New tool: How to get meaningful information out of big data

October 13 2015

Every second trillions of data bits are accumulated and stored. All these

data bits make no sense if you don´t know how to sort them. Now University of Southern Denmark (SDU) researchers present a tool that helps researchers sort data and retrieve meaningful knowledge from the data jungle, presenting their work in the journal *Nature Methods*.

Pretend for a second that you work with obesity research and that you have a trillion bits of obesity related data stored on a server: What do overweight people eat? How do they sleep? What time of day do they eat?

You suspect that the patients' lifestyle may influence their weight, and you can ask your computer to compare weight change and the number of consumed cheese sandwiches to see if there is a link. Then you can ask for another comparison. And yet another. And so you can continue for a very long time and collect a wide range of comparisons for your research.

Or you can approach your data in a way that is not only much faster, but also will discover links, you might not even have considered. Then you will not only be able to put your own suspicions about weight and lifestyle to the test - perhaps you will discover completely unexpected links, for instance that patients who are losing weight, more often eat gouda than cheddar sandwiches.

Looking for the hidden patterns

This is what clustering is about: To look for hidden patterns that we are unable to see ourselves; to ask a computer to group objects which share common traits together into groups.

In principle, it could be any kind of data: patients, proteins or maybe planets in distant galaxies.

At SDU Assistant Professor and head of the research group Practical Computer Science & Bioinformatics, Richard Röttger, and his colleagues from the Department of Mathematics and Computer Science use clustering for example to find regulatory networks in pathogenic organisms allowing for a fundamental understanding of these organisms without the dangerous and expensive need for wet-lab studies.

But clustering is a complicated way to work - even for a computer scientist and regardless of the fact that clustering is a long standing problem in computer science and one of the most fundamental data analysis procedures:

Clustering should be easy for all scientists, not just computer scientists

"Today there are hundreds of comparable but different clustering tools out there; but each of them requires very specific settings and often a deep understanding of the underlying algorithm. There is no overview of what is out there, what should be used when and there is no objective comparison of the available possibilities", explains Richard Röttger.

Therefore, he and his colleagues, Ph.D. student Christian Wiwie and Associate Professor Jan Baumbach, have now created a tool that can provide an objective overview of all available cluster tools, so that researchers get an unbiased, objective overview and suggestions to what tool to use with what parameters in which setting. "The entire process is speed-up tremendously and made more objective now", says Röttger.

The tool is called ClustEval and it is described in the journal *Nature Methods*.

More information: Christian Wiwie et al. Comparing the

performance of biomedical clustering methods, *Nature Methods* (2015). DOI: 10.1038/nmeth.3583