Advanced Analytics for Intelligent Society

● Nobuhiro Yugami ● Nobuyuki Igata ● Hirokazu Anai
● Hiroya Inakoshi

Fujitsu Laboratories is analyzing and utilizing various types of data on the behavior and actions of people and society, as well as environmental change. In this way, it is proceeding with R&D on Intelligent Society to achieve a more prosperous and secure society. This paper focuses on two new types of data. The first one is social media including blogs, Twitter, and social networking services (SNS). The second is data obtained from various types of sensors such as mobile phones, automobiles, and environmental sensors. These data are very different from business data that traditional analytic technologies deal with in business intelligence applications. To realize Intelligent Society, we are researching new and advanced technologies to analyze such data. This paper introduces three of the technologies: social media analysis, optimization and spatiotemporal data processing.

1. Introduction

This paper presents new analytic technologies required for realizing Intelligent Society together with simple applications. Fujitsu Laboratories is engaged in R&D on technologies for predicting human or social behavior by analyzing massive amounts of data such as social media, sensor data, and corporate business data. These predictions can be used as the basis for formulating business plans and solutions to social problems. To realize Intelligent Society, various technologies are required such as information extraction and integration, time-series analysis, prediction and optimization, in addition to the traditional and typical analytic technologies including statistics and data mining.

From among these technologies, this paper describes the following three technologies. First, natural language processing technology for picking out various actions of people and society from social media (e.g., Twitter) is introduced. Second, this paper describes optimization technology for formulating the optimum business plans and solutions to social problems based on the results of prediction. The third technology is spatiotemporal data processing technology for handling data on time and space. Data aggregation and retrieval in relation to time and location are essential technologies for analyzing human or social behavior and they are likely to become even more important in the future.

2. Social media analytics technology

One major purpose of the Intelligent Society project in Fujitsu Laboratories is to help to solve increasingly complex social problems by using information and communications technology (ICT). As the first step toward achieving this purpose, Fujitsu Laboratories is promoting R&D on technologies for capturing human or social behavior by analyzing social media such as blogs, micro-blogs (Twitter, etc.), and SNS (mixi, Facebook, etc.). The following presents Sentiment Analysis and Social Event Visualization which

We have been developing so far, and gives a description about the technologies required.

Sentiment Analysis that we have jointly developed with Nifty Corporation[1)] is intended for collecting consumers' opinions on specific products or services from social media and analyzing their reputation. **Figure 1** shows a result of Sentiment Analysis on a certain soft-drink product. Figure 1 indicates that, while this product is mostly negatively evaluated in terms of the taste, it is well received from the perspective of being an aid to losing weight.

Figure 2 is an example of Social Event Visualization, which we are currently developing. Social Event Visualization can help us to understand more accurately what is happening in our society by monitoring and detecting events that have already occurred. Our

