Editorial

Special Issue "MoDAT: Designing the Market of Data"

Yukio Ohsawa

School of Engineering, The University of Tokyo, Tokyo 113-8653, Japan; ohsawa@sys.t.u-tokyo.ac.jp

check for
updates

The fifth International Workshop on the Market of Data (MoDAT2017) was held on November 18th, 2017 in New Orleans, USA in conjunction with IEEE ICDM 2017. This was the last of the workshop series of MoDAT, which lasted 5 years—the first was in Dallas, USA, in 2013. A total of 38 presentations were selected out of more than 100 submissions in total by our peer review system. In the MoDAT workshop series, we discussed technical and philosophical issues for designing, creating, or improving markets where data are reasonably dealt with (i.e., sold reflecting the potential utility, freely distributed, shared after negotiations, and so on), fitting the demand-supply balance. The papers submitted and/or presented so far in the MoDAT workshops have included, but are not limited to, the following:

- Visualization of links among data, to aid the participants of the market of data in creating use scenarios of data by combining data from various domains.
- Visualization of links among data and AI tools, to aid in creating use scenarios of AI.
- Data/text mining as a pre-process step for the visualizations above.
- Data/text mining for detecting important events and concepts, in order to evaluate the importance of the requirements of data users.
- Data/text mining for finding important essential requirements of data users, in order to provide satisfactory data matching such requirements.
- Web mining for exploring information as elements for creating useful data.
- Extracting causalities/scenarios of events and actions in the data, to show the value of the data.
- Similarity modelling among data, to aid the analogical reasoning of data scientists in the market.
- Representation of the hierarchical knowledge of relevance among requirements, data usage, and variables in data, used in the communications and thought of analysts and users.
- Construction of dictionaries of variables, showing possible data connections through variables, such as time, place, human ID, and so on, and possible connections of these variables to create new data.
- Data-based communication for evaluating the value of an event (i.e., chance discovery) and data which may include such an event.
- Methods for aiding creative thought, communication, and debate.
- Visual interfaces for showing the value of data to trigger meaningful thoughts in stakeholders.

With the reference of these past workshop outcomes, we should facilitate more effective and efficient shared mechanisms in the Market of Data to move forward; that is, not only within a certain area of interest, but also across data domains, functions, and applications to yield new opportunities for integrating different applications.

In this Special Issue from MDPI, we intend to appeal to the activities of MoDAT and its future extensions by presenting a carefully selected selection of revisions of the best papers from past MoDAT workshops. Here, authors from previous MoDAT workshops were requested to extend their papers in their submissions to this journal issue by showing technical extensions, in-depth discussions and

evaluations, or additional high-impact use cases of the technologies they proposed. Submissions new to conferences or workshops, so far, were also encouraged. We had eight submissions, as a result, for this special issue. All the submitted papers were processed by the standard peer-review procedure of MDPI, where reviewers were fairly selected with the advice of the guest editors, with very careful consideration by the MDPI editorial team to avoid the influence of any personal interest from the guest editors. As a result, the five papers below were accepted and published.

In Predict Electric Power Demand with Extended Goal Graph and Heterogeneous Mixture Modeling (by Noriyuki Kushiro, Ami Fukuda, Masatada Kawatsu, and Toshihiro Mega), a method for predicting energy demand by hourly consumption data is presented for realizing an energy management system for buildings. The method combines (1) data separation, (2) a linear regression model for each partition on the heterogeneous mixture of models, and (3) an extended goal graph to extract useful variables for data partitioning and for linear regression. The method was applied to energy prediction given two years' worth of hourly consumption data for a building and was validated experimentally. This work is contributive to MoDAT in that (3) is a method to logically relate social requirements to the used variables in data to use in (1) and (2), where computational models are connected to meet the requirement for the prediction.

Tangled String for Multi-Timescale Explanation of Changes in Stock Market (Yukio Ohsawa, Teruaki Hayashi, and Takaaki Yoshino) addresses the question of explaining changes in desired timescales of the stock market. Tangled string, a sequence visualization tool wherein a sequence is compared to a string, on which trends and periods between trends are shown, is extended and applied to detect the stocks which trigger changes and explain trend shifts. From 11 years of data from the First Section of the Tokyo Stock Exchange, the authors found that the change points obtained by the tangled string coincided well with changes in the average prices of listed stocks, and the changes in stock prices came to be explained by stock analysts. The tangled string was created using a data-driven innovation platform called Innovators Marketplace on Data Jackets (IMDJ), a method for designing a practical market of data for triggering innovation, and is extended to aid data users here.

