Spatio-Temporal Dependency Analysis for Temporally-Shifted Web Sensors

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

***Abstract*—We experience or forecast various phenomena (e.g., rain, snow, earthquake) in the physical world, while we carry out various actions (e.g., blogging, searching, e-shopping) in the Web world. There have been many researches to mine the exploding Web world for knowledge about various phenomena and events in the physical world, and also Web services with the Web-mined knowledge have been made available for the public. However, there are few detailed investigations on how accurately Web-mined data reflect physical-world data. It must be socially-problematic to idolatrously utilize the Web-mined data in public Web services without ensuring their accuracy sufficiently. This paper introduces temporally-shifted Web Sensors with a temporal shift parameter δ to extract spatiotemporal numerical value about a physical phenomenon from Web documents searched by linguistic keyword(s) representing the physical phenomenon, and analyzes the spatiotemporal dependency of the temporal shift parameter δ with respect to their coefficient correlation with Japan Meteorological Agency's spatiotemporal statistics.**

I. Introduction

We experience or forecast various phenomena (e.g., rain, snow, earthquake, influenza, traffic accidents) in the physical world, while we carry out various actions (e.g., blogging, searching, e-shopping) in the Web world. Recently, many researches to mine a huge amount of Web documents in the explosively-growing Web, especially User Generated Content such as weblogs, 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 have been conducted actively. For example, opinion and reputation extraction [1], [2] of various products and services in the physical world, experience mining [3], [4] of various phenomena and events held in the physical world, and concept hierarchy (semantics) extraction such as is-a/has-a relationships [5–10] and visual appearance (look and feel) extraction [10–14] of physical objects in the physical world. Meanwhile, Web services with 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 few detailed investigations [15–17] on how accurately Web-mined data about a target phenomenon or event in the physical world reflect physical-world data. It is not difficult to mine the Web for some kind of potential knowledge data by using various text mining techniques, and it might be not problematic just to enjoy browsing Web-mined knowledge data. But while choosing better products, services, and actions in the physical world, it must be socially-problematic to immoderately utilize the Web-mined data in public Web services without ensuring their accuracy sufficiently.

The author has defined **Web Sensors** [18–23] to sense the Web (i.e., mine various actions in the Web world) for a target phenomenon in the physical world, and investigated how correlated Web-sensed spatiotemporal data are with physically-sensed spatiotemporal data as shown in Fig. 1. And also the author is integrating Web Sensors into Smart Spaces [24] and **Secure Spaces** [25–27].

This paper introduces the simplest and spatiotemporally-normalized Web Sensors and temporally-shifted Web Sensors with a temporal shift parameter δ to extract spatiotemporal numerical value about a physical phenomenon from Web documents searched by linguistic keyword(s) representing the physical phenomenon, and analyzes the spatiotemporal dependency of the temporal shift parameter δ (i.e., how the optimal value of the temporal shift parameter δ varies depending on geographical spaces and/or time periods) with respect to their coefficient correlation with Japan's rainfall, snowfall, and earthquake statistics per day by region (e.g., 47 prefectures) of Japan Meteorological Agency (JMA) [28].

The rest of this paper is organized as follows. Section II introduces temporally-shifted Web Sensors to sense the Web for spatiotemporal numerical value about physical phenomena. Section III analyzes the spatiotemporal dependency of temporal shift parameter δ with respect to their coefficient correlation with Japan's rainfall, snowfall, and earthquake statistics by Japan Meteorological Agency. Section IV concludes this paper.

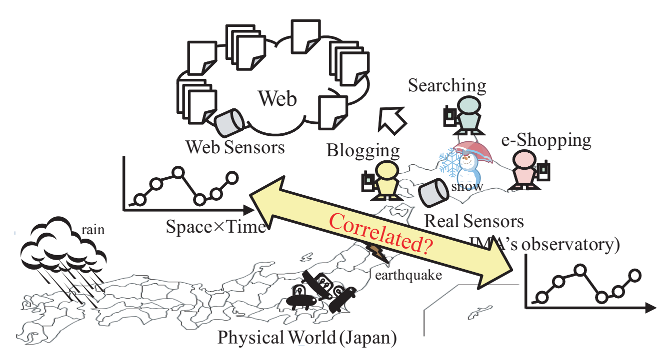


Fig. 1. Are Web Sensors Correlated with Real Sensors?

II. Temporally-Shifted Web Sensors

This section introduces the simplest and spatiotemporally-normalized Web Sensors and temporally-shifted Web Sensors to sense the Web for spatiotemporal numerical value dependent on such a space as 47 prefectures in Japan and such a time period as days and weeks in 2011 about such a physical phenomenon as rainfall, snowfall, and earthquake.

First, the simplest and spatiotemporally-normalized Web Sensor [18] by analyzing only Web documents with a geographic space s, e.g., one of 47 prefectures such as "Hokkaido" and "Kyoto," a time period t, e.g., one of 52 weeks in 2011 such as from January 1st to January 7th and from December 24th to December 30th, and a linguistic (e.g., Japanese) keyword kw representing a targeted physical phenomenon, e.g., "rain," "snow," and "earthquake," is defined as

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

where $\mathrm{df}_t(\texttt{[}q\texttt{]})$ 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 [29]. Note that the Weblog is superior to the whole Web, Twitter, Facebook, and News as a corpus of documents used by Web Sensors [19].

To investigate how temporally-shifted Web-sensed data are from real-sensed data, the temporally-shifted Web Sensor [20] with a temporal shift parameter δ $\texttt{[day]}$ is defined as

$$\mathrm{ws}_\delta(kw, s, t) := \mathrm{ws}(kw, s, t + \delta). \qquad (2)$$

As shown in Fig. 2, Shifted-to-Past Web Sensors for a physical phenomenon (e.g., earthquake) when the temporal shift parameter δ is positive (e.g., $+14$) calculate the numerical value dependent on a geographical space (e.g., Hokkaido prefecture in Japan) and a time period t (e.g., one of 52 weeks in 2011) by analyzing Web documents uploaded δ day(s) after the time period t (i.e., infer the past from the future), while Shifted-to-Future Web Sensors for a physical phenomenon when the temporal shift parameter δ is negative (e.g., -14) calculate the numerical value dependent on a geographical space and a time period t by analyzing Web documents uploaded $|\delta|$ day(s) before the time period t (i.e., infer the future from the past).

