Interpolating Lost Spatio-Temporal Data by Web Sensors

Shun Hattori
Web Intelligence Time-Space (WITS) Laboratory
College of Information and Systems, Graduate School of Engineering
Muroran Institute of Technology
27–1 Mizumoto-cho, Muroran, Hokkaido 050–8585, Japan
E-mail: hattori@csse.muroran-it.ac.jp

***Abstract*—We experience various phenomena (e.g., rain, snow, and earthquake) in the physical world, while we carry out various actions (e.g., posting, querying, and e-shopping) in the Web world. Many researches have tried to mine the Web for knowledge about various phenomena in the physical world, and also several Web services using Web-mined knowledge have been made available for the public. Meanwhile, the previous papers have introduced various kinds of "Web Sensors" with Temporal Shift, Temporal Propagation, and Geospatial Propagation to sense the Web for knowledge about a targeted physical phenomenon, i.e., to extract its spatiotemporal data sensitively by analyzing big data on the Web (e.g., Web documents, Web query logs, and e-shopping logs), and compared them based on their correlation coefficients with Japan Meteorological Agency's physically-sensed spatiotemporal statistics to ensure the accuracy of Web-sensed spatiotemporal data sufficiently. As an industrial application of Web Sensors to a problem of the loss or error of physically-sensed spatiotemporal data due to some sort of troubles (e.g., temporary faults of JMA's observatories), this paper tries to enable Web Sensors to interpolate lost spatiotemporal data of physical statistics by regression analysis.**

***Keywords*-Spatiotemporal Data Mining; Big Data Analysis;**

I. Introduction

We experience or forecast various phenomena (e.g., rainfall, snowfall, earthquake, influenza, and traffic accident) in the physical world, while we carry out various actions (e.g., posting, querying, and e-shopping) in the Web world. Recently, there have been many researches to mine a huge amount of various documents in the exploding Web world, especially User Generated Content such as blogs, microblogs (e.g., Twitter), Word-of-Mouth sites, and Social Networking Services (e.g., Facebook), for knowledge about various phenomena and events in the physical world. For instance, opinion and reputation extraction [1], [2] of various products and services in the physical world, experience mining [3], [4] of various phenomena and events in the physical world, concept hierarchy (semantics) extraction [5], [6], [7], [8], [9], [10] such as is-a/has-a relationships, and visual appearance (look and feel) extraction [10], [11], [12], [13], [14], [15] of physical objects in the physical world. Meanwhile, Web services using Web-mined knowledge have been made available for the public, and more and more ordinary people actually utilize them as important information for choosing better products, services, and actions in the physical world.

However, there are not enough investigations [16], [17], [18] on how accurately Web-mined data about a targeted phenomenon or event in the physical world reflect physical-world data. It is not so difficult to mine the Web for some kind of potential knowledge data by using various text mining techniques, and it might not be problematic only to enjoy browsing the Web-mined knowledge data. But while choosing better products, services, and actions in the physical world, it must be socially-problematic to idolatrously/immoderately utilize the Web-mined data in public Web services without ensuring their accuracy sufficiently.

The previous papers [19], [20], [21], [22], [23], [24], [25], [26], [27] have introduced various kinds of "Web Sensors" to sense the Web for knowledge about a targeted phenomenon (e.g., rainfall, snowfall, and earthquake) in the physical world, i.e., to extract its spatiotemporal numerical values by analyzing big data on the Web, i.e., various action-based data (e.g., Web documents, Web query logs, and e-shopping logs) in the Web world, and investigated how correlated Web-sensed spatiotemporal data are with physically-sensed spatiotemporal data (e.g., Japan Meteorological Agency's rainfall, snowfall, and earthquake statistics [28]) as shown in Figure 1.

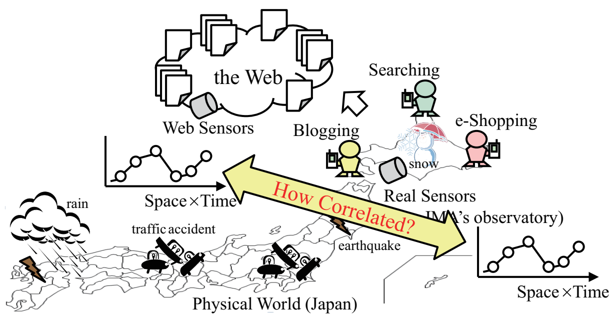

Figure 1. Can Web Sensors sense the physical world sensitively?

