Linearly-Combined Web Sensors for Spatio-Temporal Data Extraction from the Web

Shun Hattori
School of Computer Science
Tokyo University of Technology
1404-1 Katakura-machi, Hachioji, Tokyo 192-0982, Japan
Email: hattori@cs.teu.ac.jp

***Abstract*—Many researches on mining the Web, especially CGM (Consumer Generated Media) such as Weblogs, for knowledge about various phenomena and events in the physical world have been done actively, and Web services with the Web-mined knowledge have begun to be developed for the public. However, there is no detailed investigation on how accurately Web-mined data reflect real-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 defines the basic Weblog Sensor with a neutral, positive, or negative description for a target phenomenon, and their linearly-combined Weblog Sensors, and tries to validate the potential and reliability of these Weblog Sensors' spatio-temporal data by measuring the correlation with weather (precipitation) and earthquake (maximum seismic intensity and number of felt quakes) statistics per day by region of Japan Meteorological Agency as real-world data.**

***Keywords*-Web mining; Web credibility; Web sensors; knowledge extraction; search engine indices; spatio-temporal data.**

I. Introduction

In recent years, how to make physical spaces smarter has become one of the hottest topics in the research field of ubiquitous/pervasive computing. Smart Spaces are often physically isolated environments such as rooms, which are made smart by various information communication technologies. They would be much more convenient for information access in the future. Meanwhile, information security has also become very significant in any situation, especially in public places such as indoor work places, educational facilities, healthcare centers and so on. The amount of physical or virtual information resources which should be protected in the physical world grows exponentially.

Physical environments are becoming smart but not always secure. When a virtual (computational) information resource is requested to access by a user via an output device, conventional access control systems make a decision on whether the user should be granted or denied to access the resource based on its access policies and surely enforce the access decision. However, even if the requester is authorized by it, it should not be immediately offered to her via the output device, because there might be its unauthorized users as well as the authorized requester around the output device, especially in public places. A user trying to visit a physical environment might in turn be unexpectedly exposed to her unwanted information access. For example, although she does not want to know about the results of a football game that she had recorded on video to watch later, she unfortunately encounters it in her train. Meanwhile, when a user enters a physical environment, the user might hate its physical characteristics (e.g., degrees of dismal and danger) and/or be forced to access her unwanted information resources unexpectedly. This paper proposes a method to extract information for making access or entry decisions in Secure Spaces [1] from very large text corpora such as the Weblog, and improve the Secure Spaces by adding the concept of the Weblog Sensors [2], [3], in order to enable users to specify their access policies by keyword-based expressions about their unwanted physical spaces.

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 of the physical world. But in recent years, the explosively-growing Web has had more and more familiar relationship with the physical world, as the use of the Web, especially CGM (Consumer Generated Media) such as Weblogs, WOM (Word of Mouth) sites, SNSs (Social Networking Services), has become more popular with various people without distinction of age/sex.

Many researches on mining the Web, especially CGM (Consumer Generated Media) such as Weblogs, for knowledge about various phenomena and events in the physical world have been done very actively. For example, opinion and reputation extraction [4] of various products and services provided in the physical world, experience mining [5] of various phenomena and events held in the physical world, and concept hierarchy (semantics) extraction such as is-a/has-a relationships [6], [7] and appearance (look and feel) extraction [8], [9] 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 information for choosing better products, services, and actions in the physical world.

978-0-7695-4409-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDMW.2011.114

However, there is no detailed investigation on how accurately Web-mined data about a phenomenon or event held in the physical world reflect real-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 defines the basic Weblog Sensor with a neutral, positive, or negative Japanese description for a target phenomenon (e.g., rainfall and earthquake) in the physical world as follows.

- Neutral: “雨” (rain) and “地震” (earthquake)
- Positive: “強い雨” (heavy rain) and “強い地震” (strong earthquake)
- Negative: “弱い雨” (light rain) and “弱い地震” (weak earthquake)

And also this paper defines their linearly-combined Weblog Sensors to mine the Web, especially CGM such as Weblog documents for spatio-temporal data about the target phenomenon. And then this paper tries to validate the potential and reliability of these Weblog Sensors’ spatio-temporal data by measuring the correlation with weather (precipitation) and earthquake (maximum seismic intensity and number of felt quakes) statistics per day by region of Japan Meteorological Agency [10] as real-world data.

The remainder of this paper is organized as follows. Section II introduces Secure Spaces with Weblog Sensors. Section III defines the basic Weblog Sensor with a neutral, positive, or negative description for a target phenomenon, and their linearly-combined Weblog Sensors. Section IV validates the potential and reliability of these Weblog Sensors’ spatio-temporal data. Section V concludes this paper.

II. Secure Spaces

To build Secure Spaces in the real 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 Figure 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 of physical spaces as well as our unauthorized contents of physical spaces, 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.
- **Weblog 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.

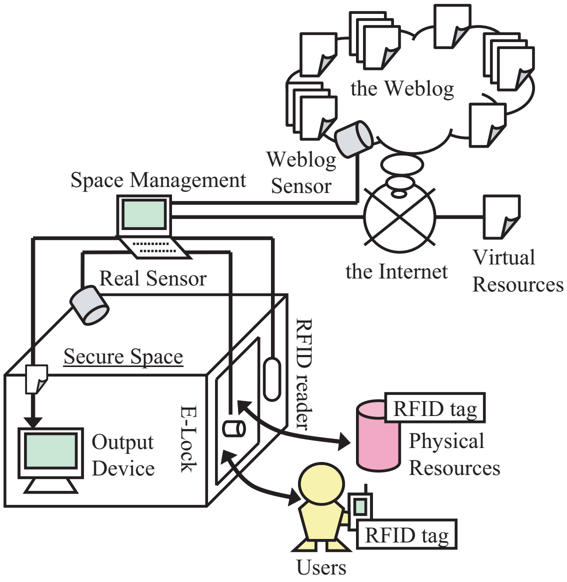

Figure 1. Secure Spaces.

