World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:7, No:2, 2013

Granularity Analysis for Spatio-Temporal Web Sensors

Shun Hattori

Abstract**—**In recent years, many researches to mine the exploding Web world, especially User Generated Content (UGC) such as weblogs, for knowledge about various phenomena and events in the physical world have been done actively, and also Web services with the Web-mined knowledge have begun to be developed for the public. However, there are few detailed investigations on how accurately Web-mined data reflect physical-world data. It must be problematic to idolatrously utilize the Web-mined data in public Web services without ensuring their accuracy sufficiently. Therefore, this paper introduces the simplest Web Sensor and spatiotemporally-normalized Web Sensor to extract spatiotemporal data about a target phenomenon from weblogs searched by keyword(s) representing the target phenomenon, and tries to validate the potential and reliability of the Web-sensed spatiotemporal data by four kinds of granularity analyses of coefficient correlation with temperature, rainfall, snowfall, and earthquake statistics per day by region of Japan Meteorological Agency as physical-world data: spatial granularity (region's population density), temporal granularity (time period, e.g., per day vs. per week), representation granularity (e.g., "rain" vs. "heavy rain"), and media granularity (weblogs vs. microblogs such as Tweets).

Keywords**—**Granularity analysis, knowledge extraction, spatiotemporal data mining, Web credibility, Web mining, Web sensor.

I. Introduction

THE former Web world did not have a familiar relationship with the physical world, and it is not too much to say that the former Web world was isolated and independent from the physical world. But in recent years, the explosively-growing Web world has had more and more familiar relationships with the physical world as the use of the World Wide Web (WWW) on the Internet, especially User Generated Content (UGC) such as weblogs, Word of Mouth (WOM) sites, and Social Networking Services (SNS), has become more popular with various people without distinction of age/sex.

Many researches to mine the exploding Web, especially the Weblog, for knowledge about various phenomena and events in the physical world have been done actively. 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 [5–10] such as is-a/has-a relationships and visual appearance (look and feel) extraction [9, 11–14] of physical objects in the physical world. Meanwhile, Web services with the Web-mined knowledge have begun to be developed for the public, and more and more ordinary people actually utilize them as very important information for choosing better products, services, and actions in the physical world.

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 (see http://www3.muroran-it.ac.jp/wits/~hattori/).

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.

This paper introduces the concept of **Web Sensors** [15–18], the simplest/spatiotemporally-normalized ones, to extract spatiotemporal data about such a target phenomenon as temperature, rainfall, snowfall, and earthquake from Web documents searched by keyword(s) representing the target phenomenon, and carries out 4 kinds of granularity analyses of coefficient correlation with 4 kinds of physical-world statistics per day by region of Japan Meteorological Agency (JMA) [19] to validate the potential and reliability of the Web-sensed spatiotemporal data for such a space as 47 prefectures and 47 prefectural capitals in Japan and such a time period as a day and a week in 2011. The four kinds of granularity analyses include

- **Space** Granularity Analysis: analyzes the spatial dependency of coefficient correlation between Web-sensed spatiotemporal data and JMA's stats on space's population density. The other examples of spatial features include population, land area, and geographic location.
- **Time** Granularity Analysis: analyzes the temporal dependency of coefficient correlation between Web-sensed spatiotemporal data and JMA's stats on time's period, e.g., per day vs. per week.
- **Representation** Granularity Analysis: analyzes the hyponymy dependency of coefficient correlation between Web-sensed spatiotemporal data and JMA's stats on a coarse keyword ("rain") vs. a fine keyword ("heavy rain") representing a target phenomenon (e.g., rainfall).
- **Media** Granularity Analysis: analyzes the media dependency of coefficient correlation between Web-sensed spatiotemporal data and JMA's stats on weblogs vs. microblogs such as Tweets. The number of Weblog documents is about 50 times more than the number of Twitter (as one of microblogging sites) documents in 2011. And Tweets are restricted up to 140 characters.

The remainder of this paper is organized as follows. Section II introduces the simplest Web Sensor and spatiotemporally-normalized Web Sensor in Secure Spaces. Section III validates the potential and reliability of the Web-sensed spatiotemporal data by granularity analyses. Section IV concludes this paper.

II. Method: Web Sensors in Secure Spaces

In public spaces, there are a number of different contents such as visitors, and physical information resources, and virtual information resources via their embedded output devices. Therefore, we might unexpectedly enter the public spaces that have our unauthorized contents and/or unwanted characteristics, i.e., they are convenient and comfortable for somebody but not always secure and safe for all of us. To solve this problem, my previous works [15, 20–23] have introduced the concept of **Secure Spaces**, physical environments in which any visitor is protected from being pushed her unwanted information resources on and also any information resource is always protected from being accessed by its unauthorized visitors, and the model and architecture for space entry control and information access control based on their dynamically changing contents.

To build Secure Spaces in the physical world by using space entry control based on their dynamically changing contents such as their visitors, physical/virtual information resources via their embedded output devices, each Secure Space requires the following facilities (as shown in Fig. 1).

- **Space Management**: is responsible for managing a Secure Space, i.e., for constantly figuring out its contents such as its visitors, its embedded physical information resources and virtual information resources outputted via its embedded output devices and also for ad-hoc making an authorization decision on whether an entry request to enter the Secure Space by a visitor or a physical/virtual information resource should be granted or denied, and for notifying the entry decisions to the Electrically Lockable Doors or enforcing entry control over virtual information resources according to the entry decisions by itself.
- **User/Object Authentication**: is responsible for authenticating what physical entity such as a user or a physical information resource requests to enter or exit the Secure Space (e.g., by using Radio Frequency IDentification or biometrics technologies) and also for notifying it to the Space Management.
- **Electrically Lockable Door**: is responsible for electrically locking or unlocking itself, i.e., for assuredly enforcing entry control over physical entities such as users and physical information resources, according to instructions by the Space Management.
- **Physically Isolating Opaque Wall**: is responsible for physically isolating inside a Secure Space from outside there with regard to information access, i.e., for validating the basic assumption that any user inside a Secure Space can access any resource inside the Secure Space while any user outside the Secure Space can never any resource inside the Secure Space.

