

PEER-REVIEWED CONTRIBUTIONS

Do we still need data compression?

By Fei Nan
(IEEE Senior Member)

I recently came across a colleague and was asked whether we still need data compression as data storage is becoming more and more affordable. The floppy drive, first invented in 1976, became a popular and ubiquitous form of data storage and exchange in the mid-1970s to the late 1990s. There were hundreds of, if not thousands of, data compression programs available back then to reduce file size and improve disk usage on the 1.44 MB floppy disk. Nowadays, people can easily purchase a terabyte scale hard drive from any electronics store down the street. Is it still necessary to use data compression programs to zip data and to save the extra few bytes? To answer the question, we have to know what data compression is about, and what it can bring to us. Then we will understand why we still need compression. We need compression a lot more than ever.

We need data compression because it can help reveal hidden structures and recognize patterns which had never been disclosed before. The fundamental theory behind data compression is to search repetition and use reference links to point at those repeats. The reference links are abstract conceptions. In implementation, it would be a dictionary like data structure such as array lists, etc, which save the starting and ending positions for each occurrence of the repeat. Each entry in the dictionary is considered as a standalone segment. By following the reference link dictionary, the compression program only records the verbatim repeat once. All the fol-

lowing occurrences are labelled with reference links, which are readily accessible for a program to look up. The reference links are merely two integer numbers which are trivial in size, thus giving positive compression gain.

The data compression helps pattern recognition by building a dictionary for fast retrieval. The dictionary is a pre-arranged management system and data exchange for those repetitious segments from the input. If one segment or one portion of the segment contains some patterns of interest, it is effortless to follow reference links in the dictionary in order to retrieve all occurrences of the pattern from the input. This is a lot more efficient than directly searching the original input to track down the patterns of interest. Because the chance of finding a match is much higher in the dictionary, the pattern matching algorithm can start off searching for patterns in the dictionary first, before extending to the remaining parts. This will significantly optimize the search operation and reduce the search latency.

During the process of compressing data, some long-range tandem structures will be



Dr. Fei Nan is an IEEE Senior member and currently a senior research engineer at the Samsung Mobile R&D lab in San Jose, CA.

revealed. Repetitious structures will have more meaning to input data. By tracing down the repetition, it will undoubtedly help expose key components. This feature is widely applied on computational biology, biomedical imaging and bioinformatics. When properly used on protein or DNA sequences, most researchers are able to uncover some biologically meaningful segments, or exons.

LEARNING HAS NO
BOUNDARIES

YOU KNOW YOUR STUDENTS NEED **IEEE** INFORMATION.
NOW THEY CAN HAVE IT. AND YOU CAN AFFORD IT.

IEEE RECOGNIZES THE SPECIAL NEEDS OF SMALLER COLLEGES,
and wants students to have access to the information that will
put them on the path to career success. Now, smaller colleges can
subscribe to the same IEEE collections that large universities
receive, but at a lower price, based on your full-time enrollment
and degree programs.

Find out more—visit www.ieee.org/learning