Rich Information Produce

By Yukio EBINO* and Koichi KONISHI†

ABSTRACT The Internet and the emerging ubiquitous network are producing large bodies of information, and their growth goes readily beyond human capability to handle. This situation calls for making much advances in technology to produce high quality knowledge from a huge amount of raw data. This paper describes the growth of the bodies of information, the limit of human capability to handle them, and the current state of such technologies pursued in the large and at NEC. This paper also discusses expectations for them and their open issues.

KEYWORDS **Machine learning, Pattern recognition, Ubiquitous network**

1. INTRODUCTION

The amount of information harvested by the ubiquitous network is growing rapidly and significantly. The vast quantity of raw information is already going beyond the human capability to utilize it and there is a strong need for machine intervention in acquiring high quality knowledge from the raw information that humans can use readily. The authors call such acquired knowledge "rich information produce (RIP)," and claim that technology for RIP generation will be increasingly important in keeping pace with the total amount of information. This paper elaborates on this perspective, clarifying the background, justifications, current state, expectations, and open issues.

2. WHAT IS RIP?

The RIP perspective has as its background the continuous, accelerating increase of publicly accessible information. There are changes not just in the quantity but also in the nature of this information. Increasingly, a part of the newly available information tends to be taken directly from around our daily life in the office as well as in the home. A natural expectation is emerging that from such a large amount of information close to our daily life, highly valuable knowledge can be extracted. On the other hand, human capability is limited in that one can directly handle and utilize only a certain amount of data. In order to meet this expectation, therefore, it is essential to develop a suitable technology to assist us in extracting knowledge from large amounts of data. The RIP perspective requires that to fully make use of data generated by the Internet and ubiquitous network, technologies for the advanced use of information such as machine learning should be extensively developed.

2.1 The Vast Amount of Data

Figure 1 shows the expected growth in storage sales in Japan. It shows that while the total capacity of sold storage grows 80-100% a year, the sales scarcely grow at all. It can be expected that more and more data is going to be stored without much increase in cost.

In the last ten years, it was mainly the growth of the World Wide Web that drove the increase in the amount and variety of publicly available information. While this trend is likely to persist, a new surge of

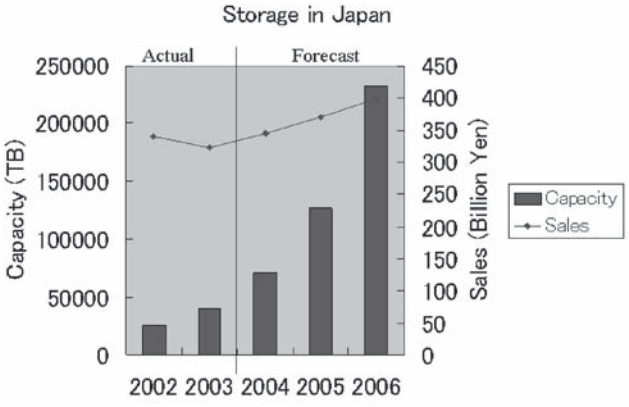

Fig. 1 The estimated growth of storage sale.

*Vice President, Central Research Laboratories
†Research Planning Division

information growth is emerging as the ubiquitous network forms and becomes fully functional.

Information on the Web was reported to consist of at least eight hundred million pages in 1999[1]. As of March 2005, Google alone claimed to have indexed eight billion pages. This volume represents probably only a fraction of the entire Web. The recent craze on blogs probably further accelerated the increase, as blogs and its surrounding technologies make it much easier for anyone to create and maintain Web pages.

A harbinger of the information blowout induced by the ubiquitous network can be seen in the changes in the usage of the cellular phone, which will play the central role in the coming ubiquitous network. Cellular phone usage is widespread and it has so many functions to collect information. For example, the cellular phone with digital camera became so popular that it is now common to always carry a digital camera, which in turn has resulted in an increase in the number of digital images on the Web. Also, cellular phones that can read two-dimensional bar codes are being rapidly deployed. In Japan, four million cellular phones capable of reading QR codes, a kind of two-dimensional code, sold by March 2004, and twenty million by a year later. A QR code can convey tens or hundreds of times more information than a conventional bar code (8,000 digits, or more than 1,800 Kanji characters). Every time somebody samples a QR code, that much information is input to the cellular phone, along with information of who showed an interest on which QR code. When the cellular phone also features GPS, the location of the sampling will also be recorded. Furthermore, electronic money based on the IC card embedded in the cellular phone has recently been introduced to the public and is being promoted intensely. When this function becomes popular, it will record most of the user's personal purchasing history.