To protect us from our unwanted characteristics (e.g., degrees of congestion, dismal, and danger) of physical spaces as well as our unauthorized contents, the following additional facilities are required.

- **Real Sensor**: is responsible for physically sensing inside a Secure Space for its physical characteristics to make access decisions in the Secure Space and also for notifying the sensor data stream to the Space Management. For example, thermometers, hygrometers, (security) cameras.
- **Web Sensor**: is responsible for logically sensing the Weblog for the approximate characteristics of each Secure Space to make access decisions in the Secure Space and also for notifying the Web-mined data to the Space Management. Note that any Secure Space does not have to equip the extra devices unlike Real Sensors.

This paper introduces two kinds of Web Sensors from my previous works [15–18], the simplest Web Sensor and spatiotemporally-normalized Web Sensor, to extract spatiotemporal data about such a target phenomenon as temperature, rainfall, snowfall, and earthquake from Web documents searched by keyword(s) representing the target phenomenon.

First, the simplest Web Sensor with a geographic space s, e.g., one of 47 prefectures such as “北海道” (Hokkaido) and 47 prefectural capitals such as “札幌市” (Sapporo City), a time period t, e.g., per day and per week in 2011, and a Japanese keyword kw representing a target phenomenon in the physical world, e.g., “暑い” (hot for temperature), “雨” (rain), “雪” (snow), and “地震” (earthquake), by analyzing a corpus c of Web documents, the Weblog or Twitter (one of microblogging sites), is defined as

$$\mathrm{ws}_0^c(kw, s, t) := \mathrm{df}_t^c(\texttt{["}kw\texttt{" \& "}s\texttt{"]}), \tag{1}$$

where $\mathrm{df}_t^c(\texttt{[}q\texttt{]})$ stands for the Frequency of Web Documents searched from the corpus c by submitting the search query q with the custom time range t to Google Web Search [24], and `&` stands for an `AND` operator.

Next, the spatiotemporally-normalized Web Sensor by the frequency $\mathrm{df}_t^c(\texttt{["}s\texttt{"]})$ of Web documents from the corpus c by submitting the geographical space s with the custom time range t to Google Web Search to clean up spatio-temporal dependency is defined as

$$\mathrm{ws}_1^c(kw, s, t) := \mathrm{ws}_0^c(kw, s, t) \ / \ \mathrm{df}_t^c(\texttt{["}s\texttt{"]}). \tag{2}$$

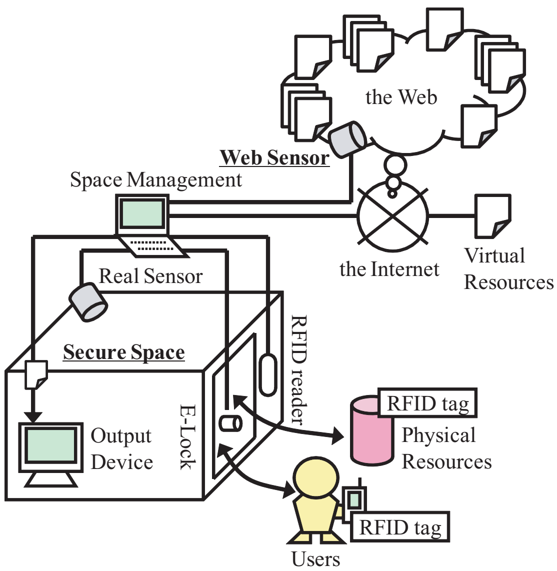


Fig. 1 Spatio-temporal Web Sensors in Secure Spaces

III. Experiment: Granularity Analyses

This section carries out 4 kinds of granularity analyses of coefficient correlation with 4 kinds of physical-world statistics per day by region of Japan Meteorological Agency (JMA) [19] to validate the potential and reliability of the Web-sensed spatiotemporal data for such a space as 47 prefectures and 47 prefectural capitals in Japan and such a time period as a day and a week in 2011. Fig. 2 shows various different features of the four kinds of target phenomena in the physical world.

1) Temperature: changes slowly in all seasons.
2) Rainfall: has spikes in any seasons.
3) Snowfall: has spikes in only winter season.
4) Earthquake: has sharper spikes anytime potentially.

Fig. 2(4) shows that the Web Sensor 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 keep on the Web Sensor as well as the physical world.

Fig. 3 to 6 show the granularity analyses of coefficient correlation between the simplest Web Sensor's spatiotemporal data and JMA's average temperature, rainfall amount, snowfall amount, and number of felt quakes, respectively.

A. Space Granularity Analysis

The right columns of 4 figures (pages) analyze the spatial dependency of coefficient correlation between Web-sensed spatiotemporal data and JMA's stats on space's population density. The smaller the space s is, the larger the deviation of coefficient correlation in the space is.

B. Time Granularity Analysis

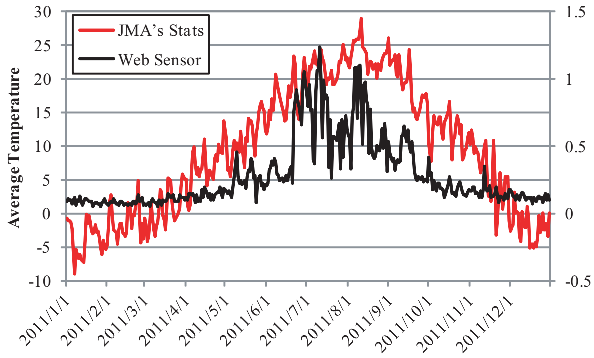
The left columns of 4 figures (pages) analyze the temporal dependency of coefficient correlation between Web-sensed spatiotemporal data and JMA's stats on time's period. The larger the time period t is, the larger the average, maximum, and deviation of coefficient correlation in the time period are.

