2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Spatio-Temporal Web Sensors by Social Network Analysis

Shun Hattori
College of Information and Systems
Muroran Institute of Technology
27–1 Mizumoto-cho, Muroran, Hokkaido 050–8585, Japan
Email: hattori@csse.muroran-it.ac.jp
Tel: +81–143-46-5498
Fax: +81–143-46-5499

***Abstract*—Many researches on mining the Web, especially Social Networking Media such as weblogs and microblogging sites which seem to store vast amounts of information about human societies, for knowledge about various phenomena and events in the physical world have been done actively, and Web applications with 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 applications without ensuring their accuracy sufficiently. Therefore, this paper defines spatio-temporal Web Sensors by analyzing Twitter, Facebook, weblogs, news sites, or the whole Web for a target natural phenomenon, and tries to validate the potential and reliability of the Web Sensors' spatio-temporal data by measuring the coefficient correlation with Japanese weather, earthquake, and influenza statistics per week by region as real-world data.**

***Keywords*-Web mining; Web credibility; Web Sensor; spatio-temporal data mining; social network analysis; microblogging;**

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 and 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 instance, although she does not want to know about a football game that she had recorded on video to watch later, she unfortunately encounters its result via an output device embedded 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 Web, and improve the Secure Spaces by adding the concept of Web Sensors [2–4], 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 world has had more and more familiar relationship with the physical world, as the use of the Web, especially UGM (User Generated Media) such as weblogs, WOM (Word of Mouth) sites, and SNS (Social Networking Services), has become much more popular with various people without distinction of age, sex, or country.

Many researches on mining the Web, especially Social Networking Media such as weblogs and microblogging sites which seems to store vast amounts of information about human societies, for knowledge about various phenomena and events in the physical world have been done very actively. For instance, opinion and reputation extraction [5], [6] of various products and services provided in the physical world, experience mining [7], [8] of various phenomena and events held in the physical world, and concept hierarchy (semantics) extraction such as is-a/has-a relationships [9–11] and appearance (look and feel) extraction [12–17] of physical objects in the physical world. Meanwhile, Web applications with Web-mined knowledge have begun to be developed for the public, and more and more ordinary people actually utilize them as vital information for choosing better

978-0-7695-4799-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ASONAM.2012.172

products, services, and actions in the physical world.

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 the Web-extracted data. 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 applications without ensuring their accuracy sufficiently.

This paper defines spatio-temporal Web Sensors by analyzing SNS sites such as Twitter and Facebook, weblogs, news sites, or the whole Web for a target natural phenomenon such as rainfall, snowfall, earthquake, and influenza in the physical world. And this paper tries to validate the potential and reliability of the Web Sensors' spatio-temporal data by measuring the coefficient correlation with Japanese weather (rainfall and snowfall) and earthquake (number of felt quakes) statistics of Japan Meteorological Agency [18], and influenza (reports of influenza virus isolation/detection) statistics of National Institute of Infectious Diseases [19] per week by region as real-world data.

The remainder of this paper is organized as follows. Section II introduces Secure Spaces with Web Sensors. Section III defines spatio-temporal Web Sensors with various kinds of Web documents. Section IV validates the potential and reliability of the Web Sensors' spatio-temporal data. Section V concludes this paper.

II. Secure Spaces

To build Secure Spaces [1] 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 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 instance, thermometers, hygrometers, (security) cameras, and so forth.
- **Web Sensor**: is responsible for logically sensing the Web, especially Social Networking Media such as weblogs and microblogging sites, 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 to use Web Sensors, any Secure Space does not have to equip the extra devices unlike Real Sensors.

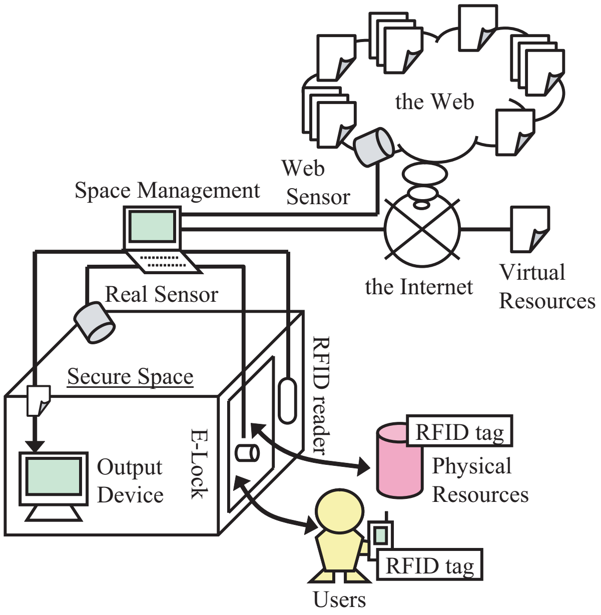

Figure 1. Secure Spaces and Web Sensors.

III. Spatio-Temporal Web Sensors

This section defines the simplest Web Sensor and its spatiotemporally-normalized Web Sensor per week by region to mine SNS sites such as Twitter and Facebook, weblogs, news sites, or the whole Web for spatio-temporal data about a target natural phenomenon such as rainfall, snowfall, earthquake, and influenza in the physical world.