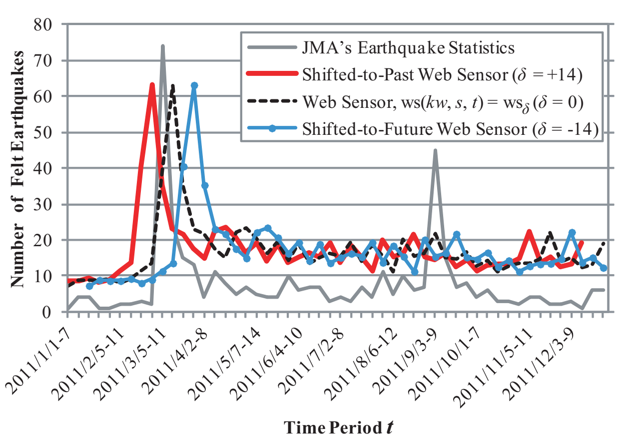


Fig. 2. Temporally-Shifted Web Sensors for Earthquake and JMA's Weekly Earthquake Statistics in Hokkaido Prefecture (ISO 3166-2:JP-01), 2011.

III. Spatio-Temporal Dependency Analysis of Temporal Shift Parameter δ

The previous section introduces Web Sensors with a temporal shift parameter δ to extract spatiotemporal numerical value about a physical phenomenon from the Weblog. To optimize the temporal shift parameter δ, this section analyzes the spatiotemporal dependency of the temporal shift parameter δ (i.e., how the optimal value of the temporal shift parameter δ varies depending on geographical spaces and/or time periods) with respect to their coefficient correlation with Japan's rainfall, snowfall, and earthquake statistics per day by region (e.g., 47 prefectures) of Japan Meteorological Agency (JMA) [28].

Various different features of three kinds of target physical phenomena in Japan are summarized as follows.

1) Rainfall: has spikes in any seasons and regions, and is forecasted in advance by JMA and others.
2) Snowfall: has spikes in only winter season, and is forecasted in advance by JMA and others.
3) Earthquake: has sharper spikes anytime potentially, and is not yet predicted well in advance.

Fig. 3, Fig. 11, and Fig. 19 show the average of coefficient correlation of daily (47 prefectures $\times$ 1 day $\times$ 364 data) vs. weekly Web Sensors (47 prefectures $\times$ 1 week $\times$ 52 data) depending on their temporal shift parameter δ for rainfall, snowfall, and earthquake, respectively. Fig. 3 shows that Not-Shifted Web Sensor whose temporal shift parameter δ is (almost) ± 0 gives the best correlation for rainfall. Meanwhile, Fig. 11 shows that Shifted-to-Future Web Sensor whose δ is negative gives the best correlation (gains avg. 5% over $\delta = 0$) for snowfall which is forecasted in advance, and Fig. 19 shows that Shifted-to-Past Web Sensor whose δ is positive gives the best correlation (gains avg. 16% over $\delta = 0$) for earthquake which cannot yet be predicted well in advance.

A. Temporal Dependency Analysis

Fig. 4, Fig. 12, and Fig. 20 analyze the optimal temporal shift parameter δ and coefficient correlation of daily Web Sensors (47 prefectures $\times$ 1 day $\times$ 29 data) depending on 49 time periods in 2011 (e.g., 29 days of 2011/1/1-1/29 or 2011/12/3-12/31) for rainfall, snowfall, and earthquake, respectively. Fig. 4 shows that the optimal δ and correlation of Web Sensors for rainfall are not much dependent on time periods except winter season (in Jan. to Mar.) when it may not rain but snow. Meanwhile, Fig. 12 shows that the optimal δ of Web Sensors for snowfall varies more widely, and Fig. 20 shows that both the optimal δ and correlation of Web Sensors for earthquake varies the most widely. Fig. 20 also shows that more shaken time periods are given higher correlation by the Great East Japan Earthquake (3.11).

B. Spatial Dependency Analysis

Figs. 5 to 10, Figs. 13 to 18, and Figs. 21 to 26 show the optimal temporal shift parameter δ and coefficient correlation of daily Web Sensors vs. weekly Web Sensors depending on 47 prefectures (geographical spaces) in Japan for rainfall, snowfall, and earthquake, respectively. They show that the optimal δ for rainfall is not much dependent on prefectures, while the optimal δ for snowfall and earthquake varies widely. And that more shaken prefectures are given higher correlation.

– *Rainfall* –

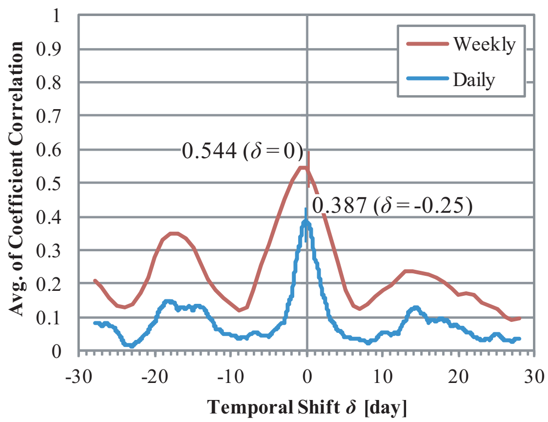


Fig. 3. Avg. Coefficient Correlation of Daily/Weekly Web Sensors on Temporal Shift Parameter δ. (Daily: 1 day × 364 data, Weekly: 1 week × 52 data)