C. Representation Granularity Analysis

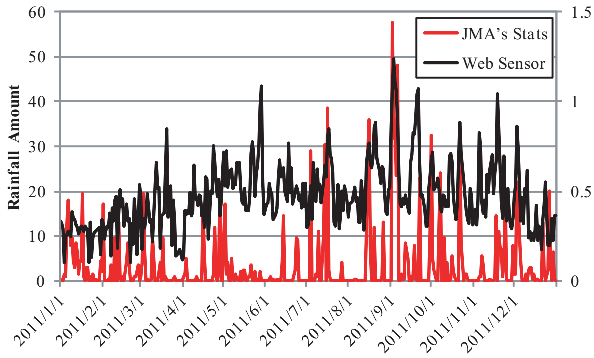
The (a) vs. (b) and (c) vs. (d) of 3 figures except Fig. 3 analyze the hyponymy dependency of coefficient correlation between Web-sensed spatiotemporal data and JMA's stats on a coarse keyword (e.g., "rain") vs. a fine keyword (e.g., "heavy rain") representing a target phenomenon (e.g., rainfall). The finer the keyword kw representing a target phenomenon is, the larger the average and maximum of coefficient correlation by Web Sensors with the keyword are.

D. Media Granularity Analysis

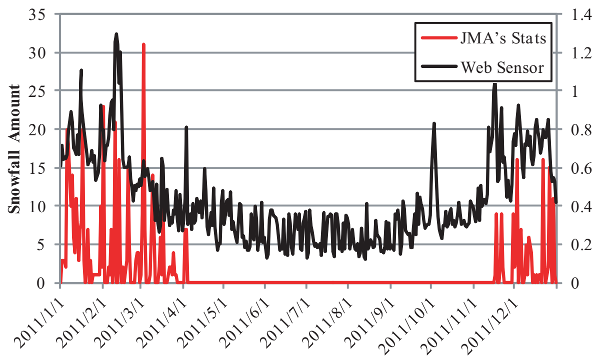
The (a) vs. (c) and (b) vs. (d) of 4 figures (pages) analyze the media dependency of coefficient correlation between Web-sensed spatiotemporal data and JMA's stats on weblogs vs. microblogs such as Tweets. Weblog documents tend to be superior to Twitter (microblog) documents for Web Sensors to extract spatiotemporal data about physical-world phenomena from the Web.



