Spatio-Temporal Web Sensors Using Web Queries vs. Documents

Shun Hattori, *Member, IEEE*

***Abstract*—In the Web world, one user creates and uploads a Web document about a phenomenon or event in the physical world, while another user retrieves and consumes Web documents by submitting a Web query. There have been many researches to mine Web documents in the exploding Web world, especially User Generated Content such as weblogs and microblogs, 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 immoderately utilize the Web-mined data in public Web services without ensuring their accuracy sufficiently. Therefore, the previous papers introduced various Web Sensors to extract spatiotemporal data about a target phenomenon from Web documents searched by keyword(s) representing the target phenomenon, and tried to validate the potential and reliability of the Web-sensed spatiotemporal data. Moreover, this paper compares spatio-temporal Web Sensors analyzing Web queries and/or Web documents.**

***Index Terms*—Knowledge extraction, Query analysis, Spatio-temporal data mining, Web credibility, Web sensor.**

I. Introduction

In the Web world, one user creates and uploads a Web document about a phenomenon or event in the physical world, while another user retrieves and consumes Web documents by submitting a Web query. Recently, there have been many researches to mine 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. For example, opinion and reputation extraction [1], [2] of various products and services provided 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 the 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 very few detailed investigations on how accurately Web-mined data about a phenomenon or event held in the physical world reflect physical-world data. It is not difficult for us to extract some kind of the potential knowledge data from the Web by using various text mining techniques, and it might be not problematic just to enjoy browsing them. But while choosing better products, services, and actions in the physical world, it must be problematic to idolatrously utilize the Web-mined data in public Web services without ensuring their accuracy sufficiently. Therefore, the previous papers [15]–[18] introduced various Web Sensors to extract spatiotemporal data about a target phenomenon from Web documents searched by keyword(s) representing the target phenomenon, and tried to validate the potential and reliability of the Web-sensed spatiotemporal data by coefficient correlation with physical-world statistics. Fig. 1 shows spatio-temporal Web Sensors used in Secure Spaces [19]–[22].

This paper defines a novel kind of spatio-temporal Web Sensors by analyzing Web queries, which are submitted to Web search engines such as Google to retrieve Web documents about a phenomenon (e.g., rainfall, snowfall, and earthquake) in the physical world, and also compares and combines them with spatio-temporal Web Sensors by analyzing Web documents, which are created and uploaded to the Web, with respect to coefficient correlation with physical-world statistics per week by region (e.g., 47 prefectures) of Japan Meteorological Agency [23].

The rest of this paper is organized as follows. Section II defines spatio-temporal Web Sensors by analyzing Web queries and/or Web documents. Section III compares them for three kinds of physical-world phenomena (e.g., rainfall, snowfall, and earthquake). Section IV concludes this paper.

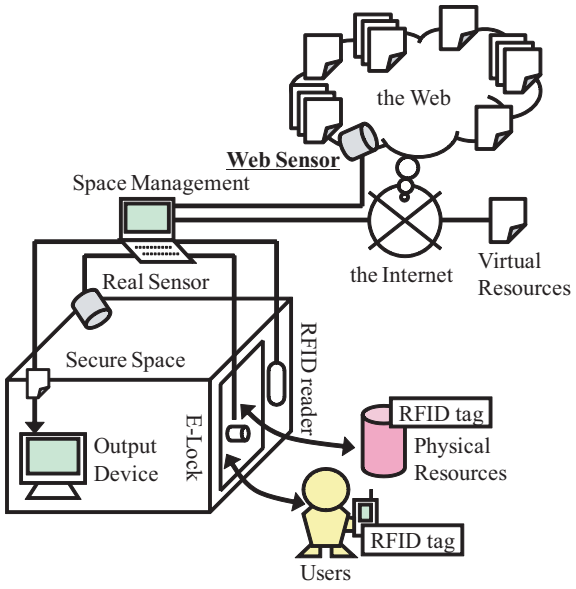

Fig. 1. Spatio-temporal Web Sensors in Secure Spaces.

S. Hattori is with the 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).

II. METHOD

This section defines several spatio-temporal Web Sensors by analyzing Web queries and/or Web documents, to sense the Web for spatiotemporal data 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 spatiotemporally-normalized Web Sensor [15] by analyzing only Web documents with a geographic space s, e.g., one of 47 prefectures such as "北海道" (Hokkaido) and "京都" (Kyoto), a time period t, e.g., one of 52 weeks in 2011 such as from January 2nd to January 8th and from December 25th to December 31st, and a Japanese keyword kw representing a targeted physical phenomenon, e.g., "雨" (rain), "雪" (snow), and "地震" (earthquake), is defined as

$$\text{ws-doc}(kw, s, t) := \frac{\text{df}_t(\texttt{["}kw\texttt{" AND "}s\texttt{"]})}{\text{df}_t(\texttt{["}s\texttt{"]})}, \quad (1)$$

where $\text{df}_t(\texttt{[}q\texttt{]})$ 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 [24]. Note that the Weblog is superior to the whole Web, Twitter, Facebook, and News for a Web corpus used by Web Sensors [16].