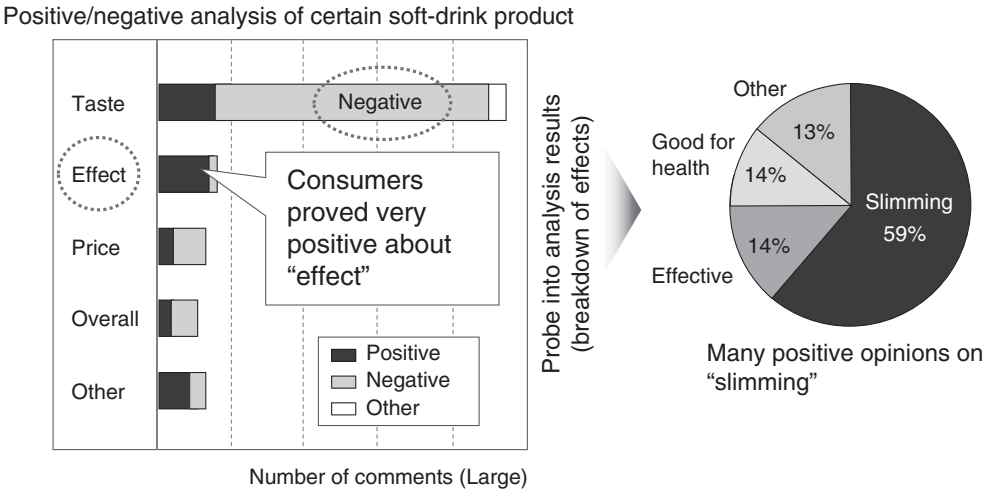
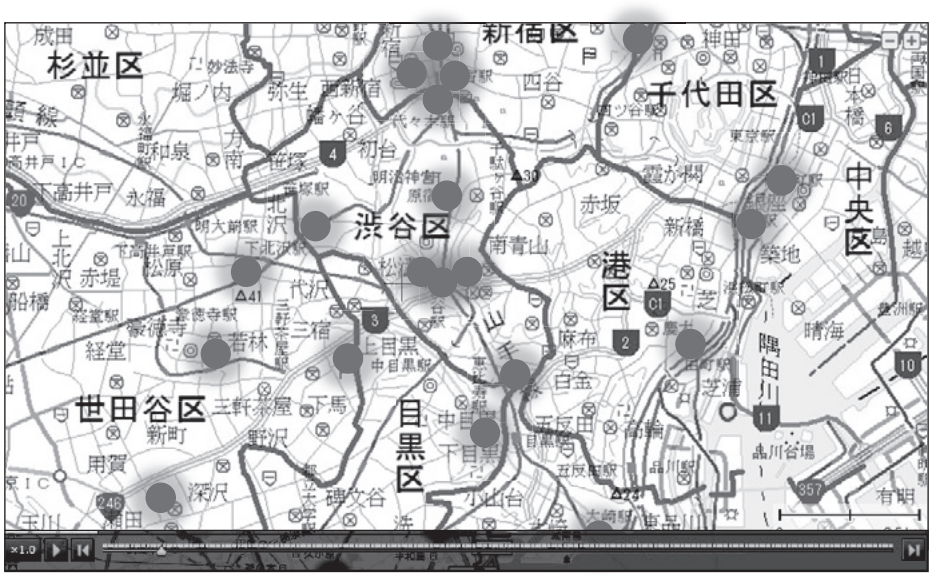


Figure 1
Example of Sentiment Analysis.



The data including the background map have been provided by the Digital Japan Web System of the Geospatial Information Authority of Japan.

Figure 2
Example of Social Event Visualization.

advanced social media analysis techniques are helping us to interpret social behavior patterns and provide beneficial new products, services and proactive support mechanisms for local communities. For example, by using Twitter data, this technology can detect crime-related information and automatically show hotspots as a clearly understandable map (Figure 2). This technology can be applied to various fields to gather residents' opinions and behavior and people's actions in society.

To realize Sentiment Analysis and Social Event Visualization as described above, Natural Language Processing is the most important technology to extract valuable information from individual entries. However, entries in social media do not have the 5W1H (when, where, who, what, why and how) clearly described but mostly use spoken expressions and often use abbreviations. This generates a need for more robust Natural Language Processing rather than processing based on traditional techniques such as Bag of Words.

To address this problem, we have been developing robust Natural Language Processing technologies based on its original Machine Learning technique.[2] Machine Learning technique can automatically generate a set of classification rules for classifying unlabeled data from a set of labeled data, or so called "training data." This allows more sophisticated classification based on the semantic of contents. For example, to see if a certain entry is information on a suspicious individual, data is not classified simply on whether or not it contains the phrase "suspicious individual," but phrases related to "suspicious individual" or context can be taken into account. An entry "I may have been mistaken for a suspicious individual..." contains the phrase "suspicious individual" but is classified as not being "information on a suspicious individual."

Extraction of information on "where" from the entry also poses a problem: various expressions other than name of places or addresses may be used to indicate locations such as store names and landmarks and their abbreviations. Since it is not practical to prepare training data for all location expressions, we have subjected large amounts of text data to statistical processing to use as parameters of judgment rules for reinforcement, thereby developing highly accurate extraction technology capable of handling various expressions.[3]

In this way, Social Event Visualization using robust natural language processing allows the detection and grasping, although still fragmentary, of "events that occurred in society."

