Kunihiko Sadakane, Associate Professor, Principles of Informatics Research Division

For equitable human intellect sharing in an information-oriented society

Technologies needed to process massive data volumes

In the current information-oriented society, efficiently processing massive volumes of data that grow daily is a key topic. Specifically, we need to develop technologies that can compress large volumes of data to make data burdens as light and manageable as possible, while also enabling high-speed search. Achieving both goals isn't easy.

Consider a duvet. A duvet can be hardened into a compact form that doesn't require much storage space by putting it into a compression bag and vacuuming out any excess air. But this simply reduces the storage space. To use the duvet once again, we need to return the air to the duvet and puff it out to its original size. Similarly, if computer data needs to be restored for use after compression, the significance of the compression tends to get lost.

Technology that decompresses only the portions of the data actually needed can help resolve these issues. One can visualize how technologies that restore just part of the data achieve both compactness and high-speed search by thinking of cups of instant noodles. These noodles have been compressed by removing the water. Without adding boiling water, if we insert our chopsticks into the desired portion, we can "decompress" this portion and eat it right away without waiting several minutes to be done. Each day my research focuses on the issue of balancing high compression rates with search speed.

Identifying truly useful information from the chaff of the Web

My research covers not just simple data structures like those of textual data, but more complex structures such as trees and graphs. A tree structure is like a family pedigree, in which a single point of data branches off into multiple other points. A graph structure is even more complex, resembling the mesh of a net, with each data point connected to multiple others. A typical example is the Web that many people use every day. As the next step in my studies, my goal is to identify compression and searching technologies compatible with this structure.

With textual data, the more precise an index one establishes, the more advanced one's search can be—for example, by searching for cases in which certain nouns are close to certain verbs, rather than simply searching for specific keywords. However, in a graph structure, in which the data structure itself is even more complex, we can also do what is known as a community search. This approach makes it possible to identify the presence of communities involving large volumes of information on specific topics based on the extent of the concentration of relevant links. Such communities are places with a high likelihood of obtaining the most useful, accurate information on a topic. This allows even users with no knowledge of search technologies to put the treasure trove of information scattered across the Internet to wide-ranging use. Without sounding too lofty, one might even say such technology helps achieve sharing human intellect.

Another of my dreams is to have the results of my research put to use in personal computers used by the general public. Since my research relates to infrastructure technologies that support the exchange of data in an information-oriented society, it is rarely seen by ordinary users in their everyday lives. But nothing would make me happier than to see the results of my research improve the lives of greater numbers of users. As a first step to making this dream real, I am releasing to the public free of charge resources such as libraries for searching compressed text strings without decompressing them. These can be used by anyone with a basic knowledge of programming, and I encourage all those interested to download them. I plan to make steady progress on preparing more detailed descriptions and other information.

Interviewed and summarized by: Emiko Nakano