And a novel Web Sensor, ws-query$(kw, s, t) \in [0.0, 1.0]$, by analyzing only Web queries is defined by Google Trends [25], that can compare search volume patterns including a keyword kw across specific regions s (e.g., 47 prefectures in Japan), categories, date ranges t (e.g., from January 2011 to December 2011), and Google's search services (e.g., Web, Image, News, and Product Searches), and analyze a portion of Google's Web queries to compute the number of searches that have been done for the user-given query terms, relative to the total number of searches done on Google over time.

Next, the temporally-shifted Web Sensors [17] with a temporal shift parameter δ [week] are defined as

$$\text{ws-doc}_\delta(kw, s, t) := \text{ws-doc}(kw, s, t + \delta), \quad (2)$$

$$\text{ws-query}_\delta(kw, s, t) := \text{ws-query}(kw, s, t + \delta). \quad (3)$$

Note that the value of temporally-shifted Web Sensor for a time period t is calculated by analyzing Web documents or queries before the time period when δ is negative, while it is calculated by analyzing Web documents or queries after the time period when δ is positive. For example, on January 7th, 2011, one user created and uploaded a Web document about yesterday, today, and tomorrow rainfall as a physical phenomenon: "Hokkaido had a record snowfall yesterday. It is snowing hard today. It will be also snowy tomorrow." Another user who lives in Canada retrieves and consumes Web documents about the Great East Japan Earthquake on March 11th, 2011 as a past phenomenon by submitting a Web query ["earthquake" AND "japan" AND "2011"] to Google.

Last, the linearly-combined Web Sensor [18] with a combination parameter $\alpha \in [0.0, 1.0]$ is defined as

$$\begin{aligned}\text{ws}^\alpha_\delta(kw, s, t) := (1 - \alpha) \cdot \text{ws-doc}_\delta(kw, s, t) \\ + \ \alpha \ \cdot \text{ws-query}_\delta(kw, s, t).\end{aligned} \quad (4)$$

III. EXPERIMENT

This section compares spatio-temporal Web Sensors analyzing Web queries and/or Web documents by coefficient correlation with three kinds of physical-world statistics per week by region of Japan Meteorological Agency (JMA) [23] to validate the potential and reliability of the Web-sensed spatiotemporal data for such a space as 47 prefectures in Japan and such a time period as weeks in 2011.

Fig. 2 to 4 show various different features of the following three kinds of target phenomena in the physical world (in Hokkaido where is the most northern prefecture of Japan).

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.

And they show the correlation and its coefficient between spatio-temporal data of Web Sensor, ws-query, using only Web queries by Google Trends and JMA's statistics for each physical phenomenon. Meanwhile, Fig. 5 to 7 show spatio-temporal data of Web Sensor, ws-doc, using only Weblog documents searched by Google.

Fig. 7 shows that the Web Sensor using Weblog documents can sense the sharpest spike caused by the Great East Japan Earthquake (M9.0) on March 11th, 2011, but cannot acutely sense the 2nd sharpest spike caused by the earthquake (M5.1) in Hokkaido on September 7th, 2011, and that for a while after the Great East Japan Earthquake, its very huge effects decreasingly kept on the Web Sensor using Weblog documents (i.e., kept people creating and providing Web documents) as well as the physical world. Meanwhile, Fig. 4 shows that people have stopped searching and consuming earlier than creating and providing Web documents about the Great East Japan Earthquake. There might exist a gap between demands and creations of Web documents about rarer earthquakes.

Next, Fig. 8 to 10 show the dependency of correlation coefficient between spatiotemporal data of temporally-shifted Web Sensor, ws-query$_\delta$, using only Web queries by Google Trends and JMA's statistics on its temporal shift parameter δ for each physical phenomenon. Meanwhile, Fig. 11 to 13 show the dependency of correlation coefficient between spatiotemporal data of temporally-shifted Web Sensor, ws-doc$_\delta$, using only Weblog documents by Google and JMA's statistics on its temporal shift parameter δ for each physical phenomenon.

They show that temporally-shifted Web Sensor, ws-query$_\delta$, using only Web queries is superior to (i.e., gains average 7.03% and 12.92% against) temporally-shifted Web Sensor, ws-doc$_\delta$, using only Weblog documents for snowfall and earthquake, while ws-query$_\delta$ is inferior to (i.e., loses average 12.39% against) ws-doc$_\delta$ for rainfall. And that not-shifted Web Sensor whose temporal shift parameter δ is ± 0 gives the best correlation for rainfall, shifted-to-future Web Sensor whose δ is negative gives the best correlation for snowfall which can be forecasted in advance by JMA and others, and shifted-to-past Web Sensor whose δ is positive gives the best correlation for earthquake which cannot yet be predicted well in advance.

