PHYS ORG

# Data mining made faster: New method eases analysis of 'multidimensional' information

July 22 2010

To many big companies, you aren't just a customer, but are described by multiple "dimensions" of information within a computer database. Now, a University of Utah computer scientist has devised a new method for simpler, faster "data mining," or extracting and analyzing massive amounts of such data.

"Whether you like it or not, Google, Facebook, Walmart and the government are building profiles of you, and these consist of hundreds of attributes describing you" - your online searches, purchases, shared videos and recommendations to your Facebook friends, says Suresh Venkatasubramanian, an assistant professor of computer science.

"If you line them up for each person, you have a line of hundreds of numbers that paint a picture of a person: who they are, what their interests are, who their friends are and so forth," he says. "These strings of hundreds of attributes are called high-dimensional data because each attribute is called one dimension. Data mining is about digging up interesting information from this high-dimensional data."

A group of data-mining methods named "multidimensional scaling" or MDS first was used in the 1930s by psychologists and has been used ever since to make data analysis simpler by reducing the "dimensionality" of the data. Venkatasubramanian says it is "probably one of the most important tools in data mining and is used by countless researchers everywhere."

Now, Venkatasubramanian and colleagues have devised a new method of multidimensional scaling that is faster, simpler, can be used universally for numerous problems and can handle more data, basically by "squashing things [data] down to size."

He is scheduled to present the new method on Wednesday, July 28 in Washington at the premier meeting in his field, the Conference on Knowledge Discovery and Data Mining sponsored by the Association for Computing Machinery.

"This problem of dimensionality reduction and data visualization is fundamental in many disciplines in natural and social sciences," says Venkatasubramanian. "So we believe our method will be useful in doing better data analysis in all of these areas."

"What our approach does is unify into one common framework a number of different methods for doing this dimensionality reduction" to simplify high-dimensional data, he says. "We have a computer program that unifies many different methods people have developed over the past 60 or 70 years. One thing that makes it really good for today's data - in addition to being a one-stop shopping procedure - is it also handles much larger data sets than prior methods were able to handle."

He adds: "Prior methods on modern computers struggle with data from more than 5,000 people. Our method smoothly handles well above 50,000 people."

Venkatasubramanian conducted the research with University of Utah computer science doctoral student Arvind Agarwal and postdoctoral fellow Jeff Phillips. It was funded by the National Science Foundation.

## The Curse of Dimensionality

When analyzing long strings of attributes describing people, "you are looking at not just the individual variables but how they interact with each other," he says. "For example, if you describe a person by their height and weight, these are individual variables that describe a person. However, they have correlations among them; a person who is taller is expected to be heavier than someone who is shorter."

The high "dimensionality" of data stems from the fact "the variables interact with each other. That's where you get a [multidimensional] space, not just a list of variables."

"Data mining means finding patterns, relationships and correlations in high-dimensional data," Venkatasubramanian says. "You literally are digging through the data to find little veins of information."

He says uses of data mining include Amazon's recommendations to individual customers based not only on their past purchases, but on those of people with similar preferences, and Netflix's similar method for recommending films. Facebook recommends friends based on people who already are your friends, and on their friends.

"The challenge of data mining is dealing with the dimensionality of the data and the volume of it. So one expression common in the data mining community is 'the curse of dimensionality,'" says Venkatasubramanian.

"The curse of dimensionality is the observed phenomenon that as you throw in more attributes to describe individuals, the data mining tasks you wish to perform become exponentially more difficult," he adds. "We are now at the point where the dimensionality and size of the data is a big problem. It makes things computationally very difficult to find these patterns we want to find."

Multidimensional scaling to simplify multidimensional data is an attempt "to reduce the dimensionality of data by finding key attributes defining most of the behavior," says Venkatasubramanian.

## Universal, Fast Data Mining

Venkatasubramanian's new method is universal - "a new way of abstracting the problem into little pieces, and realizing many different versions of this problem can be abstracted the same way." In other words, one set of instructions can be used to do a wide variety of multidimensional scaling that previously required separate instructions.

The new method can handle large amounts of data because "rather than trying to analyze the entire set of data as a whole, we analyze it incrementally, sort of person by person," Venkatasubramanian says. That speeds data mining "because you don't need to have all the data in front of you before you start reducing its dimensionality"

Venkatasubramanian and colleagues performed a series of tests of their new method with "synthetic data" - data points in a "high-dimensional space."

The tests show the new way of data mining by multidimensional scaling "can be faster and equally accurate - and usually more accurate" than existing methods, he says.

The method has what is known as "guaranteed convergence," meaning that "it gets you a better and better and better answer, and it eventually will stop when it gets the best answer it can find," Venkatasubramanian says. It also is modular, which means parts of the software are easily swapped out as improvements are found.

## Privacy and Data Mining

What of concerns that we are sacrificing our privacy to marketers?

"The issue of privacy in data mining is like any set of potentially negative consequences of scientific advances," says Venkatasubramanian, adding that much research has examined how to mine data in a manner that protects individual privacy.

He cites Netflix's movie recommendations, for example, noting that "if you target advertising based on what people need, it becomes useful. The better the advertising gets, the more it becomes useful information and not advertising."

"And the way we are being inundated with all forms of information in today's world, whether we like it or not we have no choice but to allow machines and automated systems to sift through all this to make sense of the deluge of information passing our eyes every day."

Provided by University of Utah

Citation: Data mining made faster: New method eases analysis of 'multidimensional' information (2010, July 22) retrieved 18 April 2024 from https://phys.org/news/2010-07-faster-method-eases-analysis-multidimensional.html