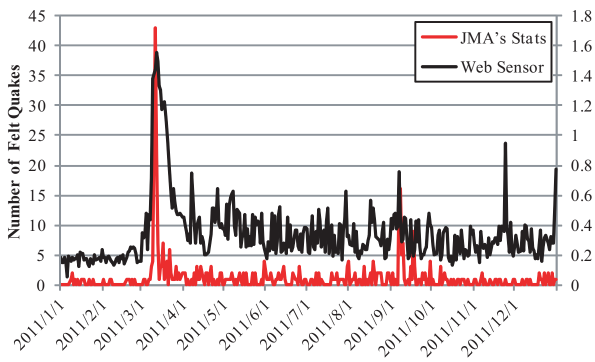
(1) Temperature



(2) Rainfall

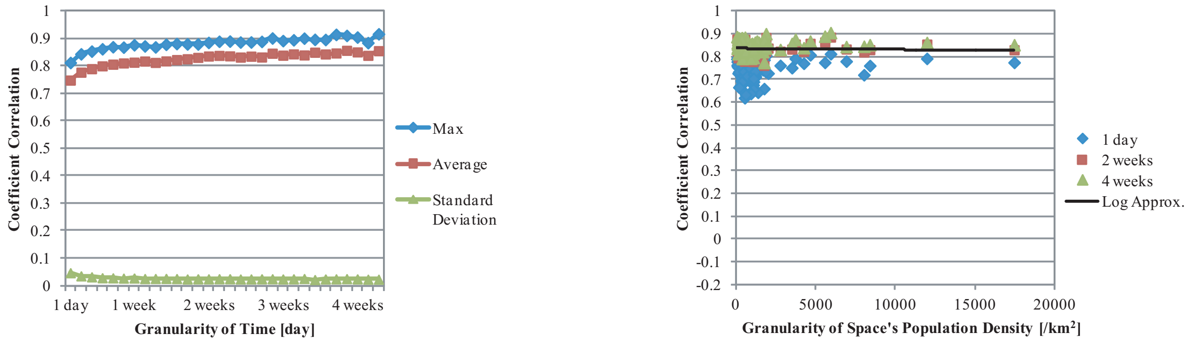


(3) Snowfall

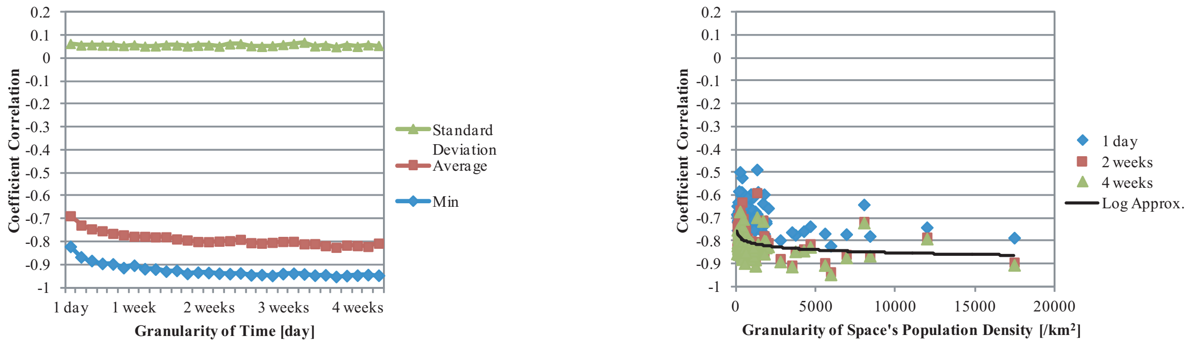


(4) Earthquake

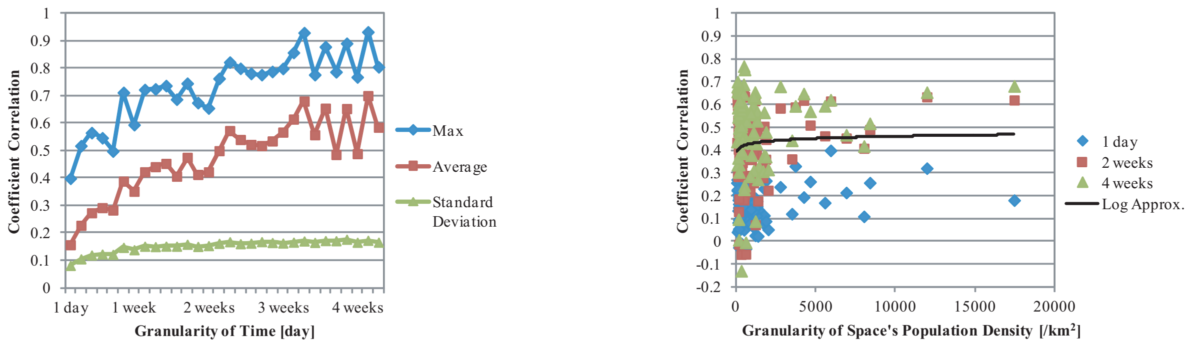
Fig. 2 JMA's daily statistics and Web Sensor's spatiotemporal data for four physical-world phenomena in Hokkaido prefecture, 2011



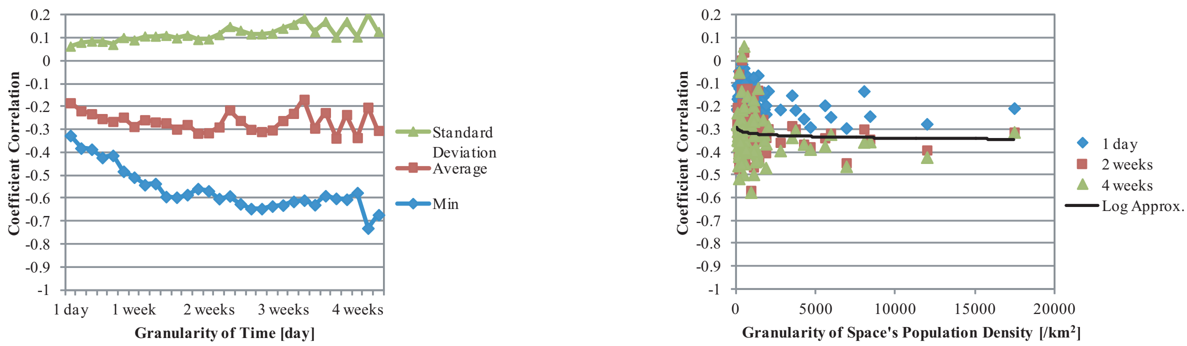
(a) using Blog documents searched by a positive keyword kw = “暑い” (hot)



(b) using Blog documents searched by a negative keyword kw = “寒い” (cold)

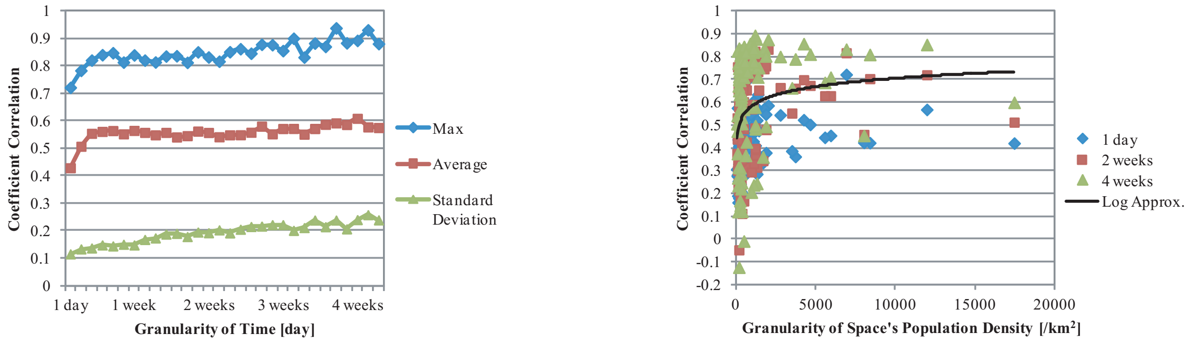


(c) using Twitter (Microblog) documents searched by a positive keyword kw = “暑い” (hot)

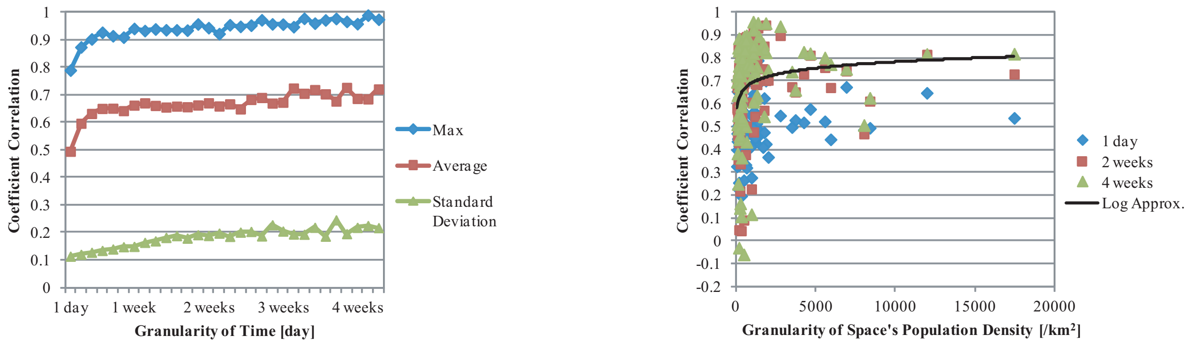


(d) using Twitter (Microblog) documents searched by a negative keyword kw = “寒い” (cold)