III. Method

This section constructs the basic (unnormalized) Weblog Sensors with a neutral, positive, or negative Japanese description for a target phenomenon (e.g., rainfall and earthquake) in the physical world, the spatially-normalized Weblog Sensors, and their linearly-combined Weblog Sensors to mine the Web, especially Weblog documents for spatio-temporal data about the target phenomenon.

First, I define the basic Weblog Sensors with a geographical space s, e.g., 47 prefectural capitals in Japan such as "東京" (Tokyo) and "京都" (Kyoto), a time period t, e.g., a day such as "2011/1/1" and "2011/6/30", and a Japanese phrase, e.g., "雨" (ame; rain), "強い雨" (tuyoi-ame; heavy-rain), and "弱い雨" (yowai-ame; light-rain) for such a target phenomenon as rainfall:

$$\begin{aligned}
\text{ws-rain}^{0}_{0}(s,t) &:= \mathrm{bf}_t(\texttt{["}s\texttt{" \& "雨"]}),\\
\text{ws-rain}^{0}_{+}(s,t) &:= \mathrm{bf}_t(\texttt{["}s\texttt{" \& "雨" \& "強い"]}),\\
\text{ws-rain}^{0}_{-}(s,t) &:= \mathrm{bf}_t(\texttt{["}s\texttt{" \& "雨" \& "弱い"]}),
\end{aligned}$$

where $\mathrm{bf}_t(\texttt{[}q\texttt{]})$ stands for the Frequency of weBlog documents searched by submitting the query q with the custom time range t to Google Blog Search [11], and `&` stands for an `AND` operator. And also I define the basic Weblog Sensors with a geographical space s, e.g., 47 prefectural capitals in Japan such as "東京" (Tokyo) and "京都" (Kyoto), a time period t, e.g., a day such as "2011/1/1" and "2011/6/30", and a Japanese phrase, e.g., "揺れ" (yure; quake), "強い揺れ" (tuyoi-yure; strong-quake), and "弱い揺れ" (yowai-yure; weak-quake) for such a target phenomenon as earthquake:

$$\begin{aligned}
\text{ws-quake}^{0}_{0}(s,t) &:= \mathrm{bf}_t(\texttt{["}s\texttt{" \& "揺れ"]}),\\
\text{ws-quake}^{0}_{+}(s,t) &:= \mathrm{bf}_t(\texttt{["}s\texttt{" \& "揺れ" \& "強い"]}),\\
\text{ws-quake}^{0}_{-}(s,t) &:= \mathrm{bf}_t(\texttt{["}s\texttt{" \& "揺れ" \& "弱い"]}).
\end{aligned}$$

Next, I define the normalized Weblog Sensors by the frequency $\mathrm{bf}_t(\texttt{["}s\texttt{"]})$ of Weblogs searched by submitting each geographical space s with the custom time range t to Google Blog Search to clean up spatio-temporal dependency:

$$\begin{aligned}
\text{ws-rain}^{1}_{x}(s,t) &:= \frac{\text{ws-rain}^{0}_{x}(s,t)}{\mathrm{bf}_t(\texttt{["}s\texttt{"]})},\\
\text{ws-rain}^{2}_{x}(s,t) &:= \frac{\text{ws-rain}^{0}_{x}(s,t)}{\sqrt{\mathrm{bf}_t(\texttt{["}s\texttt{"]})}},\\
\text{ws-quake}^{1}_{x}(s,t) &:= \frac{\text{ws-quake}^{0}_{x}(s,t)}{\mathrm{bf}_t(\texttt{["}s\texttt{"]})},\\
\text{ws-quake}^{2}_{x}(s,t) &:= \frac{\text{ws-quake}^{0}_{x}(s,t)}{\sqrt{\mathrm{bf}_t(\texttt{["}s\texttt{"]})}},
\end{aligned}$$

where x stands for 0 (neutral), $+$ (positive), or $-$ (negative).

Last, I define their linearly-combined Weblog Sensors to make the above-mentioned Weblog Sensors more robust:

$$\begin{aligned}
\text{ws-rain}^{y}_{\pm}(s,t) &:= (1-\alpha)\cdot\text{ws-rain}^{y}_{+}(s,t)\\
&\quad + \alpha\cdot\text{ws-rain}^{y}_{-}(s,t),\\
\text{ws-rain}^{y}_{0\pm}(s,t) &:= (1-\alpha-\beta)\cdot\text{ws-rain}^{y}_{0}(s,t)\\
&\quad + \beta\cdot\text{ws-rain}^{y}_{+}(s,t)\\
&\quad + \alpha\cdot\text{ws-rain}^{y}_{-}(s,t),\\
\text{ws-quake}^{y}_{\pm}(s,t) &:= (1-\alpha)\cdot\text{ws-quake}^{y}_{+}(s,t)\\
&\quad + \alpha\cdot\text{ws-quake}^{y}_{-}(s,t),\\
\text{ws-quake}^{y}_{0\pm}(s,t) &:= (1-\alpha-\beta)\cdot\text{ws-quake}^{y}_{0}(s,t)\\
&\quad + \beta\cdot\text{ws-quake}^{y}_{+}(s,t)\\
&\quad + \alpha\cdot\text{ws-quake}^{y}_{-}(s,t),
\end{aligned}$$

where y stands for 0 (basic), 1, or 2 (normalized).