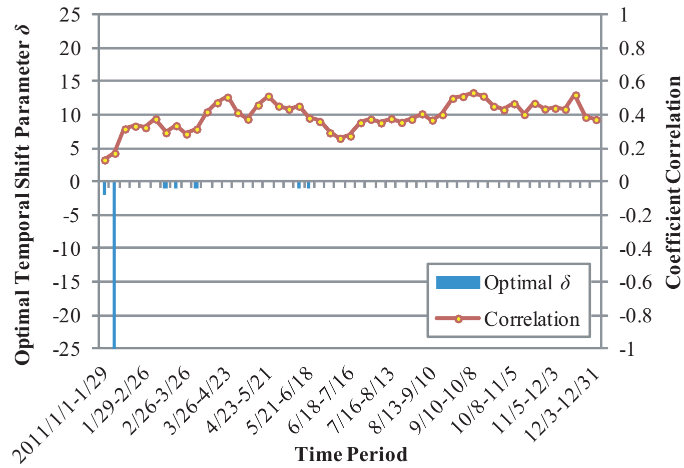


Fig. 4. Temporal Dependency Analysis of Optimal Temporal Shift Param δ and Coefficient Correlation for Daily Web Sensors. (Daily: 1 day × 29 data)

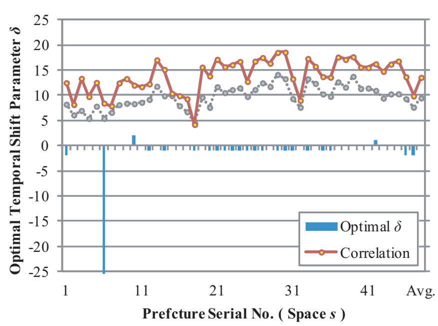


Fig. 5. Spatial Dependency Analysis of Optimal Temporal Shift Parameter δ and Coefficient Correlation for Weekly Web Sensors. (Weekly: 1 week × 52 data)

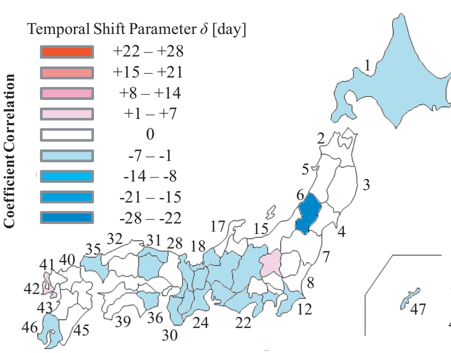


Fig. 6. Spatial Distribution of Optimal Temporal Shift Parameter δ for Weekly Web Sensors. (Weekly: 1 week × 52 data)

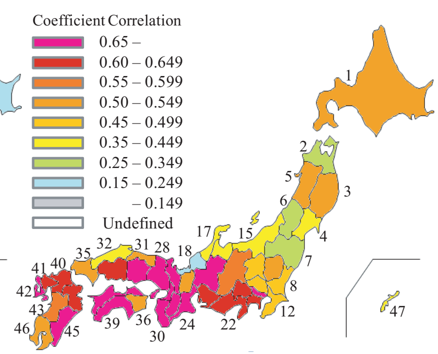


Fig. 7. Spatial Distribution of Coefficient Correlation for Weekly Web Sensors. (Weekly: 1 week × 52 data)

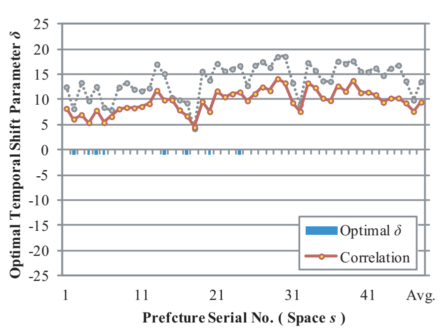


Fig. 8. Spatial Dependency Analysis of Optimal Temporal Shift Parameter δ and Coefficient Correlation for Daily Web Sensors. (Daily: 1 day × 364 data)

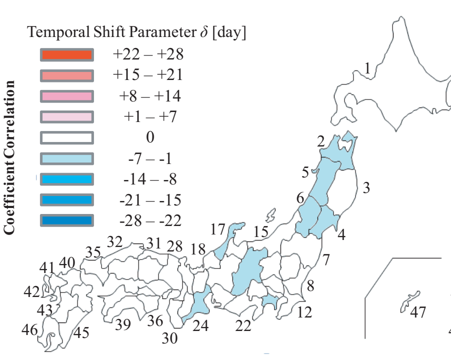


Fig. 9. Spatial Distribution of Optimal Temporal Shift Parameter δ for Daily Web Sensors. (Daily: 1 day × 364 data)