First, the simplest Web Sensor with a geographical space s, e.g., one of 47 prefectures in Japan such as "東京都" (Tokyo) and "北海道" (Hokkaido), a time period t, e.g., a week such as "2012/1/2 – 2012/1/8" (1st week) and "2012/4/30 – 2012/5/6" (18th week), and a Japanese keyword kw representing a target phenomenon in the physical world, e.g., "雨" (rain), "雪" (snow), "地震" (earthquake), and "インフルエンザ" (influenza), by analyzing a corpus c of Web documents, e.g., Twitter, Facebook, weblogs, news sites, and the whole Web:

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

where $\mathrm{df}_t^c(\texttt{[}q\texttt{]})$ stands for the Frequency of Web Documents retrieved from the corpus c by submitting the search query q with the custom time range t to Google Web Search [20], 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:

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

IV. Experiment

This section shows several experimental results to investigate the potential and reliability of Web Sensors' spatio-temporal data by measuring the coefficient correlation with Japanese weather (rainfall and snowfall) and earthquake (number of felt quakes) statistics of Japan Meteorological Agency [18], and influenza (reports of influenza virus isolation/detection) statistics of National Institute of Infectious Diseases [19] per week by region as real-world data.

Table I
Coefficient Correlation of Simplest Web Sensor's Spatio-Temporal Data with Statistics.
(Web Corpus c: Twitter.com)

Stats kw	Rain	Snow	Earthquake	Influenza
雨 (rain)	0.23662 ±0.30633	0.08918 ±0.61574	-0.06320 ±0.17826	-0.48524 ±0.08673
雪 (snow)	0.16155 ±0.20154	0.22374 ±0.56503	-0.07155 ±0.19502	-0.12197 ±0.19852
地震 (earthquake)	0.18220 ±0.28792	0.05189 ±0.64425	-0.03678 ±0.19762	-0.55301 ±0.10032
インフルエンザ (influenza)	0.12026 ±0.19316	**0.22577** ±0.53940	-0.00470 ±0.20210	-0.05639 ±0.22805
iPad (iPad)	**0.24009** ±0.31883	0.07906 ±0.62598	-0.07354 ±0.19038	-0.53527 ±0.08530
オリンピック (Olympic)	0.18799 ±0.29944	0.10988 ±0.60963	-0.06083 ±0.19541	-0.48652 ±0.11060
節電 (power-saving)	0.22158 ±0.36260	0.13265 ±0.58592	-0.01995 ±0.19466	-0.40714 ±0.08649

Table II
Coefficient Correlation of Normalized Web Sensor's Spatio-Temporal Data with Statistics.
(Web Corpus c: Twitter.com)

Stats kw	Rain	Snow	Earthquake	Influenza
雨 (rain)	**0.26642** ±0.30635	0.09992 ±0.61168	-0.05636 ±0.18292	-0.46133 ±0.11744
雪 (snow)	0.10202 ±0.29691	**0.37743** ±0.49326	-0.04459 ±0.20590	**0.17167** ±0.31227
地震 (earthquake)	0.15882 ±0.29948	0.05840 ±0.64293	0.00862 ±0.23269	-0.51825 ±0.19808
インフルエンザ (influenza)	0.07595 ±0.21618	0.32434 ±0.50819	0.03146 ±0.20590	0.13096 ±0.30780
iPad (iPad)	0.25559 ±0.31306	0.09372 ±0.62301	-0.06644 ±0.21744	-0.51100 ±0.12538
オリンピック (Olympic)	0.17916 ±0.29932	0.15635 ±0.59256	-0.05279 ±0.21885	-0.39658 ±0.22361
節電 (power-saving)	0.20620 ±0.37025	0.15835 ±0.57762	0.01477 ±0.23811	-0.35154 ±0.14936

Table III
Coefficient Correlation of Simplest Web Sensor's Spatio-Temporal Data with Statistics.
(Web Corpus c: Facebook.com)

Stats kw	Rain	Snow	Earthquake	Influenza
雨 (rain)	**0.21670** ±0.33216	0.15997 ±0.56654	-0.01841 ±0.19186	-0.38273 ±0.09798
雪 (snow)	0.15266 ±0.36949	0.18868 ±0.55368	-0.01217 ±0.23438	-0.27948 ±0.14262
地震 (earthquake)	0.19107 ±0.34634	0.18168 ±0.55805	-0.00518 ±0.24029	-0.32333 ±0.12727
インフルエンザ (influenza)	0.11445 ±0.27793	**0.27713** ±0.51909	-0.04556 ±0.20378	-0.05711 ±0.21462
iPad (iPad)	0.20459 ±0.35018	0.18198 ±0.54743	-0.01551 ±0.22547	-0.31515 ±0.09223
オリンピック (Olympic)	0.18384 ±0.32788	0.19791 ±0.55462	-0.02841 ±0.21757	-0.26471 ±0.14440
節電 (power-saving)	0.19600 ±0.34995	0.18563 ±0.54833	0.01399 ±0.22550	-0.33381 ±0.10203

Table IV
Coefficient Correlation of Normalized Web Sensor's Spatio-Temporal Data with Statistics.
(Web Corpus c: Facebook.com)