Document-based Web Sensors with "Temporal Shift" [20], [25] showed that

1) The optimized temporal shift parameter δ of Web Sensors depends on physical phenomena: Not-Shifted Web Sensor whose temporal shift parameter δ is ± 0 gives the highest correlation coefficient (i.e., the Web runs parallel to the physical world)

978-1-5090-4119-0/16 $31.00 © 2016 IEEE
DOI 10.1109/EUROSIM.2016.89

for rainfall, Shifted-to-Future Web Sensor whose temporal shift parameter δ is negative gives the highest correlation coefficient (i.e., the Web leads the physical world) for snowfall, and Shifted-to-Past Web Sensor whose temporal shift parameter δ is positive gives the highest correlation coefficient (i.e., the Web follows the physical world) for earthquake,

2) The optimized temporal shift parameter δ and correlation coefficient for rainfall are not much dependent on geographical spaces (e.g., 47 prefectures in Japan) and time periods, while the optimized temporal shift parameter δ for snowfall and earthquake varies more widely, and
3) More shaken geographical spaces and time periods are given higher correlation coefficient between Web-sensed spatiotemporal data and physically-sensed spatiotemporal data by the Great East Japan Earthquake (3.11).

Query-based Web Sensors using Web search query logs [24] are superior to Document-based Web Sensors using Web documents such as blogs for snowfall and earthquake, while Query-based Web Sensors are inferior to Document-based Web Sensors for rainfall. In addition, the best combined Web Sensor using both Web search query logs and Web documents is superior to uncombined Web Sensors using only Web search query logs or Web documents.

This paper introduces a novel method to interpolate the loss of physically-sensed spatiotemporal data about a targeted physical phenomenon (e.g., Japan Meteorological Agency's rainfall, snowfall, and earthquake statistics) by regression analysis between physically-sensed spatiotemporal data and Web-sensed spatiotemporal data about the targeted physical phenomenon, as an industrial application of variously defined "Web Sensors" with Temporal Shift, Temporal Propagation, and Geospatial Propagation to sense the Web for knowledge about a targeted physical phenomenon, i.e., to extract its spatiotemporal data sensitively by analyzing big data on the Web (e.g., Web documents, Web queries, and e-shopping logs).

The rest of this paper is organized as follows. Section II shows various definitions of Web Sensors, and Section III introduces a novel method of interpolating lost spatiotemporal data of physical statistics by Web Sensors and regression analysis. And Section IV concludes this paper.

II. Web Sensors

This section shows various definitions of Web Sensors with Temporal Shift, Temporal Propagation, and Geospatial Propagation to sense the Web for spatiotemporal numerical values dependent on a geographic space (e.g., one of 47 prefectures in Japan) and a time period (e.g., days and weeks in 2011) about a physical phenomenon (e.g., rainfall, snowfall, and earthquake).

First, the simplest and spatiotemporally-normalized Web Sensor [19], [23] by using only Web documents (not Web search query logs [24]) with a linguistic name of a geographic space s, e.g., one of 47 prefectures in Japan such as "Hokkaido," a time period t, e.g., one of 365 days or 52 weeks in 2011 such as January 1st (1st day) or from January 1st to 7th (1st week) and from December 24th to 30th (52nd week), and a linguistic keyword kw representing a targeted physical phenomenon, e.g., "rain," "snow," and "earthquake," is defined as

$$\mathrm{ws}(kw, s, t) := \frac{\mathrm{df}_t(\texttt{["}kw\texttt{" AND "}s\texttt{"]})}{\mathrm{df}_t(\texttt{["}s\texttt{"]})}, \tag{1}$$

where $\mathrm{df}_t(\texttt{["}s\texttt{"]})$ stands for the Frequency of Web Documents searched from the Web, especially the Weblog, by submitting the search query q with the custom time range t to Google Web Search. Note that the Weblog is superior to the whole Web, Twitter, Facebook, and News as a corpus of Web Sensors [22].