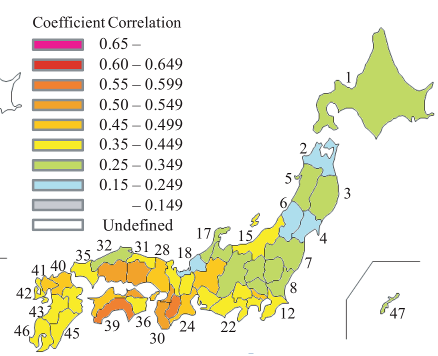
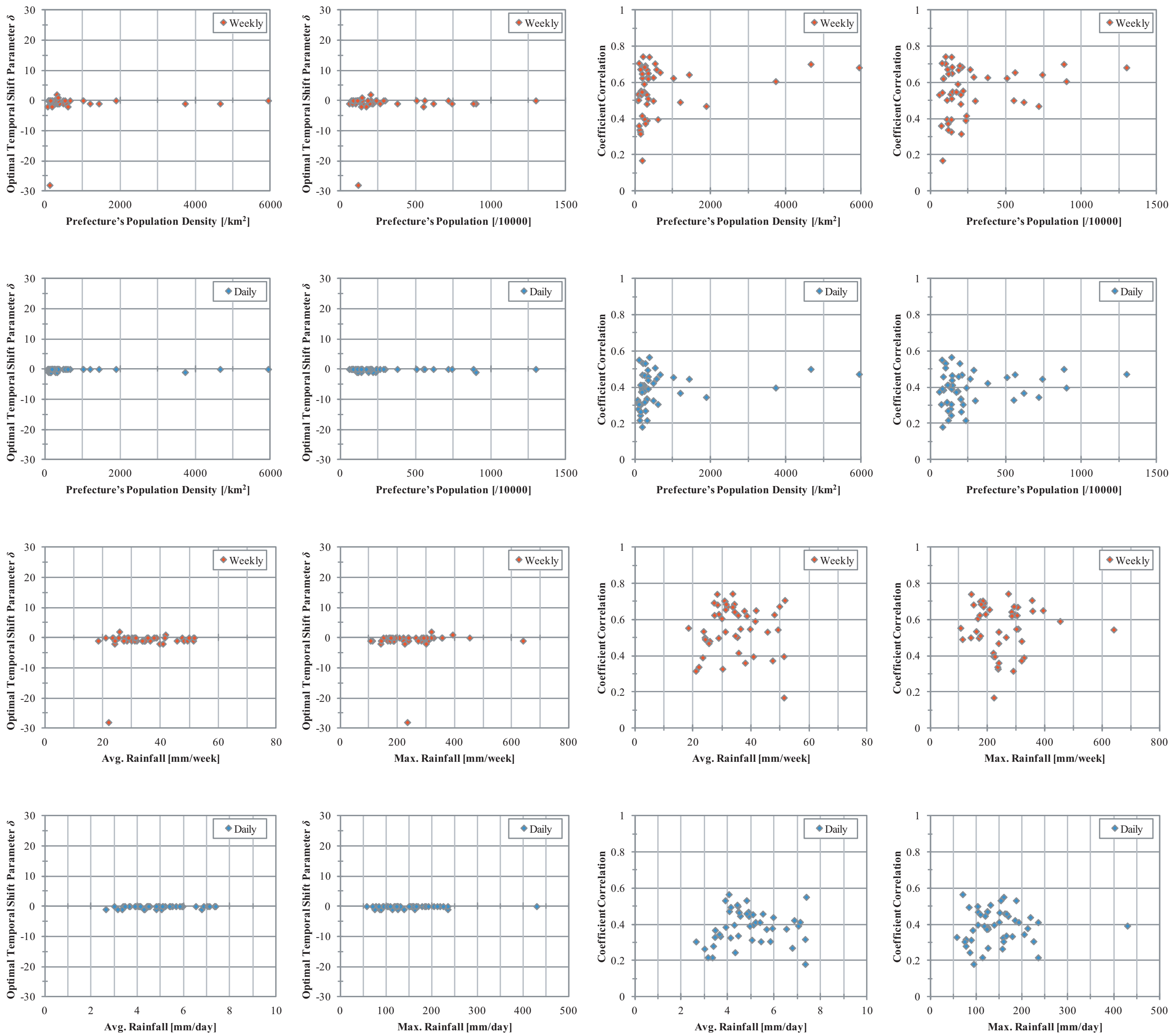


Fig. 10. Spatial Distribution of Coefficient Correlation for Daily Web Sensors. (Daily: 1 day × 364 data)

– *Rainfall* –


Weekly
Daily
Optimal Temporal Shift Parameter δ
Coefficient Correlation
Prefecture's Population Density [/km²]
Prefecture's Population [/10000]
Avg. Rainfall [mm/week]
Max. Rainfall [mm/week]
Avg. Rainfall [mm/day]
Max. Rainfall [mm/day]

– Snowfall –

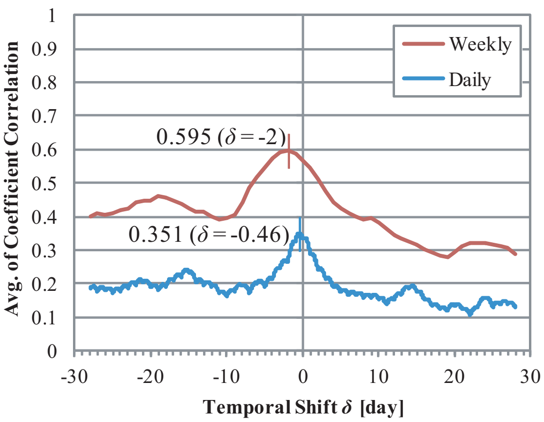


Fig. 11. Avg. Coefficient Correlation of Daily/Weekly Web Sensors on Temporal Shift Parameter δ. (Daily: 1 day $\times$ 364 data, Weekly: 1 week $\times$ 52 data)

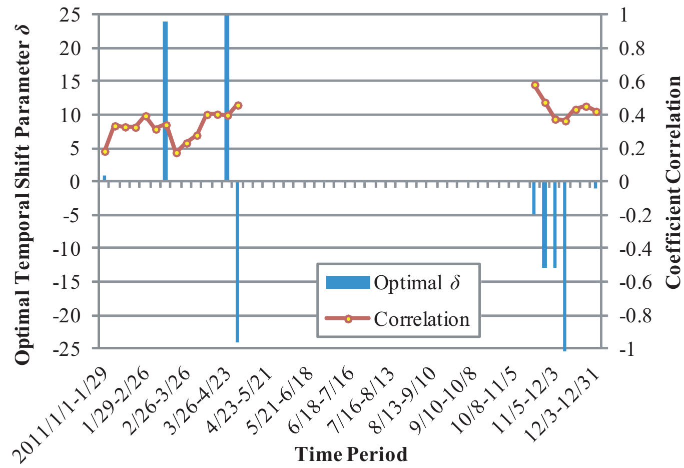


Fig. 12. Temporal Dependency Analysis of Optimal Temporal Shift Param δ and Coefficient Correlation for Daily Web Sensors. (Daily: 1 day $\times$ 29 data)

