Spatio-Temporal Propagation for Web Sensors

Shun HATTORI
Muroran Institute of Technology
27–1 Mizumoto-cho, Muroran, Hokkaido 050–85858, JAPAN
hattori@csse.muroran-it.ac.jp

ABSTRACT

We experience or forecast various phenomena (e.g., rain, snow, and earthquake) in the physical world, while we carry out various actions (e.g., blogging, searching, and e-shopping) in the Web world. Many researchers have tried 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 investigations on how accurately Web-mined data reflect physical-world data. It is socially-problematic to utilize Web-mined data in public Web services without ensuring their accuracy sufficiently. The previous papers have introduced "Web Sensors" to extract spatiotemporal numerical values about a physical phenomenon from various kinds of Web documents (e.g., news, blogs, and tweets) searched by linguistic keyword(s) representing the physical phenomenon, and extended Web Sensors with temporal shift and propagation. This paper appends "Spatial Propagation" to Web Sensors, and compares Web sensors with spatial propagation and/or temporal propagation by calculating their correlation coefficients with Japan Meteorological Agency's physically-sensed spatiotemporal statistics.

KEYWORDS

Web Sensor, Web Mining, Knowledge Extraction, Web Credibility, Spatiotemporal Data Mining.

1 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., blogging, searching, and e-shopping) in the Web world. Recently, there have been many researches to mine a huge amount of various documents in the explosively-growing Web, 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 [5–10] such as is-a/has-a relationships, and visual appearance (look and feel) extraction [10–15] 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 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 [19–25] have introduced "Web Sensors" that sense the Web to extract spatiotemporal numerical values about a target phenomenon in the physical world from various action-based data (e.g., blogs, search query logs, and e-shopping history) in the Web world, and investigated how correlated Web-sensed spatiotemporal data are with physically-sensed spatiotemporal data as shown in Figure 1.

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

- The optimal 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 Weblog runs parallel to the physical world) for rainfall, Shifted-to-Future Web Sensor whose temporal shift parameter δ is negative gives the highest correlation coefficient (i.e., the Weblog 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 Weblog follows the physical world) for earthquake,
- The optimal 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 optimal temporal shift parameter δ for snowfall and earthquake varies more widely, and
- More shaken geographical spaces and time periods are given higher correlation coefficient between Web-sensed data and physically-sensed data by the Great East Japan Earthquake (3.11).

Query-based Web Sensor using Web search query logs [24] is superior to Blog-based Web Sensor using Web documents such as blogs for snowfall and earthquake, while Query-based Web Sensor is inferior to Blog-based Web Sensor 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.

Meanwhile, these Web Sensors are being integrated into Smart Spaces [26] and Secure Spaces [19], [20], [27–29] as shown in Figure 2. Secure Spaces assuredly enforce space entry control and information access control based on Web-sensed spatiotemporal data as their approximate characteristics and their changing contents such as visitors, physical information resources, and virtual information resources via their embedded output devices, to protect any visitor from her/his unwanted information resources and also to protect any information resource from its unauthorized visitors.

This paper appends the novel concept of "Spatial Propagation" to Web Sensors with temporal shift and "Temporal Propagation" in Section 2, and compares them by using Japan's rainfall, snowfall, and earthquake statistics [30] per day by region of Japan Meteorological Agency as physically-sensed data in Section 3.

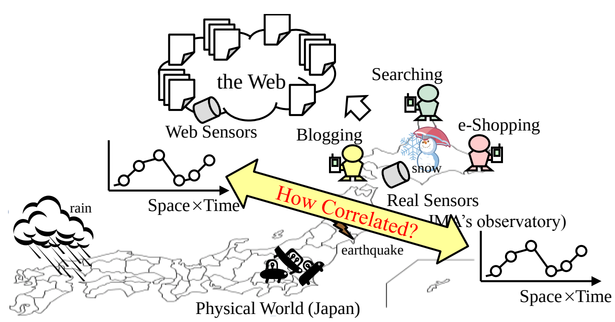

Figure 1. Web Sensors correlate with Real Sensors?

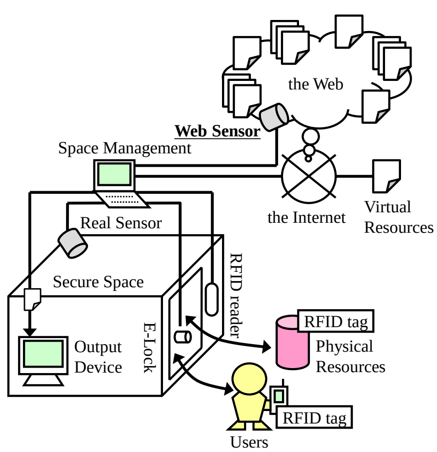

Figure 2. Web Sensors in Secure Spaces.

