New Generation Computing, 28(2010)207-215
Ohmsha, Ltd. and Springer

Preface

Special Issue on Information Explosion

Masaru KITSUREGAWA
The University of Tokyo
4-6-1 Komaba Meguro-ku, Tokyo, 153-8505, JAPAN
kitsure@tkl.iis.u-tokyo.ac.jp
Toyoaki NISHIDA
Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, JAPAN
nishida@i.kyoto-u.ac.jp

Received 1 April 2010

The "Infrastructure for the Information-explosion Era" (Info-plosion) project was launched for July 2005 - March 2011, aiming at establishing the fundamental technologies in the information-explosion era.[1)] It addresses efficient and trustworthy information retrieval from explosively growing and heterogeneous information resources; stable, secure, and scalable information systems for managing rapid information growth; and information utilization with harmonized human-system interaction. The purpose of this special issue is to provide the reader with glimpses of the project by giving a brief introduction in the preface followed by a showcase of outstanding work resulting from the project. Although the five papers alone might not be enough for the reader to grasp the project as a whole, which really covers a broad spectrum of research being conducted in the Info-plosion project, they will definitely enable her/him to feel the prominent achievements obtained from unceasing challenges underlying the project and the shared atmosphere of innovation toward the unprecedented increase of information on the globe.

The Aim of Info-plosion Project in the Age of Zettabyte Data

Our research project is motivated by the advent of the Info-plosion era. When this project was proposed in autumn 2004, the volume of created information was measured in the unit of Exabyte. Four years later, Google was processing 20 petabyte data on MapReduce in 2008.[2)] In 2010, the volume of information human creates in a single year will exceed 1 zettabyte, according

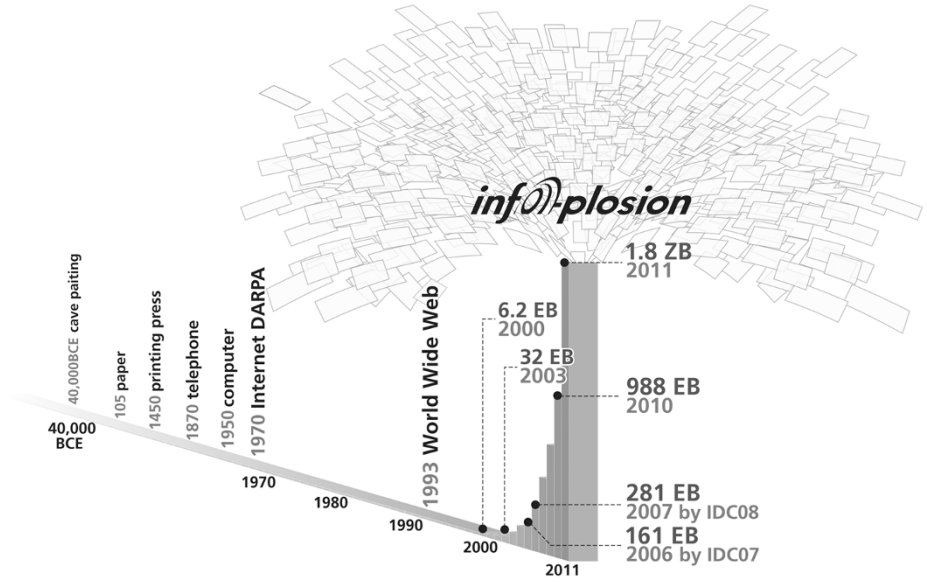

Fig. 1 Information Explosion, or "Info-plosion"[4)]

to IDC[3)] (Fig. 1). We are in the midst of Information explosion (Info-plosion). What our project aims at is, in short, focusing on Info-plosion as a phenomenon and taking it as a new challenge for researchers in the information field: 1) We identify various problems that arise in the Info-plosion era, and tackle them with pioneer work. 2) Instead of regarding Info-plosion as headaches, we challenge unprecedented research works in a positive attitude to create values from the explosive information. Generally speaking, research projects in this field usually aim at a clearly defined specific target. This project, on the other hand, has wider focus on the phenomenon of Info-plosion as a whole. Thinking it over, it is researchers in the information field who gave the bases of this Info-plosion phenomenon. We are responsible to overcome this phenomenon.