IV. Experiment

This section shows several experimental results to validate the basic Weblog Sensor with a neutral, positive, or negative Japanese description for a target phenomenon (e.g., rainfall and earthquake) in the physical world, and their linearly-combined Weblog Sensors to mine the Web, especially Weblog documents for spatio-temporal data about the target phenomenon. The whole experiments evaluate the correlation coefficient between real statistics by JMA (Japan Meteorological Agency) [10] as a physical sensor and spatio-temporal data mined by my proposed Weblog Sensors.

Table I shows the average, maximum, minimum, and standard deviation of correlation coefficient between the JMA's precipitation and my basic or normalized Weblog Sensor $\text{ws-rain}^{y}_{x}(s,t)$ using a Japanese keyword "雨" (ame; rain) for each space $s \in$ 47 prefectural capitals in Japan and each day $t \in$ January 1st, 2011 to June 30th, 2011. It shows that my normalized Weblog Sensors $\text{ws-rain}^{1}_{x}(s,t)$ seem to be superior to my basic (unnormalized) Weblog Sensors $\text{ws-rain}^{0}_{x}(s,t)$, and also that my Weblog Sensors $\text{ws-rain}^{y}_{-}(s,t)$ using a negative description seem to be superior to the other Weblog Sensors $\text{ws-rain}^{y}_{0}(s,t)$ or $\text{ws-rain}^{y}_{+}(s,t)$ using a neutral or positive description.

Table I
COEFFICIENT CORRELATION OF BASIC/NORMALIZED WEBLOG SENSOR $\text{ws-rain}^{y}_{x}(s,t)$ WITH JMA'S PRECIPITATION.

	Avg.	Max.	Min.	Dev.
$\text{ws-rain}^{0}_{0}(s,t)$	−0.0017	0.1941	−0.2110	0.1089
$\text{ws-rain}^{0}_{+}(s,t)$	0.0711	0.3106	−0.1370	**0.0898**
$\text{ws-rain}^{0}_{-}(s,t)$	0.1487	0.4972	−0.1153	0.1268
$\text{ws-rain}^{1}_{0}(s,t)$	0.2317	0.4897	**−0.1043**	0.1651
$\text{ws-rain}^{1}_{+}(s,t)$	0.1999	0.4851	−0.1305	0.1522
$\text{ws-rain}^{1}_{-}(s,t)$	**0.2332**	**0.6842**	−0.1763	0.1931
$\text{ws-rain}^{2}_{0}(s,t)$	0.0413	0.2213	−0.1194	0.0906
$\text{ws-rain}^{2}_{+}(s,t)$	0.1164	0.3437	−0.1467	0.1009
$\text{ws-rain}^{2}_{-}(s,t)$	0.1943	0.5474	−0.1684	0.1575

Figures 2 to 4 show the α-dependency of the average, maximum, minimum, and standard deviation of correlation coefficient between the JMA's precipitation and my 2-linearly-combined Weblog Sensor ws-rain$^y_\pm(s, t)$ using positive and negative phrases for each space $s \in 47$ prefectural capitals in Japan and each day $t \in$ 2011/1/1 to 2011/6/30.

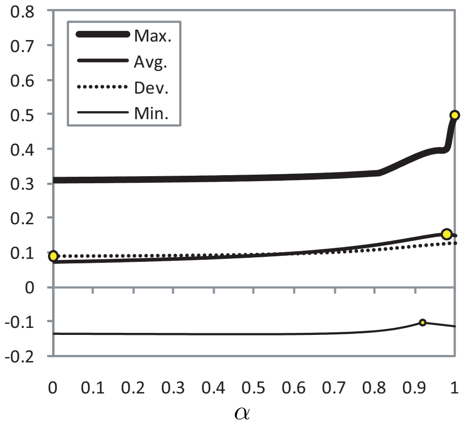

Figure 2. Dependency of 2-Linearly-Combined Basic Weblog Sensor ws-rain$^0_\pm(s, t)$ on α (the maximum Avg. 0.1527 when $\alpha = 0.98$).

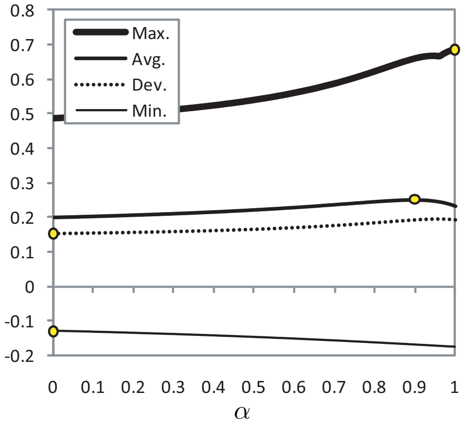

Figure 3. Dependency of 2-Linearly-Combined Normalized Weblog Sensor ws-rain$^1_\pm(s, t)$ on α (the maximum Avg. 0.2510 when $\alpha = 0.90$).

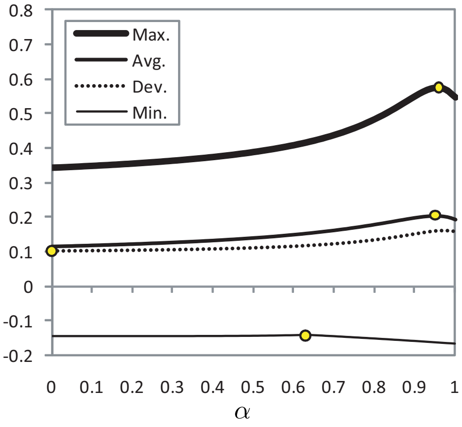

Figure 4. Dependency of 2-Linearly-Combined Normalized Weblog Sensor ws-rain$^2_\pm(s, t)$ on α (the maximum Avg. 0.2049 when $\alpha = 0.95$).

