TRN News

April 17/20, 2002

File compressor ID’s authors

By Kimberly Patch, Technology Research News

Using computers to find language patterns like the forms that make up the grammar of a particular language or the turns of phrase a certain author tends to use is both useful and tricky.

The main challenge is a version of the classic chicken-and-egg conundrum: if the patterns are unknown, how you know what to tell a program to look for?

Researchers from the University of Rome in Italy are using differences in the way text files are compressed as points of comparison. Using these patterns, the researchers can compare how close a string of unknown text is to a block of text written in a specific language, or even by a specific author. Data compression techniques are used to shrink files so they take up less storage space and travel more quickly over networks.

The researchers tested the method using gzip, a well-known compression program, and 90 texts written by 11 authors. They included a short string of unidentified text at the end of each of a group of longer, known texts, and used the compression process to compare the short string of text to the group of known texts.

This type of research could make it easier to automatically identify languages, identify authorship of a document, or adapt computer tools like spell checkers to different languages. It could also be used to automatically classify other types of data like DNA sequences and seismic information.

Compression algorithms, or zippers, save space by rooting out redundancy. They replace an original file with a group of building blocks plus instructions for putting the building blocks together to reconstruct the original file. "A zipper is... explicitly conceived to take a file and try to transform it into the shortest possible file," said Vittorio Loreto, a physics researcher at the University of Rome.

The researchers measured how similar the short, unknown string of text was to the file it was compressed with by noting how concisely the zippers for each known file compressed the unknown string.

"If we take a long English text and... apend to it an Italian text, then zip [the resulting file], the zipper begins reading the file starting from the English text," said Loreto. "So after a while it is able to encode optimally the English file. [But] when the Italian part begins, the zipper starts encoding it in a way which is optimal for the English," he said.

As long as the Italian portion of the file is short enough, the zipper will compress the Italian using the English rules, Loreto said. This makes it possible to use the compression rate as a measure of how close the two languages are, he said. "If the length of the Italian file is small enough, the difference between the length in bits of the zipped Italian plus English text and the length in bits of the English text zipped alone will give a measure of the distance between the two texts," he said.

This distance, or entropy between two languages is the difficulty a native speaker of one language encounters in attempting to read text written in another, said Loreto.

This distance between known and unknown texts is a tool that computer algorithms can use to recognize the context of a given sequence -- things like its language, subject and author, Loretto said.

For automatic language recognition, for example, a piece of unknown text can be compared with many texts in many different languages. The procedure will show the language that is the closest to the unknown file. If the language collection includes the language in question, the closest file will be the same language, said Loreto.

In a similar way, unknown files can be automatically sorted by subject and author as well, he said. "If you are able to guess automatically the subject a given text is about without the need to read it, you can exploit this feature for automatic classification," he said.

The distance data can be used to automatically build tree-type representations of relationships among texts, said Loreto. "This way one also has a graphical representation of the corpus structure."

A key problem plaguing computer linguists is finding patterns even when you don't know exactly what you're looking for, said Patrick Juola, an assistant professor of computer science at Duquesne University. The University of Rome researchers have tackled the problem by adapting the "best-known pattern-finding machines -- compression algorithms -- to the task of learning patterns useful for categorizing language," he said.

There are several groups working on similar methods, Juola added. "Language has interesting patterns that can be found automatically. The question that remains is what's the best way to find and exploit these patterns, and what kinds of problems can we solve with them," he said.

This type of research is useful for things like Web searches and finding authorship, said Juola. "The major implication... is that this kind of technique makes it much easier to analyze linguistic documents without having a detailed knowledge of the language in question," he said.

The method could help adapt existing language tools like Web search engines, and spelling and grammar checkers to new languages that don't have a long scientific and analytical tradition, said Juola. "If I were the University of Nepal, I'd be very interested in seeing how this could help with the development of a modern information processing system for Nepali documents."

One language recognition challenge is gaining information by using just a few characters. The ability to identify the language a document is written in in as few as 20 characters may allow the researchers' method to identify the language of picture captions and table headings, which is hard to do right now, Juola said. One possible drawback to the method is that because it needs a lot of text in various languages to make a comparison, it may take a long time to do so, Juola added.

The researchers have tested the method by classifying large law and literature collections, said Loreto. The method can also automatically classify strings of characters that represent non-text coding, [like] time sequences, genetic sequences... stock market data, and medical monitoring data, he said.

The next step in the work is applying it in one of these areas, said Loreto. The researchers are looking at "the analysis of biological data -- DNA or protein sequences -- and the investigation of phenomena described in terms of time-series: earthquake sequences, stock-market data, or other kinds of signals," he said.

In both cases, "we expect to be able to implement procedures for automatic recognition and classification [to identify] specific significant patterns like the sequences codifying the genes in DNA or some premonitory pattern in geological or medical data," he said.

Loreto’s research colleagues were Dario Benedetto of the University of Rome and Emanuele Chaglioti of the University of Rome and the Center for Statistical Candidates and Complexity (INFM) in Italy. They published the research in the January 28, 2002 issue of Physical Review Letters. The research was funded by the University.

Timeline: Now
Funding: University
TRN Categories: Applied Technology; Pattern Recognition
Story Type: News
Related Elements: Technical paper, "Language Trees and Zipping," Physical Review Letters, January 28, 2002.

Find out about the TRN Newswire for Web sites, print publications and newsletters.

Help TRN Grow: suggestions, email headlines and T-shirts.

© Copyright Technology Research News, LLC 2000-2002. All rights reserved.