Secondly, the temporally-shifted Web Sensor [20], [25] with a "Temporal Shift" parameter δ [day], a geographic space s, a time period t, and a linguistic keyword kw representing a targeted physical phenomenon is defined as

$$\text{ws-ts}_\delta(kw, s, t) := \mathrm{ws}(kw, s, t + \delta). \tag{2}$$

As shown in Figure 2, Shifted-to-Past Web Sensor for a targeted physical phenomenon (e.g., earthquake) when its Temporal Shift parameter δ is positive (e.g., +14) calculates a numerical value dependent on a geographic space s (e.g., "Hokkaido" prefecture in Japan) and a time period t (e.g., one of 52 weeks in 2011) by using Web documents uploaded δ day(s) after the time period t (i.e., infers the past from the future), while Shifted-to-Future Web Sensor when its Temporal Shift parameter δ is negative (e.g., -14) calculates a numerical value dependent on a geographic space s and a time period t by using Web documents uploaded $|\delta|$ day(s) before the time period t (i.e., infers the future from the past).

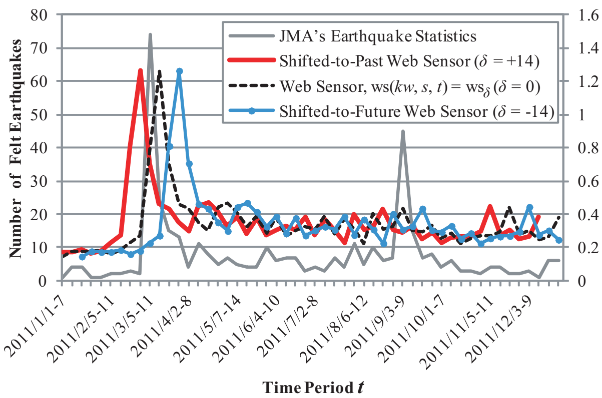

Figure 2. Temporally-shifted Web Sensors for earthquake and JMA's weekly earthquake statistics in Hokkaido prefecture, 2011.

Thirdly, the temporally-propagated Web Sensor [20] with a "Temporal Propagation" parameter σ_t^2, a geographic space s, a time period t, and a linguistic keyword kw representing a physical phenomenon is defined by integrating the surrounding time periods as

$$\text{ws-tp}^{\sigma_t^2}(kw, s, t) := \sum_{\forall\delta} \text{ws-ts}_{\delta}(kw, s, t) \cdot p^{\sigma_t^2}(\delta) \quad (3)$$

$$p^{\sigma_t^2}(\delta) := \frac{1}{\sqrt{2\pi\sigma_t^2}} \cdot \exp\Big(-\frac{\delta^2}{2\sigma_t^2}\Big) \quad (4)$$

where $p^{\sigma_t^2}(\delta)$ stands for a Normal Distribution $N(0, \sigma_t^2, \delta)$ with a mean 0 and a variance σ_t^2. In this paper, $\forall\delta$ is restricted to $[-30, 30]$.

Next, the geospatially-propagated Web Sensor [26], [27] with a "Spatial Propagation" parameter σ_s^2, a geographic space s, a time period t, and a linguistic keyword kw representing a targeted physical phenomenon is defined by integrating the surrounding geographic spaces as

$$\text{ws-sp}^{\sigma_s^2}(kw, s, t) := \sum_{\forall s_i} \text{ws}(kw, s_i, t) \cdot p^{\sigma_s^2}(\text{distance}(s, s_i)) \quad (5)$$

$$p^{\sigma_s^2}(d) := \frac{1}{\sqrt{2\pi\sigma_s^2}} \cdot \exp\Big(-\frac{d^2}{2\sigma_s^2}\Big) \quad (6)$$

where distance(s, s_i) stands for the geographic distance [km] between geographic spaces s and s_i and is calculated based on their latitude and longitude. In this paper, $\forall s_i$ is restricted to 47 prefectures in Japan, and the latitude and longitude of its prefectural capital are used for calculating distance(s, s_i) by using the Survey Calculation API of Geospatial Information Authority of Japan (GSI) [29]. In pairs of 47 prefectures in Japan, the pair of Hokkaido pref. (Sapporo city) and Okinawa pref. (Naha city) has the longest distance, 2243.9 [km], while the pair of Shiga pref. (Otsu city) and Kyoto pref. (Kyoto city) has the shortest distance, 10.5 [km].