Estimating Spatiotemporal Information from Behavioral Sensing Data of Wheelchair Users by Machine Learning Technologies (by Ikuko Eguchi Yairi, Hiroki Takahashi, Takumi Watanabe, Kouya Nagamine, Yusuke Fukushima, Yutaka Matsuo, and Yusuke Iwasawa) introduces a new methodology to estimate road accessibility from the acceleration data collected by users by a smartphone attached onto a wheelchair seat, in order to realize a system to provide road accessibility visualization services to users by pattern matching, while gradually learning to improve service accuracy on the deep convolutional neural network. Here, a CNN learns the state of the road surface from the acceleration data. The results show that the features can capture the difference of the road surface conditions in more detail than the label attached by authors and are effective as the means for quantitatively expressing the conditions. This paper developed and evaluated a prototype system that estimated types of ground surfaces by focusing on knowledge extraction and visualization. The impact of this paper is that the data are designed from the aspect of service improvement with the efforts of humans working on and using the service.

Matrix-Based Method for Inferring Elements in Data Attributes Using a Vector Space Model (by Teruaki Hayashi and Yukio Ohsawa) addresses the task of inferring elements in the attributes of desired data. So far, it has been difficult to obtain data that accurately correspond to users' real requirements, because the users might not express their objects of interest using the exact terms (variables, outlines of data, and so on) used in the data or metadata for two reasons. The first is that the latent interest of data users is not easy to verbalize; the second is that the vocabulary of the user typically differs from that of the data creators. In this paper, the authors propose a method to enable useful elements of data (type, format, and variable given in data jackets) to be inferred for a free-text query. The experimental results indicate the proposed method outperforms those obtained from string matching and word embedding. The impact of this work is that participants of the market of data can get desired data for developing

the interest of users or find vocabulary gaps between segments of participants; that is, the providers of data or metadata (including data jackets) and data users.

In Related Stocks Selection with Data Collaboration Using Text Mining (by Masanori Hirano, Hiroki Sakaji, Shoko Kimura, Kiyoshi Izumi, Hiroyasu Matsushima, Shintaro Nagao, and Atsuo Kato), the authors propose an extended scheme for selecting related stocks for themed mutual funds. In the preliminary experiments, building a themed mutual fund was found to be quite difficult. This scheme is a type of natural language processing, based on words extracted using their similarity to a theme by word2vec and the original similarity measure on co-occurrences in the information about companies. They used data such as investor relations and official websites as company information data. The scheme achieved higher accuracy than a standard method. Finally, the possibility is shown that official websites are not necessary for the proposed new scheme. This point means "trendy" data, of which the utility tends to be trusted, may be replaced by data by or for experts if they that are selected on the requirements of analysts, who are an essential participant in the market of data.

All in all, in these papers, we find reasons for choosing or creating data and/or tools for fulfilling social requirements. This can be positioned as the pre-processing of all kinds of data-use processes, regardless of whatever methods or tools we may employ in the "main" process. In this sense, the market of data should be regarded as the root of the big tree of data science and data engineering, with positioning algorithms and AI technologies at the leaves. So, we are calling. and will continue to call, for the attention shift of data owners and data users to the root.

Let me also confess, here, that there were other very interesting submissions, which were, unfortunately, not accepted for this special issue. They showed true cutting-edge insight towards the re-design of the market of data into a platform of data-interactive innovations. This accident occurred in this special issue, where the reviewers were selected not by guest editors who are leading experts of the design of data market, but in the manner of the scientific journal, where criteria used in computer science, so far, for evaluating results or theories were strictly applied. For example, some contributions to the creativity of people in the market of data, from cognitive scientific viewpoints, tend to be rejected in this manner. However, we sincerely appreciate the editorial team of the *Information* journal, because a strict cross-disciplinary review trains a rising domain. The papers detailed above, which are now finally published, came to be reinforced as a result of facing the requirements of scientific society as well as those in the market of data. In the next step, we shall reinforce MoDAT to be a further strengthened scientific field of thought and communication. The result should appear in the forthcoming special issue on CDEC, Cross-disciplinary Data Exchange and Collaboration; again, from the journal *Information*.

Conflicts of Interest: The author declares no conflict of interest.