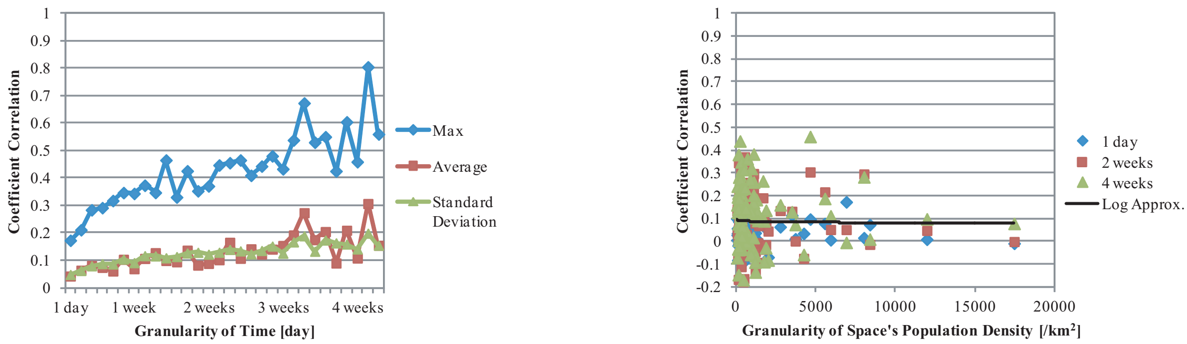
Fig. 3 Granularity analyses of coefficient correlation between Web Sensor’s spatiotemporal data and JMA’s average temperature



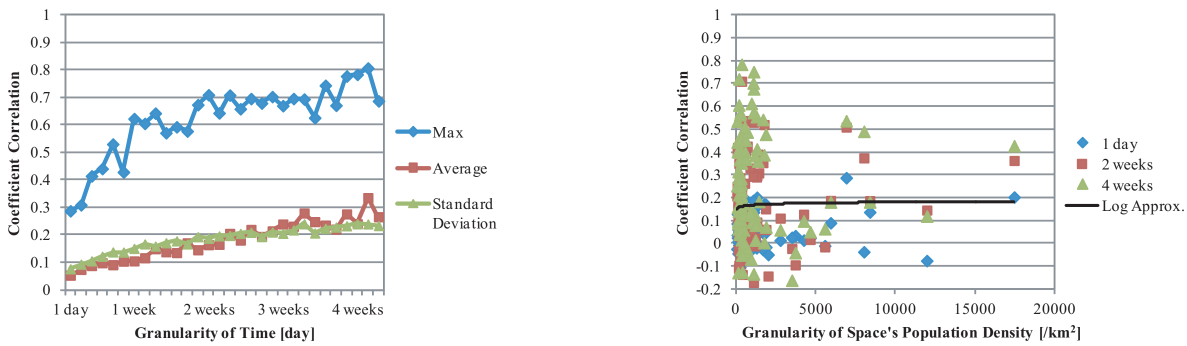
(a) using Blog documents searched by a coarse keyword kw = “雨” (rain)



(b) using Blog documents searched by a fine keyword kw = “大雨” (heavy rain)

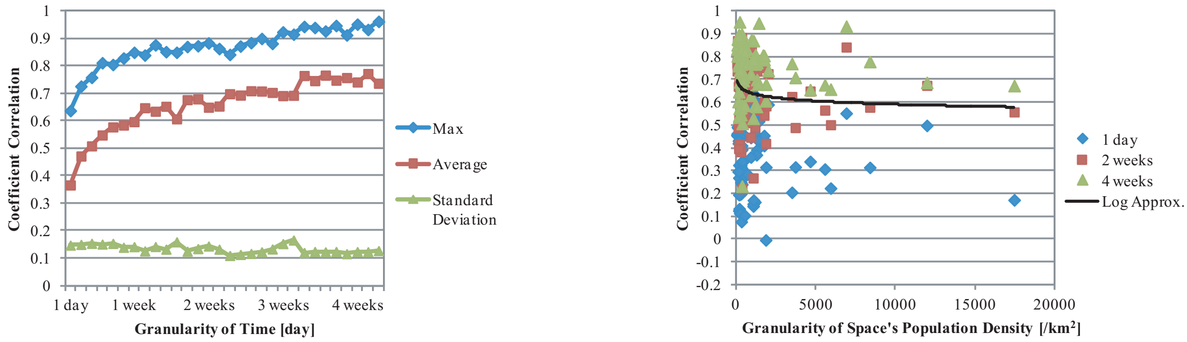


(c) using Twitter (Microblog) documents searched by a coarse keyword kw = “雨” (rain)

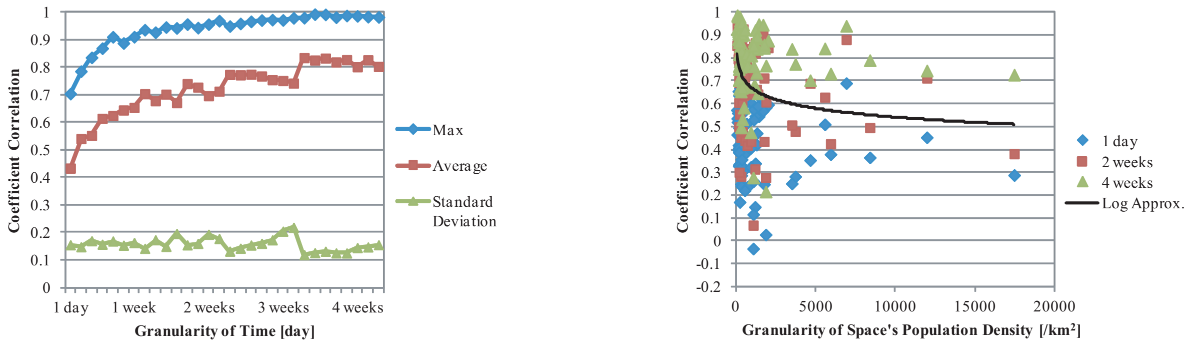


(d) using Twitter (Microblog) documents searched by a fine keyword kw = “大雨” (heavy rain)