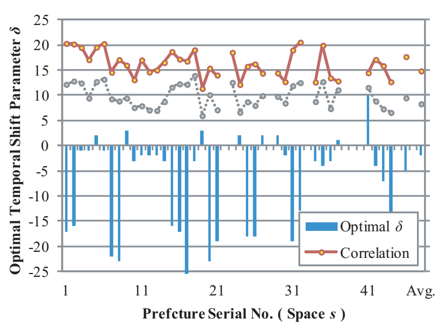


Fig. 13. Spatial Dependency Analysis of Optimal Temporal Shift Parameter δ and Coefficient Correlation for Weekly Web Sensors. (Weekly: 1 week $\times$ 52 data)

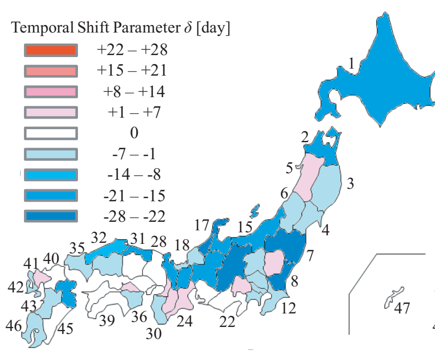


Fig. 14. Spatial Distribution of Optimal Temporal Shift Parameter δ for Weekly Web Sensors. (Weekly: 1 week $\times$ 52 data)

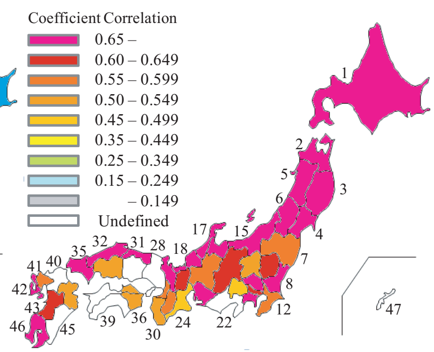


Fig. 15. Spatial Distribution of Coefficient Correlation for Weekly Web Sensors. (Weekly: 1 week $\times$ 52 data)

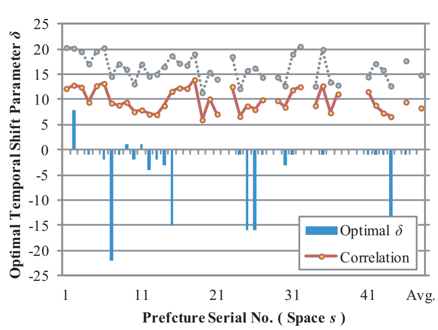


Fig. 16. Spatial Dependency Analysis of Optimal Temporal Shift Parameter δ and Coefficient Correlation for Daily Web Sensors. (Daily: 1 day $\times$ 364 data)

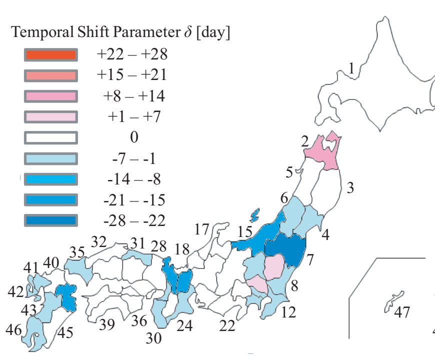


Fig. 17. Spatial Distribution of Optimal Temporal Shift Parameter δ for Daily Web Sensors. (Daily: 1 day $\times$ 364 data)

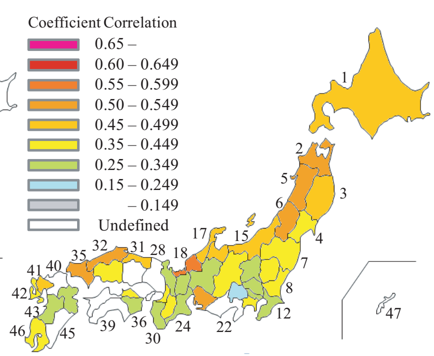
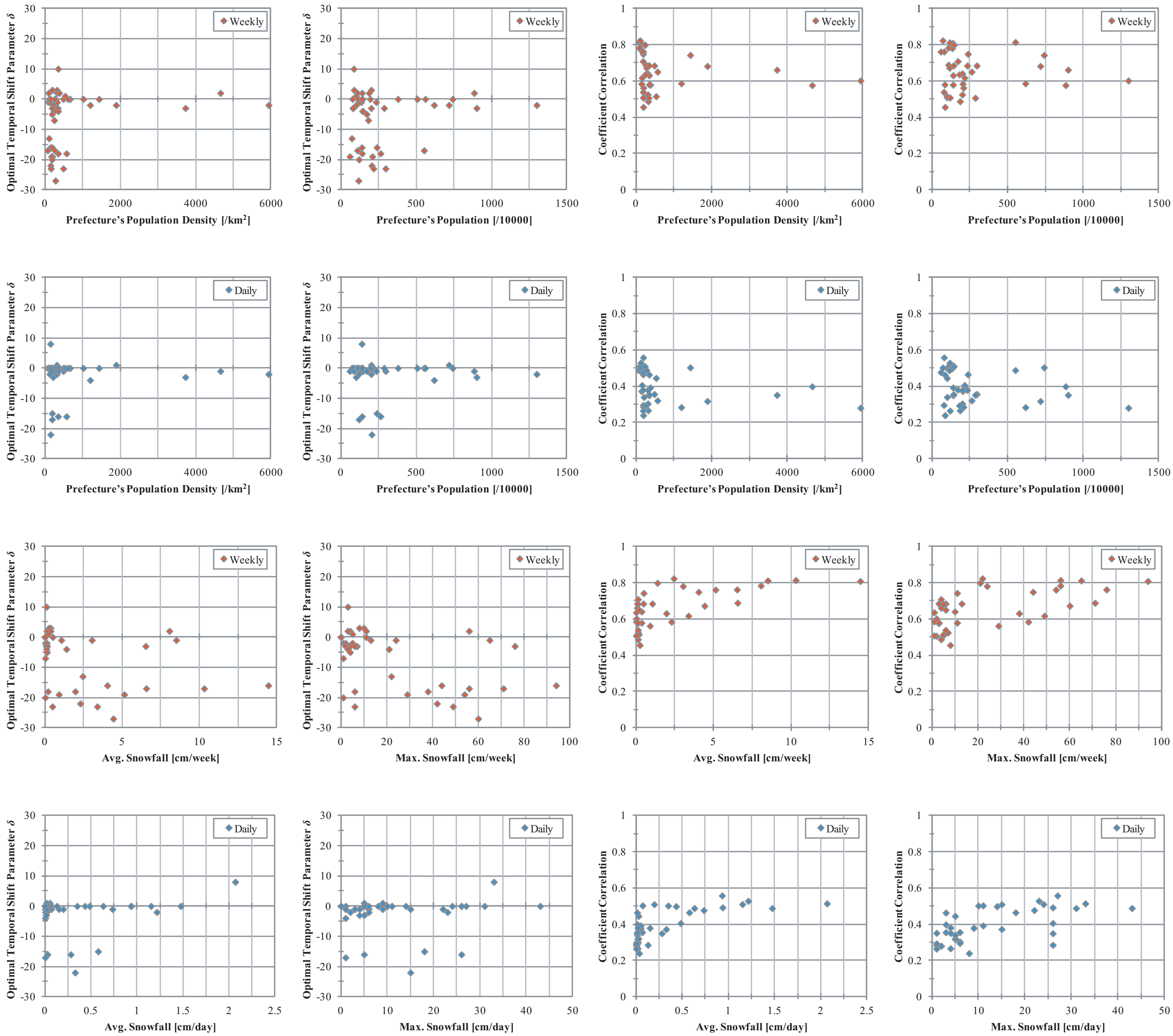