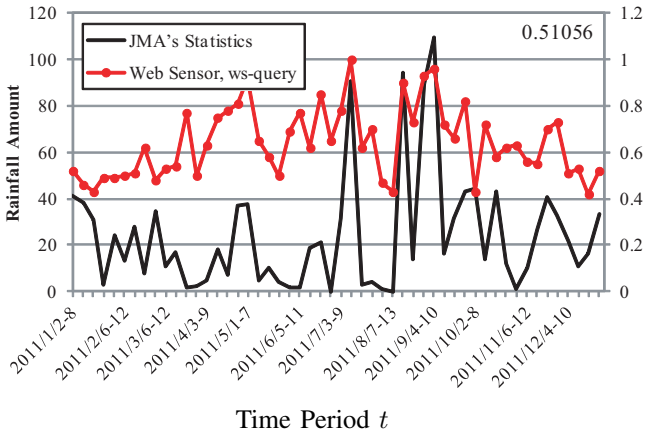

Fig. 2. JMA's weekly statistics and Web Sensor's data using Web queries collected by Google Trends for rainfall in Hokkaido prefecture, 2011.

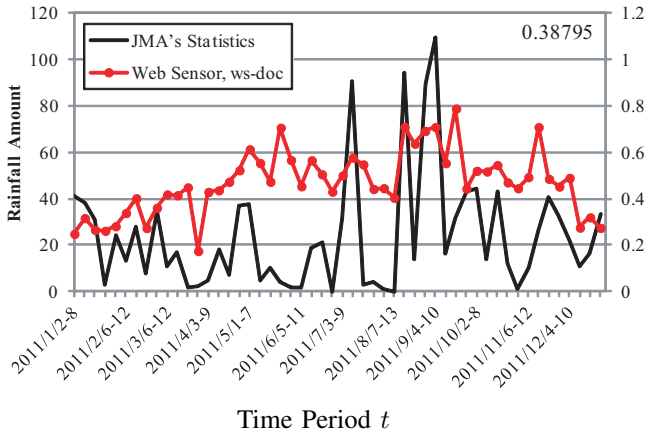

Fig. 5. JMA's weekly statistics and Web Sensor's data using Web documents searched by Google for rainfall in Hokkaido prefecture, 2011.

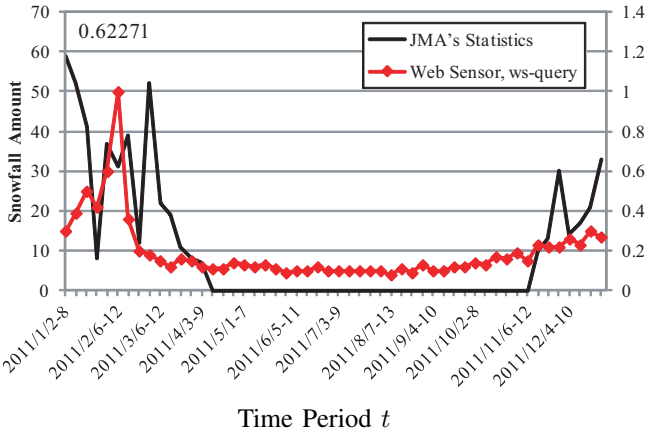

Fig. 3. JMA's weekly statistics and Web Sensor's data using Web queries collected by Google Trends for snowfall in Hokkaido prefecture, 2011.

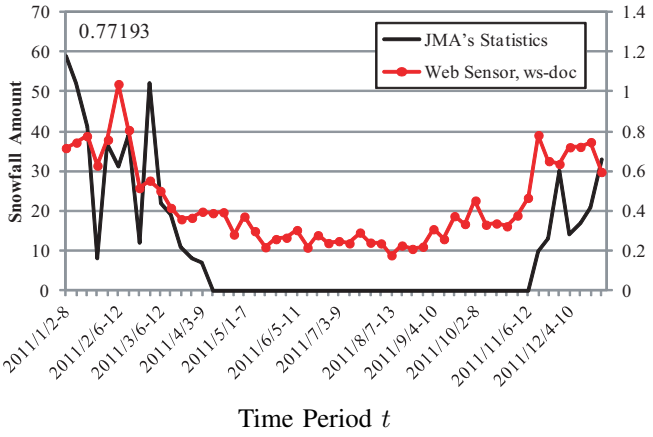

Fig. 6. JMA's weekly statistics and Web Sensor's data using Web documents searched by Google for snowfall in Hokkaido prefecture, 2011.