It is true not only of cellular phones that using a terminal will make a record of the user's action. The desktop PC in an office, for example, contains lots of records of work done by its user, such as what file was opened when and what mail was sent to whom. To make use of this kind of information, Microsoft, Google, and other companies have developed search tools for desktop PCs and offer them free to users. Applications of this kind of information, however, do not stop at searching desktop PCs; activity reports may be able to be compiled semi-automatically from the information.

As security concern is recently rising, installing surveillance cameras in public locations is also becoming common. Increasingly, the camera will be digital and connected to a network to constantly provide a large amount of image data.

2.2 Acquiring Useful Knowledge

Acquiring useful knowledge from a large amount of raw data is actually quite a common practice. The explosive increase in information, however, goes beyond human ability, making computer-aided data acquisition essential. Technologies such as search, pattern recognition and machine learning play the central role in this trend. In general, the role is to figure out rules unknown to us or too complex for us to describe.

Conventional searches, typically performed in databases for enterprise systems, look for a target whose description can be clearly given. An example would be the "top 10 most sold items last month in the store No 23." On the other hand, when you search in the Web, you are likely to be looking for an article that informs you of concise knowledge on a specific topic in easy words. However to express "concise" or "in easy words" as search conditions, is usually difficult. It was Google that made a breakthrough with regard to this problem area. Based on the linked structure of the large number of Web pages, it assigned a numeric value to each page representing the degree in which the page is regarded as useful collectively by the authors of the other Web pages. With these values Google can show at the top of the search result the pages most highly valued by many people. This was possible because opinion of the Web authors on how useful was the page is subtly embedded in the linked structure of the Web pages.

Likewise, rules for recognizing human figures or faces in images from surveillance cameras are quite hard to describe. In pattern recognition tasks, recognition targets are usually well understood; they are, for example, faces, handwritten letters, and spoken words. What is hard for us to achieve is to satisfactorily describe rules that explain why some images are faces (or other items) while others are not. With machine learning technology, the system can learn the rules from a lot of sample images labeled either as "It is a face" or "It is not a face." Once the system has learned the rules, it can repeat recognition of the target without learning it again.

2.3 Why Now?

It has been a goal pursued for a long time to extract information from large amounts of data. The reason for emphasizing the RIP perspective now is that changes in human needs and in the computer's capability have just begun to coincide.

With but one head, a single human being can

handle only so much information. Although information beyond that limit has been handled by specialists, this was achieved with the help of a large computer. For example, the elementary particle physicist handles a tremendous amount of data produced by experiments with the help of supercomputers. Now, ordinary people are faced with much more information than the amount they can handle satisfactorily by themselves. To process this much information requires some way of levering its quantity in order to obtain quality.

An estimate indicates that the maximum number of elements a human being can grasp well is tens of thousands. As a circumstantial evidence supporting this, there is an episode about the maximum size of well-functioning enterprise organizations; Reference [2] estimates its number at 150. When the organization size is within this limit, each member can keep a detailed knowledge about relationships between the members. As the number of possible relationships between members is the number of members squared, 150 to 1,500 as the size limit translates into tens of thousands as the limit of the number of relationships.

Another circumstantial evidence is the number of elements in LSI designs. Despite the fact that LSI grew more and more complex during the last two decades, the number of elements the designer directly deals with was kept to around the tens of thousands. This was possible because the design of increasingly more complex modules was automated so that they could be regarded as elemental units in the design process.

In contrast to the human situation, computers are gaining more and more computational capacity. As **Fig. 2** shows, the performance of the world fastest computer became 300 times faster during the last ten years. Performance of the desktop PCs have been more than sufficient for several years if they are to be used only for e-mails, Web browsing, and preparing documents.

2.4 Current Examples

Currently the main field of RIP is the Web, as it is the largest and fastest growing accumulation of information. From now on, however, the streams of context information generated by the ubiquitous network will prove to be a fertile field for RIP and eventually become the largest field.

Following the success of Google, Yahoo!, Amazon, Microsoft, and others are eagerly developing search technologies. They aim to acquire and make use of valuable knowledge from the large amount of information on the Web. Besides, they are expanding the searchable sources of information. First of all they tap the usage history from their large customer base. Then they turn to a variety of traditional information pools. Several companies now offer desktop search tools, which present the seamlessly combined results of searches in a PC and from the Web. Amazon now lets you search through the content of a large number of books. Google has recently started to search through academic papers as a beta service. This process has a predecessor. NEC Laboratories America, then NEC Research Institute, used to be running CiteSeer[3], which let you search through academic papers following the citing of relationships between them.