Stats kw	Rain	Snow	Earthquake	Influenza
雨 (rain)	**0.20736** ±0.30370	0.20409 ±0.54480	0.01078 ±0.19348	-0.30857 ±0.19523
雪 (snow)	0.08581 ±0.35320	0.27189 ±0.51001	0.01377 ±0.21756	-0.13029 ±0.21534
地震 (earthquake)	0.17640 ±0.32238	0.23735 ±0.54338	0.01862 ±0.23211	-0.21980 ±0.20192
インフルエンザ (influenza)	0.08867 ±0.26080	**0.31717** ±0.51409	-0.03570 ±0.21035	0.02592 ±0.23787
iPad (iPad)	0.14091 ±0.34892	0.22204 ±0.53201	0.00769 ±0.23840	-0.23436 ±0.17717
オリンピック (Olympic)	0.12751 ±0.30789	0.23357 ±0.55150	-0.02222 ±0.21688	-0.17130 ±0.20456
節電 (power-saving)	0.16572 ±0.33132	0.22429 ±0.53446	0.04081 ±0.25896	-0.27228 ±0.16126

Table V
Coefficient Correlation of Simplest Web Sensor's Spatio-Temporal Data with Statistics.
(Web Corpus c: Blog)

Stats kw	Rain	Snow	Earthquake	Influenza
雨 (rain)	**0.29395** ±0.31520	0.13594 ±0.63432	-0.04185 ±0.21323	-0.35697 ±0.28011
雪 (snow)	-0.18492 ±0.27629	**0.55921** ±0.31716	0.09311 ±0.21320	0.55156 ±0.17017
地震 (earthquake)	0.06482 ±0.21913	0.25580 ±0.53980	0.04428 ±0.28048	-0.12773 ±0.34642
インフルエンザ (influenza)	-0.14449 ±0.24131	0.54445 ±0.36887	**0.14408** ±0.25318	**0.63923** ±0.22427
iPad (iPad)	0.05968 ±0.30502	0.19241 ±0.57391	0.00877 ±0.20037	-0.19017 ±0.31285
オリンピック (Olympic)	-0.05190 ±0.24784	0.29997 ±0.51207	0.07716 ±0.24843	0.01624 ±0.31001
節電 (power-saving)	0.09883 ±0.23307	0.15892 ±0.59217	-0.00333 ±0.19940	-0.23072 ±0.23964

Table VI
Coefficient Correlation of Normalized Web Sensor's Spatio-Temporal Data with Statistics.
(Web Corpus c: Blog)

Stats kw	Rain	Snow	Earthquake	Influenza
雨 (rain)	**0.25657** ±0.30347	0.15614 ±0.61916	-0.03172 ±0.21172	-0.28262 ±0.33312
雪 (snow)	-0.13444 ±0.24341	0.49374 ±0.38649	0.08073 ±0.20615	0.48383 ±0.20563
地震 (earthquake)	0.07327 ±0.27358	0.24780 ±0.54817	0.04055 ±0.25169	-0.10453 ±0.36255
インフルエンザ (influenza)	-0.14369 ±0.23352	**0.543416** ±0.36638	**0.14304** ±0.24788	**0.64430** ±0.22469
iPad (iPad)	0.05945 ±0.31249	0.19711 ±0.56959	0.00652 ±0.20168	-0.17697 ±0.33550
オリンピック (Olympic)	-0.02743 ±0.26224	0.29705 ±0.51754	0.06779 ±0.21366	0.01830 ±0.30728
節電 (power-saving)	0.09566 ±0.25163	0.16036 ±0.59003	-0.00526 ±0.19422	-0.21141 ±0.26796

Table VII
Coefficient Correlation of Simplest Web Sensor's Spatio-Temporal Data with Statistics.
(Web Corpus c: News)

Stats kw	Rain	Snow	Earthquake	Influenza
雨 (rain)	**0.13344** ±0.33860	0.13382 ±0.58220	-0.05997 ±0.18446	-0.43215 ±0.06097
雪 (snow)	0.06941 ±0.31387	**0.14044** ±0.57729	-0.06593 ±0.15844	-0.40719 ±0.06200
地震 (earthquake)	0.09941 ±0.31280	0.13538 ±0.58130	-0.06301 ±0.16567	-0.42320 ±0.05861
インフルエンザ (influenza)	0.08328 ±0.33843	0.13419 ±0.58251	-0.06458 ±0.17277	-0.42377 ±0.06083
iPad (iPad)	0.06112 ±0.32254	0.13425 ±0.58221	-0.06143 ±0.16809	-0.41898 ±0.05938
オリンピック (Olympic)	0.11054 ±0.33833	0.13548 ±0.58129	-0.05789 ±0.17500	-0.42661 ±0.05977
節電 (power-saving)	0.11020 ±0.32801	0.13920 ±0.57802	-0.05730 ±0.17297	-0.41990 ±0.05753

Table VIII
Coefficient Correlation of Normalized Web Sensor's Spatio-Temporal Data with Statistics.
(Web Corpus c: News)