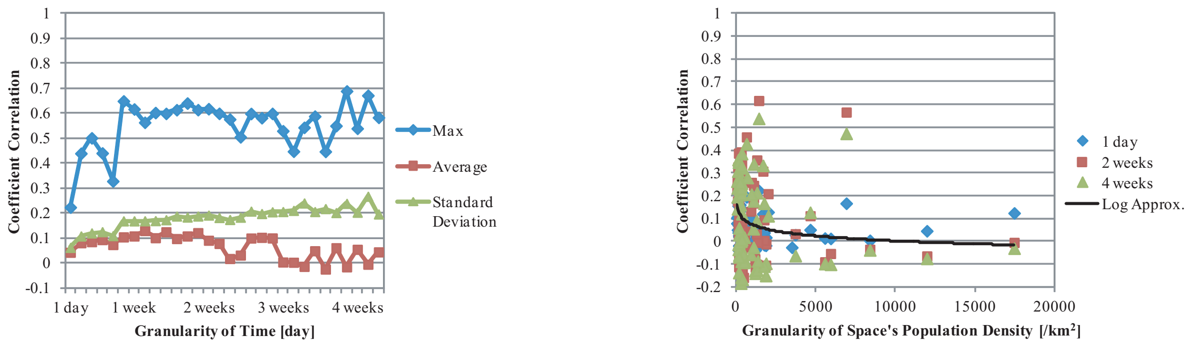
Fig. 4 Granularity analyses of coefficient correlation between Web Sensor’s spatiotemporal data and JMA’s rainfall amount



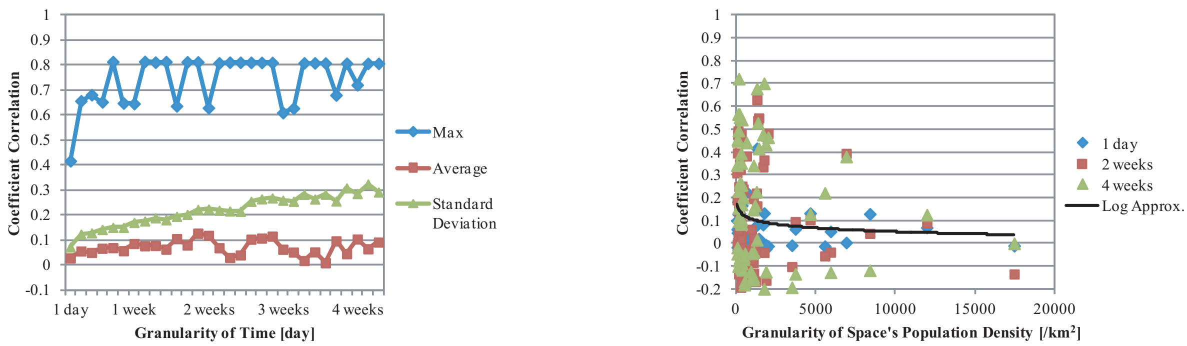
(a) using Blog documents searched by a coarse keyword kw = “雪” (snow)



(b) using Blog documents searched by a fine keyword kw = “大雪” (heavy snow)

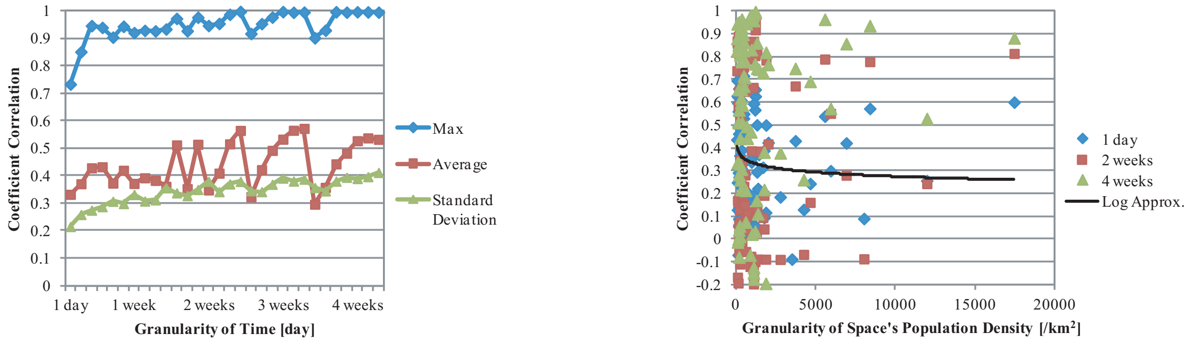


(c) using Twitter (Microblog) documents searched by a coarse keyword kw = “雪” (snow)

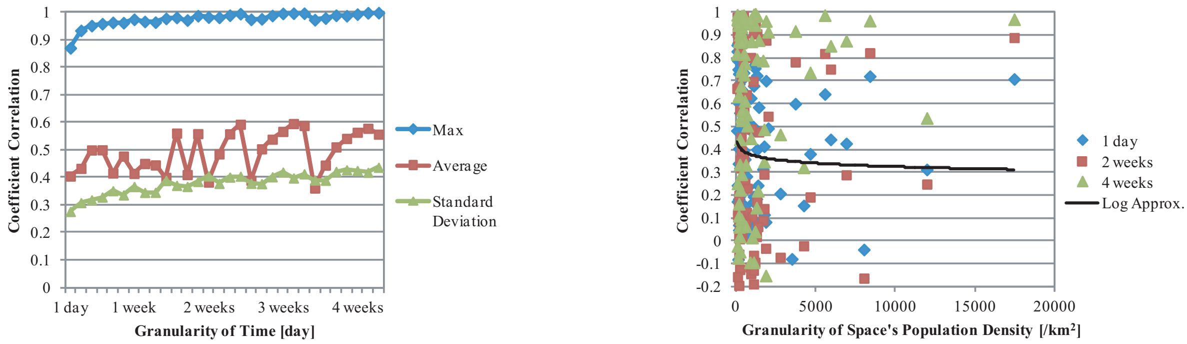


(d) using Twitter (Microblog) documents searched by a fine keyword kw = “大雪” (heavy snow)