Figures 5 to 7 show the α/β-dependency of the average, maximum, minimum, and standard deviation of correlation coefficient between the JMA's precipitation and my linearly-combined Weblog Sensor ws-rain$^y_{0\pm}(s, t)$ using neutral, positive and negative Japanese phrases for each space $s \in 47$ prefectural capitals in Japan and each day $t \in$ January 1st, 2011 to June 30th, 2011. These figures show that my 3-linearly-combined Weblog Sensor ws-rain$^1_{0\pm}(s, t)$ when $\alpha = 0.91$, $\beta = 0.05$ gives the best performance and is superior to my 2-linearly-combined Weblog Sensors ws-rain$^y_\pm(s, t)$.

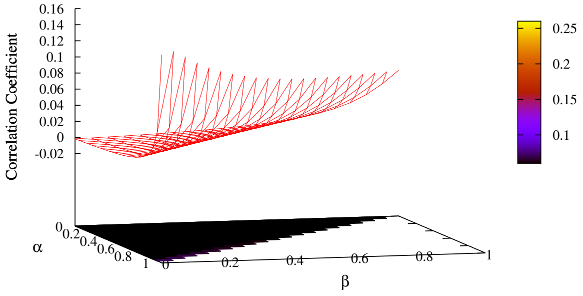

Figure 5. Dependency of 3-Linearly-Combined Basic Weblog Sensor ws-rain$^0_{0\pm}(s, t)$ on α, β (the maximum Avg. 0.1527 when $\alpha = 0.98$, $\beta = 0.02$).

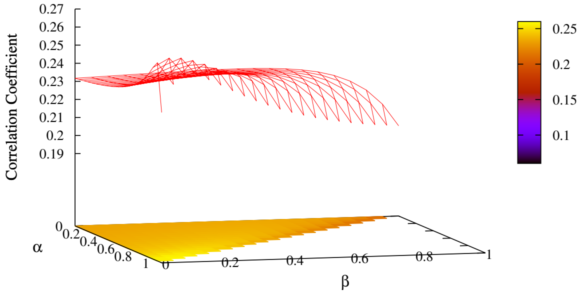

Figure 6. Dependency of 3-Linearly-Combined Normalized Weblog Sensor ws-rain$^1_{0\pm}(s, t)$ on α, β (the maximum Avg. 0.2614 when $\alpha = 0.91$, $\beta = 0.05$).

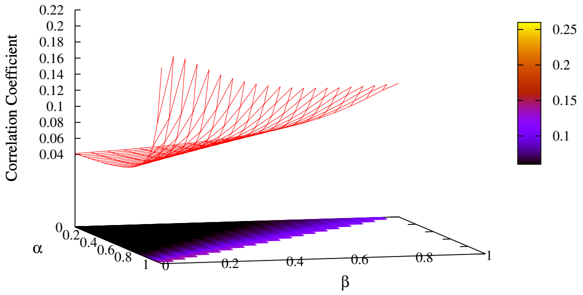

Figure 7. Dependency of 3-Linearly-Combined Normalized Weblog Sensor ws-rain$^2_{0\pm}(s, t)$ on α, β (the maximum Avg. 0.2049 when $\alpha = 0.95$, $\beta = 0.05$).

Table III shows the average, maximum, minimum, and standard deviation of correlation coefficient between the JMA's maximum seismic intensity and my basic or normalized Weblog Sensor ws-quake$^y_x(s,t)$ using a Japanese keyword "揺れ" (yure; quake) for each space $s \in 47$ prefectural capitals in Japan and each day $t \in$ January 1st, 2011 to June 30th, 2011. Here, JMA's seismic intensity scale is not always numerical, e.g., $5-$, $5+$, $6-$, $6+$. So, it is converted to numerical data as shown in Table II.

Table II
JMA's Seismic Intensity Scale to Numerical Value.

Scale	0	1	2	3	4	5−	5+	6−	6+	7
Value	0	1	2	3	4	4.75	5.25	5.75	6.25	7

Table IV also shows the average, maximum, minimum, and standard deviation of correlation coefficient between the JMA's number of felt quakes and my basic or normalized Weblog Sensor ws-quake$^y_x(s,t)$ using a Japanese keyword "揺れ" (yure; quake) for each space $s \in 47$ prefectural capitals in Japan and each day $t \in$ January 1st, 2011 to June 30th, 2011. The JMA's maximum seismic intensity and the JMA's number of felt quakes are derived from quite the same natural phenomenon "earthquake".

Table III
Coefficient Correlation of Basic/Normalized Weblog Sensor ws-quake$^y_x(s,t)$ with JMA's Seismic Intensity.

	Avg.	Max.	Min.	Dev.
ws-quake$^0_0(s,t)$	0.2695	0.6763	−0.0684	0.2705
ws-quake$^0_+(s,t)$	0.2877	0.7291	**−0.0353**	0.2716
ws-quake$^0_-(s,t)$	0.2845	0.6964	−0.0555	0.2716
ws-quake$^1_0(s,t)$	0.3113	0.6864	−0.0506	0.2596
ws-quake$^1_+(s,t)$	**0.3260**	0.7592	−0.0356	0.2708
ws-quake$^1_-(s,t)$	0.3234	0.7082	−0.0460	0.2658
ws-quake$^2_0(s,t)$	0.3025	0.6727	−0.0619	**0.2159**
ws-quake$^2_+(s,t)$	0.3040	**0.7690**	−0.0404	0.2212
ws-quake$^2_-(s,t)$	0.3086	0.7105	−0.0563	0.2210

Table IV
Coefficient Correlation of Basic/Normalized Weblog Sensor ws-quake$^y_x(s,t)$ with JMA's Number of Felt Quakes.