Stats kw	Rain	Snow	Earthquake	Influenza
雨 (rain)	**0.13676** ±0.33802	0.13235 ±0.58344	-0.06355 ±0.18571	-0.43286 ±0.06084
雪 (snow)	0.07290 ±0.31723	**0.13810** ±0.57925	-0.06938 ±0.16178	-0.41107 ±0.06270
地震 (earthquake)	0.10355 ±0.31363	0.13335 ±0.58297	-0.06605 ±0.16973	-0.42620 ±0.05915
インフルエンザ (influenza)	0.08924 ±0.34016	0.13298 ±0.58350	-0.06687 ±0.17499	-0.42356 ±0.06196
iPad (iPad)	0.06563 ±0.32589	0.13253 ±0.58352	-0.06478 ±0.17136	-0.42132 ±0.06067
オリンピック (Olympic)	0.11397 ±0.34068	0.13353 ±0.58279	-0.06107 ±0.17831	-0.42994 ±0.06060
節電 (power-saving)	0.11354 ±0.33133	0.13690 ±0.57985	-0.05955 ±0.17704	-0.42432 ±0.05853

Table IX
Coefficient Correlation of Simplest Web Sensor's Spatio-Temporal Data with Statistics.
(Web Corpus c: Web)

Stats kw	Rain	Snow	Earthquake	Influenza
雨 (rain)	-0.03267 ±0.28756	0.24338 ±0.53230	0.02747 ±0.24844	-0.05891 ±0.26059
雪 (snow)	0.07769 ±0.24018	0.30493 ±0.49643	0.03006 ±0.22265	0.17868 ±0.27185
地震 (earthquake)	**0.12454** ±0.30459	0.14421 ±0.60138	-0.08739 ±0.22066	-0.32519 ±0.27176
インフルエンザ (influenza)	-0.03929 ±0.18869	**0.47917** ±0.38286	0.09176 ±0.24970	**0.50688** ±0.21110
iPad (iPad)	-0.00767 ±0.27006	0.28349 ±0.50922	-0.00584 ±0.23194	-0.14336 ±0.32602
オリンピック (Olympic)	0.08059 ±0.28171	0.32244 ±0.46942	-0.02021 ±0.20856	0.10205 ±0.29141
節電 (power-saving)	0.01929 ±0.25848	0.20801 ±0.55460	-0.00943 ±0.22373	-0.24271 ±0.20521

Table X
Coefficient Correlation of Normalized Web Sensor's Spatio-Temporal Data with Statistics.
(Web Corpus c: Web)

Stats kw	Rain	Snow	Earthquake	Influenza
雨 (rain)	-0.16111 ±0.28865	0.36915 ±0.46733	0.01460 ±0.23873	0.14595 ±0.23775
雪 (snow)	-0.11592 ±0.28484	0.39546 ±0.44067	0.01305 ±0.22484	0.26101 ±0.23662
地震 (earthquake)	-0.04259 ±0.17815	0.24796 ±0.53549	-0.09056 ±0.25741	-0.12469 ±0.26405
インフルエンザ (influenza)	-0.16293 ±0.24857	**0.51123** ±0.36023	0.05399 ±0.23129	**0.51224** ±0.16315
iPad (iPad)	-0.14144 ±0.23778	0.37823 ±0.45093	-0.01702 ±0.26398	0.04613 ±0.28411
オリンピック (Olympic)	-0.07747 ±0.30861	0.40524 ±0.42514	-0.02750 ±0.17989	0.22656 ±0.20725
節電 (power-saving)	-0.13365 ±0.29216	0.31067 ±0.49778	-0.01837 ±0.22062	-0.03894 ±0.26432

Tables I to X show the coefficient correlation between the simplest or spatiotemporally-normalized Web Sensor's spatio-temporal data for a Japanese keyword kw with a Web corpus c and each of four kinds of Japanese statistics. They show that weblogs are the most appropriate corpus of Web documents for Web Sensors because the Web Sensors by analyzing weblogs give the highest coefficient correlation with any kind of Japanese statistics, while SNS sites such as Twitter and Facebook which seem to store vaster amounts of information about human societies are not so appropriate for Web Sensors to extract spatio-temporal data about natural phenomena, and that the simplest Web Sensor by analyzing weblogs is slightly superior to the spatiotemporally-normalized Web Sensor, while in contrast the spatiotemporally-normalized Web Sensor by analyzing SNS sites is superior to the simplest Web Sensor. And also the Web Sensors by analyzing SNS sites show negative correlation coefficient for negative phenomena such as earthquake and influenza.

