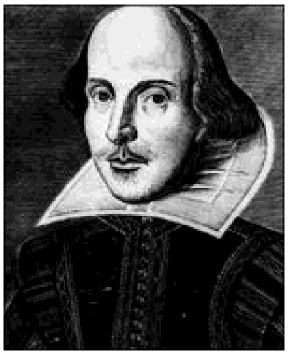
January 30, 2002

HOMEPAGE
NEWS SUMMARY
US
INTERNATIONAL
MONEYScope
WEATHER.com
LOCAL NEWS
ENTERTAINMENT
ESPN SPORTS
SCI / TECH
POLITICS
HEALTH
TRAVEL

FEATURED SERVICES
SHOPPING
DOWNLOADS
WIRELESS

INTERACT
VIDEO & AUDIO
WEBCASTS
BOARDS
CHAT
NEWS ALERTS
CONTACT ABC

SCI/TECH



Researchers say that a common computer file compression program could help determine if unidentified bits of text were written by Shakespeare or not. (AP Photo)

Shrink and Identify

Researchers Discover a Novel Way to Identify Authors of Text

By Paul Eng
abcNEWS.com

Jan. 30 - *To be or not to be: That is the question.*

ALSO ON abcNEWS.com

* Bush Declares State of Union Strong

* Report: Stephen King Ponders Retirement

* Lewis Has Doubts Over Tyson Fight

That notable quotable is instantly recognizable among people as a line from Shakespeare's *Hamlet*. But soon, even computers may be able to automatically identify strings of text with their appropriate authors - and by using a free computer program already available on the Internet.

According to a report to be published in the *Physical Review Letters* magazine, researchers at La Sapienza University in Rome have found that a computer file compression program called Gzip provided an unusual means of analyzing strings of data.

Typically, computer compression programs such as Gzip shrink large computer files - text files, for instance - by searching for repetitive strings of information. By finding and identifying those patterns, the compression program can reduce the original file to a smaller one that contains just the basic "building blocks" of data and instructions on how to use those blocks to recreate the original,

SEAR
Find

SCI / T
HEADI
* Athle Fear' T Olymp
* Progr Shrink
* Teen
* Navy Warshi
* Silico When I Get Pro

Sear

and instructions on how to use those blocks to recreate the original, larger file.

But Emanuele Caglioti, an associate professor of mathematics at the university and one of the report's authors says that the program's compression process is also the key that helps identify files of unknown data.

Learning Language

When a program such as Gzip shrinks or "zips" a file, "it is learning something about the file," says Caglioti. Specifically, it is learning the file's so-called entropy, or the minimum number of bits needed to encode the file. Files of similar content would share similar entropies since they share the same common "building blocks."

"If you zip a file - say one composed of English text - while [the Gzip program] is reading the file, it's learning the statistics of English," says Caglioti. "The more it reads it, the more it can compress it." And adding additional English files wouldn't produce a great change in the file's size since the basic pattern - its entropy - is already known.

But, if the second file turns out to be Italian, Caglioti says the process has to start all over again and a new entropy is created. "It has to learn [the] Italian," says Caglioti. And "This effort has a cost in terms of bits. It takes more space to incorporate the Italian file because it's a different language."

And Caglioti and his team of researchers discovered that this same process and principle can be used to "identify' works by author. In their research, the Italian scientists collected 90 texts by 11 Italian authors and in 93 percent of the cases, the method correctly matched small text samples with the authors.

Other Possibilities?

"It's pretty clever what they did," said James Riordon, a physicist with the American Institute of Physics, the group that publishes the *Physical Review Letters*. "Effectively, it's like you're training someone in a language to identify it."

And Caglioti say that there's no reason to believe that the compression process couldn't be used in other means. "Aside from text reconginition, it can be used to compare Web pages and find ones that are similar," he says. In addition to creating a better Web search engine, Caglioti notes "there is the challenge of biological DNA sequencing." He said genetic researchers have already reported in *Bioinformatics* of using similar zipper approaches to map the human genome.

Needs Further Research

Mark Adler, the programmer who co-created Gzip in early 1990 as an alternative to other file compression programs, said he was surprised someone had used his program in such a manner. "It is impressive and a little surprising that simply comparing the length of the compressed output from concatenated known and unknown text provides such high accuracy," he says.

But he remains skeptical that the Italians' research paves the way to fool-proof text identifiers - at least until more studies are done.

"At some point using entropy as a measure may not be fine enough to distinguish between authors with similar styles or use of words and phrases," he says. "I'd wonder how well it would work for author recognition if you tried to distinguish between a thousand authors instead of a dozen."

"Up to now, this is more theoretical than practical," Caglioti conceeds. But he says he and his team will continue to work with the program and see what else turns up. "We ought to try and see where it can work." ■

PRINT THIS PAGE | SEND THIS TO A FRIEND | VIEW MOST EMAILED