2 SPATIO-TEMPORAL PROPAGATION

This section appends the novel concept of "Spatial Propagation" to Web Sensors with temporal shift and temporal propagation that 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 target kind of physical phenomenon (e.g., rainfall, snowfall, and earthquake).

First, the simplest and spatiotemporally-normalized Web Sensor [19] by using only Web documents (not Web search query logs) 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 52 weeks in 2011 such as from January 1st to 7th and from December 24th to 30th, and a linguistic keyword kw representing a targeted physical phenomenon, e.g., "rain," "snow," and "earthquake," is defined as

$$\mathrm{ws}(kw, s, t) \coloneqq \frac{\mathrm{df}_t([\,"kw"\,\&\,"s"\,])}{\mathrm{df}_t([\,"s"\,])} \tag{1}$$

where $\mathrm{df}_t([q])$ 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

$$\mathrm{ws}_\delta(kw, s, t) \coloneqq \mathrm{ws}(kw, s, t+\delta) \tag{2}$$

As shown in Figure 3, 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).

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

$$\mathrm{ws}^{\sigma_t^2}(kw, s, t) \coloneqq \sum_{\forall\delta} \mathrm{ws}_\delta(kw, s, t) \cdot p^{\sigma_t^2}(\delta) \tag{3}$$

$$p^{\sigma_t^2}(\delta) \coloneqq N(0, \sigma_t^2, \delta) \tag{4}$$

$$N(\mu, \sigma_t^2, \delta) \coloneqq \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{(\delta-\mu)^2}{2\sigma_t^2}\right) \tag{5}$$

where $N(\mu, \sigma_t^2, \delta)$ stands for a Normal Distribution with a mean μ and a variance σ_t^2. In the experiment, $\forall\delta$ is restricted to $[-30, 30]$.

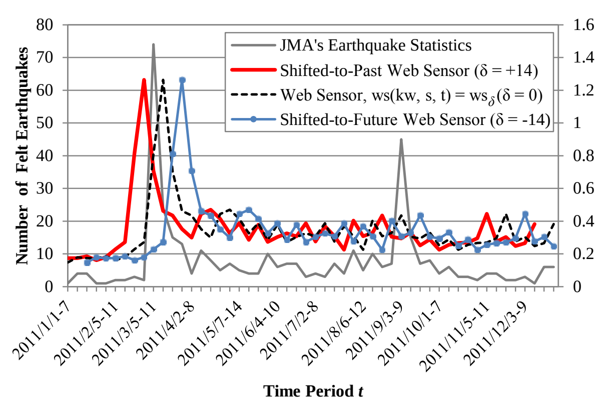

Figure 3. Three kinds of Temporally-Shifted Web Sensors for earthquake and JMA's weekly earthquake statistics in Hokkaido prefecture, 2011.

Next, the novel spatially-propagated Web Sensor 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

$$\mathrm{ws}^{\sigma_s^2}(kw, s, t) \coloneqq \sum_{\forall s_i} \mathrm{ws}(kw, s_i, t) \cdot p^{\sigma_s^2}(\mathrm{distance}(s, s_i)) \quad (6)$$

$$p^{\sigma_s^2}(d) \coloneqq N(0, \sigma_s^2, d) \quad (7)$$

$$N(\mu, \sigma_s^2, d) \coloneqq \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp\left(-\frac{(d-\mu)^2}{2\sigma_s^2}\right) \quad (8)$$

where $\mathrm{distance}(s, s_i)$ stands for the geographic distance [km] between geographic spaces s and s_i and is calculated based on their latitude and longitude. In the experiment, $\forall s_i$ is restricted to 47 prefectures in Japan, and the latitude and longitude of its prefectural capital are used for calculating $\mathrm{distance}(s, s_i)$ by using the Survey Calculation API of Geospatial Information Authority of Japan (GSI) [31]. 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], as shown in Figure 4.

Figure 4. The longest distance between Hokkaido (Sapporo) and Okinawa (Naha), and the shortest distance between Shiga (Otsu) and Kyoto (Kyoto).