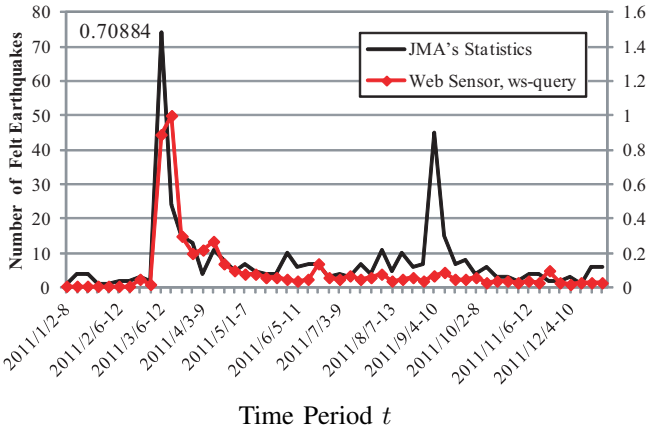

Fig. 4. JMA's weekly statistics and Web Sensor's data using Web queries collected by Google Trends for earthquake in Hokkaido prefecture, 2011.

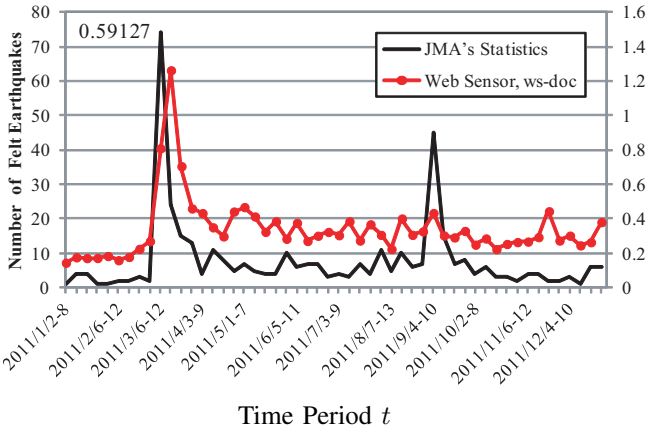

Fig. 7. JMA's weekly statistics and Web Sensor's data using Web documents searched by Google for earthquake in Hokkaido prefecture, 2011.

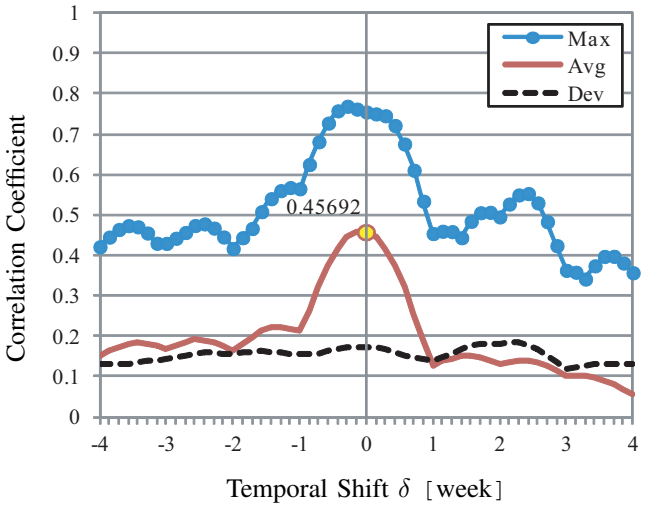

Fig. 8. Dependency of temporally-shifted Web Sensor, ws-query$_\delta$, using Web queries collected by Google Trends on temporal shift δ for rainfall.

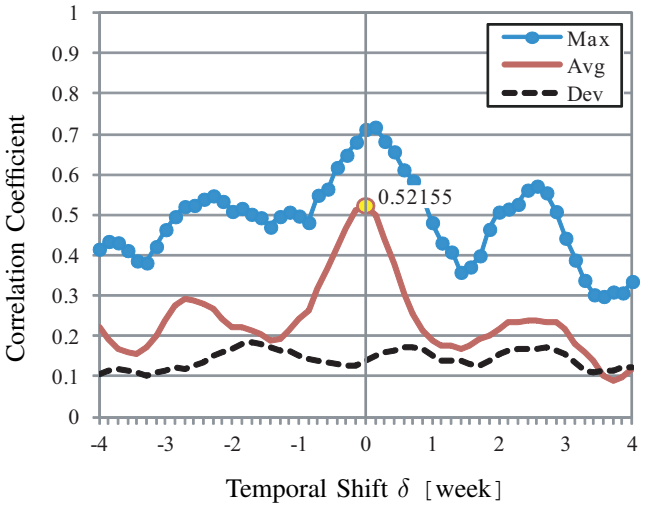

Fig. 11. Dependency of temporally-shifted Web Sensor, ws-doc$_\delta$, using Weblog documents searched by Google on temporal shift δ for rainfall.