Concerning network analysis technologies intended for social media, social media can be seen as massive networks built by people and containing posts and topics. That means there is a need for technologies to analyze large networks, especially technologies capable of promptly analyzing network structures and their temporal changes. We have cooperated with Carnegie Mellon University to work on a technique to quickly detect characteristic subnetworks in a large network.[4] This technique is capable of conducting analysis in a time proportional to the size of a network (the number of sides) by focusing on the distribution of feature quantities such as node degree and hub score. This is an essential requirement for analyzing extremely large networks such as social media. By using this technology, characteristic communities and topics in social media can be extracted, and they may be used for detecting changes in awareness in society and changes in behavior based on them.

In the future, we intend to develop technologies that can extract more abundant and accurate knowledge by expanding analyzable fields and combining technologies with spatiotemporal data processing technologies, which will be described later.

3. Dynamic optimization technology

In this section we briefly explain the direction of optimization technologies required to achieve Intelligent Society and also mention our activities on optimization.

Optimization technologies have come to be utilized in many fields including logistics and manufacturing in line with the drastic improvement in the computational performance of computers and progress of efficient algorithms. Recently, optimization technologies have been attracting much attention again: they are seen as fundamental technologies for finding valuable knowledge and designing the best action plan out of data that have enormously increased in amount. Such increase has accompanied the development of the Internet such as Web information, social media and sensor data.

We have been engaged in developing optimization technologies and conducting research on their practical applications in various fields including logistics and manufacturing.[5] One of them is an optimization technology that uses a parameter space approach and it aims to improve design performance and efficiency in design policy decision-making under more complicated constraint conditions. This optimization technology allows solution properties such as optimality and robustness to be easily grasped by capturing and visualizing all possible design parameter regions (exactly) that satisfy the design specifications. It has been realized by applying symbolic and algebraic computation technologies developed by Fujitsu Laboratories (**Figure 3**). The optimization technology has been applied to some manufacturing design process (such as hard disks and semiconductor memory), and has reduced the period of a design process for a hard disk from 14 days to 1 day.

For optimization problems appearing in community energy management and market quality management targeted by Intelligent Society, new challenges must be addressed in addition to the optimization technologies mentioned above. For example, community energy management involves predicting electric power demand at homes and in the office, and energy output by photovoltaic and other power generation means. Then, the optimum supply and demand control plans are created based on such predicted results so as to balance electric power supply and demand. However, it is impossible to perfectly observe or predict people's actions and the prediction itself includes considerable uncertainty. Optimization of planning on supply and demand must take this uncertainty into account. In addition, new data are constantly observed regarding the actual power consumption at homes, offices, and photovoltaic cell output, so supply and demand

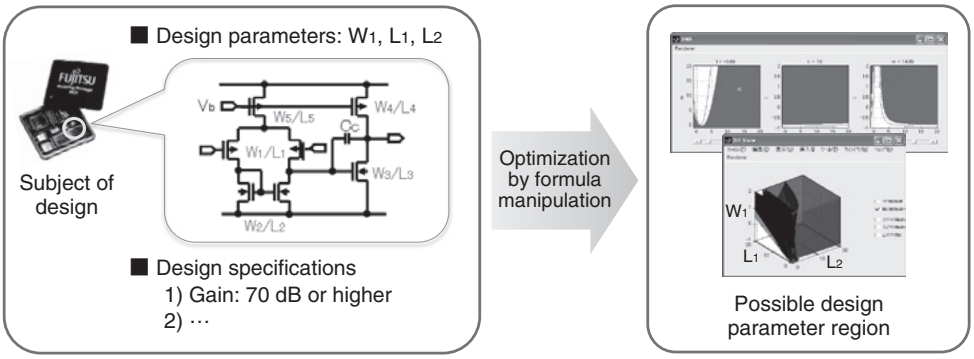


Figure 3
Optimization by symbolic computation.

plans must be changed accordingly.

We are proceeding with research on "dynamic" optimization to address these two challenges, i.e., achieving appropriate responses to uncertainty and changes in the situation. To deal with uncertainty in prediction, the focus is placed on two approaches to conducting research: robust optimization and probabilistic optimization. Robust optimization is optimization in which the range of uncertainty is estimated and specified in advance and the worst-case scenario is assumed. Probabilistic optimization is based on prediction models that include probabilistic factors. In the example of energy management, the former addresses uncertainty by incorporating the range of variation of the demand and output in the optimization problem. To respond to a range of variation that could be extremely large, the latter makes multiple predictions with probabilities rather than adopting a single demand or output prediction. It then incorporates the predictions into the optimization problem so as to formulate supply and demand plans. Regarding response to changes in the situation, we aim to establish a framework of circulating optimization, which allows changes in the situation to be dynamically reflected in problem setting and efficient optimum planning decision each time.

