2 Dutch researchers analyze striking behavior of Web surfers

October 29 2008

What behaviour do website visitors exhibit? Do they buy a specific product mainly on Mondays? Do they always return at a certain time of day? Being able to recognise and make use of such patterns is lucrative business for companies. Edgar de Graaf discovered that interesting patterns often contain a time aspect. Jeroen De Knijf developed methods to detect relevant patterns quicker.

In subject jargon it is called data mining: looking for interesting relationships within large quantities of data. Many data-mining programs produce a flood of potentially interesting patterns: as a user, how can you then find what you are looking for? Furthermore, the files are not always set up for such search actions, as is the case on the Internet or for instance in bioinformatics. It usually concerns semi-structured files: they often contain, for example, hyperlinks to other files, and contain (partial) information in a range of formats, such as text, images and sound.

Edgar de Graaf and Jeroen De Knijf both worked within the NWO-funded MISTA project (Mining in Semi-Structured Data) on methods to find patterns more quickly and effectively within large quantities of semi-structured data. De Graaf discovered that some patterns are interesting because they occur in quick succession. Other patterns are striking because, for example, they occur weekly. According to De Graaf, this time aspect merits further investigation.

The patterns can best be presented visually so that the user can find the

information sought at a single glance. To realise this De Graaf described various ways of presenting different types of information.

De Knijf demonstrated that the number of patterns can be drastically reduced by allowing the user to indicate in advance the minimum requirements that a pattern must satisfy. This allows the data-mining program to find the interesting patterns much faster.

A second method De Knijf devised to reduce the number of results is the compression of the entire collection of documents (for example, Wikipedia pages) into a single document. By building accurate models that only make use of the compressed document, De Knijf was able to demonstrate that this summary does indeed contain the essential information from the entire collection.

Source: Netherlands Organization for Scientific Research