III. Data Interpolation

As an industrial application of variously above-defined "Web Sensors" with Temporal Shift, Temporal Propagation, and Geospatial Propagation to the loss or error of physically-sensed spatiotemporal data due to some sort of troubles (e.g., temporary faults of Japan Meteorological Agency's observatories), this section proposes a novel method to interpolate lost spatiotemporal data about a targeted physical phenomenon (e.g., Japan Meteorological Agency's rainfall, snowfall, and earthquake statistics).

For a lost spatiotemporal numerical value ps(s, t, kw) about a targeted physical phenomenon (which is represented by a linguistic keyword kw, e.g., "rain," "snow," and "earthquake") in a geographic space s, e.g., one of 47 prefectures in Japan such as "Hokkaido" over a time period t, e.g., one of 365 days or 52 weeks in 2011 such as January 1st (1st day) or from January 1st to 7th (1st week) and from December 24th to 30th (52nd week), the proposed method interpolates it by regression analysis with its surrounding N physically-sensed spatiotemporal data, their corresponding N Web-sensed spatiotemporal data, and its corresponding Web-sensed spatiotemporal data ws(s, t, kw) or ws-XX(s, t, kw) (where XX $\in$ {"ts," "tp," "sp"}). In this paper, N is restricted to [1,30]. The variety of N physically-sensed spatiotemporal data surrounding a lost physically-sensed spatiotemporal numerical value ps(s, t, kw) has:

1) N physically-sensed spatiotemporal data followed by it (i.e., only N past data),
$$\text{ps}(s, t-N, kw), \cdots, \text{ps}(s, t-1, kw),$$
2) N physically-sensed spatiotemporal data following it (i.e., only N future data),
$$\text{ps}(s, t+1, kw), \cdots, \text{ps}(s, t+N, kw),$$
3) $\lfloor N/2 \rfloor$ physically-sensed spatiotemporal data followed by it and $\lceil N/2 \rceil$ physically-sensed spatiotemporal data following it (i.e., both $\lfloor N/2 \rfloor$ past data and $\lceil N/2 \rceil$ future data, future-preferred when N is odd-numbered),
4) $\lceil N/2 \rceil$ physically-sensed spatiotemporal data followed by it and $\lfloor N/2 \rfloor$ physically-sensed spatiotemporal data following it (i.e., both $\lceil N/2 \rceil$ past data and $\lfloor N/2 \rfloor$ future data, past-preferred when N is odd-numbered).

The generalization of the above-mentioned examples is m ($\in [0, N]$) physically-sensed spatiotemporal data followed by it and $N - m$ physically-sensed spatiotemporal data following it (i.e., m past data and $N - m$ future data) as shown in Figure 3.

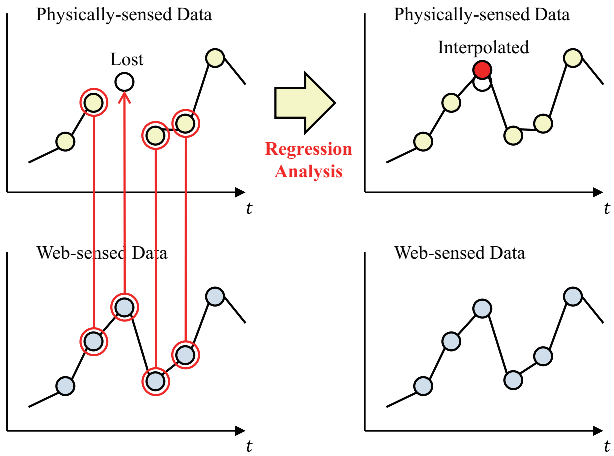

Figure 3. Interpolating a lost physically-sensed datum by Web Sensors and regression analysis using not only physically-sensed data but also Web-sensed data (when $N = 3$ and $m = 1$).

