NewScientist Archive

NewScientistArchive NewScientist.com | NewScientistJobs.com

This Week Frontiers

A gift for language

15 Dec 01

Bog-standard PC software has an unsuspected talent

YOUR PC has a hidden flair for recognising foreign languages. WinZip, the software routinely used to compress PC files, can help you figure out the language a document is written in, say Italian scientists. It could even be used to guess a document's author.

The technique, which can identify the language of a segment of text as little as 20 characters long, will be described in a future edition of the journal *Physical Review Letters*. Vittorio **Loreto**, a physicist at Rome's La Sapienza University, and his colleagues Dario Benedetto and Emanuele Caglioti have applied for a patent on their idea.

Anybody with WinZip software on their PC can use the technique. Take a page of text written in unknown language X. Compress a large piece of text in a known language, say a whole e-book in French, and then repeat the task with the page written in X added to the end of the e-book. Note the difference in the sizes of the two compressed versions.

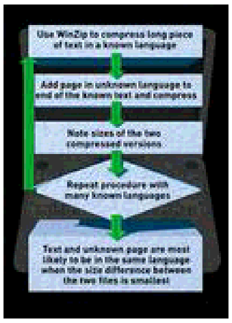
Next, take an e-book in another language, say German or Spanish, and do it all again. Then repeat the procedure with as many more languages as you can. Finally, find which e-book gives you the smallest difference between the two compressed files, says **Loreto**. The e-book and the text written in X will be the same language (provided one of the languages you've used is the right one, of course). But how come?

The method works because WinZip's compression routine records the occurrence and positions of duplicated words and phrases. As it analyses text, WinZip does an increasingly efficient packing job-because it encounters fewer and fewer words it hasn't already seen. But if, when it reaches the added page, it runs into text that's in a different language, nearly all the extra words will be unfamiliar ones. Since these cannot be recorded as incidences of words WinZip has already seen, the compressed version will be larger than if language X contained only familiar words.

Identifying languages with off-the-shelf compression software like this is "kinda cute", says Bob Moore, head of Microsoft's Natural Language Processing group at the firm's campus in Redmond, Washington.

But file compression can do more than merely identify the language, according to **Loreto** . "More ambitiously, we're interested in guessing the author of a piece of text and more importantly, the subject under discussion," he says. This works because different authors have distinctive turns of phrase and different subjects have distinctive vocabularies. You could use the technique to determine whether the author of a newly discovered Elizabethan play was by Shakespeare, for example.

And since the strategy relies on sequences of characters, rather than actual words, it will work with any data held in the form of a character string. **Loreto** speculates that it could one day be put to work helping to classify proteins, or relating patterns in genes to their biological role.



WinZip used to recognise foreign languages

Andrew Watson

From **New Scientist magazine**, vol 172 issue 2321, 15/12/2001, page 25

back to top back to search results