4. Spatiotemporal data processing technology

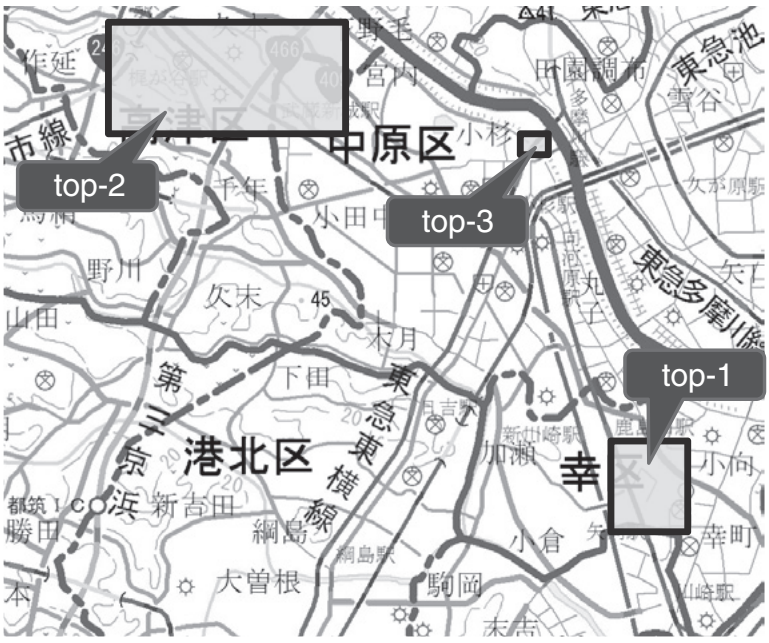
Recently, massive amounts of data from various devices including mobile phones, cameras and car navigation systems have become available for real-time collection. These sensor data include information on the times and locations of observation in addition to observed values. Processing this information allows immediate recognition of the actions of people and society with unprecedented accuracy. Finding out the actions of people and society is an element essential to various solutions targeted by Intelligent Society. For that reason, spatiotemporal data processing for accumulating large amounts of data on time and space such as sensor data to retrieve and aggregate data required for analysis is fundamental technology that forms the basis of Intelligent Society. As technologies to support spatiotemporal data processing, the following presents the technologies for discovering optimal areas with respect to score and pattern matching over compressed data.

One typical method of handling spatial information is to divide space into a mesh for aggregation and analysis. For example, demographic statistics analysis using an administrative district mesh and precipitation data aggregation using a fixed-length mesh have the benefits of simplicity, ease of use and easy-to-understand results. However, aggregating the same data may produce greatly differing results depending on how the mesh is segmented. Our approach to spatiotemporal analysis, on the other hand, is to discover optimum areas in the sense that they have highest aggregated values (scores) based on individual location data. For that reason, the need to specify the mesh is eliminated and the same data always produce the same results. In relation to this technology, research on algorithms for finding one region with the maximum score such as density and probability has been conducted up to now. In real-world applications, however, more than one region often needs to be detected when there is more than one region to pay attention to. To address this issue, we have developed SplitRegionSearch algorithm, which promptly detects multiple regions that do not cross each other in the order of score.[6] This technique offers tens to hundreds of times higher speed compared with an existing technique naively extended to detect multiple regions, and the effect increases as the number of data to be covered increases. For example, in a problem to find areas with the highest population growth that extend over multiple towns based on the demographic statistics of Kawasaki City, the top

three regions have been successfully extracted at a speed more than 500 times higher compared with the existing techniques (**Figure 4**). At present, we are working on algorithms capable of handling regions with complicated shapes including unevenness and various scores.