Figures 2 to 9 show the spatial dependency of correlation coefficient of the simplest Web Sensor by analyzing weblogs or the spatiotemporally-normalized Web Sensor by analyzing Tweets with four kinds of Japanese statistics. They show that the simplest Web Sensor by analyzing weblogs has less prefectures with too low coefficient correlation for snowfall and influenza stats, while two kinds of Web

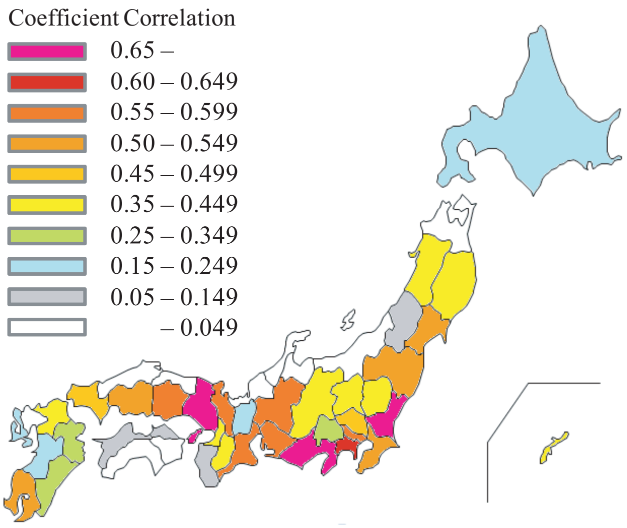

Figure 2. Spatial Dependency of Coefficient Correlation between the Simplest Web Sensor with Blogs and Rainfall Stats.

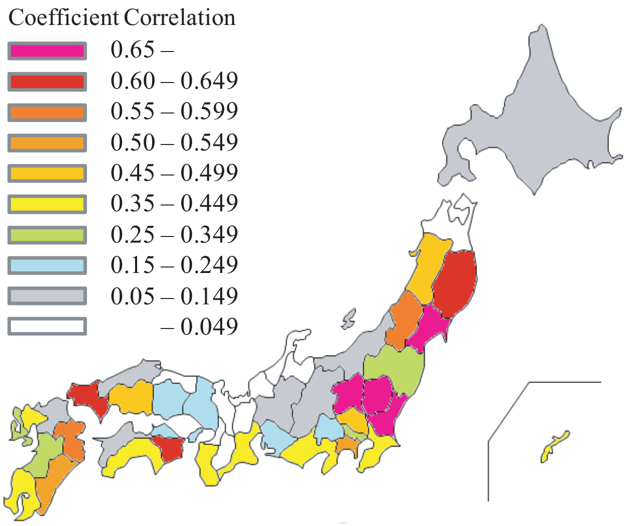

Figure 3. Spatial Dependency of Coefficient Correlation between the Normalized Web Sensor with Twitter and Rainfall Stats.

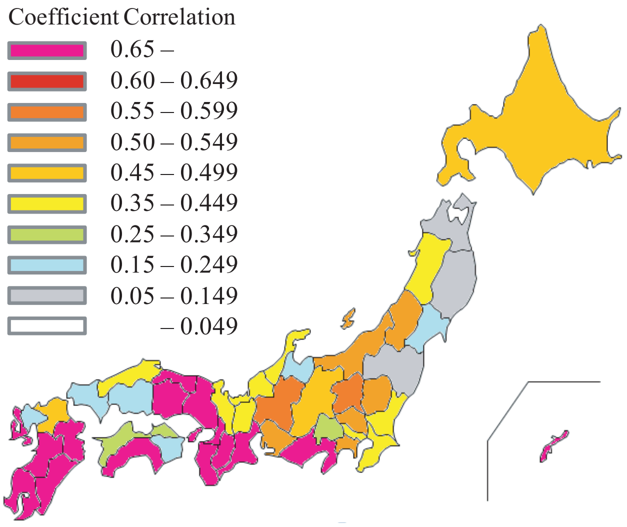

Figure 4. Spatial Dependency of Coefficient Correlation between the Simplest Web Sensor with Blogs and Snowfall Stats.

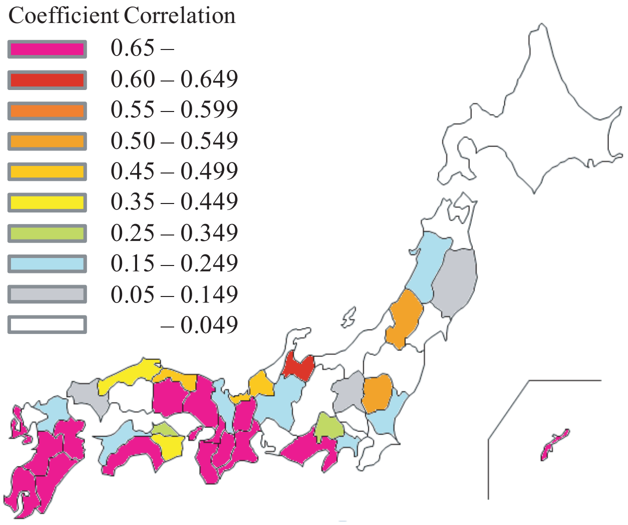

Figure 5. Spatial Dependency of Coefficient Correlation between the Normalized Web Sensor with Twitter and Snowfall Stats.

Sensors are not much different for rainfall and earthquake stats. Figures 4 and 5 show that more southern prefectures where it snows very little have higher coefficient correlation with snowfall stats. And also Figure 8 and 9 show that the simplest Web Sensor by analyzing weblogs fortunately gives enough high coefficient correlation with influenza stats to almost all prefectures in Japan, while the spatiotemporally-normalized Web Sensor by analyzing Tweets unfortunately gives too low coefficient correlation to numerous prefectures.

