SBD - Social and Big Data

Special Track in association with MMEDIA 2017, April 23-27, 2017 – Venice, Italy
https://www.iaria.org/conferences2017/MMEDIA17.html

Hiroshi Ishikawa
Faculty of System Design
Tokyo Metropolitan University
Tokyo, Japan
e-mail: ishikawa-hiroshi@tmu.ac.jp

***Abstract*—Analyzing both physical real world data and social data by relating them to each other, is called social big data science/mining or social big data (SBD) for short In a special track on Social and Big Data, held as a part of MMEDIA 2017 conference in Venice, Italy, seven papers were presented ranging from conceptual model, data mining, and data management to applications in SBD.**

Keywords*- *social data; big data; science data; data model; data management; data mining.

I. SOCIAL BIG DATA

In the present age, large amounts of data are produced continuously in science, on the internet, and in physical systems. Such data are collectively called data deluge. According to researches carried out by International Data Corporation, or IDC for short [1], the size of data which are generated and reproduced all over the world every year is estimated to be 161 Exa bytes. The total amount of data produced in 2011 exceeded 10 or more times the storage capacity of the storage media available in that year.

Experts in scientific and engineering fields produce a large amount of data by observing and analyzing the target phenomena. Even ordinary people voluntarily post a vast amount of data via various social media on the internet. Furthermore, people unconsciously produce data via various actions detected by physical systems in the real world. It is expected that such data can generate various values.

In the above-mentioned research report of IDC, data produced in science, the internet, and in physical systems are collectively called big data. The features of big data can be summarized as follows:

- The quantity (Volume) of data is extraordinary, as the name denotes.
- The kinds (Variety) of data have expanded into unstructured texts, semi-structured data such as XML, and graphs (i.e., networks).
- As is often the case with Twitter and sensor data streams, the speed (Velocity) at which data are generated is very high.

Therefore, big data is often characterized as V^3 by taking the initial letters of these three terms Volume, Variety, and Velocity. Big data are expected to create not only knowledge in science but also derive values in various commercial ventures. "Variety" implies that big data appear in a wide variety of applications. Big data inherently contain "vagueness" such as inconsistency and deficiency. Such vagueness must be resolved in order to obtain quality analysis results.

Based on the origins of where big data are produced, they can be roughly classified into physical real world data (i.e., heterogeneous data such as science data, event data, and transportation data) and social data (i.e., social media data such as Twitter articles and Flickr photos). Most of the physical real world data are generated by customers who leave their behavioral logs in the information systems. For example, data about the customers' check-in and check-out are inserted into the databases in the transportation management systems through their IC cards. Data about the customers' use of facilities are also stored in the facility management databases. Further, the customers' behaviors are recorded as sensor data and video data. In other words, real world physical data mostly contain only latent or implicit semantics because the customers are unconscious of their data being collected.

On the other hand, the customers consciously record their behaviors in the physical real world as social data on their own. For example, they post photos and videos, taken during events or trips, to sharing services and post various information (e.g., actions and sentiments) about the events or trips to microblogs. In a word, unlike physical real world data, social data contain explicit semantics because the customers voluntarily create the data.

Furthermore, there are bidirectional interactions between the physical real world data and social data through users. That is, if one direction of such interactions is focused on, it will be observed that events which produce physical real world data affect the users and make them describe the events in social data. Moreover, if attention is paid to the reverse direction of such interactions, it will turn out that the contents of social data affect other users' actions (e.g., consumer behaviors), which, in turn, produce new physical real world data. If such interactions can be analyzed in an integrated fashion, it is possible to apply the results to a wide range of application domains including business and science. That is, if interactions are analyzed paying attention to the direction from physical real world data to social data, for example, the following can be accomplished.

- Measurement of effectiveness of marketing such as promotions of new products

- Discovery of reasons for sudden increase in product sales
- Awareness of need of measures against problems about products or services

Moreover, the following may be predicted if interactions are analyzed paying attention to the reverse direction of such interactions.

- Customer behaviors of the future
- Latent customer demands

Analyzing both physical real world data and social data by relating them to each other based on common information, such as space and time, is called social big data science/mining or social big data for short [2]. To the knowledge of the author, there are few modeling frameworks which allow the end user or analyst to describe hypotheses spanning across data mining, quantitative analysis, and qualitative analysis. In other words, conceptualizing social big data [3] is particularly required which allows the user to describe hypotheses for social big data in an integrated manner at the conceptual layer and translate them for execution by existing techniques such as multivariate analysis and data mining at the logical layer if needed.