Last, the linearly-combined Web Sensor [21] with a combination parameter $\alpha \in [0.0, 1.0]$ by combining the spatially/temporally-propagated Web Sensors is defined as

$$\mathrm{ws}_\alpha^{\sigma_s^2, \sigma_t^2}(kw, s, t) \coloneqq \alpha \cdot \mathrm{ws}^{\sigma_s^2}(kw, s, t) + (1-\alpha) \cdot \mathrm{ws}^{\sigma_t^2}(kw, s, t) \quad (9)$$

3 EXPERIMENT

This section compares the spatially-propagated Web Sensor with the novel concept of "Spatial Propagation" and the temporally-propagated Web Sensor [20] with "Temporal Propagation," by calculating correlation coefficients between their Web-sensed spatiotemporal data and Japan's rainfall, snowfall, and earthquake statistics [30] per day by region of Japan Meteorological Agency (JMA) as physically-sensed spatiotemporal data.

Figures 5 to 7 show the average of correlation coefficients of the temporally-propagated Web Sensor with JMA's daily stats in 2011 for rainfall, snowfall, and earthquake. The best temporally-propagated Web Sensor, which integrates the very near surrounding time periods, is superior to the simplest Web Sensor.

Figures 8 to 10 show the average of correlation coefficients of the spatially-propagated Web Sensor with JMA's daily stats in 2011 for rainfall, snowfall, and earthquake. The best spatially-propagated Web Sensor is superior to both the simplest Web Sensor and the best temporally-propagated Web Sensor.

For rainfall and snowfall, the spatially-propagated Web Sensor draws very similar curves of the average of correlation coefficients with JMA's daily statistics and also its optimal Spatial Propagation parameter σ_s^2 is similar. Meanwhile, the spatially-propagated Web Sensor draws quite a different curve and its optimal Spatial Propagation parameter σ_s^2 is much huger for earthquake. This is caused by the difference between more local physical phenomena (e.g., rainfall and snowfall) and more global ones (e.g., earthquake).

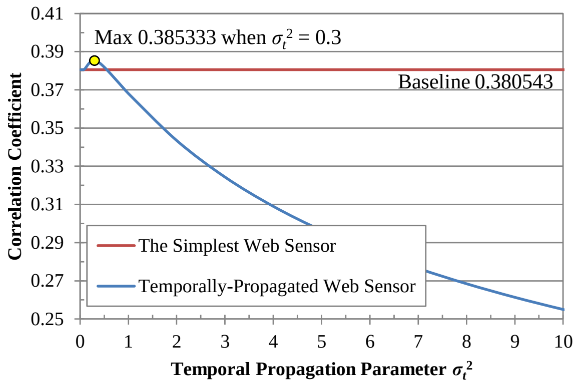

Figure 5. Temporally-Propagated Web Sensor with σ_t^2 vs. the simplest Web Sensor (baseline) for rainfall.

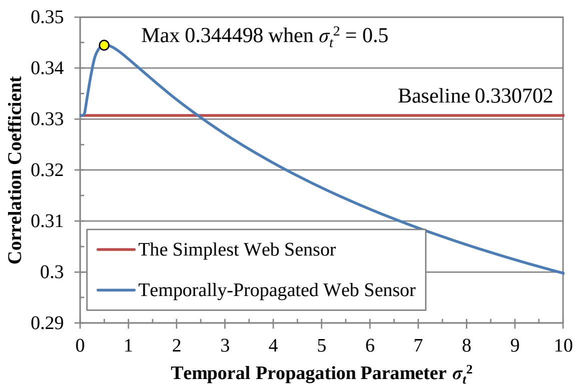

Figure 6. Temporally-Propagated Web Sensor with σ_t^2 vs. the simplest Web Sensor (baseline) for snowfall.

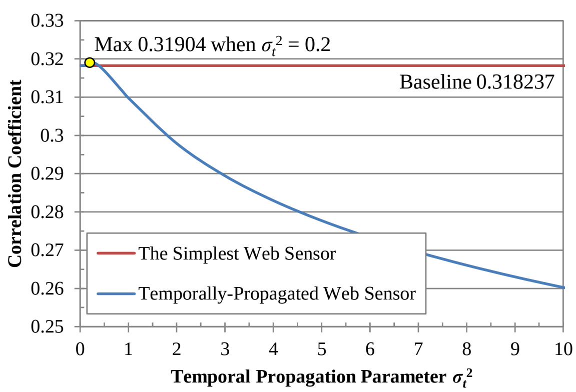

Figure 7. Temporally-Propagated Web Sensor with σ_t^2 vs. the simplest Web Sensor (baseline) for earthquake.