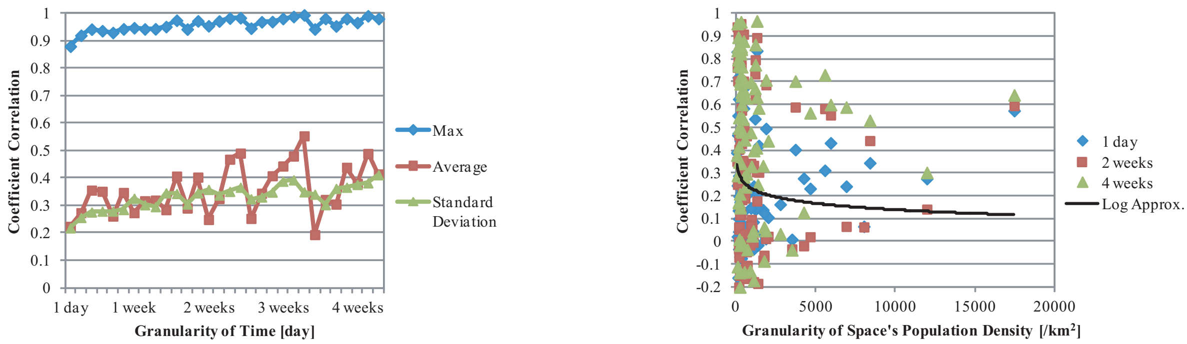
Fig. 5 Granularity analyses of coefficient correlation between Web Sensor’s spatiotemporal data and JMA’s snowfall amount



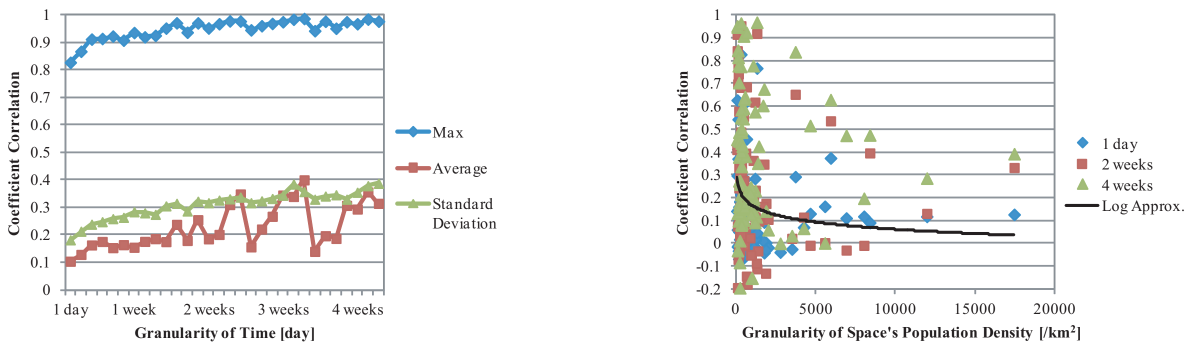
(a) using Blog documents searched by a coarse keyword kw = “地震” (earthquake)



(b) using Blog documents searched by a fine keyword kw = “大地震” (huge earthquake)



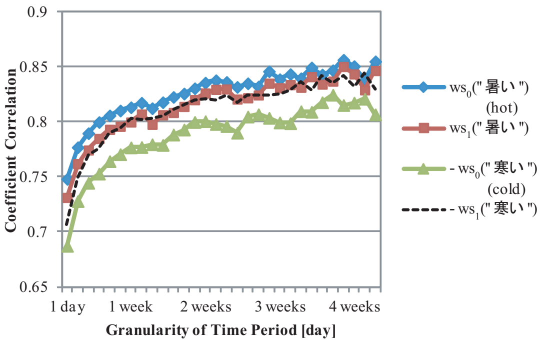
(c) using Twitter (Microblog) documents searched by a coarse keyword kw = “地震” (earthquake)



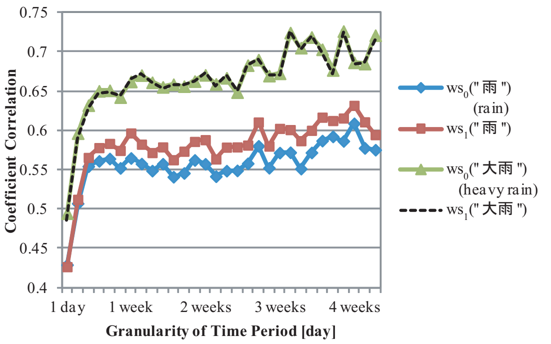
(d) using Twitter (Microblog) documents searched by a fine keyword kw = “大地震” (huge earthquake)

Fig. 6 Granularity analyses of coefficient correlation between Web Sensor’s spatiotemporal data and JMA’s number of felt earthquakes

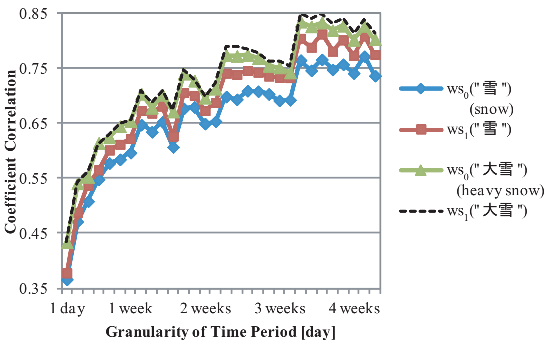
Fig. 7 compares the simplest and spatiotemporally-normalized Web Sensors with weblogs for four physical-world phenomena by Time and Representation granularity analyses.



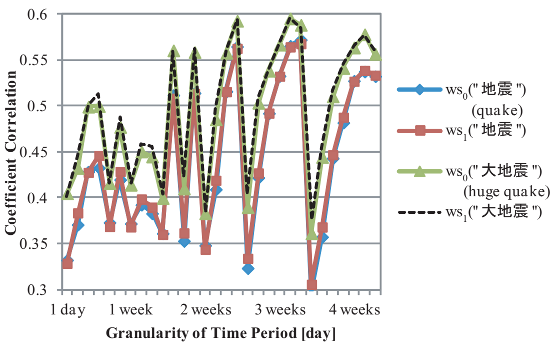
(1) with JMA's Average Temperature



(2) with JMA's Rainfall Amount



(3) with JMA's Snowfall Amount



(4) with JMA's Number of Felt Quakes

Fig. 7 Comparison of simple and normalized Web Sensors

It shows that the spatiotemporally-normalized Web Sensor is slightly superior to the simplest Web Sensor, and that both Web Sensors give better performance for a longer (coarser) time period and/or with a finer keyword. And it also shows that spatio-temporal Web Sensors indicate periodically for number of felt earthquakes, but increase gradually for the other physical-world phenomena.