In general, a database management system, which is often used to store target data for mining, consists of three layers: the conceptual, logical, and physical layers. Following the three-layered architecture of the database management system, the reference architecture of the integrated system for social big data science can be depicted in the following way. At the conceptual layer, the system allows the user (i.e., analyst) to describe integrated hypotheses relating to social big data. At the logical layer, the system converts the hypotheses defined at the conceptual layer in order for the user to actually confirm them by applying individual techniques such as data mining and multivariate analysis. At the physical layer, the system performs further analysis efficiently by using both software and hardware frameworks, such as parallel distributed processing.

II. TRACK OVERVIEW

In a special track on Social and Big Data, held as a part of MMEDIA 2017 conference in Venice, Italy [4], seven papers were presented ranging from conceptual model, data mining, and data management to SBD applications.

Ishikawa et al [5] propose a conceptual model for integrated analysis of SBD. Endo et al [6] discuss a method for estimating best-time of enjoying cherry blossoms using Tweets and Open Data as an application of SBD. Kikuchi et al [7] and Kato et al [8] apply and evaluate methods for mining science big data. Tanaka et al [9] propose a precise cost estimation method for efficient query processing of data management essential in SBD. Hayashi et al [10] discuss an application of behavior analysis of users at wedding community sites. Kusu et al [11] propose an efficient method for dynamic dependency analysis.

REFERENCES

[1] IDC, *The Digital Universe In 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East* (2012). http://www.emc.com/leadership/digital-universe/iview/index.htm accessed 2017.03

[2] H. Ishikawa, Social Big Data Mining, CRC Press, 2015.

[3] E. Olshannikova, T. Olsson,J. Huhtamäki and H. Kärkkäinen, “Conceptualizing Big Social Data,” Journal of Big Data, Springer, vol.4, no.3, 2017.

[4] MMEDIA 2017, The The Ninth International Conferences on Advances in Multimedia, April 23-27, 2017, Venice, Italy, https://www.iaria.org/conferences2017/MMEDIA17.html

[5] H. Ishikawa, and R. Chbeir, “A data model for integrating data management and data mining in social big data,” A Special Track on Social and Big Data (SBD2017) in association with MMEDIA2017, April 23-27, 2017, Venice, Italy, https://www.iaria.org/conferences2017/MMEDIA17.html

[6] M. Endo, S. Ohno, M. Hirota, Y. Shoji, and H. Ishikawa, “Examination of best-time estimation using interpolation for geotagged tweets,” A Special Track on Social and Big Data (SBD2017) in association with MMEDIA2017, April 23-27, 2017, Venice, Italy, https://www.iaria.org/conferences2017/MMEDIA17.html

[7] S. Kikuchi, R. Yamada, Y. Yamamoto, M. Hirota, S. Yokoyama, and H. Ishikawa, “Classification of Unlabeled Deep Moonquakes Using Machine Learning,” A Special Track on Social and Big Data (SBD2017) in association with MMEDIA2017, April 23-27, 2017, Venice, Italy, https://www.iaria.org/conferences2017/MMEDIA17.html

[8] K. Kato, R. Yamada, Y. Yamamoto, M. Hirota, S. Yokoyama, and H. Ishikawa, “Analysis of spatial and temporal features to classify the deep moonquake sources using balanced random forest,” A Special Track on Social and Big Data (SBD2017) in association with MMEDIA2017, April 23-27, 2017, Venice, Italy, https://www.iaria.org/conferences2017/MMEDIA17.html

[9] T. Tanaka, and H.Ishikawa, “Measurement-based Cost Estimation Method of a Join Operation for an In-Memory Database,” A Special Track on Social and Big Data (SBD2017) in association with MMEDIA2017, April 23-27, 2017, Venice, Italy, https://www.iaria.org/conferences2017/MMEDIA17.html

[10] T. Hayashi, Y. Wang, Y. Kawai, and K. Sumiya, “A Proposal of Activation Mechanism for User Communication based on User Behavior analysis on Wedding Community Sites,” A Special Track on Social and Big Data (SBD2017) in association with MMEDIA2017, April 23-27, 2017, Venice, Italy, https://www.iaria.org/conferences2017/MMEDIA17.html

[11] K. Kusu, I. Kume, and K. Hatano, “A Node Access Frequency based Graph Partitioning Technique for Efficient Dynamic Dependency Analysis,” A Special Track on Social and Big Data (SBD2017) in association with MMEDIA2017, April 23-27, 2017, Venice, Italy, https://www.iaria.org/conferences2017/MMEDIA17.html