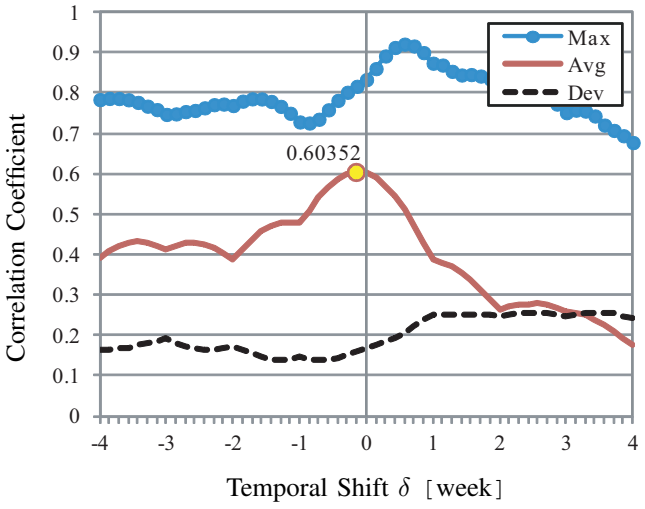

Fig. 9. Dependency of temporally-shifted Web Sensor, ws-query$_\delta$, using Web queries collected by Google Trends on temporal shift δ for snowfall.

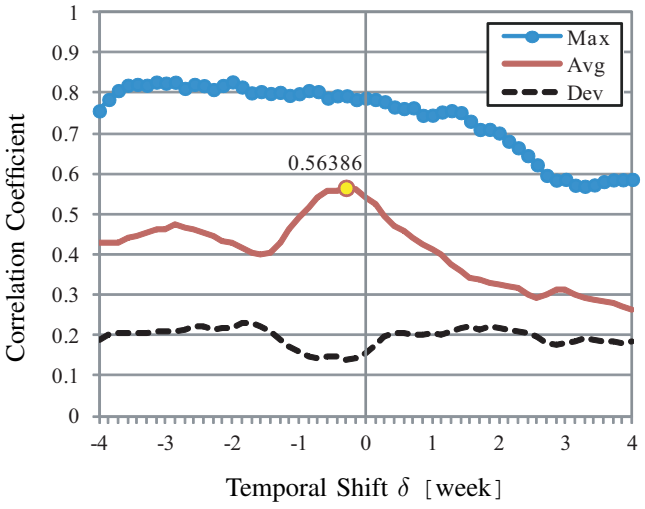

Fig. 12. Dependency of temporally-shifted Web Sensor, ws-doc$_\delta$, using Weblog documents searched by Google on temporal shift δ for snowfall.

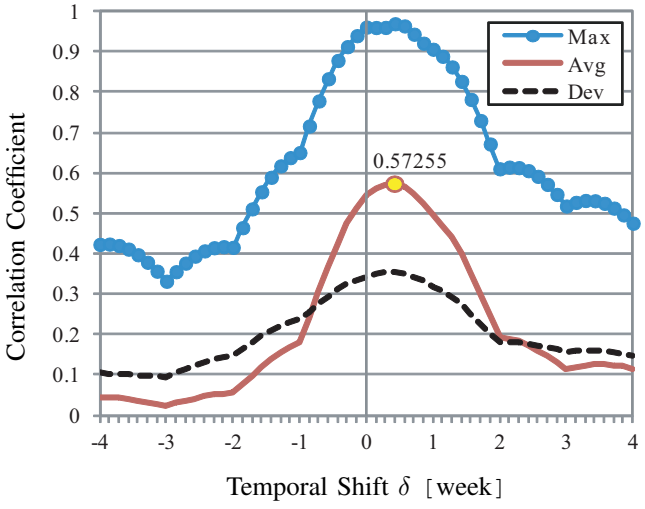

Fig. 10. Dependency of temporally-shifted Web Sensor, ws-query$_\delta$, using Web queries collected by Google Trends on temporal shift δ for earthquake.

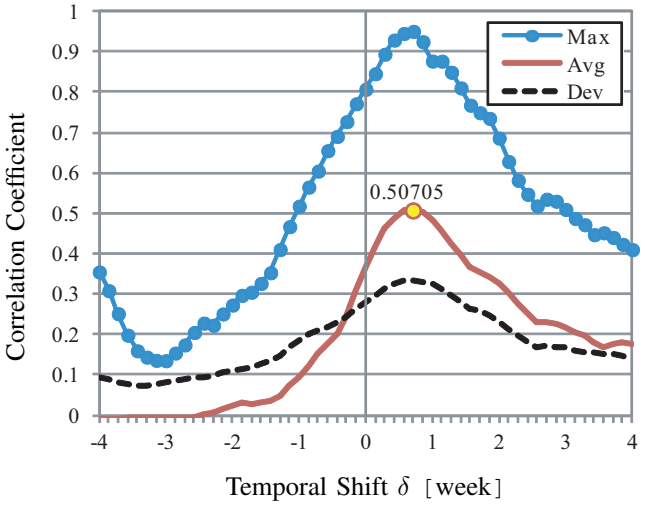

Fig. 13. Dependency of temporally-shifted Web Sensor, ws-doc$_\delta$, using Weblog documents searched by Google on temporal shift δ for earthquake.