Pattern matching over compressed data is, as the name indicates, a technology for searching a large amount of compressed data as they are without decompressing them. The amount of sensor data keeps growing continuously and compressing them so as to reduce their size is preferable when storing or transferring them. However, compressed data must be decompressed before they can be processed and the costs relating to the computational time and memory required for decompression pose a problem. Pattern matching over compressed data, which achieves both data size reduction by compression and high-speed retrieval, may solve this problem. To that end, we need to construct dictionaries for compression to smaller sizes and be able to conduct faster searching of compressed data. The following describes the former, the construction of dictionaries.

The pattern matching technology over compressed data[7] developed by us jointly with Hokkaido University uses a compression method called VF coding. VF coding uses a pregenerated dictionary to replace variable-length character strings with fixed-length codes for compression (**Figure 5**). For that reason, the compression ratio depends on which character strings are assigned to the limited number of codes, or how the dictionary is built. To realize a high compression ratio, the dictionary must be constructed in view of the occurrences of character strings in data before compression. With the existing technologies, however, all of data before compression must be stored in the memory for optimizing the dictionary, which could not be applied to large-scale files. In



The data including the background map have been provided by the Digital Japan Web System of the Geospatial Information Authority of Japan.

Figure 4
Discovery of optimal areas with respect to score function
(e.g., regions with highest population growth in Kawasaki).

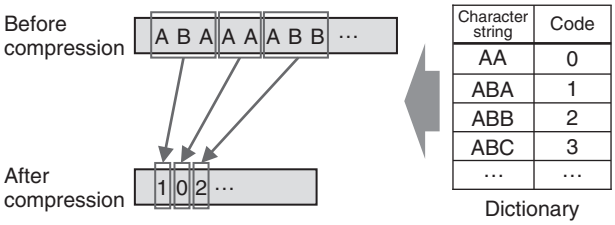


Figure 5
Text compression by VF coding.

contrast, with our technology, the features of data are learned by scanning files several times to realize a compression ratio comparable to that of a typical compression tool gzip without storing all data in the memory, or without limit to the file size.

5. Conclusion

This paper presented three analysis technologies for realizing Intelligent Society: social media analytics, dynamic optimization and spatiotemporal data processing technology. We intend to proceed with R&D so as to establish these technologies as the basis of Intelligent Society through their application to various fields including energy management, traffic, lifelog and risk management.

References

1) NIFTY Corporation. *http://www.nifty.co.jp/english/index.htm*
2) T. Iwakura et al.: An AdaBoost Using a Weak-Learner Generating Several Weak-Hypotheses for Large Training Data of Natural Language Processing. (in Japanese), *Transactions of the Institute of Electrical Engineers of Japan. C*, Vol. 130. No. 1, pp. 83–91 (2010).
3) T. Iwakura et al.: Japanese Named Entity Extraction by Augmenting Features with Unlabeled Data. *IPSJ Journal*, Vol. 49, No. 10, pp. 3657–3669 (2008).
4) K. Maruhashi et al.: Spotting Connection Patterns and Outliers in Large Graphs. (in Japanese), ICDM Workshops 2010, 2010, pp. 1328–1337.
5) H. Anai et al.: Design Technology Based on Symbolic Computation. (in Japanese), *FUJITSU*, Vol. 60, No. 5, pp. 514–521 (2009).
6) H. Morikawa et al.: Effective Method of Detection of the Optimum Region Set out of Large-scale Spatial Data. (in Japanese), Proceedings of the 73rd National Convention of IPSJ, Vol. 1, 2011, pp. 561–562.
7) T. Uemura et al.: Training Parse Trees for Efficient VF Coding. 17th edition of the Symposium on String Processing and Information Retrieval (SPIRE 2010), LNCS 6393, 2010, pp. 179–184.

Nobuhiro Yugami
Fujitsu Laboratories Ltd.
Dr. Yugami is currently engaged in research on data mining and optimization technologies.

Hirokazu Anai
Fujitsu Laboratories Ltd.
Dr. Anai is currently engaged in research on symbolic-numeric hybrid computation, optimization and control theory.

Nobuyuki Igata
Fujitsu Laboratories Ltd.
Mr. Igata is currently engaged in research on social media analysis by natural language processing.

Hiroya Inakoshi
Fujitsu Laboratories Ltd.
Mr. Inakoshi is currently engaged in research on big data processing technologies including character string matching and spatiotemporal data processing.