	Avg.	Max.	Min.	Dev.
ws-quake$^0_0(s,t)$	0.1826	0.4633	−0.0928	0.1560
ws-quake$^0_+(s,t)$	0.2290	0.4692	−0.0311	**0.1502**
ws-quake$^0_-(s,t)$	0.3100	**0.8231**	−0.0578	0.2266
ws-quake$^1_0(s,t)$	0.1967	0.4572	−0.0508	0.1504
ws-quake$^1_+(s,t)$	0.2378	0.4608	**−0.0126**	0.1529
ws-quake$^1_-(s,t)$	0.3010	0.7263	−0.0567	0.2073
ws-quake$^2_0(s,t)$	0.1955	0.4745	−0.0696	0.1558
ws-quake$^2_+(s,t)$	0.2388	0.4779	−0.0239	0.1533
ws-quake$^2_-(s,t)$	**0.3129**	0.8142	−0.0584	0.2239

Figures 8 to 10 show the α-dependency of the average, maximum, minimum, and standard deviation of correlation coefficient between the JMA's maximum seismic intensity and my 2-linearly-combined Weblog Sensor ws-quake$^y_\pm(s,t)$ for each space $s \in 47$ prefectural capitals in Japan and each day $t \in$ 2011/1/1 to 2011/6/30.

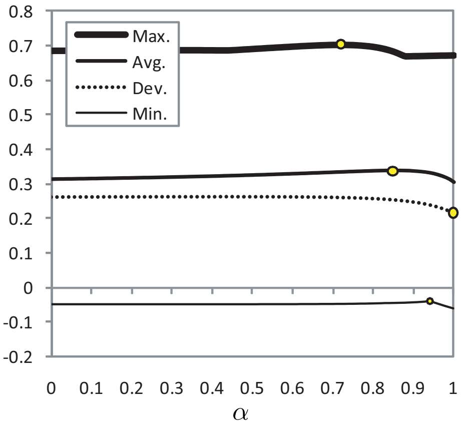

Figure 8. Dependency of 2-Linearly-Combined Basic Weblog Sensor ws-quake$^0_\pm(s,t)$ on α (the maximum Avg. 0.3369 when $\alpha = 0.85$).

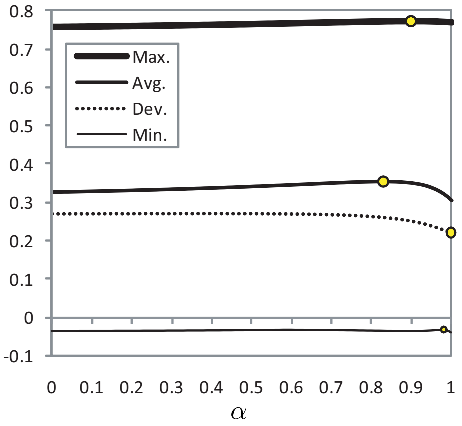

Figure 9. Dependency of 2-Linearly-Combined Normalized Weblog Sensor ws-quake$^1_\pm(s,t)$ on α (the maximum Avg. 03539. when $\alpha = 0.83$).

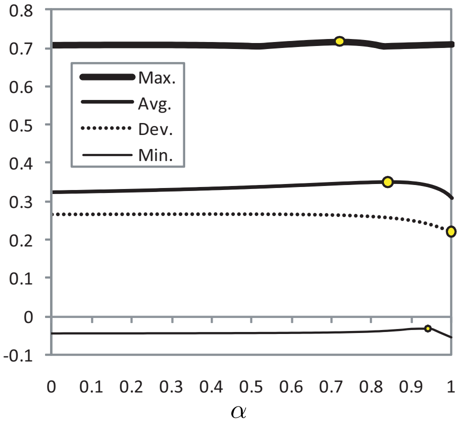

Figure 10. Dependency of 2-Linearly-Combined Normalized Weblog Sensor ws-quake$^2_\pm(s,t)$ on α (the maximum Avg. 0.3492 when $\alpha = 0.84$).

Figures 11 to 13 show the α/β-dependency of the average, maximum, minimum, and standard deviation of correlation coefficient between the JMA's maximum seismic intensity and my 3-linearly-combined Weblog Sensor ws-quake$^{y}_{0\pm}(s,t)$ for each space $s \in 47$ prefectural capitals in Japan and each day $t \in$ January 1st, 2011 to June 30th, 2011. These figures show that my 3-linearly-combined Weblog Sensor ws-quake$^{1}_{0\pm}(s,t)$ when $\alpha = 0.83$, $\beta = 0.15$ gives the best performance and is superior to my 2-linearly-combined Weblog Sensors ws-quake$^{y}_{\pm}(s,t)$.

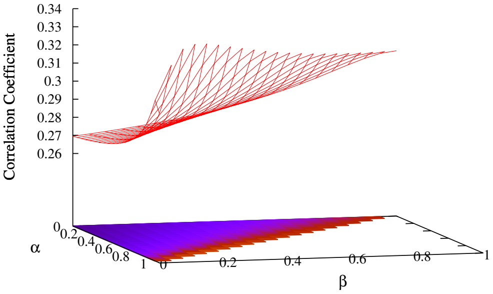

Figure 11. Dependency of 3-Linearly-Combined Basic Weblog Sensor ws-quake$^{0}_{0\pm}(s,t)$ on α, β (the maximum Avg. 0.3369 when $\alpha = 0.85$, $\beta = 0.15$).