Web Search-engine Infrastructure Based on NLP In the Info-plosion era, one of the most serious problems is overloading of information. Because the volume of flooding information exceeds the ability of acceptance of individuals, the relative distance to reach necessary information becomes longer rather than it previously was. Therefore, a method to find exact information has become much more important. Although searching information sounds a classical topic, the situation has been dramatically changed in the Info-plosion era. One of the Info-plosion we have to tackle is textual information on the Web. The existing search engines function well for a so-called navigational search, locating the top pages of companies, shops, and restaurants. However, they cannot support an informational search in a satisfactory way, such as "problems in child-rearing" in general, "a decline in the children's physical strength," or "measures against chil-

Fig. 2 An Open Search-engine Infrastructure: TSUBAKI[5)]

dren's physical strength decline." We need much more powerful, next-generation search facilities to exploit the accumulated information on the Web. Our Infoplosion Project puts stress on research infrastructures, and has constructed an open search-engine infrastructure, called TSUBAKI developed by Kurohashi and Shinzato (Kyoto University), for next-generation search technologies based on deep natural language processing (NLP).[5)] TSUBAKI is equipped with 100 million Japanese Web pages, cleaned up and marked up with sentence boundaries for each page, and preprocessed with the morphological analyses, syntactic analyses and synonymous expression analyses for sentences. The resultant enriched data is stored in an XML-based format. In addition to words, synonymous expressions and dependency relations are registered in the index, which enables TSUBAKI to handle complex natural language expressions in queries and documents and to identify synonymous expressions (Fig. 2).

TSUBAKI is not only used through ordinary Web browsers, but it also provides an open access to its Web corpus and search APIs. A large text corpus is essential to knowledge extraction and enhancement of NLP, and the XML-based 100-million-page corpus has been widely used in the research community. Furthermore, white-boxed TSUBAKI APIs are supporting the next-generation search research activities, such as clustering-based Web information summarization, or opinion and sentiment overview.

Innovative Search Technologies As Web is expected to be a huge library which collects human knowledge, users expect more convenience of search. This requires thorough localization for each language, as TSUBAKI has realized, which is considered to be a weak point of existing commercial search engines developed in the United States. As advanced search technologies, in addition to search under complex conditions, another difficult problem to solve on the Web is to grasp different opinions. Generally, there is a wide variety of opinions for social problems. Examples of such problems include: "baby post (anonymous commission system of baby whom parents cannot raise)," "medical care for elderly people," and "hospital managed by an incorporated company." However, because current commercial search engines display major opinions by ranking methods, it is difficult to find minor opinions. Even the existence of such minor opinions might be overlooked, despite their fresh points of view of the problem. A system called "OpinionReader" developed by Fujii (University of Tsukuba) focuses on this point. Also, it is impossible to know what we don't know on the Web. In order to help finding this "unknown unknown," Torisawa (NICT) et al. have developed "TORISHIKI-KAI" in this project.[9] They have constructed a million-word-scale semantic network which can be browsed as a Web search directory. This has a function such as showing hidden problems and generalization of topics, which is useful for finding "unknown unknowns."

Development of Research Platforms to Support Info-plosion Experiments For challenging such a deep analysis of complex queries and comprehension of long-tailed information, it is not enough to put an original processing system on top of current commercial search engines, because they only display top 1000 pages which may not include long-tailed information. That is to say, we must crawl and parse a huge volume of Web pages by ourselves in order to achieve such fundamental research works. Since it is extremely troublesome for individual researchers to manage such things all by themselves, we have provided co-developed research platforms in our project. Such a strategy is becoming more important for computer science. One of the platforms is a large-scale distributed computing environment called "InTrigger"[*1] developed by Taura (University of Tokyo) et al., which collects more than 1000 CPU cores. Recently, middleware suitable for a large-scale cluster were developed such as MapReduce, Hadoop, and PigLatin. They have features to process a huge volume of data, although the amount of calculation for each data is relatively small. This is completely different from conventional scientific calculation performed on a supercomputer. Development of such a computing environment is interesting for system researchers, while this is also an exciting platform for researchers who want to process a huge volume of data in them. Various research systems including "TSUBAKI" have been developed using InTrigger. Other co-developed research platforms include "IMADE" developed by Sumi, Bono, and Nishida (Kyoto University).[6] This is a sensor room equipped with latest devices which