Enterprise application software vendors in the United States have been advocating, for a number of years, business intelligence, which essentially consists of observing activities within and close to the enterprise, thus making analyses based on observations to discover problems or opportunities, and properly handling such discoveries to improve the business operation. This obviously originates in data mining; the difference is that the observation and the analysis used to be made on accounting information while now they are performed on data that was rarely

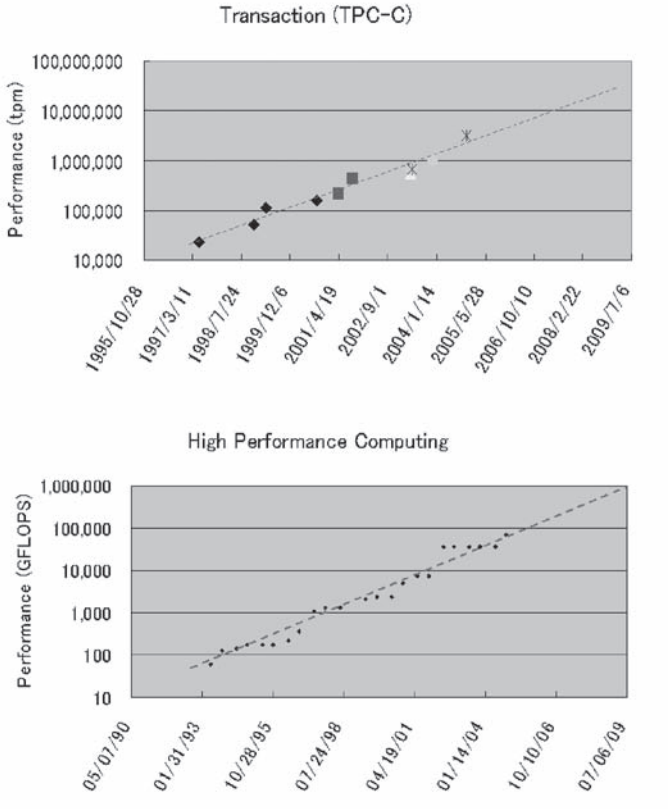

Fig. 2 Computing performance growths.

measured quantitatively until recently, such as customer and e-mail traffic.

Ubiquitous computing facilitates the measurement and distribution of information that traditionally was not subject to measurement. Equipment and facilities required for this purpose are being deployed steadily in Japan. The wide availability of broadband and mobile networks guarantees connectivity and high bandwidth at the office, in home, and in transit. Making digital images is a trivial task even for non-technology oriented people as most of the cellular phones in Japan are now equipped with a digital camera. RF tags, equipment for producing information on the location of goods, are being put to practical use to implement the traceability of enterprise assets and merchandise.

3. TECHNOLOGY FOR INFORMATION UTILIZATION AT NEC

The NEC Intellectual Asset R&D Unit has pursued research activities on machine learning with a long history and achievement and is now engaged in a variety of first-class machine learning technologies. Two major branches among these activities are pattern recognition and data mining. Their details are described in the other articles of this issue.

Research on pattern recognition, performed at the Media and Information Research Laboratories has produced various original feature selection and other preprocessing methods as well as GLVQ[4], an original machine learning algorithm. It can thus boast of a high performance in a variety of industrial applications. NEC products incorporating GLVQ include letter sorting systems used in postal offices and a face recognition system. The letter sorting system recognizes the recipient's address written on envelopes. The face recognition system uses the algorithm to detect areas showing a face in the target image.

The research on data mining[5], conducted by Internet Systems Research Laboratories, maintains a rich arsenal of learning algorithms based on the information theoretic learning theory, and has pioneered a wide variety of applications in domains such as CRM, open mission-critical systems, and system security. In the system security domain, for example, the research output has been used to automatically detect an outbreak of Internet worms, even if the worms are of a hitherto unknown kind. This is a function that conventional systems could not perform, as they identify worms based on data describing the traits of known worms. The new detection system learns the normal behavior of the mail system so that it can notice any anomalous behavior, which is possibly caused by worms.

Prof. Vapnik, a world renowned machine learning research scientist since the 1960s, is leading a team of scientists in exploring the frontiers of machine learning algorithms at the Princeton office of NEC Laboratories America. Support Vector Machine[6], one of the innovations made by the Prof. Vapnik, has created a whole new research domain that has been actively pursued by algorithm and application researchers worldwide.

4. EXPECTATIONS FOR THE RIP PERSPECTIVE

As RIP becomes more commonly used, there will be day-to-day occurrences that involve fewer uncertainties, where people can start acting earlier and with more confidence than before. The RIP will also facilitate summarizing the large amount of data on the behavior of people and help them to communicate their experiences and findings to others in a more lively and comprehensive way. The primary source of RIP so far has been from Web information, but will gradually expand first into context information and then to real-world sensory information.

4.1 Organizing the Web Information

Google has presented and made available to everybody the information which Web site users worldwide collectively regard on average as valuable.