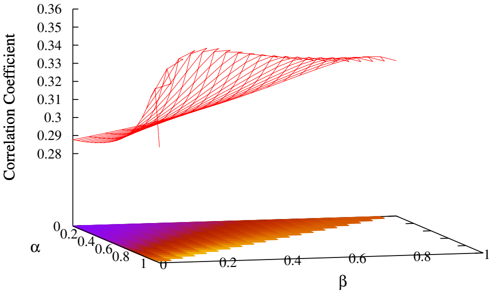

Figure 12. Dependency of 3-Linearly-Combined Normalized Weblog Sensor ws-quake$^{1}_{0\pm}(s,t)$ on α, β (the maximum Avg. 0.3558 when $\alpha = 0.83$, $\beta = 0.15$).

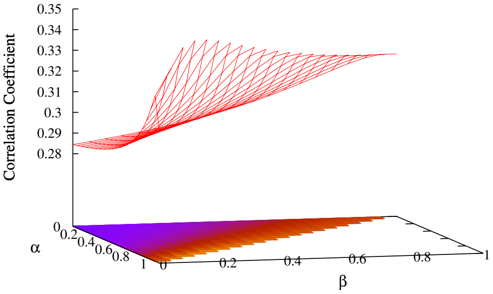

Figure 13. Dependency of 3-Linearly-Combined Normalized Weblog Sensor ws-quake$^{2}_{0\pm}(s,t)$ on α, β (the maximum Avg. 0.3492 when $\alpha = 0.84$, $\beta = 0.16$).

Figure 14 shows the spatial dependency of correlation coefficient between the JMA's maximum seismic intensity and my 3-linearly-combined normalized Weblog Sensor ws-quake$^{1}_{0\pm}(s,t)$ when $\alpha = 0.83$, $\beta = 0.15$. Figure 15 shows the distribution of JMA's maximum seismic intensity by the Great East Japan Earthquake on March 11th, 2011. These figures show that my 3-linearly-combined normalized Weblog Sensor ws-quake$^{1}_{0\pm}(s,t)$ seems to give more correlation coefficient with the JMA's maximum seismic intensity for regions with more seismic intensity.

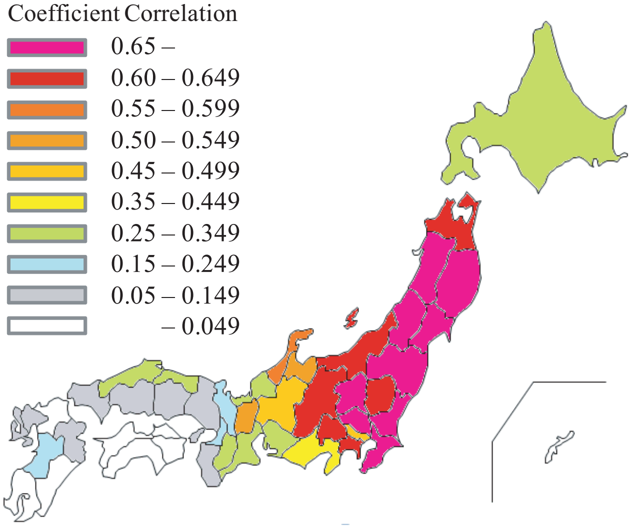

Figure 14. Spatial Dependency of 3-Linearly-Combined Normalized Weblog Sensor ws-quake$^{1}_{0\pm}(s,t)$ when $\alpha = 0.83$, $\beta = 0.15$.

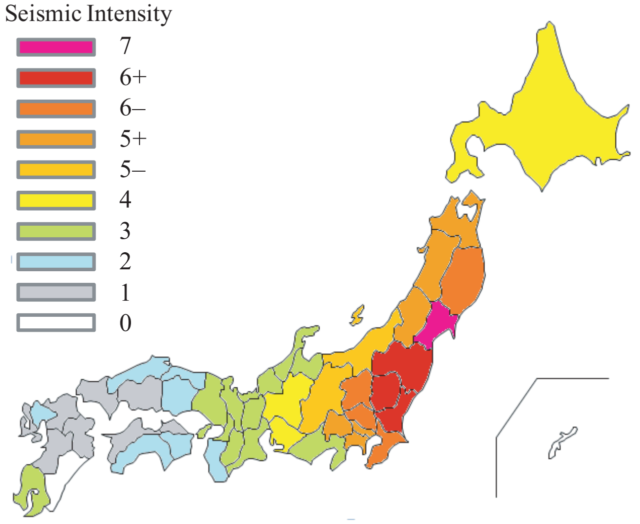

Figure 15. Distribution of Seismic Intensity on March 11th, 2011 in Japan.

Figures 16 to 18 show the α-dependency of the average, maximum, minimum, and standard deviation of correlation coefficient between the JMA's number of felt quakes and my 2-linearly-combined Weblog Sensor ws-quake$^{y}_{\pm}(s,t)$ for each space $s \in 47$ prefectural capitals in Japan and each day $t \in$ January 1st, 2011 to June 30th, 2011.