[*1] http://www.intrigger.jp/

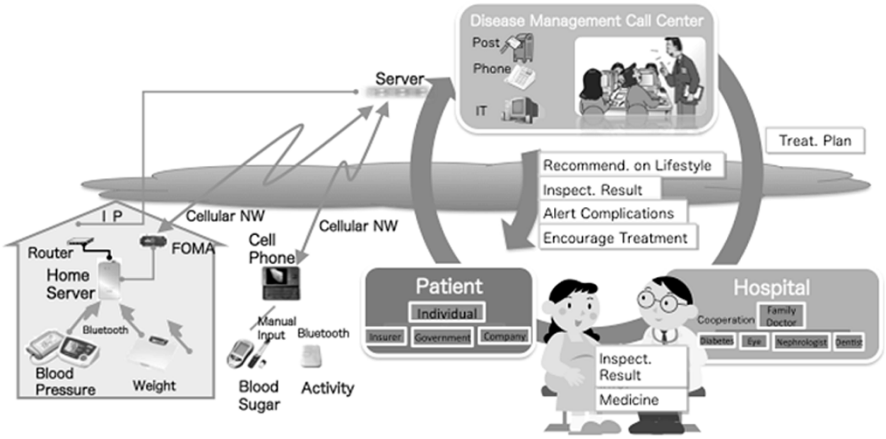

Fig. 3 Preventive Healthcare with Vital Sensor Network[8)]

trace sensitive interaction among multiple people. "SlothLib" is a toolkit for making own search applications developed by Oshima, Nakamura, and Tanaka (Kyoto University).[7)] Web applications can be built up upon SlothLib instantly without elaborate programming, realized by using commercial search engines.

Creation of Value Extracting Information from Info-plosion While we are in the age of Info-plosion, the situation was completely different 25 years ago – we were in hunger for information in those days. Since an explosive volume of data is available today, we should make use of the data aggressively. The increase of the volume of data is largely due to the progress of sensor technologies. There are a number of innovative projects using sensors such as analyzing soil conditions for cultivation, improving efficiency of logistics, PLM for safety of products, and so on. They seem to possess business scenarios to create new values by extracting precious information from Info-plosion. They have great possibilities in the coming Info-plosion era. Our project focuses on sensors for human activities, and makes researches into preventive healthcare with a vital sensor network developed by Inoue (Kyushu Institute of Technology) and Nakashima (Kyushu University) (Fig. 3).[8)]

Information Can be Medicine The Ministry of Health, Labour and Welfare of Japan reported that the number of diabetics increased to 8.2 million. On the other hand, the number of diabetic specialists is only less than 4 thousand, which means one specialist must treat over two thousand patients. This leads to the national increase of medical cost by causing atherosclerotic diseases or by causing renal diseases. Accordingly, our goals are to prevent and control diabetes mellitus, using a sensor network to capture vital information and record lifestyles of patients.[*2] Sensor networks will optimize disease management, since observation

[*2] The issues are discussed in disease management in more general.

must be made in the daily life of patients out of hospitals. We conducted two verification studies: one for metabolic syndrome, which was shown as the main cause of diabetes in recent studies as obesity with visceral fat accumulation, and another for diabetes.

The first study aimed at managing metabolic syndromes. We asked one hundred monitors to wear activity sensors and pulse rate sensors, and to use wireless weight scales and blood pressure sensors for 11 days. On the 6th day, health instructors suggested lifestyle problems, upon the sensor data, to the monitors. The sensed data were automatically uploaded to the server via Bluetooth or wireless LAN. The result showed that 91.6 % of the monitors increased their amount of exercise (9.55 % increase in average ($n = 83$)) after the healthcare instruction. In the analysis of the kinds of activities, "standing" and "walking" were increased, and "using elevator" was apparently decreased in daily behaviors. Additionally, we could estimate the optimization of health instructions in both time and accuracy, which should lead to enormous cost reduction of Particular Health Checkup System held all over Japan.

