Investigating documents in depth

February 9 2006

Keyword searches in text databases are a standard procedure today. Related content in different documents can now be analyzed on numerous levels using the software tool SWAPit. Researchers will be demonstrating at CeBIT how football news can be evaluated.

Does Ballack actually play any better now that he has signed a lucrative advertising contract? Or has his performance deteriorated instead? Has the disagreement between Kahn and Lehmann improved the two goalkeepers' performance, or are they tending to stop fewer balls than before? And what effect does this have on their clubs? If scoop-hungry reporters are to assess these issues on a founded basis, rather than just relying on their gut feeling, they need to square up the news in sports magazines with up-to-date statistics, club communications and articles in the tabloids.

Such multi-layered analyses can now be prepared semi-automatically, using the software tool SWAPit developed by scientists at the Fraunhofer Institute for Applied Information Technology FIT in Sankt Augustin near Bonn. This tool makes it possible to discover related content in textual data at a glance, revealing any associated additional information.

“The name SWAPit is derived from the verb ‘to swap’,” explains Andreas Becks of the FIT. "The program challenges users to look at textual information from alternative points of view, enabling them to compare supplementary information related to the documented topics.” To make this possible the tool presents collections of texts as a kind of map, in which similar texts are grouped into clusters. When a user clicks on one of these clusters, the shared features are displayed on the monitor in a field immediately adjacent to the map. “These additional ways of looking at information allow users to analyze their data much more fully. They can compile statistics and discern patterns that were not evident before,” Becks emphasizes.

Press research is just one possible application of the method known as integrated text and data mining. Other ways of using this software might be to analyze patents for research planning, examine documents on segments of the market or evaluate inquiries at service centers. “But at one point we even had an interdisciplinary cultural project in which SWAPit solved communication problems,” Becks reports. “It showed us how differently various disciplines define the same term.”

The researchers have already tested their prototype with industrial partners in a wide range of sectors. It is compatible with standard text formats such as doc, pdf and html, but could easily be extended to cover other formats if required for concrete marketing purposes, Becks assures us. Interested parties can learn more details at CeBIT in Hanover from March 9 to 15.

Source: Fraunhofer-Gesellschaft

Citation: Investigating documents in depth (2006, February 9) retrieved 1 May 2024 from https://phys.org/news/2006-02-documents-depth.html