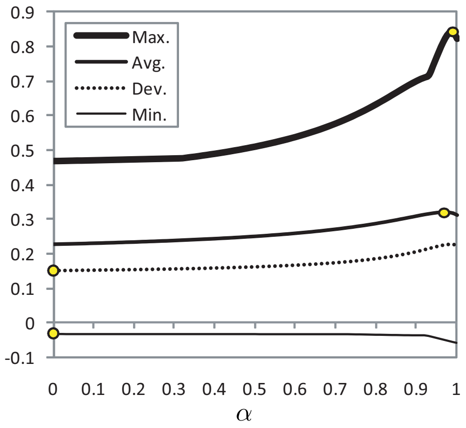

Figure 16. Dependency of 2-Linearly-Combined Basic Weblog Sensor ws-quake$^{0}_{\pm}(s,t)$ on α (the maximum Avg. 0.3179 when $\alpha = 0.97$).

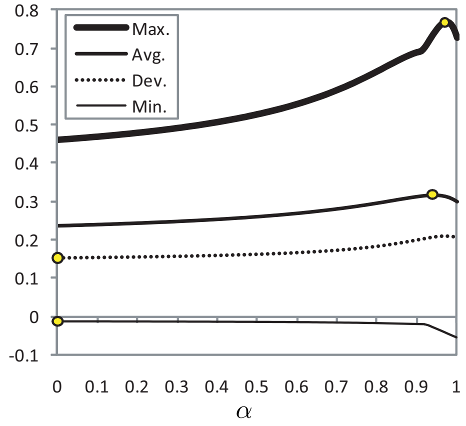

Figure 17. Dependency of 2-Linearly-Combined Normalized Weblog Sensor ws-quake$^{1}_{\pm}(s,t)$ on α (the maximum Avg. 0.3179 when $\alpha = 0.94$).

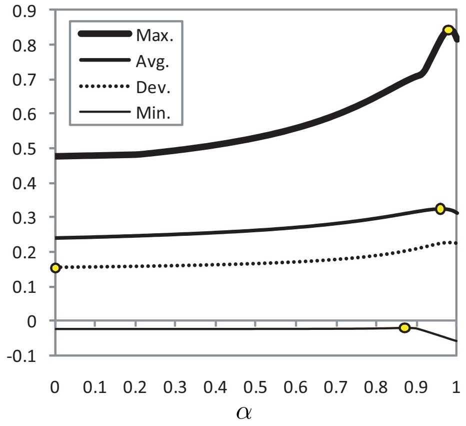

Figure 18. Dependency of 2-Linearly-Combined Normalized Weblog Sensor ws-quake$^{2}_{\pm}(s,t)$ on α (the maximum Avg. 0.3252 when $\alpha = 0.96$).

Figures 19 to 21 show the α/β-dependency of the average, maximum, minimum, and standard deviation of correlation coefficient between the JMA's number of felt quakes and my 3-linearly-combined Weblog Sensor ws-quake$^{y}_{0\pm}(s,t)$ for each space $s \in 47$ prefectural capitals in Japan and each day $t \in$ January 1st, 2011 to June 30th, 2011. These figures show that my 3-linearly-combined Weblog Sensor ws-quake$^{2}_{0\pm}(s,t)$ when $\alpha = 0.94$, $\beta = 0.06$ gives the best performance but is equal to my 2-linearly-combined Weblog Sensor ws-quake$^{2}_{\pm}(s,t)$ when $\alpha = 0.94$.

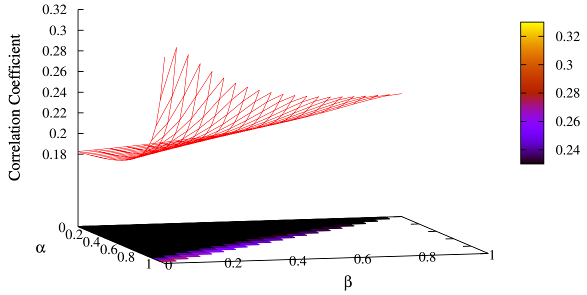

Figure 19. Dependency of 3-Linearly-Combined Basic Weblog Sensor ws-quake$^{0}_{0\pm}(s,t)$ on α, β (the maximum Avg. 0.3179 when $\alpha = 0.97$, $\beta = 0.03$).

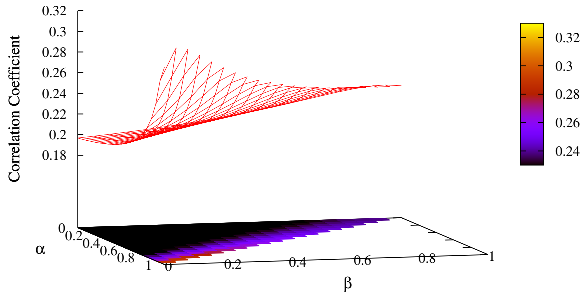

Figure 20. Dependency of 3-Linearly-Combined Normalized Weblog Sensor ws-quake$^{1}_{0\pm}(s,t)$ on α, β (the maximum Avg. 0.3179 when $\alpha = 0.94$, $\beta = 0.06$).

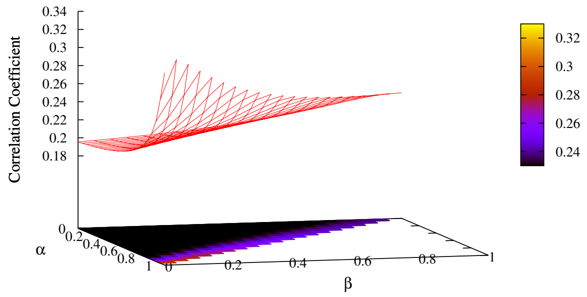

Figure 21. Dependency of 3-Linearly-Combined Normalized Weblog Sensor ws-quake$^{2}_{0\pm}(s,t)$ on α, β (the maximum Avg. 0.3252 when $\alpha = 0.94$, $\beta = 0.06$).