Fig. 18. Spatial Distribution of Coefficient Correlation for Daily Web Sensors. (Daily: 1 day $\times$ 364 data)

– Snowfall –


Optimal Temporal Shift Parameter δ
Coefficient Correlation
Prefecture's Population Density [/km²]
Prefecture's Population [/10000]
Avg. Snowfall [cm/week]
Max. Snowfall [cm/week]
Avg. Snowfall [cm/day]
Max. Snowfall [cm/day]
Weekly
Daily

– *Earthquake* –

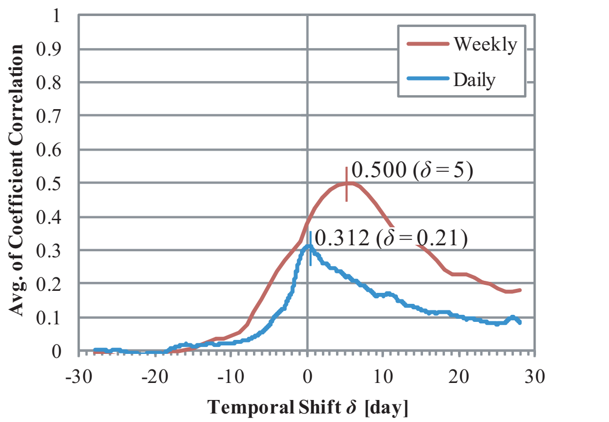


Fig. 19. Avg. Coefficient Correlation of Daily/Weekly Web Sensors on Temporal Shift Parameter δ. (Daily: 1 day $\times$ 364 data, Weekly: 1 week $\times$ 52 data)

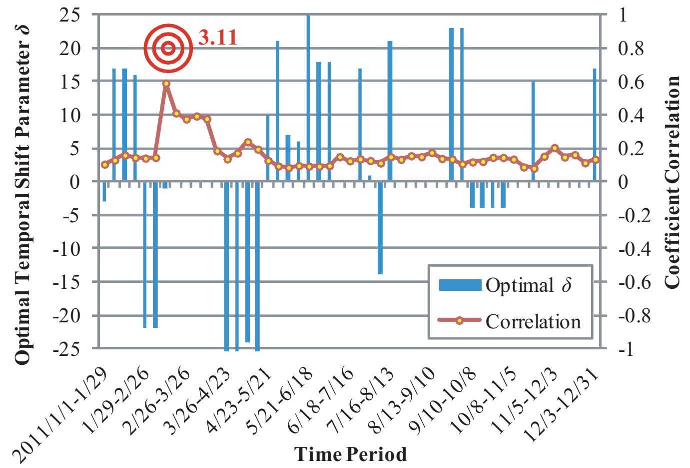


Fig. 20. Temporal Dependency Analysis of Optimal Temporal Shift Param δ and Coefficient Correlation for Daily Web Sensors. (Daily: 1 day $\times$ 29 data)

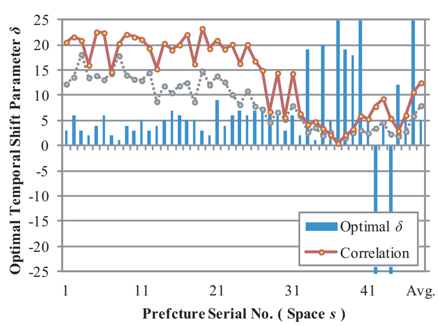


Fig. 21. Spatial Dependency Analysis of Optimal Temporal Shift Parameter δ and Coefficient Correlation for Weekly Web Sensors. (Weekly: 1 week $\times$ 52 data)

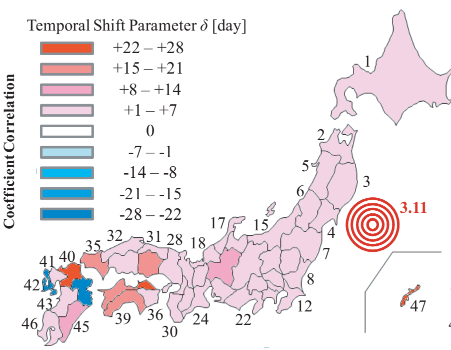


Fig. 22. Spatial Distribution of Optimal Temporal Shift Parameter δ for Weekly Web Sensors. (Weekly: 1 week $\times$ 52 data)