Figures 10 to 17 show the spatio-temporal data of a Japanese prefecture s per week t for four kinds of natural phenomena by the simplest or spatiotemporally-normalized Web Sensor with weblogs and four kinds of Japanese statistics as real-world databases. They show that the simplest and the spatiotemporally-normalized Web Sensors with weblogs are not much different, but rather the latter decays for some cases. Figures 16 and 17 show the most definite coefficient correlation between Web Sensors' spatio-temporal data and influenza stats' real-world data. Figures 12 and 13 also show more definite coefficient correlation with snowfall stats' real-world data, but unnaturally steep rising for only 16th to 18th weeks. And the other figures show that the Web Sensors' spatio-temporal data exhibit less volatility than rainfall and earthquake stats' real-world data. In the near future, the author has to investigate causes of these problems in more detail to improve the functions of Web Sensors.

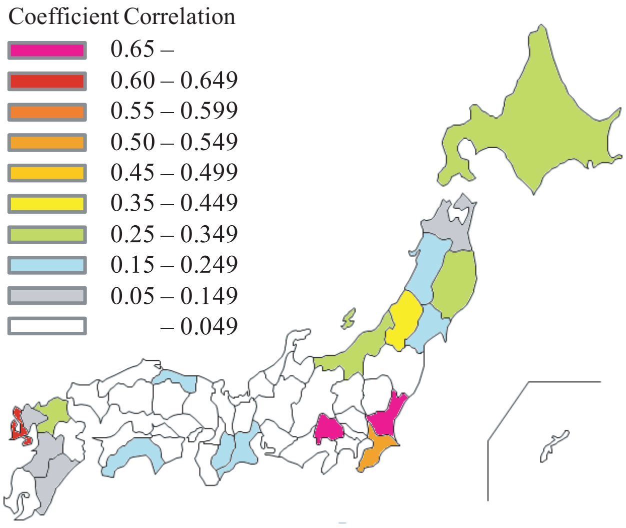

Figure 6. Spatial Dependency of Coefficient Correlation between the Simplest Web Sensor with Blogs and Earthquake Stats.

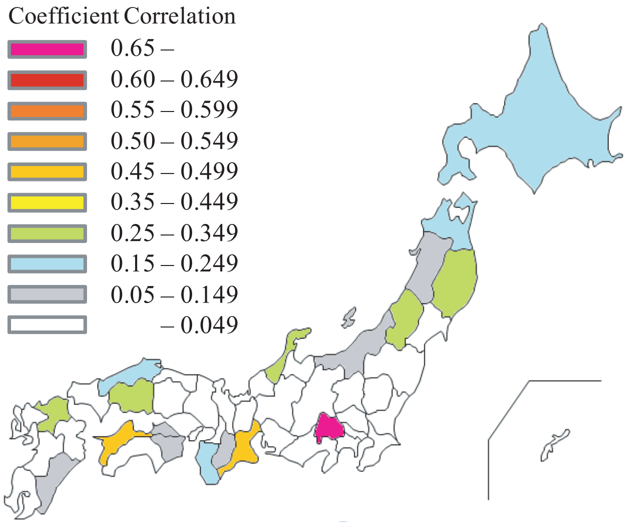

Figure 7. Spatial Dependency of Coefficient Correlation between the Normalized Web Sensor with Twitter and Earthquake Stats.

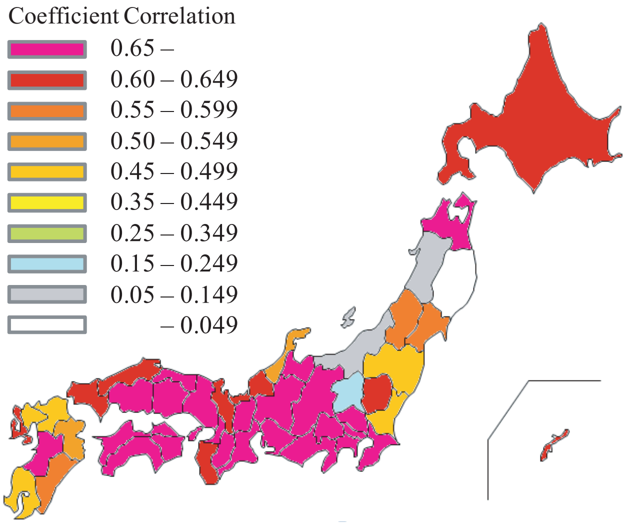

Figure 8. Spatial Dependency of Coefficient Correlation between the Simplest Web Sensor with Blogs and Influenza Stats.

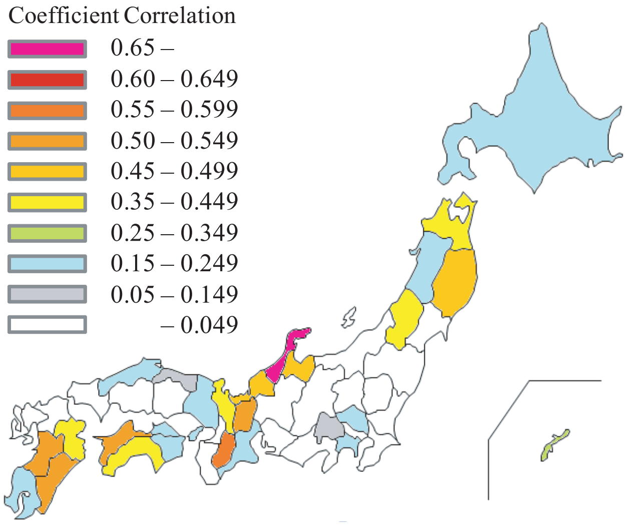

Figure 9. Spatial Dependency of Coefficient Correlation between the Normalized Web Sensor with Twitter and Influenza Stats.

