Free dataset archive helps researchers quickly find a needle in a haystack

July 18 2019

Ahmed Eldawy. Credit: UC Riverside

Let's say you're doing research that requires millions of geotagged tweets. Or perhaps you're a journalist who wants to map murders in Chicago from 2001 to the present. You need to find large spatio-temporal datasets—but where?

While there are hundreds of publicly available datasets, locating them can take months of searching. When potential sources are found, they rarely provide enough information for a researcher to decide if the set actually contains the kind of data they need without downloading the often huge file and sorting through it first.

Thanks to a computer scientist at the University of California, Riverside, finding the right dataset is now as easy as bookmarking a website, and it costs absolutely nothing.

Ahmed Eldawy, an assistant professor of computer science in the Marlan and Rosemary Bourns College of Engineering, and his group spent the last three years combing the internet for public spatio-temporal datasets, studying their attributes, and summarizing the results for each set on interactive maps that show the user exactly what they're getting.

"People who work on data science need datasets but can spend a lot of time finding them," Eldawy said. "I wanted to build an archive they can find easily."

Called the UCR Spatio-temporal Active Repository, or UCR STAR, the archive is made available as a service to the research community to provide easy access to large spatio-temporal datasets through an interactive exploratory interface. Users can search and filter those datasets as if shopping for their research, except that everything is free.

"The map interface visualizes the data, so you can see if it's a good fit," Eldawy said. "It's like a catalog for datasets."

At the heart of UCR STAR, the map provides an interactive exploratory interface for the dataset. Similar to Google Maps or other web maps, users can zoom in and out and pan around to get a quick overview of the data distribution, coverage, and accuracy.

Important details are displayed once a dataset is selected, such as the original homepage, a link to the original download source, size in bytes, number of records, file format, and other useful information. The subset download feature allows users to quickly download the data in a given geographical region, which reduces the download size. They can also

embed their customized view on a webpage or share the link via social media and bookmark it to revisit later.

UCR STAR contains 102 datasets and 5 billion records. The datasets were mapped using Da Vinci, an open source framework built on top of Apache Spark that Eldawy designed to work with spatial data. The UCR STAR website is best accessed through a desktop browser but also has a limited mobile-friendly interface.

More information: UCR STAR: star.cs.ucr.edu/

Provided by University of California - Riverside