whether the count was 176,000 or 216,000 or some number in between.

"In recent work, my lab has shown that hashing algorithms that were originally designed to do search can also be used as adaptive samplers that precisely mitigate the high variance associated with random sampling," Shrivastava said.

"Resolving every duplicate seems very appealing," he said, "but it is the harder way of estimating the number of unique entities. The new theory of adaptive sampling with hashing allows us to directly estimate unique entity counts efficiently, with high confidence, without resolving the duplicates."

"At the end of the day, it's been phenomenal to make methodological and algorithmic progress motivated by such an important problem," Steorts said. "HRDAG has paved the way. Our goal and hope is that our efforts will prove useful to their work."

Shrivastava and Steorts said they are planning future research to apply

the hashing technique for unique entity approximation to other types of datasets.

More information: Unique Entity Estimation with Application to the Syrian Conflict. arxiv.org/abs/1710.02690

Provided by Rice University