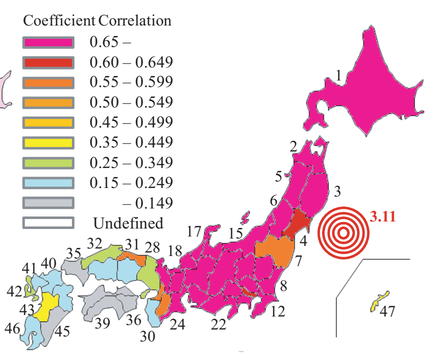


Fig. 23. Spatial Distribution of Coefficient Correlation for Weekly Web Sensors. (Weekly: 1 week $\times$ 52 data)

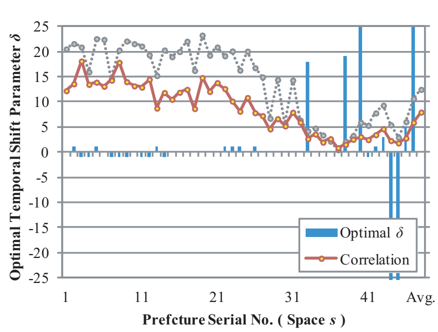


Fig. 24. Spatial Dependency Analysis of Optimal Temporal Shift Parameter δ and Coefficient Correlation for Daily Web Sensors. (Daily: 1 day $\times$ 364 data)

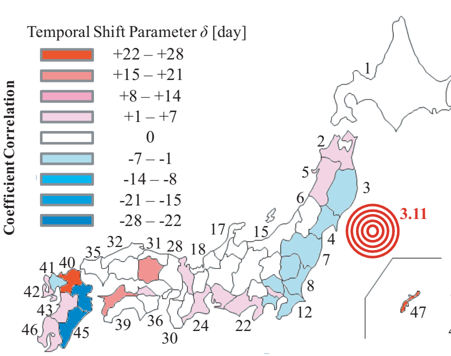


Fig. 25. Spatial Distribution of Optimal Temporal Shift Parameter δ for Daily Web Sensors. (Daily: 1 day $\times$ 364 data)

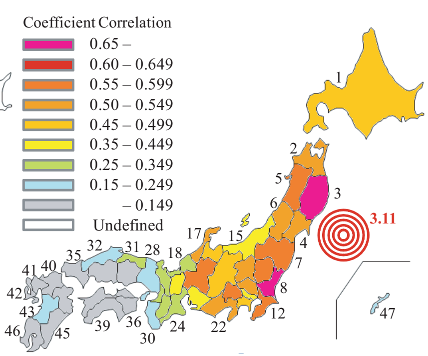
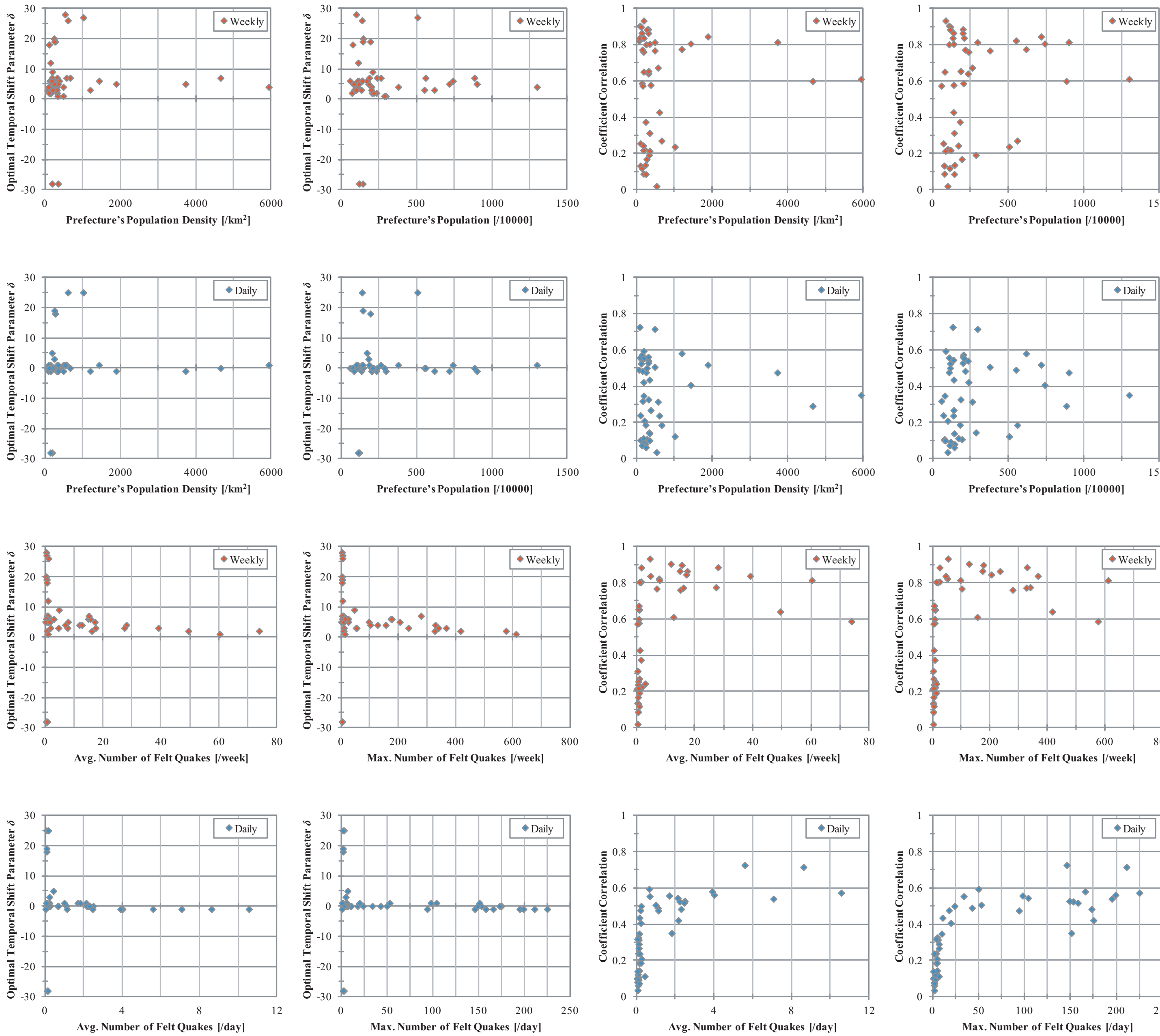