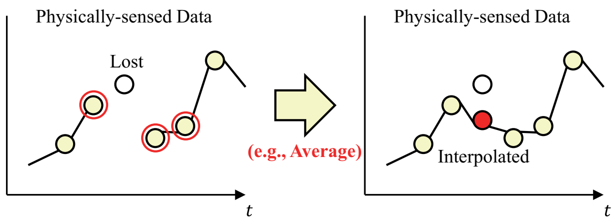

Figure 4. Interpolating a lost physically-sensed datum by average function using only physically-sensed data (adopted as a baseline in the experiment).

IV. Conclusion

This paper has introduced a novel method to interpolate the loss of physically-sensed spatiotemporal data about a targeted physical phenomenon (e.g., Japan Meteorological Agency's rainfall, snowfall, and earthquake statistics) by regression analysis between physically-sensed spatiotemporal data and Web-sensed spatiotemporal data about the targeted physical phenomenon, as an industrial application of variously defined "Web Sensors" with Temporal Shift, Temporal Propagation, and Geospatial Propagation to sense the Web for knowledge about a targeted physical phenomenon, i.e., to extract its spatiotemporal data sensitively by analyzing big data on the Web (e.g., Web documents, Web queries, and e-shopping logs).

The future work has to perform experiments to validate the introduced method of interpolating lost spatiotemporal data of physical statistics by Web Sensors and regression analysis, and also will try to apply the other kinds of physical phenomena to the proposed interpolation. In addition, Web Sensors will be able to forecast future data about a targeted physical phenomenon and to alert falsified data of real statistics.

Acknowledgment

This research project was partially supported by a grant-in-aid for scientific research from the Japan Society for Promotion of Science (15K00329).

References

[1] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," Proceedings of the 12th International World Wide Conference (WWW'03), pp.519–528 (2003).

[2] S. Fujimura, M. Toyoda, and M. Kitsuregawa, "A Reputation Extraction Method Considering Structure of Sentence," Proceedings of the 16th IEICE Data Engineering Workshop (DEWS'05), 6C-i8 (2005).

[3] T. Tezuka, T. Kurashima, and K. Tanaka, "Toward Tighter Integration of Web Search with a Geographic Information System," Proceedings of the 15th International World Wide Web Conference (WWW'06), pp.277–286 (2006).

[4] K. Inui, S. Abe, H. Morita, M. Eguchi, A. Sumida, C. Sao, K. Hara, K. Murakami, and S. Matsuyoshi, "Experience Mining: Building a Large-Scale Database of Personal Experiences and Opinions from Web Documents," Proceedings of the 7th IEEE/WIC/ACM International Conference on Web Intelligence (WI'08), pp.314–321 (2008).

[5] M. A. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," Proceedings of the 14th International Conference on Computational Linguistics (COLING'92), vol.2, pp.539–545 (1992).

[6] M. Ruiz-Casado, E. Alfonseca, and P. Castells, "Automatising the Learning of Lexical Patterns: An Application to the Enrichment of WordNet by Extracting Semantic Relationships from Wikipedia," Data & Knowledge Engineering, vol.61, no.3, pp.484–499 (2007).

[7] S. Hattori, H. Ohshima, S. Oyama, and K. Tanaka, "Mining the Web for Hyponymy Relations based on Property Inheritance," Proceedings of the 10th Asia-Pacific Web Conference (APWeb'08), LNCS vol.4976, pp.99–110 (2008).

[8] S. Hattori and K. Tanaka, "Extracting Concept Hierarchy Knowledge from the Web based on Property Inheritance and Aggregation," Proceedings of the 7th IEEE/WIC/ACM International Conference on Web Intelligence (WI'08), pp.432–437 (2008).

[9] S. Hattori, "Hyponym Extraction from the Web based on Property Inheritance of Text and Image Features," Proceedings of the 6th International Conference on Advances in Semantic Processing (SEMAPRO'12), pp.109–114 (2012).