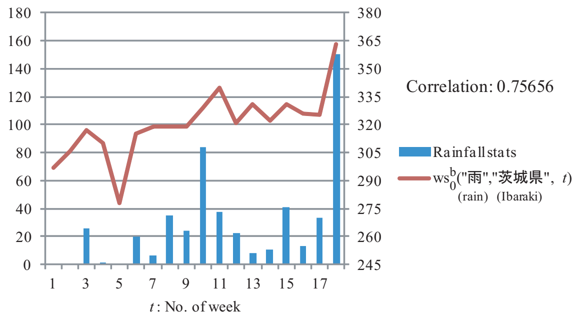

Figure 10. Simplest Web Sensor with Blogs and Rainfall Stats.

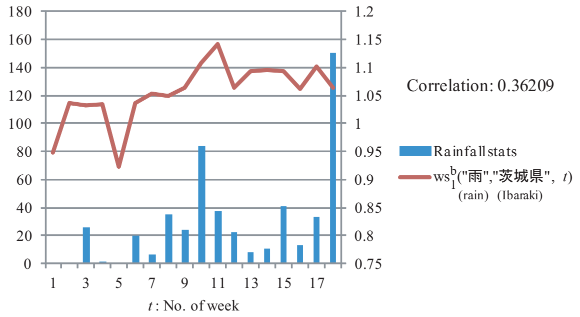

Figure 11. Normalized Web Sensor with Blogs and Rainfall Stats.

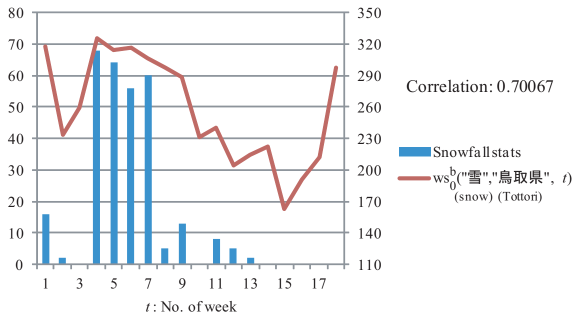

Figure 12. Simplest Web Sensor with Blogs and Snowfall Stats.

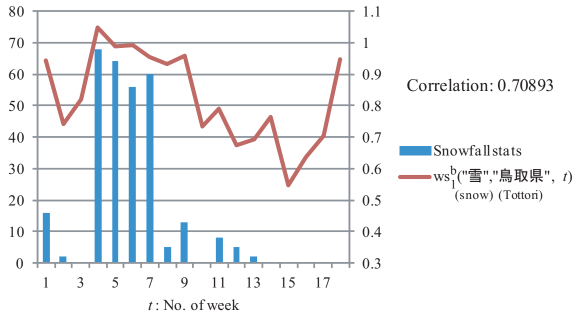

Figure 13. Normalized Web Sensor with Blogs and Snowfall Stats.

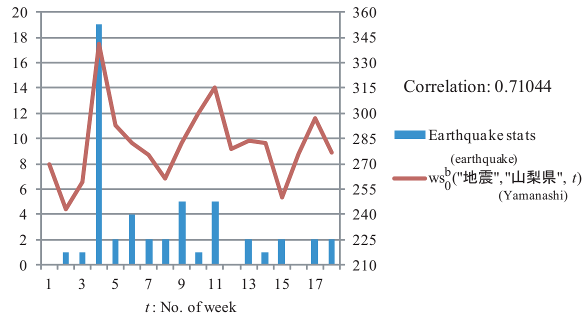

Figure 14. Simplest Web Sensor with Blogs and Earthquake Stats.

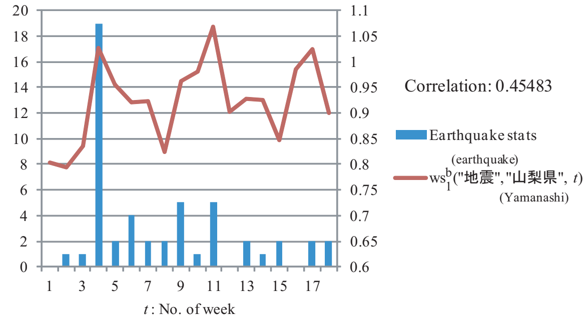

Figure 15. Normalized Web Sensor with Blogs and Earthquake Stats.

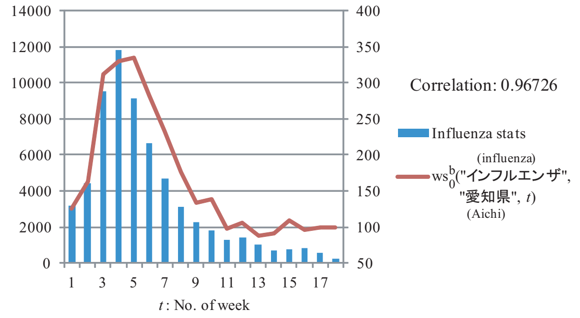

Figure 16. Simplest Web Sensor with Blogs and Influenza Stats.

