April 30, 2002

Fun With Your Zip Program: Sort Through Texts, and More

By BRUCE SCHECHTER

One of the basic truths of the digital age is that almost anything - the plays of Shakespeare, the genetic sequence of DNA, or the twitching of a seismograph needle - can be reduced to a sequence of ones and zeroes. More striking is the discovery that these sequences are largely full of hot air - redundancies that add nothing to their meaning. Clever computer programs can "zip" or compress these files, streamlining them for speedier transmission. Zipping programs have long been a boon to computer users with slow modem connections.

But now a group of Italian physicists has shown how these same programs can be used to analyze and categorize text quickly. Using little more than the zipping programs found on most personal computers, they can easily distinguish between texts written in 10 different languages and almost unfailingly tell which of a large group of texts were written by the same author.

Writing in the January issue of Physical Review Letters, the scientists - Dr. Dario Benedetto, Dr. Emanuele Caglioto and Dr. Vittorio Loreto - explain their work with an analogy to Morse code.

To keep the number of dits and dahs to a minimum, Samuel Morse considered how often each letter was used in an average English message. The letter e is the most common in English so Morse encoded it as a single dit. The next most common letter is t, so he assigned that a single dah. A relatively uncommon letter like Q takes four taps to encode: dah dah dit dah.

Compression programs work in a similar fashion, except they

Compression programs work in a similar fashion, except they invent a new code for each message based on patterns unique to that message. The program might, for example, find that a text uses the word "compression" frequently and save space by substituting a two-letter abbreviation.

The Italian physicists understood that a compression scheme invented to compress a text written in English would do a poor job on one written in Italian.

"Transmitting an Italian text with a Morse code optimized for English will result in the need of transmitting an extra number of bits," they wrote. They conjectured that just how many extra bits it takes would be a measure of the distance between English and Italian.

To demonstrate this, the researchers used a zip program to compress a text written in one language. They then appended to the original text some text written in Italian or another language and compressed that document. As predicted, the compression program did not do as good a job when the languages of the two texts were different. They tried the same trick on a group of texts all written in Italian, but by a variety of different authors. They found they could distinguish between the authors more than 90 percent of the time.

The scientists performed a further test of their technique by analyzing a single text that has been translated into many different languages - in this case the Universal Declaration of Human Rights. The researchers used their method to measure the linguistic "distance" between more than 50 translations of this document. From these distances, they constructed a family tree of languages that is virtually identical to the one constructed by linguists.

The researchers say linguistics is just a "playground" for them to sharpen their techniques. The same methods, they say, might help create order out of the rapidly accumulating libraries of DNA and protein sequences, earthquake catalogs and other geophysical data. They might even lead to a solution to one of the most troublesome problems of modern computer science: filtering the junk e-mail from your in box.