Last, Fig. 14 to 16 show the dependency of combined Web Sensor using both Web queries and Weblog documents on temporal shift δ [week] and combination parameter α for a target physical phenomenon, rainfall, snowfall, and earthquake, respectively. And Fig. 17 to 19 show the dependency on only combination parameter α. They show that the best combined Web Sensor for each of three physical phenomena prefers Weblog documents rather than Web queries (i.e., the best combination parameter $0.0 < \alpha < 0.5$), and is superior to (i.e., gains average 5.92% or 8.62% against) uncombined Web Sensor using only Web queries or Weblog documents.

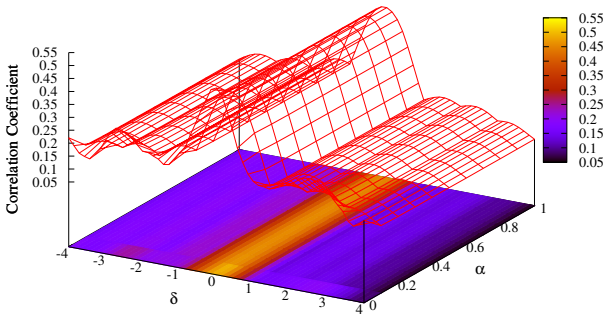

Fig. 14. Dependency of combined Web Sensor, $\mathrm{ws}_\delta^\alpha$, using Web queries and documents on temporal shift δ and combination parameter α for rainfall. (Max 0.53993 when $\delta = \pm 0$ [week] $= \pm 0$ [day] and $\alpha = 0.00211$)

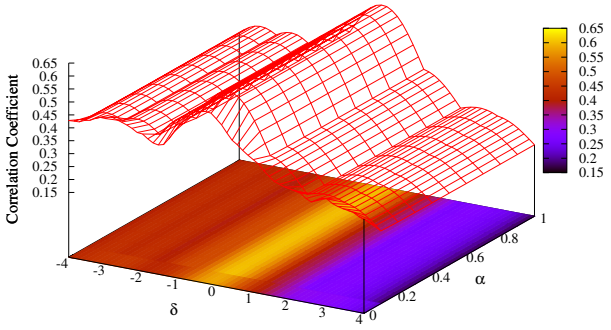

Fig. 15. Dependency of combined Web Sensor, $\mathrm{ws}_\delta^\alpha$, using Web queries and documents on temporal shift δ and combination parameter α for snowfall. (Max 0.61716 when $\delta = -0.14286$ [week] $= -1$ [day] and $\alpha = 0.00907$)

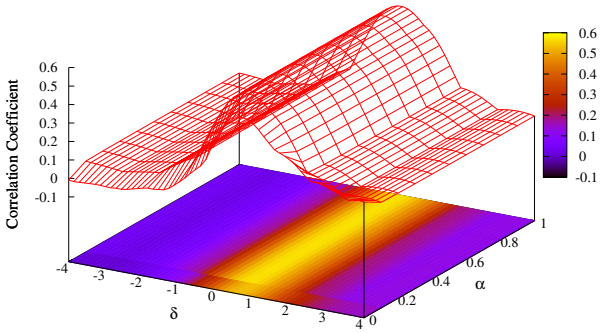

Fig. 16. Dependency of combined Web Sensor, $\mathrm{ws}_\delta^\alpha$, using Web queries and documents on temporal shift δ and combination parameter α for earthquake. (Max 0.57258 when $\delta = +0.42857$ [week] $= +3$ [day] and $\alpha = 0.37330$)

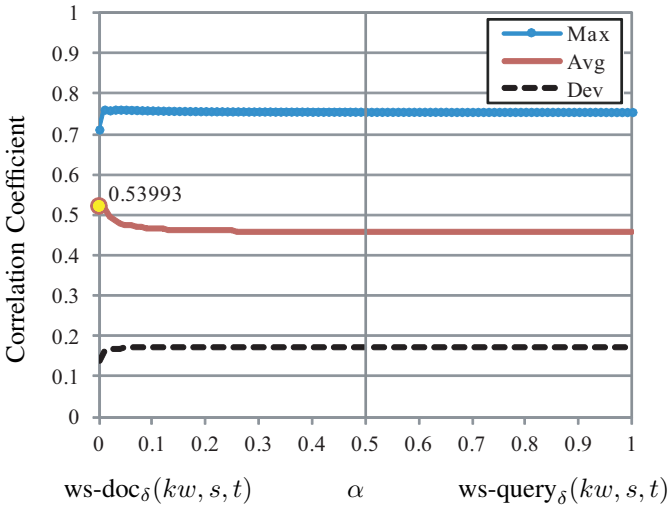

Fig. 17. Dependency of combined Web Sensor, $\mathrm{ws}_\delta^\alpha$, using Web queries and Weblog documents on combination parameter α for rainfall.