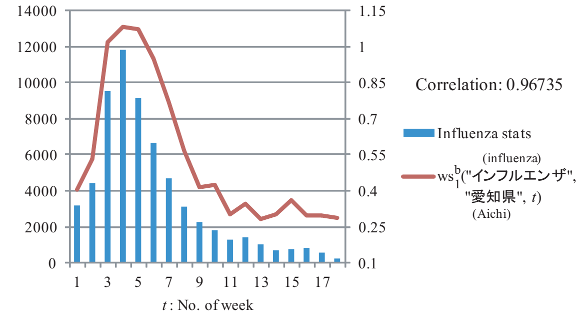

Figure 17. Normalized Web Sensor with Blogs and Influenza Stats.

V. Conclusion

This paper has defined the simplest Web Sensor and its spatiotemporally-normalized Web Sensor per week by region to mine SNS sites such as Twitter and Facebook, weblogs, news sites, or the whole Web for spatio-temporal data about a target phenomenon such as rainfall, snowfall, earthquake, and influenza in the physical world. And this paper has shown several experimental results to investigate the potential and reliability of these Web Sensors' spatio-temporal data by measuring the correlation coefficient with Japanese weather, earthquake, and influenza statistics per week by region as real-world data. As a result, this paper has found that weblogs are the most appropriate corpus of Web documents for Web Sensors, while SNS sites such as Twitter and Facebook are not so appropriate. The Web Sensors with SNS sites show negative correlation coefficient for negative phenomena such as earthquake and influenza.

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," Proceeding of the Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on advanced Intelligent Systems (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," Proceedings of the 2011 First IRAST International Conference on Data Engineering and Internet Technology (DEIT'11), pp.1042–1047 (2011).

[4] Hattori, S.: "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).

[5] Dave, K., Lawrence, S., and Pennock, D.M.: "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," Proceeding of the 12th International World Wide Web Conference (WWW'03), pp.519–528 (2003).

[6] Fujimura, S., Toyoda, M., and Kitsuregawa, M.: "A Reputation Extraction Method Considering Structure of Sentence," Proceeding of Data Eingineering Workshop (DEWS'05), 6C-i8 (2005).

[7] Tezuka, T., Kurashima, T., and Tanaka, K.: "Toward Tighter Integration of Web Search with a Geographic Information System," Proceeding of the 15th International World Wide Web Conference (WWW'06), pp.277–286 (2006).

[8] Inui, K., Abe, S., Morita, H., Eguchi, M., Sumida, A., Sao, C., Hara, K., Murakami, K., and Matsuyoshi, S.: "Experience Mining: Building a Large-Scale Database of Personal Experiences and Opinions from Web Documents," Proceeding of the 7th IEEE/WIC/ACM International Conference on Web Intelligence (WI'08), pp.314–321 (2008).

[9] Hattori, S., Ohshima, H., Oyama, S., and Tanaka, K.: "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).

[10] Hattori, S., and Tanaka, K.: "Extracting Concept Hierarchy Knowledge from the Web based on Property Inheritance and Aggregation," Proceeding of the 7th IEEE/WIC/ACM International Conference on Web Intelligence (WI'08), pp.432–437 (2008).

[11] Hattori, S.: "Object-oriented Semantic and Sensory Knowledge Extraction from the Web," Web Intelligence and Intelligent Agents, In-Tech, pp.365–390 (2010).

[12] Hattori, S., Tezuka, T., and Tanaka, K.: "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] Hattori, S.: "Peculiar Image Search by Web-extracted Appearance Descriptions," Proceedings of the 2nd International Conference on Soft Computing and Pattern Recognition (SoCPaR'10), pp.127–132 (2010).

[14] Shun Hattori: "Cross-Language Peculiar Image Search Using Translation between Japanese and English," Proceedings of the 2011 First IRAST International Conference on Data Engineering and Internet Technology (DEIT'11), pp.418–424 (2011).

[15] Shun Hattori: "Searching the Web for Peculiar Images based on Hand-made Concept Hierarchies," Proceedings of the 7th International Conference on Next Generation Web Services Practices (NWeSP'11), pp.152–157 (2011).

[16] Shun Hattori: "Query Expansion for Peculiar Images by Web-extracted Hyponyms," Proceedings of the 5th International Conference on Advances in Semantic Processing (SEMAPRO'11), pp.69–74 (2011).

[17] Shun Hattori: "Peculiar Image Retrieval by Cross-Language Web-extracted Appearance Descriptions," International Journal of Computer Information Systems and Industrial Management (IJCISIM), Vol.4, pp.486–495 (2012).

[18] Japan Meteorological Agency, http://www.jma.go.jp/jma/indexe.html (2012).

[19] National Institute of Infectious Diseases, http://www.nih.go.jp/niid/en/ (2012).

[20] Google, http://www.google.co.jp/ (2012).