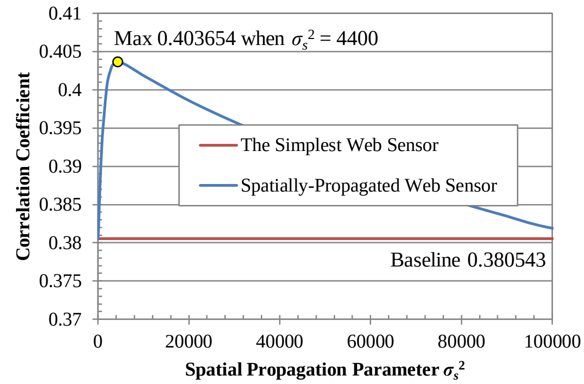

Figure 8. Spatially-Propagated Web Sensor with σ_s^2 vs. the simplest Web Sensor (baseline) for rainfall.

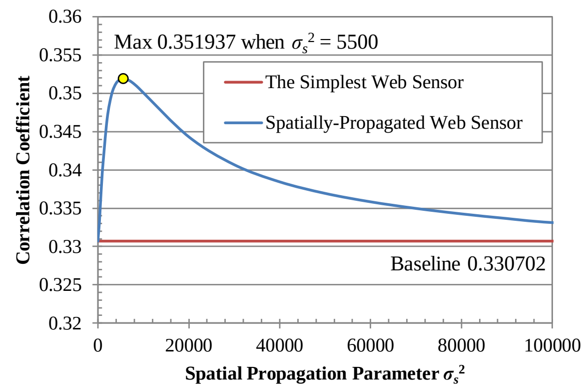

Figure 9. Spatially-Propagated Web Sensor with σ_s^2 vs. the simplest Web Sensor (baseline) for snowfall.

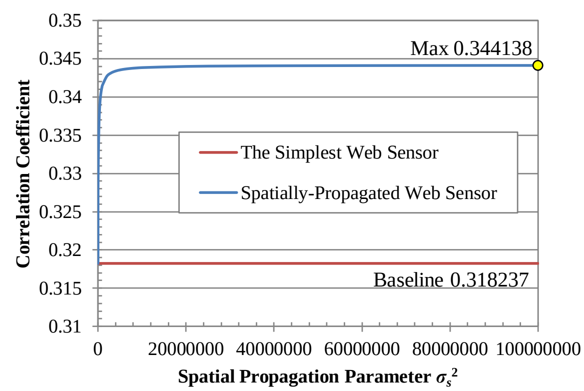

Figure 10. Spatially-Propagated Web Sensor with σ_s^2 vs. the simplest Web Sensor (baseline) for earthquake.

Figures 11 to 13 show the average of correlation coefficients of the linearly-combined Web Sensor of the optimized spatially-propagated Web Sensor and the optimized temporally-propagated Web Sensor with JMA's daily statistics in 2011 for rainfall, snowfall, and earthquake. Only for snowfall, the best linearly-combined Web Sensor when $\alpha = 0.991$ is slightly superior to the spatially-propagated Web Sensor (when $\alpha = 1.0$).

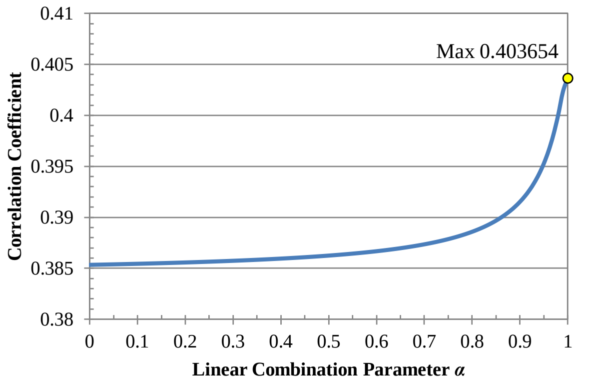

Figure 11. Linearly-Combined Web Sensor with α of Spatially-Propagated Web Sensor with the optimized $\sigma_s^2 = 4400$ and Temporally-Propagated Web Sensor with the optimized $\sigma_t^2 = 0.3$ for rainfall.