Fig. 8 and 9 show the spatial distribution (on 47 prefectures in Japan) of coefficient correlation between Web Sensor's spatiotemporal data and JMA's daily statistics for rainfall amount and number of felt earthquakes, respectively. They show that the spatial distribution for rainfall amount is more uniform than for number of felt earthquakes, and that the farther the space (prefecture) is from the Great East Japan Earthquake on March 11th, 2011 (or the less felt earthquakes the space has), the lower the coefficient correlation between Web Sensor's spatiotemporal data and JMA's daily earthquake stats for the space is.

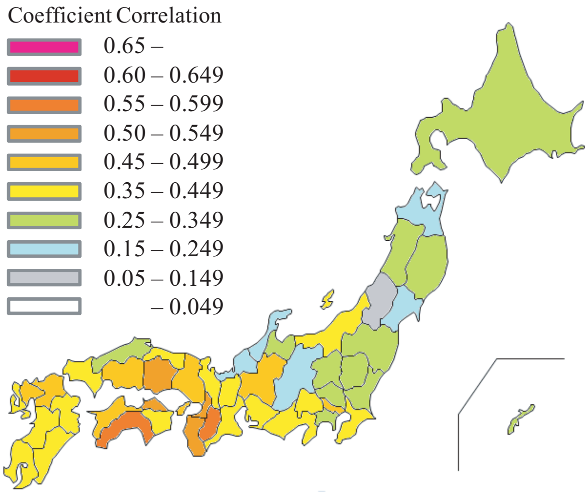


Fig. 8 Spatial distribution of coefficient correlation between Web Sensor's spatiotemporal data and JMA's daily rainfall statistics

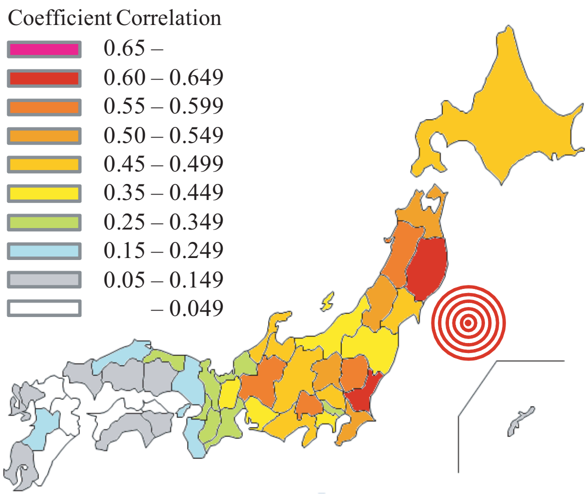


Fig. 9 Spatial distribution of coefficient correlation between Web Sensor's spatiotemporal data and JMA's daily earthquake statistics

IV. Conclusion

This paper has introduced the simplest Web Sensor and spatiotemporally-normalized Web Sensor to extract spatiotemporal data about a target phenomenon in the physical world from Weblog documents searched by keyword(s) representing the target phenomenon. And also this paper has tried to validate the potential and reliability of the Web-sensed spatiotemporal data by carrying out 4 kinds of granularity analyses of coefficient correlation with temperature, rainfall, snowfall, and earthquake statistics per day by region of Japan Meteorological Agency (JMA) as physical-world data:

- Spatial granularity analysis (region's population density),
- Temporal granularity analysis (time period, e.g., per day vs. per week vs. per month),
- Representation granularity analysis (e.g., a coarse keyword "rain" vs. a fine keyword "heavy rain"), and
- Media granularity analysis (weblogs vs. microblogs such as Tweets).

The four kinds of granularity analyses conclude that

- The smaller the space is, the larger the deviation of coefficient correlation in the space is,
- The larger the time period is, the larger the average, maximum, and deviation of coefficient correlation in the time period are,
- The finer the keyword representing a target phenomenon is, the larger the average and maximum of coefficient correlation by Web Sensors with the keyword are, and
- Weblog documents tend to be superior to microblog documents for Web Sensors to extract spatiotemporal data about physical-world phenomena from the Web.

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 International World Wide Web Conference (WWW'03)*, Hungary, pp. 519–528, 2003.

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

[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 Conference (WWW'06)*, Scotland, 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," in *Proc. 7th IEEE/WIC/ACM International Conference on Web Intelligence (WI'08)*, Australia, pp. 314–321, 2008.

[5] M. A. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," in *Proc. 14th International Conference on Computational Linguistics (COLING'92)*, France, 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, 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'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," in *Proc. 7th IEEE/WIC/ACM International Conference on Web Intelligence (WI'08)*, Australia, pp. 432–437, 2008.

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

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

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

[12] S. Hattori, T. Tezuka, and K. Tanaka, "Mining the Web for Appearance Description," in *Proc. 18th International Conference on Database and Expert Systems Applications (DEXA'07)*, Germany, 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, Dec. 2011.

[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, June 2012.

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

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

[17] 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'11)*, Canada, pp. 897–904, 2011.

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

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

[20] 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.

[21] S. Hattori and K. Tanaka, "Secure Spaces: Protecting Freedom of Information Access in Public Places," in *Proc. 5th International Conference on Smart Homes and Health Telematics (ICOST'07)*, Japan, LNCS vol. 4541, pp. 99–109, 2007.

[22] 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.

[23] S. Hattori, "Ability-Based Expression Control for Secure Spaces," *Proc. Joint 6th International Conference on Soft Computing and Intelligent Systems and 13th International Symposium on advanced Intelligent Systems (SCIS&ISIS'12)*, Japan, F1-54-3, pp. 1298–1303, 2012.

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

Shun Hattori was born in Amagasaki, Japan on September 1st, 1981. He 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 Computational Intelligence Unit, 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. He is a member of the IPSJ, IEICE, DBSJ, and IEEE.