Fig. 26. Spatial Distribution of Coefficient Correlation for Daily Web Sensors. (Daily: 1 day $\times$ 364 data)

– *Earthquake* –


Weekly
Daily
Optimal Temporal Shift Parameter δ
Coefficient Correlation
Prefecture's Population Density [/km²]
Prefecture's Population [/10000]
Avg. Number of Felt Quakes [/week]
Max. Number of Felt Quakes [/week]
Avg. Number of Felt Quakes [/day]
Max. Number of Felt Quakes [/day]

IV. Conclusion

To investigate how correlated/temporally-shifted Web-sensed data are with/from real-sensed data, this paper has introduced Web Sensors with a temporal shift parameter δ to extract spatiotemporal numerical value about a physical phenomenon from the Weblog. And to optimize the temporal shift parameter δ, this paper has analyzed the spatiotemporal dependency of the temporal shift parameter δ (i.e., how the optimal value of δ varies depending on geographical spaces and/or time periods) with respect to their coefficient correlation with Japan's rainfall, snowfall, and earthquake statistics per day by region (e.g., 47 prefectures) of Japan Meteorological Agency. The spatiotemporal dependency analysis shows that

- The optimal temporal shift parameter δ of Web Sensors depends on physical phenomena: Not-Shifted Web Sensor whose δ is ± 0 gives the best correlation (i.e., the Weblog runs parallel to the physical world) for rainfall, Shifted-to-Future Web Sensor whose δ is negative gives the best (i.e., the Weblog leads the physical world) for snowfall, and Shifted-to-Past Web Sensor whose δ is positive gives the best (i.e., the Weblog follows the physical world) for earthquake.
- The optimal temporal shift parameter δ and coefficient correlation for rainfall are not much dependent on geographical spaces and time periods, while the optimal δ for snowfall and earthquake varies more widely.
- More shaken geographical spaces and time periods are given higher correlation between Web-sensed and real-sensed data by the Great East Japan Earthquake.

In the future, various Web actions, e.g., not only blogging and searching but also e-shopping, will be combined to construct more high-sensitive Web Sensors. For example, spatiotemporal e-shopping logs of umbrellas on e-commerce sites such as Amazon, Yahoo! Shopping, and Rakuten Market might be useful to sense the Web for spatiotemporal numerical value about rainfall phenomenon. And also Web Sensors try to forecast future data about a target phenomenon, interpolate lost data of real statistics, and alert falsified data of real statistics.

References

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

[2] S. Fujimura, M. Toyoda, and M. Kitsuregawa, "A Reputation Extraction Method Considering Structure of Sentence," Proc. 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," Proc. 15th Int'l World Wide Web Conf. (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," Proc. 7th IEEE/WIC/ACM Int'l Conf. on Web Intelligence (WI'08), pp.314–321 (2008).

[5] M. A. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," Proc. 14th Int'l Conf. 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," Proc. 10th Asia-Pacific Web Conf. (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," Proc. 7th Int'l Conf. 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," Proc. 6th Int'l Conf. 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 Conf. (APWeb'06), LNCS vol.3841, pp.955–960 (2006).

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

[13] S. Hattori, "Peculiar Image Retrieval by Cross-language Web-extracted Appearance Descriptions," Int'l Journal of Computer Information Systems and Industrial Management, MIR Labs, vol.4, pp.486–495 (2012).

[14] 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).

[15] 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).

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

[17] E. Aramaki, S. Maskawa, and M. Morita, "Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter," Proc. Conf. on Empirical Methods in Natural Language Processing, pp.1568–1576 (2011).

[18] S. Hattori and K. Tanaka, "Mining the Web for Access Decision-Making in Secure Spaces," Proc. Joint 4th Int'l Conf. on Soft Computing and Intelligent Systems and 9th Int' Symp. on advanced Intelligent Systems (SCIS&ISIS'08), TH-G3-4, pp.370–375 (2008).

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

[20] S. Hattori, "Secure Spaces and Spatio-Temporal Weblog Sensors with Temporal Shift and Propagation," Proc. 1st IRAST Int'l Conf. 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," Proc. the 6th Int'l Workshop on Spatial and Spatiotemporal Data Mining (SSTDM'11), pp.897–904 (2011).

[22] S. Hattori, "Granularity Analysis for Spatio-Temporal Web Sensors," Proc. WASET Int'l Conf. on Knowledge Management (ICKM'13), pp.192–200 (2013).

[23] 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).

[24] L. Rosenthal and V. Stanford, "NIST Smart Space: Pervasive Computing Initiative," Proc. 9th IEEE Int'l Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, pp.6–11 (2000).

[25] S. Hattori and K. Tanaka, "Towards Building Secure Smart Spaces for Information Security in the Physical World," Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Fuji Technology Press, vol.11, no.8, pp.1023–1029 (2007).

[26] S. Hattori, "Context-Aware Query Control for Secure Spaces," Journal of Computer Technology and Application (JCTA), David Publishing, vol.3, no.2, pp.130–139 (2012).

[27] S. Hattori, "Ability-Based Expression Control for Secure Spaces," Proc. Joint 6th Int'l Conf. on Soft Computing and Intelligent Systems and 13th Int'l Symp. on advanced Intelligent Systems (SCIS&ISIS'12), F1-54-3, pp.1298–1303 (2012).

[28] Japan Meteorological Agency, http://www.jma.go.jp/jma/ (2013).

[29] Google Web Search, http://www.google.co.jp/ (2013).