Zip Programs Can Identify Language Of Any Document

Data compression routines can accurately identify the language, and even the author, of a document without requiring anyone to bother reading the text.

The key to the analysis is the measurement of the compression efficiency that a program achieves when an unknown document is appended to various reference documents.

This was demonstrated by Dario Benedetto and Emanuele Caglioti of La Sapienza University in Rome and Vittorio Loreto of the Center for Statistical Mechanics and Complexity, also in Rome.

Zipping programs typically compress data by searching for repeated strings of information in a file. The programs record a single copy of the information and note the locations of subsequent instances of the string.

Unzipping a file consists of replacing various bits of information at the locations recorded by the zipped file. Such file compression routines work better on long files because programs are, in effect, learning about the type of information they are encoding as they move through the data.

Add a page of Italian text to an Italian document, and a zipping program achieves good efficiency because it finds words and phrases that appear earlier in the file. If, however, Italian text is appended to an English document, the program is forced to learn a new language on the fly, and compression efficiency is reduced.

The researchers found that file compression analysis worked well in identifying the language of files as short as twenty characters in length, and could correctly sort books by author more than 93% of the time.

Because subject matter often dictates vocabulary, a program based on the analysis could automatically classify documents by semantic content, leading to sophisticated search engines. The technique also provides a rigorous method for various linguistic applications, such as the study of the relationships among various languages.

Although they are currently focusing on text files, the researchers note that their analysis should work equally well for any information string, whether it records DNA sequences, geological processes, medical data or stock market fluctuations.

(Reference: D. Benedetto, E. Caglioti, and V. Loreto, Physical Review Letters, 28 January 2002; text at this URL.)

(Editor's Note: This article, with editing, is based on PHYSICS NEWS UPDATE, the American Institute of Physics Bulletin of Physics News Number 575, January 30, 2002, by Phillip F. Schewe, Ben Stein, and James Riordon.)

[Contact: Emanuele Caglioti]

04-Feb-2002

Click for Related Stories

Add the UniSci Daily Java News Ticker to Your Site or Desktop.
Click for a demo and more information.

Copyright © 1995-2002 UniSci. All rights reserved.