Following this trail, Microsoft and "Yahoo!" have started to tell us how special interest communities such as regional ones evaluate information on the Web. As credible information becomes available about what kinds of people are interested in what to what extent, companies will be able to launch new products and services for niche targets with more confidence.

4.2 Making Full Use of the Context Information

Corporate executives will be able to become aware of and respond to symptoms of unattended problems in their companies or to recognize new business opportunities much earlier than at present. This trend will spread beyond the circle of executives to managers and employees in various positions, thus bringing the real-time enterprise into reality.

4.3 Analyzing Real-World Sensory Information

IT systems that are close to individuals will acquire information from their environments that can be captured by the human senses such as images and

sounds and analyze it so that the system will be able to detect and warn the user of changes and imminent dangers, while also being able to interact with the user in more natural ways. For example an automobile with an on-board computer, through cameras and microphones all over the vehicle, will sense pedestrians and obstacles around the car, warn the driver of them and keep a constant distance from the preceding car. A voice-based conversational interface will also be available so that users will find it easier to operate, even when their hands are occupied with the steering wheel.

5. OPEN ISSUES

Turning the RIP perspective into reality involves progress in ubiquitous networking technology for gathering, distributing, and accumulating a huge amount of data as well as machine learning technology for acquiring useful knowledge from massive bodies of information. As issues concerning the former have been much discussed for some years this section deals with ones that concern the latter.

5.1 Organizing Utilization Schemes of Prior Knowledge

Although at the core of the machine learning process there are learning algorithms, these are not the only components of machine learning applications. Nor are they the most important in building high performance applications. The key to a successful machine learning application is an ingenuous incorporation of prior knowledge on the nature of the input data. With regard to face identification, for example, based on the prior knowledge that the human face is three-dimensional, it is possible to obtain a much improved identification performance. This is done by matching the target image with views on known faces from various directions that has been computed beforehand using three-dimensional face models. How prior knowledge can be incorporated in machine learning varies a lot depending on the nature of applications. As of now, personal experiences and intuitive insights are the only guides in the new application domains. An issue to be addressed is turning these empirical sets of methods into organized knowledge that can be generally applicable to new domains.

5.2 Robustness to Changes of the Environment

Conventional pattern recognition has been developed to show a better performance under the circumstances that favor the system, such as under lighting of constant brightness from a fixed direction. In order for pattern recognition, however, to be used in various scenes of people's life, it is necessary for pattern recognition technology to maintain a certain level of performance under a wide variety of conditions. A vehicle-mounted image recognition system, for example, should correctly recognize pedestrians and obstacles around the car irrespective of whether it is day or night, or sunny or raining.

5.3 Smaller and Faster Implementation

Machine learning is a rather computation intensive task. Even taking into account the ongoing rapid increase in computing power, a smaller and faster implementation of machine learning algorithms and applications is quite important for realizing the RIP perspective on portable terminals or vehicle-mounted devices.

6. CONCLUSION

The RIP perspective advocates the development of machine learning technology by aiming at making full use of the huge amount of data that the ubiquitous network will generate. This strategy is derived from such observations as the limited human ability in information utilization and the ever-increasing computing power. The US IT industry is leading the machine-assisted utilization of the Web information, but the focus is moving on to the utilization of context information and real-world sensory information. NEC has a proven record of excellent research in pattern recognition and machine learning and will take it to full advantage in realizing the RIP perspective. The challenges are organized schemes of prior knowledge utilization, more robust recognition, and smaller and faster implementations.

REFERENCES

[1] S. Lawrence and C. L. Giles, "Accessibility of Information on the Web," *Nature,* 400, p.107, 1999.
[2] M. Gladwell, "The Tipping Point: How Little Things Can Make a Big Difference," Little, Brown, 2000.
[3] S. Lawrence, C. L. Giles, and K. Bollacker, "Digital Libraries and Autonomous Citation Indexing," *IEEE Computer,* **32**, 6, pp.67-71, 1999.
[4] A. Sato and K. Yamada, "Generalized learning vector quantization, Advances in Neural Information Processing Systems 8," MIT press, Cambridge, MA, pp.423-429, 1996.
[5] K. Yamanishi, "Text and Data Mining," Iwanami Statistics Frontier series 10: Statistics on Language and Psychology, pp.179-242, 2003.
[6] V. N. Vapnik, "Statistical Learning Theory," Wiley, 1998.

Received April 8, 2005

Yukio EBINO joined NEC Corporation in 1967 and was engaged in the development of operating systems and middleware. He is now Vice President of NEC, overseeing the Central Research Laboratories.

Koichi KONISHI joined NEC Corporation in 1989 and was engaged in research and development of parallel and distributed systems. From 1995 to 1997, he was a visiting researcher at the NEC Research Institute. He is currently a manager of the Research Planning Division.

* * * * * * * * * * * * * * * * *