LearnedSketch AI system for frequency estimation improves estimates of trending search queries

4 April 2019, by Adam Conner-Simons

Credit: Stuart Miles/Freerange

If you look under the hood of the internet, you'll find lots of gears churning along that make it all possible.

For example, take a company like AT&T. They have to intimately understand what internet data are going where so that they can better accommodate different levels of usage. But it isn't practical to precisely monitor every packet of data, because companies simply don't have unlimited amounts of storage space. (Researchers actually call this the "Britney Spears problem," named for search engines' long-running efforts to tally trending topics.)

Because of this, tech companies use special algorithms to roughly estimate the amount of traffic heading to different IP addresses. Traditional frequency-estimation algorithms involve "hashing," or randomly splitting items into different buckets. But this approach discounts the fact that there are patterns that can be uncovered in high volumes of data, like why one IP address tends to generate more internet traffic than another.

Researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) have devised a new way to find such patterns using machine learning.

Their system uses a neural network to automatically predict if a specific element will appear frequently in a data stream. If it does, it's placed in a separate bucket of so-called "heavy hitters" to focus on; if it doesn't, it's handled via hashing.

"It's like a triage situation in an emergency room, where we prioritize the biggest problems before getting to the smaller ones," says MIT Professor Piotr Indyk, co-author of a new paper about the system that will be presented in May at the International Conference on Learning Representations in New Orleans, Louisiana. "By learning the properties of heavy hitters as they come in, we can do frequency-estimation much more efficiently and with much less error."

In tests, Indyk's team showed that their learning-based approach had upwards of 57 percent fewer errors for estimating the amount of internet traffic in a network, and upwards of 71 percent fewer errors for estimating the number of queries for a given search term.

The team calls their system "LearnedSketch," because they view it as a method of "sketching" the data in a data stream more efficiently. To their knowledge, it's the world's first machine learning-based approach for not just frequency-estimation itself, but for a broader class of so-called "streaming" algorithms that are used in everything

from security systems to natural language processing.

LearnedSketch could help tech companies more effectively crunch all kinds of meaningful data, from trending topics on Twitter to spikes in web traffic that might suggest future distributed denial-of-service attacks. E-commerce companies could use it to improve product recommendations: If LearnedSketch found that customers tend to do more comparative shopping for household electronics than for toys, it could automatically devote more resources to ensuring the accuracy of its frequency counts for electronics.

"We're all familiar with consumer-facing applications of machine learning like natural language processing and speech translation," says Sergei Vassilvitskii, a computer scientist who studies algorithmic machine learning and was not involved in the project. "This line of work, on the other hand, is an exciting example of how to use machine learning to improve the core computing system itself."

What's also surprising about LearnedSketch is that, as it learns how to count items, the structure it learns can be generalized even to unseen items. For example, to predict which internet connections have the most traffic, the model learns to cluster different connections by the prefix of their destination IP. This is because places that generate large traffic, like big companies and universities, tend to share a particular prefix.

"We combine the model with classical algorithms so that our algorithm inherits worst-case guarantees from the classical algorithms naturally," says Ph.D. student Chen-Yu Hsu, co-author of the new paper. "These kinds of results show that machine learning is very much an approach that could be used alongside the classic algorithmic paradigms like 'divide and conquer' and dynamic programming."

More information: Learning-Based Frequency Estimation Algorithms: openreview.net/pdf?id=r1lohoCqY7

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology