Internet Measurement and Big Data

In the years to come data analysis will be of increasing importance in every field.
The ability to solve problems by fully utilizing statistics and data analysis is crucial.

3.1 Internet Measurement

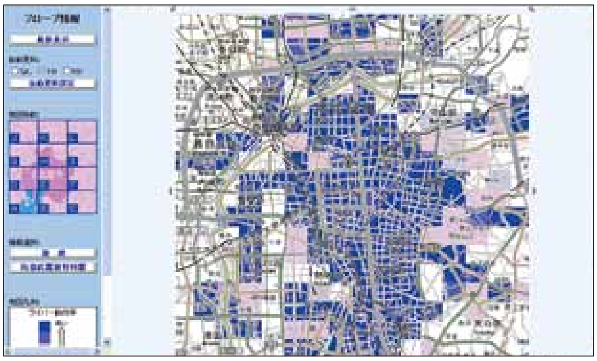
The Internet is an open system in a constant state of flux. Being a distributed autonomous system, the Internet has no central hub or representative point, and is observed differently depending on the time or place it is measured. Because it is so hard to accurately assess the Internet, a variety of Internet measurement initiatives have been implemented.

Internet measurement typically involves measuring the volume and makeup of traffic, as well as topology measurement for investigating logical network connections. This also includes observations such as spam ratios, virus infections, and security attacks that we report on in each volume of this report. Recently, many types of online service measurement have emerged, including observation of peer-to-peer systems, investigation of social network usage, and observation of the ways that the users are connected. Here we will broadly define Internet measurement as the measurement of the Internet, including Internet-based services and their usage, as well as application of this data. Services familiar to users such as spam detection, search rankings, and online recommendation systems can also be considered as applications of Internet measurement technology.

Many Internet measurement efforts share the common approach of attempting to obtain useful information from large volumes of incomplete data. This is in stark contrast to conventional engineering approaches to measurement. Conventional measurement attempts to obtain precise data by improving accuracy, but because data for Internet measurement is inherently inaccurate, it is necessary to infer the actual situation by comparing indefinite information.

For example, there is no way to accurately measure the total number of PCs and other devices connected to the Internet. However, by comparing multiple pieces of data such as Internet address usage, access numbers for major websites, Internet user number surveys from various countries, and shipment numbers for PCs and mobile devices, it is possible to estimate approximate numbers. Currently there are thought to be between three and five billion “connected” devices, depending on the definition of the word.

As another example, if information on vehicle positions and whether their wipers are in use can be gathered, it is possible to gain a detailed picture of localized torrential rain. Individual information on wiper usage is unreliable, but if a large amount of wiper information is collected, it is possible to gain a more detailed picture than can be produced with meteorological sensors, which are placed a dozen or more kilometers apart (Figure 1).



In Internet vehicle experiments conducted by the WIDE Project in Nagoya in 2001, location, speed, and wiper usage data was collected from 1,570 taxis. The blue parts of the map indicate areas with a high ratio of wiper usage, showing rainfall in detail.

Figure 1: Vehicle Wiper Information

Technology Trends

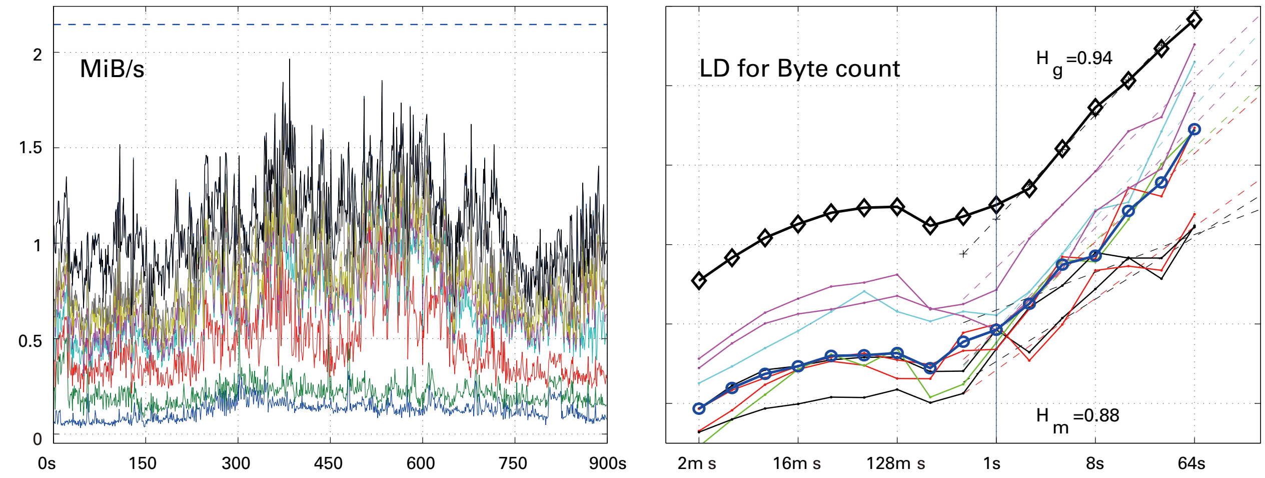
Statistical methods, including multivariate analysis for analyzing the relationship between multiple elements, are often used to uncover information hidden within data (Figure 2). Methods like this were applied to social sciences such as psychology and behavioral science as well as medical and pharmaceutical science before the advent of Internet measurement. However, the situation has changed significantly through the automation of data acquisition and analysis and advances in systemization due to the Internet and information technology. This has made previously difficult tasks possible, such as access to vast amounts of data, analysis of data that is constantly being updated, and application of data to non-linear models. The analysis of large quantities of data is now an indispensable research method in all areas of science and technology.

3.2 Big Data

The phrase “big data” has recently been showing up in a variety of places. Big data is a general term for technology that extracts information of value hidden within large volumes of unconventional, unstructured data. The concept is to use this information to formulate new business models and management innovation by collecting and analyzing large volumes of data. Behind the rise of big data is the fact that in recent years the environment necessary for its implementation has come into being, in particular with the emergence of cloud services, allowing everyone to have access to it. The use of the online behavior history of users for marketing purposes is one form of big data business that has attracted attention recently, but in the future a variety of developments are anticipated.

From a technological standpoint, big data essentially incorporates Internet measurement. The construction of online data collection systems and systems for data storage or sharing, as well as engineering applications of statistical processing technology for extracting information from large volumes of fragmentary data, have been carried out since the dawn of the Internet. The Internet itself consists of components designed based on engineering principles, but because its behavior involves the interaction of countless elements, when viewed as a whole it is representative of a complex system that displays independent behavior beyond the sum of its individual parts. Additionally, because it reflects the behavior of users, it is also affected by social, economic, and political factors. Though Internet measurement is based on engineering, it also encompasses aspects of natural and social sciences.

Data collection has changed dramatically due to the Internet. With more and more information available on the Internet, it is now possible for anyone to access a wide variety of information with ease. With the addition of sensor information such as time and position, previously difficult analysis of relationships is becoming possible. Also, with the spread of information over channels such as social media, fundamental changes are taking place in the propagation and sharing of information that was once the domain of mass media. It has also become possible to collect data on the transmission of information via methods such as the tracking of keyword popularity.



Statistical information extracted from network traffic (left) can be compared (right) to detect abnormal behavior or failures and their symptoms.

Figure 2: Detection of Abnormal Behavior using Statistical Methods

The amount of data that can be stored is increasing dramatically due to growing storage capacity and lowering prices. The processing power of computers has also improved significantly. In the past it was necessary to save and access data efficiently due to both storage capacity and processing ability constraints, so databases were tailored to their intended usage. But today it is possible to collect and store a miscellany of data including documents and images, and find information within that data later.

Analysis tools have also become easier to use, with a wide range of available tools such as data mining, machine learning, and statistical processing. Large-scale distributed processing has also become possible through technologies such as MapReduce*[1].

That said, before cloud services this kind of activity was limited to organizations that could collect, manage, and analyze data in-house. Now package tools that collect and analyze the online behavior history of customers have appeared. When cloud services and package tools are used, anyone can easily use big data with minimal initial investment.

This demonstrates that there are more and more opportunities for business use, such as data-based marketing and management decisions. At the same time, technological innovations known as the data revolution are occurring in each and every field. In March 2012 the United States government announced they were investing heavily in big data research and development, showing their commitment to proceed with a big data strategy as a nation.

3.3 Data Analysis is merely a Tool

We have dedicated ourselves to Internet measurement, taking great pains to promote wider understanding of the need for data collection and analysis and related labor and costs. With recognition for the concept of big data growing, it is becoming easier to gain understanding regarding these issues. Meanwhile, we are left with the impression that recent discussion about big data is focusing solely on tools and techniques. Data analysis is merely a tool. If you simply collect a large volume of data and dedicate massive amounts of CPU resources to analyzing it without a purpose, all you will end up with is useless numbers.

If you instead clarify what you want to gauge from the data, you will begin to see the path forward. It is important to always consider what information is useful for what purpose, as well as how, to identify problems, and to question the results. Data analysis is nothing but a means to an end. Data analysis is an iterative process that involves forming a hypothesis and then verifying it with data. If the results differ from those expected, you can identify new questions using them. By repeating this process you can uncover useful information and interesting facts.

Information technology is changing the nature of the thought process, with ideas now based on and verified using data. Of course, it has always been important to base ideas on data. However, the quality and volume of data handled and the methods of expressing it have evolved to a new level, so it is possible to consider matters while visualizing and literally interacting with data.

3.4 Issues in the Data Age

In the future, the importance of data analysis will increase in every field. Professional data scientists with field-specific knowledge and data analysis skills are now needed in each area. Those who are only able to use statistics and perform data analysis cannot identify problems. This means there is a need for specialists with domain knowledge in relevant fields, as well as the ability to challenge conventional thinking and interpretations, establish issues clearly, and use statistics and data as tools to resolve them. There is a severe shortage of specialists with these abilities, so the training of personnel poses a significant challenge.

*1 Distributed data processing technology developed by Google. It is widely used for big data analysis.

In the age of data, the collection and aggregation of data creates assets. Long-term data that makes retroactive analysis possible is particularly valuable. The quality of data is also important, even when handling large volumes of ambiguous data. If everyone based their analysis on the same data, the ability to extract beneficial information from that data would determine the degree of success. However, when the quality of the data varies, those with better quality data have the advantage. In practice, most data such as Internet traffic details and the behavior history of users of online services is not disclosed publicly. Consequently, access to information on popular services provides an enormous advantage. Put simply, companies with actual data that other companies do not have are in a position of strength.

Meanwhile, advances in the sharing of data are beneficial to society as a whole. At the same time, taking privacy into account when sharing data will become a major task in the future. Comparing multiple sets of data and the association and analysis of various data will be of increasing importance from now on. Sharing large amounts of associated data as widely as possible will be a key part of this. Third-party verification is a fundamental scientific tenet. Sharing data makes third-party verification possible, and serves as a basis for developing technology as a science.

The sharing of data must be balanced against online privacy. Social media makes it possible to create personal relationships with a wide range of people by sharing personal information with friends and acquaintances. Online shopping is also automatically customized to your tastes as you use it, providing increased convenience. Meanwhile, as technology for associating data evolves, unexpected speculation becomes possible. Private details could even be guessed from slight changes in user behavior. When it comes to online privacy, even among information technology experts opinions range from those who are hypersensitive to those who take a more relaxed approach. It is even more difficult for the average person to assess the potential risk, and it is likely it will be some time before we have a social consensus. At the end of the day, it comes down to finding a balance between the benefits of publishing and sharing information, and the risk of privacy breaches.

It will be an important social issue to form consensus with regard to online privacy, such as the degree that commercial enterprises or non-commercial public institutions will be permitted to track individuals, and the ways that information such as personal medical records will be shared for social benefit.

3.5 Recipient Literacy

It is also crucial for recipients of information to understand and question data. After all, it is possible to interpret the same data differently, and there are of course many ways to interpret data from the perspective of associating multiple pieces. Additionally, as demonstrated by how many books there are on the topic of “statistical lies,” with more emphasis being placed on data there will be more and more questionable data and dubious theories based on data. There is in fact a deluge of contrived statistical data and manipulated information originating from biased sources.

In the future it will be necessary for recipients of information to also have the ability to understand and question statistical data. We tend to want to see things as either black or white, but they are usually grey. We merely draw a line in the grey to determine black or white for the sake of expediency. When the recipient of information seeks a black or white answer, they are avoiding using their own judgment and instead placing that responsibility with the originator. However, with the multitude of information at our fingertips in contemporary society, it is necessary for recipients to accept grey as grey, and make the call on whether to judge something as black or white themselves. The same applies to online privacy, and while I believe that a certain degree of social consensus is needed, ultimately in today's society it is becoming necessary for individuals to make their own decisions and take responsibility for their actions.

Author:
Kenjiro Cho
Dr. Cho is Research Director of Research Laboratory at IIJ Innovation Institute Inc. He is engaged in Internet research such as traffic measurement and data analysis. He is a guest professor at Keio University, Faculty of Environmental Studies. He is also an adjunct professor of Information Science at Japan Advanced Science and Technology.