Shrinking 'ridiculous' data sets to manageable size

14 May 2009, By Bill Steele

Two decades ago a renowned statistician described a computer data set of 1 billion bytes as "huge" and 10 trillion bytes as "ridiculous."

Today, thanks to the use of computers to collect and generate data, such ridiculously large data sets are common, from genome databases to search engine logs to Wal-Mart sales data. But the ability to monitor and process the data has not kept up with the ability to create it.

With a new three-year, $551,508 Young Investigator Award from the U.S. Office of Naval Research (ONR), Ping Li, Cornell assistant professor of statistical science, is taking a new mathematical approach. His goal: to "shrink" massive data sets into manageable approximations that can be processed in a reasonable length of time to detect such anomalies as denial-of-service attacks on the Internet or to enable computers to learn from experience for such applications as natural language processing, Web searching and computer vision.

"Instead of storing the whole data, we compute and store a sketch of the data, which is small enough to fit in the memory and still contains enough information to recover crucial relationships of the data," Li explained.

From the resulting sketch, Li says that it is possible, for example, to compute a quantity known as the Shannon entropy, which is, roughly, a measure of the degree of uncertainty in a body of information. A change in this would warn engineers of an anomaly such as a network failure, a large transfer of money or perhaps terrorist chatter. Li also plans to develop and publicly distribute software that can be used as part of machine-learning applications on massive and high-dimensional data sets.

The ONR Young Investigator Program identifies and supports academic scientists and engineers who have received a doctorate or equivalent degrees within the past five years and who show exceptional promise for doing cutting-edge research.

Provided by Cornell University (news : web)