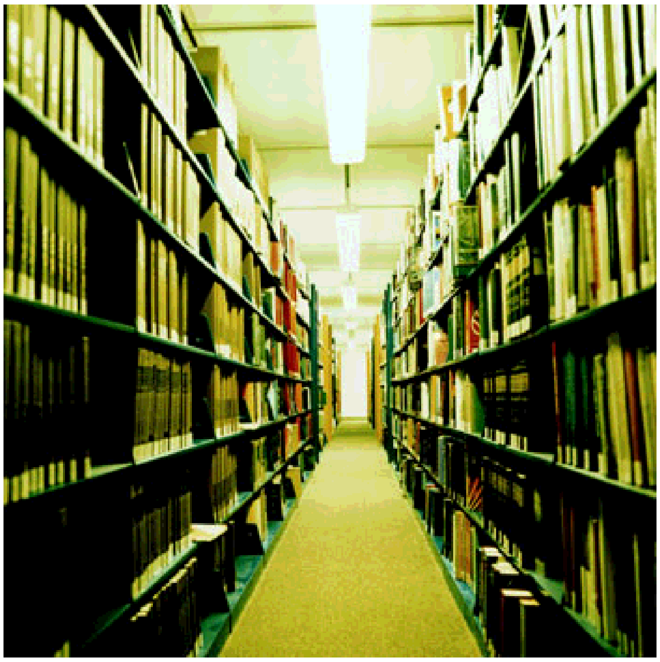
24 January 2002

Reading the Bits of Shakespeare

It may not tell you whether Francis Bacon really authored Shakespeare's plays, but a common computer program designed to compress large files can sort out who wrote what with greater than 90% accuracy.



Just compress it. A common computer program can identify authors by compressing text.
CREDIT: RYAN MCVAY/PHOTODISC

To a computer, Hamlet's first soliloquy is just a string of characters--but that string still contains information. Just how much information is what determines the string's "entropy," essentially the minimum number of bits needed to encode the string. Unless a string is infinitely long, it's impossible to calculate it's exact entropy. But a program that compresses files provides a convenient estimate: the length of the compressed file containing the string. By estimating entropy, sophisticated compression programs can identify the language and even the author of unfamiliar prose. Now, mathematicians Dario Benedetto and Emanuele Caglioti and physicist Vittorio Loreto of the University of Rome have shown that freely available, off-the-shelf software can do the trick too.

The researchers employed a common program called gzip. Gzip replaces the original file with a catalog of building blocks a few characters long, and instructions for putting the blocks back together. The trick to sleuthing texts is to compress a file containing a longer known text followed by a shorter unidentified text. If the known and unidentified texts are similar, such as a Shakespeare play and a sonnet, gzip will do a slightly better job of compressing the composite file because both require roughly similar building blocks.

To test the program, the researchers collected 90 texts by 11 Italian authors and measured their length when compressed. They used a short piece of one text as the "unidentified" sample. Then they appended this sample to each of the other 89 files and measured the compressed length of each. When the length of a composite file changed little from that of its original compression, the file was "recognizing" the unidentified text, as it needed relatively few additional building blocks. The researchers repeated this process 90 times, taking the "identified" sample from a different text each time, and in 93% of the cases, the method correctly revealed whether the same author had written both the known and unidentified texts, the researchers report in the 28 January issue of *Physical Review Letters.*

The results demonstrate the power of compression programs to classify language, says William Teahan, a computer scientist at the University of Wales in Bangor, U.K. Compression programs might someday serve as the basis for software that automatically categorize huge numbers of documents or accurately mine enormous troves of data--such as the World Wide Web--for documents discussing a particular topic, Teahan says.

--ADRIAN CHO

Related sites
The gzip home page

PAGE TOP

Next Story