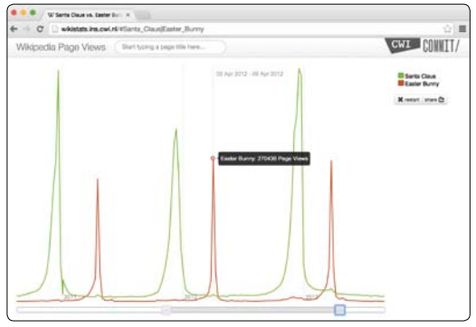
26. Rapidly visualizing Wikipedia page views



Many sectors in our modern society are producing more and more data: science, medicine, finance, business, transportation, retail and telecommunication, to name a few. Visualization is an effective way to interpret the meaning of these data.
We develop techniques that greatly speed up the statistical processing of large amounts of data. We use these techniques to rapidly visualize the statistical results.
Our demo shows how the interest in a Wikipedia page changes over time. A user can select any set of pages to compare, and also find pages with similar interest over time. This can for example be used to judge interest in certain topics from society in general.

ICT science question

How can we speed up the statistical processing of large amounts of data? What are the best visualization techniques for the statistical analysis of large data sets?
Complex statistics are usually limited by the amount of data, since statistical tools are not built to handle massive amounts of data. We embed a statistical processor into a high-performance relational database (MonetDB). This combination is unique, as the translation between the two systems is minimal and thus one hundred times faster than comparable systems. This system has the potential to deliver new insight into massive amounts of data.

Application

Our application shows the interest in Wikipedia pages over time and is already available online: http://wikistats.ins.cwi.nl.

Our database contains almost sixty million measurements from nine hundred thousand Wikipedia pages between January 2008 and July 2014. The data are taken from the Page view statistics for Wikimedia projects. (http://dumps.wikimedia.org/other/pagecounts-raw/)

Alternative Application

Any company or research institute that analyzes large data sets through statistical processes is a potential customer. Applications range from health care over retail to government.

Nice to know

According to our statistics, people visit the Wikipedia page about ‘Love’ four times more often than the one about ‘Money’. There is some hope after all…

Quote

“If you start analyzing big data with R + MonetDB, you will no longer have to wait around long enough to take a coffee break after running each analysis step.” − Anthony Damico, Statistical Analyst at the Henry J. Kaiser Family Foundation

We produce statistical analysis results 100 times faster than before.

If the amount of data you have to handle for statistical analyses exceeds what your tools can handle, and you are already using R, we can help.

Our high-performance combination of R and MonetDB is published as Open Source software and can be used for your projects without licensing fees.

We study how the combination of statistical analysis tools and analytical data management systems can yield highly improved performance.

Hannes Mühleisen
Hannes.Muehleisen@cwi.nl
Try your own Wikipedia page view visualizations on: http://wikistats.ins.cwi.nl
Watch a video about our work on: http://vimeo.com/groups/amsterdamdatascience/videos/100491517

COMMIT/ project
TimeTrails Spatiotemporal Data Warehouses for Trajectory Exploitation

CWI monetdb

This research was supported by the Dutch national program COMMIT/