The second study targeted on diabetics. We used activity sensors, blood sugar sensors, weight scales, and blood pressure sensors. The former two were mobile, and all the data were uploaded via mobile phones or wireless LAN. The mobile phone had an application to show the recommendation message either triggered by the sensor data or sent daily. As the result, we found a critical case of reverse white coat hypertension, where the blood pressure is high only at out of hospital. Moreover, according to the analysis of recommendation effect, the frequency of measuring vital information increased after a recommendation, and several messages certainly increased the amount of activities, and several decreased. These results have not been found in our studies so far without such an environment. In addition, we developed a collecting system of activity information using smart phones, and gathered well-labeled 3D accelerometer data of 150 people. Since smart phones have been deployed tens of millions in the world, they can be a global platform for progress of activity recognition.

Papers in This Special Issue

This special issue contains five papers. The first paper[9] addresses automated construction of a semantic network by making use of the huge information resource of the Web. The challenge of this work is to go beyond the conventional keyword-based Web search, to enable the user to find information that one might not even know. The main contribution of this work is the development of the collective body of knowledge acquisition methods for automatically generating TORISHIKI-KAI, a machinery that maps out the context of use or handling of the topic the user inputs, and classifies topically related search terms according to semantic categories.

The second paper[10] presents a model-based approach to fault localization that allows the human analyst to narrow down the fault localization in large-scale computing environments. Localization is regarded as anomaly detection problems in system behaviors by automatically analyzing data resulting from

recording all the function call traces and identifies anomalous behaviors. The key challenge here is the development of a scalable, automated technique of detecting anomalies that can effectively localize faults in large-scale environments. Their contribution is twofold. The pre-failure model derivation permits to build an execution model that reflects the function-calling behaviors of the target system, based on function-call traces collected from all processes. The post-failure model-based anomaly detection allows for identifying the most deviant behaviors in the failed run by comparing the failure traces with the derived model.

The third paper[11] addresses a high performance shared file system for a large-scale PC cluster system to share data and support distributed data-intensive computing. Key issues include maximizing the distributed file I/O bandwidth, avoiding access concentration, and supporting fault tolerance. The contribution of this paper is a global distributed file system called Gfarm Grid that federates local file systems on compute nodes and allows multiple file replicas in any location. Gfarm Grid features data-location-aware process scheduling that allows the user to manage file partitioning and file replica placement for better file I/O performance.

The fourth paper[12] addresses image annotation and retrieval. The problem is challenging because of various appearances of objects and a wide variety of categories. The key idea introduced in this paper is to combine the higher-order local auto-correlation features and a framework of probabilistic canonical correlation analysis. The highly compressed nature of the intrinsic space arising from conceptual learning between images and labels allows for faster and more accurate image annotation and retrieval results.

The last paper[13] aims at building a speech processing technology that is robust against the intrinsic variations, differences of speaker, microphone, environment, etc. They derive speech structures as completely transform-invariant features and discuss their linguistic and psychological validity. The idea is applied to automatic speech recognition and computer-aided language learning.

Concluding Remarks

Wikipedia defines information explosion as follow: "The information explosion is the rapid increase in the amount of published information and the effects of this abundance of data. As the amount of available data grows, the problem of managing the information becomes more difficult, which can lead to information overload." Wikipedia and others address its negative aspect such as management difficulty of information and information overload. On the other hand, we can consider this phenomenon a yet another opportunity for this century. Thirty years ago, there were no web, no emails. Students were waiting for new issue of academic journals. Airfare was so expensive and a quite small number of professors could attend international conferences whose number was also very few. In those days, people starved for information. Now information is explosively increasing and its overload becomes a serious problem. In other words, this is the first experience for human beings to face such overwhelming information.

In Info-plosion project, we are challenging two aspects of information explosion. First, in order to overcome the negative aspects of information explosion, we have been building the system TSUBAKI. The goal is 'Beyond Search.' It is much more powerful compared with the current commercial search engines. We do not pursue its quick response time at all which is extremely important for current advertisement-based commercial search engines; to the contrary, we are focusing on the quality of the results. It can answer far more complicated queries than the customary keyword-based simple queries. Second, we have been challenging value extraction from explosive information. Information was scarce before, now information is so abundant. This implies that we are able to utilize such information to extract various values, which was very difficult in old centuries. Google announced they found one trillion pages in 2008. The amount of surface web information space, however, is not that large. Nowadays, people can express their opinions easily in their blog and twitter. This enables us to sense the world's emotion. IOT (Internet of things) is expected to generate far more information, which potentially sense phenomena happening in the world much more in depth. In the Info-plosion project, we conducted various system development and experiments on sensor-based health care solution. Using accelerometer, we could measure the activities of each individual patient quantitatively, which led to personalized medication.