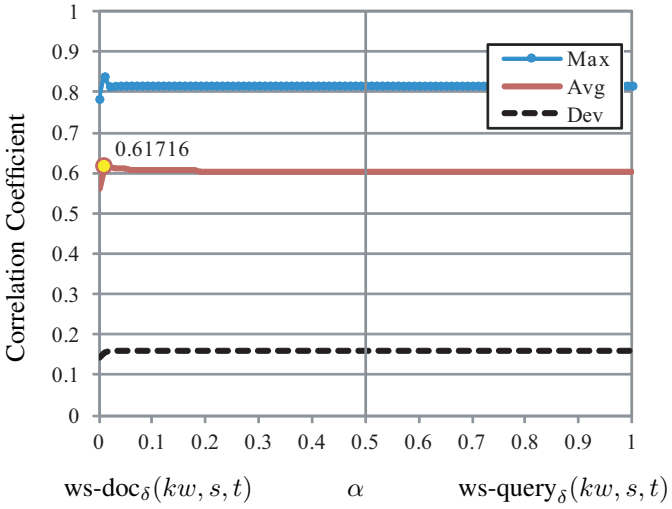

Fig. 18. Dependency of combined Web Sensor, $\mathrm{ws}_\delta^\alpha$, using Web queries and Weblog documents on combination parameter α for snowfall.

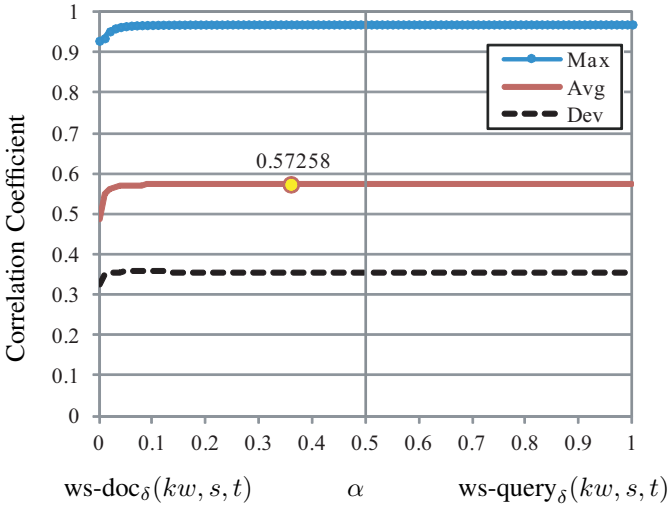

Fig. 19. Dependency of combined Web Sensor, $\mathrm{ws}_\delta^\alpha$, using Web queries and Weblog documents on combination parameter α for earthquake.

IV. Conclusion

In the exploding Web world, one user creates and uploads a Web document about a phenomenon or event in the physical world, while another user retrieves and consumes Web documents by submitting a Web query. The previous papers [15]–[18] introduced various Web Sensors, ws-doc(kw, s, t), to extract spatiotemporal data about a target phenomenon (e.g., rainfall, snowfall, and earthquake) from Web documents searched by keyword(s) kw representing the target phenomenon for such a space s as 47 prefectures in Japan and such a time period t as a day and a week.

This paper has defined the novel kind of spatio-temporal Web Sensors, ws-query(kw, s, t), by analyzing not Web documents but Web queries, which are submitted to Web search engines such as Google to retrieve Web documents about a target phenomenon in the physical world and can be analyzed by Google Trends [25], and also compared and combined them with spatio-temporal Web Sensors by analyzing Web documents, which are created and uploaded to the Web and can be searched by Google Web Search [24], with respect to coefficient correlation with physical-world statistics per week by region of Japan Meteorological Agency (JMA) [23].

The comparison has showed that ws-query is superior to ws-doc for snowfall and earthquake, but inferior for rainfall. And that the best combined Web Sensor prefers Web documents rather than Web queries, and is superior to (i.e., gains average 5.92% or 8.62% against) uncombined Web Sensor using only Web queries or Web documents.

The future work will try to optimize the combined Web Sensor's parameters α and δ depending on physical phenomena, and complement lost data of JMA's statistics by Web Sensors.

Acknowledgment

This work was supported in part by JSPS Grant-in-Aid for Young Scientists (B) "A research on Web Sensors to extract spatio-temporal data from the Web" (#23700129, Project Leader: Shun Hattori, 2011-2012).

References

[1] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," in *Proc. 12th Int'l World Wide Web Conf. (WWW)*, Hungary, 2003, pp. 519–528.

[2] S. Fujimura, M. Toyoda, and M. Kitsuregawa, "A reputation extraction method considering structure of sentence," in *Proc. 16th IEICE Data Engineering Workshop (DEWS)*, Japan, 2005, 6C-i8.

[3] T. Tezuka, T. Kurashima, and K. Tanaka, "Toward tighter integration of web search with a geographic information system," in *Proc. 15th Int'l World Wide Web Conf. (WWW)*, Scotland, 2006, pp. 277–286.

[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," in *Proc. 7th IEEE/WIC/ACM Int'l Conf. on Web Intelligence (WI)*, Australia, 2008, pp. 314–321.