Figure 22 shows the spatial dependency of correlation coefficient between the JMA's number of felt quakes and my 3-linearly-combined normalized Weblog Sensor ws-quake$_{0\pm}^{2}(s,t)$ when $\alpha = 0.94$, $\beta = 0.06$. Figure 23 shows the distribution of JMA's number of felt quakes by the Great East Japan Earthquake on March 11th, 2011. These figures show that my 3-linearly-combined normalized Weblog Sensor ws-quake$_{0\pm}^{2}(s,t)$ seems to give more correlation coefficient with the JMA's maximum seismic intensity for regions with more felt quakes, but there are some irregular regions with higher correlation but less felt quakes.

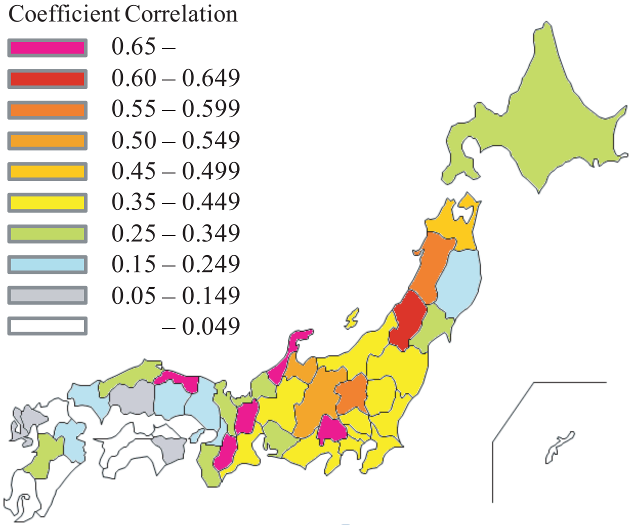

Figure 22. Spatial Dependency of 3-Linearly-Combined Normalized Weblog Sensor ws-quake$_{0\pm}^{2}(s,t)$ when $\alpha = 0.94$, $\beta = 0.06$.

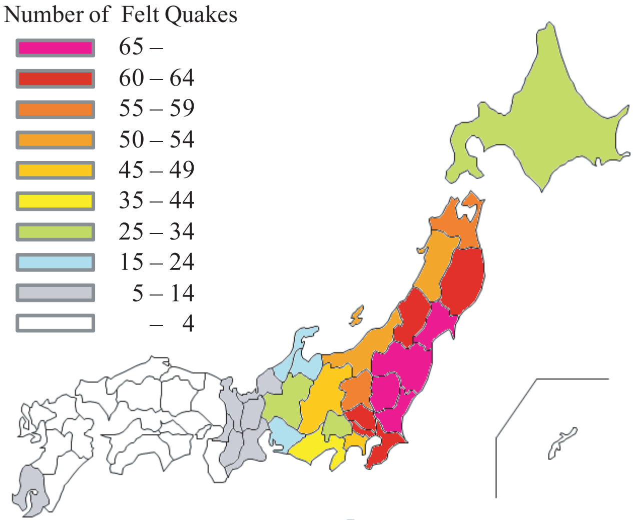

Figure 23. Distribution of JMA's Number of Felt Quakes on March 11th, 2011 in Japan.

V. Conclusion

This paper has defined the basic Weblog Sensor with a neutral, positive, or negative description for a target phenomenon (e.g., rainfall and earthquake) in the physical world, and their linearly-combined Weblog Sensors to mine the Web, especially CGM such as Weblog documents for spatio-temporal data about the target phenomenon. And also this paper has validated some potential and reliability of these Weblog Sensors' spatio-temporal data by measuring the correlation coefficient with weather (precipitation) and earthquake (maximum seismic intensity and number of felt quakes) statistics per day by region of Japan Meteorological Agency as real-world data.

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

[2] Hattori, S. and Tanaka, K.: "Mining the Web for Access Decision-Making in Secure Spaces," Proc. of SCIS&ISIS'08, TH-G3-4, pp.370–375 (2008).

[3] Hattori, S.: "Secure Spaces and Spatio-Temporal Weblog Sensors with Temporal Shift and Propagation," Proc. of DEIT'11, pp.1042–1047 (2011).

[4] Dave, K., Lawrence, S., and Pennock, D.M.: "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," Proc. of WWW'03, pp.519–528 (2003).

[5] Tezuka, T., Kurashima, T., and Tanaka, K.: "Toward Tighter Integration of Web Search with a Geographic Information System," Proc. of WWW'06, pp.277–286 (2006).

[6] Hattori, S., Ohshima, H., Oyama, S., and Tanaka, K.: "Mining the Web for Hyponymy Relations based on Property Inheritance," Proc. of APWeb'08, LNCS Vol.4976, pp.99–110 (2008).

[7] Hattori, S., and Tanaka, K.: "Extracting Concept Hierarchy Knowledge from the Web based on Property Inheritance and Aggregation," Proc. of WI'08, pp.432–437 (2008).

[8] Hattori, S., Tezuka, T., and Tanaka, K.: "Mining the Web for Appearance Description," Proc. of DEXA'07, LNCS Vol.4653, pp.790–800 (2007).

[9] Hattori, S.: "Peculiar Image Search by Web-extracted Appearance Descriptions," Proc. of SoCPaR'10, pp.127–132 (2010).

[10] Japan Meteorological Agency, http://www.jma.go.jp/jma/indexe.html (2011).

[11] Google Blog Search, http://blogsearch.google.co.jp/ (2011).