When we started the Info-plosion project, the problem caused by information explosion was addressed. As far as we knew, there were no academic projects that focused on the issue. Thus, we decided to launch it in Japan. The style of the project is rather unusual. In contrast to the clearly defined targets of systems development aimed at by other ongoing projects, such as GRID, ITS or Supercomputer, information explosion is just a phenomenon. We considered that we had to do something with information explosion, although we did not know exactly what we had to do. We have called for proposals three times during these five year project period in order to identify the key issues in the information explosion era. So far, we have reasonably succeeded in the sense that our project have derived numerous interesting results. Now we have come to believe that Info-plosion will be a key enabler for the next generation IT systems.

Acknowledgements

We thank members of the guest editorial board: Hiroshi Nakagawa, Akiko Aizawa, Satoshi Matsuoka, Shigeru Chiba, Takashi Matsuyama, and Hiroshi G. Okuno for their support throughout the whole publication processes, anonymous reviewers for their high-quality review, Masato Oguchi for his assistance in writing this preface, and Takashi Chikayama, an associate editor of the New Generation Computing journal, for kind and useful advices. The research reported in this special issue is supported by Grant-in-aid for Scientific Research on Priority Areas: "Cyber Infrastructure for the Information Explosion Era" of The Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. We are grateful for the continued support.

References

1) Kitsuregawa, M., "Creating Vital Information Technologies for the Info-plosion Era: Info-plosion Project: Current Reports (in Japanese)," *IPSJ Magazine, 49, 8*, pp.881–888, Aug. 2008.

2) Dean, J. and Ghemawat, S., "MapReduce: Simplified Data Processing on Large Clusters," *CACM, 51, 1*, Jan. 2008.

3) IDC white paper: "The Diverse and Exploding Digital Universe," 2008. http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

4) Moore, F., "Storage New Game, New Rules," *Horison Information Strategies*, 2003. http://www.horison.com/topics/2004/08/newrules pg34.pdf

5) Shinzato, K., Shibata, T., Kawahara, D., Hashimoto, C. and Kurohashi, S., "TSUBAKI: An Open Search Engine Infrastructure for Developing New Information Access Methodology," in *Proc. of Third International Joint Conference on Natural Language Processing (IJCNLP2008)*, pp.189–196, Hyderabad, India, 2008.

6) Sumi, Y., Nishida, T., Bono, M. and Kijima, H., "IMADE: Research Environment of Real-World Interactions for Structural Understanding and Content Extraction of Conversation (in Japanese)," *IPSJ Magazine, 49, 8*, pp.945–949, Aug. 2008.

7) Ohshima, H., Nakamura, S. and Tanaka, K., "SlothLib: A programming library for research on web search," *DBSJ letters, 6, 1*, pp.113–116, Jun. 2007.

8) Nakashima, N., Inoue, S., Tsuruta, H., Sudo, O., Kobayashi, K., Inoguchi, T., "INFO-MEDICINE CONCEPT, Information can be a Medicine if it is Provided in a Timely and Appropriate Manner," in *Proc. of the 12nd China-Japan-Korea Medical Informatics Conference*, pp.22–25, 2009.

9) Torisawa, K. et al, "Organizing the Web's Information Explosion to Discover Unknown Unknowns," *New Gener. Comput.* (in this issue).

10) Maruyama, N. and Matsuoka, S., "Model-Based Fault Localization: Finding Behavioral Outliers in Large-Scale Computing Systems," *New Gener. Comput.* (in this issue).

11) Tatebe, N., Hiraga, K. and Soda, N., "Gfarm Grid File System," *New Gener. Comput.* (in this issue).

12) Harada, T. et al, "Image Annotation and Retrieval for Weakly Labeled Images using Conceptual Learning," *New Gener. Comput.* (in this issue).

13) Minematsu, N., Asakawa, S. and Suzuki, M., "Speech Structure and Its Application to Robust Speech Processing," New Gener. Comput. (in this issue).