Digging Deep Into Compression

By Mark K. Anderson

2:00 a.m. Feb. 6, 2002 PST

Unless tea leaves or crystal balls are involved, predicting the future is typically a matter of finding patterns in the past.

While there are many approaches today to pattern recognition and matching, two clever techniques have recently found new applications from hurricane and earthquake forecasting to analyzing authorship of texts and making sophisticated search engines.

The first involves the seemingly unrelated task of file compression -- as performed in applications such as WinZip and StuffIt -- while the other borrows its lessons from the world of chaos, complexity theory and fractals.

TECHNOLOGY

Today's Headlines
2:00 a.m. Feb. 6, 2002 PST

They Want Their ID Chips Now

How to Turn On a Gene

Olympics Mantra: It's About Time

Seal Cams Probe Antarctic

Digging Deep Into Compression

The Passion for Designing Macs

Absorbing Apple's Aesthetic

Trillian Won't Heed AOL's Message

He Writes the Songs: Mac Songs

Europe's Scientists Discover Gold

See also:

* Firm Touts 'Perfect Compression'
* Videophone Gets a Booster Shot
* Read more Technology news

In the Jan. 28 issue of the journal *Physical Review Letters*, three Italian scientists used the Unix compression program gzip on text files to address such pattern-matching issues as language of composition and authorship.

Since data compression entails recognizing and tagging repeated strings, the more repeated internal patterns that a file or collection of files has, the more it can be compressed. Thus, if one wants to know the language in which file X was written, just compress it with files whose language is known and then compare how efficiently each operation is carried out.

If, by comparing raw and compressed file sizes, one finds that X plus an Italian text file zips tighter than X plus a French text or X plus an English text or X plus one's other linguistic reference texts, then *congratulazioni*! You've likely just found the language of X without even opening it.

The scientists -- Dario Benedetto, Emanuele Caglioti and Vittorio Loreto of Rome's La Sapienza University -- used this technique to discern the language of mystery texts as small as 20 characters. Furthermore, using a database of 90 texts from 11 different authors, they found their method could even pick out individual authors with a success rate of 93 percent.

Search engines, they say, could use this simple technique to categorize their quarry by semantic content and more qualitative categories such as style and readership level.

"I would like to see if this method could distinguish the music of John Lennon from Paul McCartney," Caglioti said.

Ming Li, a professor of computer science at the University of California in Santa Barbara, developed the file-compression technique for categorizing genetic datasets. He said he’s impressed by Benedetto et al.’s work, but he cautioned that the "zip" format leaves much to be desired.

"For some rough purposes, it’s OK," he said. "But for many applications you need a better compression algorithm."

1 of 2 Next >>

*Have a comment on this article? **Send it.***
*Printing? Use **this** version.*
***E-mail** this to a friend.*