[10] S. Hattori, "Object-oriented Semantic and Sensory Knowledge Extraction from the Web," Web Intelligence and Intelligent Agents, In-Tech, ch.18, pp.365–390 (2010).

[11] T. Tezuka and K. Tanaka, "Visual Description Conversion for Enhancing Search Engines and Navigational Systems," Proceedings of the 8th Asia-Pacific Web Conference (AP-Web'06), LNCS vol.3841, pp.955–960 (2006).

[12] S. Hattori, T. Tezuka, and K. Tanaka, "Mining the Web for Appearance Description," Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA'07), LNCS vol.4653, pp.790–800 (2007).

[13] S. Hattori and K. Tanaka, "Object-Name Search by Visual Appearance and Spatio-Temporal Descriptions," Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication (ICUIMC'09), pp.63–70 (2009).

[14] S. Hattori, "Peculiar Image Retrieval by Cross-Language Web-Extracted Appearance Descriptions," International Journal of Computer Information Systems and Industrial Management (IJCISIM), MIR Labs, vol.4, pp.486–495 (2012).

[15] S. Hattori, "Hyponymy-based Peculiar Image Retrieval," International Journal of Computer Information Systems and Industrial Management (IJCISIM), MIR Labs, vol.5, pp.79–88 (2013).

[16] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting Influenza Epidemics Using Search Engine Query Data," Nature vol.457, pp.1012–1014 (2009).

[17] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors," Proceedings of the 19th International World Wide Web Conference (WWW'10), pp.851–860 (2010).

[18] E. Aramaki, S. Maskawa, and M. Morita, "Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11), pp.1568–1576 (2011).

[19] S. Hattori and K. Tanaka, "Mining the Web for Access Decision-Making in Secure Spaces," Proceedings of the Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on advanced Intelligent Systems (SCIS&ISIS'08), TH-G3-4, pp.370–375 (2008).

[20] S. Hattori, "Secure Spaces and Spatio-Temporal Weblog Sensors with Temporal Shift and Propagation," Proceedings of the 1st IRAST International Conference on Data Engineering and Internet Technology (DEIT'11), LNEE vol.157, pp.343–349 (2011).

[21] S. Hattori, "Linearly-Combined Web Sensors for Spatio-Temporal Data Extraction from the Web," Proceedings of the 6th International Workshop on Spatial and Spatiotemporal Data Mining (SSTDM'11), pp. 897–904 (2011).

[22] S. Hattori, "Spatio-Temporal Web Sensors by Social Network Analysis," Proceedings of the 3rd International Workshop on Business Applications of Social Network Analysis (BASNA'12), pp.1020–1027 (2012).

[23] S. Hattori, "Granularity Analysis for Spatio-Temporal Web Sensors," Proceedings of the WASET International Conference on Knowledge Management (ICKM'13), pp.192–200 (2013).

[24] S. Hattori, "Spatio-Temporal Web Sensors Using Web Queries vs. Documents," Journal of Automation and Control Engineering (JOACE), Engineering and Technology Publishing, vol.1, no.3, pp.192–197 (2013).

[25] S. Hattori, "Spatio-Temporal Dependency Analysis for Temporally-Shifted Web Sensors," Proceedings of the 2nd SDIWC International Conference on Informatics & Applications (ICIA'13), pp.30–35 (2013).

[26] S. Hattori, "Spatio-Temporal Propagation for Web Sensors," Proceedings of the SDIWC International Conference on Computer Science, Computer Engineering, and Social Media (CSCESM'14), pp.69–76 (2014).

[27] Shun Hattori, "Deflection Analysis for Spatially Propagated Web Sensors," Proceedings of the SDIWC International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC'15), pp.20–28 (2015).

[28] Japan Meteorological Agency, "Weather, Climate & Earthquake Information," http://www.jma.go.jp/jma/en/menu.html (2016).

[29] Geospatial Information Authority of Japan, "Sokuchi Survey Calculation API No.2," http://vldb.gsi.go.jp/sokuchi/surveycalc/main.html (2016).