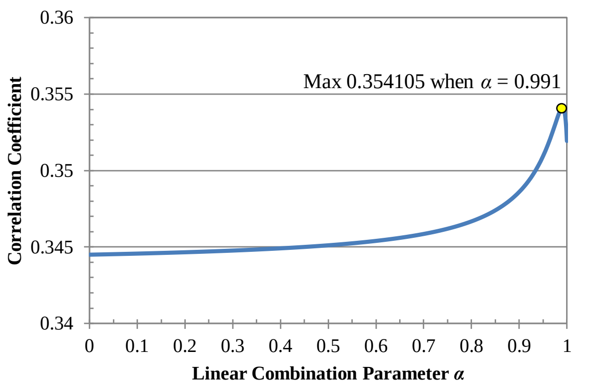

Figure 12. Linearly-Combined Web Sensor with α of Spatially-Propagated Web Sensor with the optimized $\sigma_s^2 = 5500$ and Temporally-Propagated Web Sensor with the optimized $\sigma_t^2 = 0.5$ for snowfall.

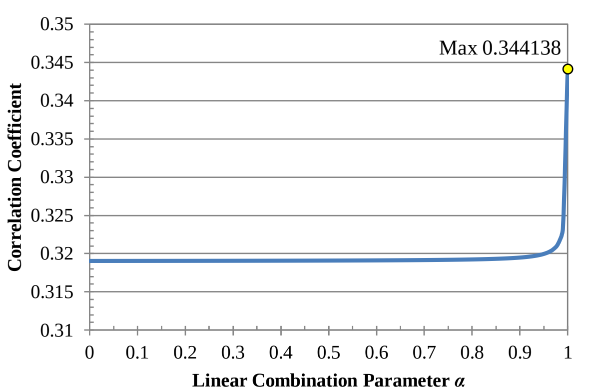

Figure 13. Linearly-Combined Web Sensor with α of Spatially-Propagated Web Sensor with $\sigma_s^2 = 1.0 \cdot 10^8$ and Temporally-Propagated Web Sensor with the optimized $\sigma_t^2 = 0.2$ for earthquake.

4 CONCLUSION

This paper has defined the novel kind of “Web Sensor” to extract spatiotemporal numerical values dependent on a geographic space (e.g., “Hokkaido” as one of 47 prefectures in Japan) and a time period about a targeted physical phenomenon from the Web, especially the Weblog, by integrating the effects from the surrounding geographic spaces (e.g., mainly “Aomori” and “Iwate” for “Hokkaido”), that is, the spatially-propagated Web Sensor with an additional “Spatial Propagation” parameter, and compared various kinds of Web Sensors with spatial propagation and/or temporal propagation by calculating correlation coefficients between their Web-sensed spatiotemporal data and Japan’s rainfall, snowfall, and earthquake statistics per day by region of Japan Meteorological Agency (JMA) as physically-sensed spatiotemporal data.

The comparison shows that for any physical phenomena (rainfall, snowfall, and earthquake), the optimized spatially-propagated Web Sensor is superior to both the simplest Web Sensor and the optimized temporally-propagated Web Sensor. It also shows that only for snowfall, the linearly-combined Web Sensor of the optimized spatially-propagated Web Sensor and the optimized temporally-propagated Web Sensor when $\alpha = 0.991$ is slightly superior to the optimized spatially/temporally-propagated Web Sensor (which is equivalent to the linearly-combined Web Sensor when $\alpha = 1.0$ or 0.0).

The future work will try to apply the other physical phenomena to Web Sensors with Temporal Shift, Temporal Propagation, and Spatial Propagation, and to combine various Web actions, e.g., not only blogging and searching but also e-shopping to construct more high-sensitive Web Sensors. In addition, Web Sensors will be able to forecast future data about a targeted physical phenomenon, to interpolate lost data of real statistics, and to 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," 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 (APWeb'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 Sympposium 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] L. Rosenthal and V. Stanford, “NIST Smart Space: Pervasive Computing Initiative,” Proceedings of the 9th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, pp. 6-11 (2000).

[27] 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).

[28] 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).

[29] S. Hattori, “Ability-Based Expression Control for Secure Spaces,” Proceedings of the Joint 6th International Conference on Soft Computing and Intelligent Systems and 13th International Symposium on advanced Intelligent Systems (SCIS&ISIS’12), F1-54-3, pp. 1298-1303 (2012).

[30] Japan Meteorological Agency, “Weather, Climate & Earthquake Information,” http://www.jma.go.jp/jma/en/menu.html (2014).

[31] Geospatial Information Authority of Japan, “Sokuchi Survey Calculation API No.2,” http://vldb.gsi.go.jp/sokuchi/surveycalc/main.html (2014).