[5] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proc. 14th Int'l Conf. on Computational Linguistics (COLING)*, France, 1992, vol. 2, pp. 539–545.

[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, June 2007.

[7] S. Hattori, H. Ohshima, S. Oyama, and K. Tanaka, "Mining the web for hyponymy relations based on property inheritance," in *Proc. 10th Asia-Pacific Web Conf. (APWeb)*, China, 2008, LNCS vol. 4976, pp. 99–110.

[8] S. Hattori and K. Tanaka, "Extracting concept hierarchy knowledge from the web based on property inheritance and aggregation," in *Proc. 7th IEEE/WIC/ACM Int'l Conf. on Web Intelligence (WI)*, Australia, 2008, pp. 432–437.

[9] S. Hattori, "Hyponym extraction from the web based on property inheritance of text and image features," in *Proc. 6th Int'l Conf. on Advances in Semantic Processing (SEMAPRO)*, Spain, 2012, pp. 109–114.

[10] S. Hattori, "Object-oriented semantic and sensory knowledge extraction from the web," in *Web Intelligence and Intelligent Agents*, In-Tech, 2010, ch. 18, pp. 365–390.

[11] T. Tezuka and K. Tanaka, "Visual description conversion for enhancing search engines and navigational systems," in *Proc. 8th Asia-Pacific Web Conf. (APWeb)*, China, 2006, LNCS vol. 3841, pp. 955–960.

[12] S. Hattori, T. Tezuka, and K. Tanaka, "Mining the web for appearance description," in *Proc. 18th Int'l Conf. on Database and Expert Systems Applications (DEXA)*, Germany, 2007, LNCS vol. 4653, pp. 790–800.

[13] S. Hattori, "Peculiar image retrieval by cross-language web-extracted appearance descriptions," *Int'l Journal of Computer Information Systems and Industrial Management (IJCISIM)*, MIR Labs, vol. 4, pp. 486–495, December 2011.

[14] S. Hattori, "Hyponymy-based peculiar image retrieval," *Int'l Journal of Computer Information Systems and Industrial Management (IJCISIM)*, MIR Labs, vol. 5, pp. 79–88, June 2012.

[15] S. Hattori and K. Tanaka, "Mining the web for access decision-making in secure spaces," in *Proc. Joint 4th Int'l Conf. on Soft Computing and Intelligent Systems and 9th Int'l Symp. on advanced Intelligent Systems (SCIS&ISIS)*, Japan, 2008, TH-G3-4, pp. 370–375.

[16] S. Hattori, "Spatio-temporal web sensors by social network analysis," in *Proc. 3rd Int'l Workshop on Business Applications of Social Network Analysis (BASNA)*, Turkey, 2012, pp. 1020–1027.

[17] S. Hattori, "Secure spaces and spatio-temporal weblog sensors with temporal shift and propagation," in *Proc. 1st IRAST Int'l Conf. on Data Engineering and Internet Technology (DEIT)*, Indonesia, 2011, LNEE vol. 157, pp. 343–349.

[18] S. Hattori, "Linearly-combined web sensors for spatio-temporal data extraction from the web," in *Proc. 6th Int'l Workshop on Spatial and Spatiotemporal Data Mining (SSTDM)*, Canada, 2011, pp. 897–904.

[19] 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, September 2007.

[20] S. Hattori and K. Tanaka, "Secure spaces: protecting freedom of information access in public places," in *Proc. 5th Int'l Conf. on Smart Homes and Health Telematics (ICOST)*, Japan, 2007, LNCS vol. 4541, pp. 99–109.

[21] 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, February 2012.

[22] S. Hattori, "Ability-based expression control for secure spaces," in *Proc. Joint 6th Int'l Conf. on Soft Computing and Intelligent Systems and 13th Int'l Symp. on advanced Intelligent Systems (SCIS&ISIS)*, Japan, 2012, F1-54-3, pp. 1298–1303.

[23] Japan Meteorological Agency, http://www.jma.go.jp/jma/indexe.html.

[24] Google Web Search, http://www.google.co.jp/.

[25] Google Trends, http://www.google.co.jp/trends/.

Shun Hattori received his B.E., M.I., and Ph.D. degrees in Informatics from Kyoto University, Japan, in 2004, 2006, and 2009, respectively. From April to September 2009, he was a Researcher at Geosphere Research Institute of Saitama University (GRIS), Japan, where he was involved in development of an earthquake report system "ZiSyn." From October 2009 to February 2012, he was an Assistant Professor at School of Computer Science, Tokyo University of Technology, Japan. In March 2012, he joined College of Information and Systems, Muroran Institute of Technology, Japan, where he is an Assistant Professor and has Web Intelligence Time-Space (WITS) Laboratory currently. His research interests include Web search, Web mining, information security (access control), and educational engineering, especially in mobile/ubiquitous computing environments. He is a member of the IPSJ, IEICE, DBSJ, and IEEE.