Interpolating Lost Spatio-Temporal Data by Web Sensors

Shun Hattori

Web Intelligence Time-Space (WITS) Laboratory, College of Information and Systems, Graduate School of Engineering, Muroran Institute of Technology, 27–1 Mizumoto-cho, Muroran, Hokkaido 050–8585, Japan,
hattori@csse.muroran-it.ac.jp

Abstract

We experience various phenomena (e.g., rain, snow, and earthquake) in the physical world, while we carry out various actions (e.g., posting, querying, and e-shopping) in the Web world. Many researches have tried to mine the Web for knowledge about various phenomena in the physical world, and also several Web services using Web-mined knowledge have been made available for the public. Meanwhile, the previous papers have introduced various kinds of "Web Sensors" with Temporal Shift, Temporal Propagation, and Geospatial Propagation to sense the Web for knowledge about a targeted physical phenomenon, i.e., to extract its spatiotemporal data sensitively by analyzing big data on the Web (e.g., Web documents, Web query logs, and e-shopping logs), and compared them based on their correlation coefficients with Japan Meteorological Agency's physically-sensed spatiotemporal statistics to ensure the accuracy of Web-sensed spatiotemporal data sufficiently. As an industrial application of Web Sensors to a problem of the loss or error of physically-sensed spatiotemporal data due to some sort of troubles (e.g., temporary faults of JMA's observatories), this paper tries to enable Web Sensors to interpolate lost spatiotemporal data of physical statistics by regression analysis

Keywords: spatiotemporal data mining, big data analysis, web sensors, regression analysis

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., posting, querying, and e-shopping) in the Web world. Recently, there have been many researches to mine a huge amount of various documents in the exploding Web world, 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 (Dave et al., 2003; Fujimura et al., 2005) of various products and services in the physical world, experience mining (Tezuka et al., 2006; Inui et al., 2008) of various phenomena and events in the physical world, concept hierarchy (semantics) extraction (Hearst, 1992; Ruiz-Casado et al., 2007; Hattori et al., 2008; Hattori and Tanaka, 2008a; Hattori, 2010, 2012a) such as is-a/has-a relationships, and visual appearance (look and feel) extraction (Hattori, 2010; Tezuka and Tanaka, 2006; Hattori et al., 2007; Hattori and Tanaka, 2009; Hattori, 2012b, 2013a) of physical objects in the physical world. Meanwhile, Web services using 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 not enough investigations (Ginsberg et al., 2009; Sakaki et al., 2010; Aramaki et al., 2011) 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 (Hattori and Tanaka, 2008b; Hattori, 2011a,b, 2012c, 2013b,c,d, 2014, 2015) have introduced various kinds of "Web Sensors" to sense the Web for knowledge about a targeted phenomenon (e.g., rainfall, snowfall, and earthquake) in the physical world, i.e., to extract its spatiotemporal numerical values by analyzing big data on the Web, i.e., various action-based data (e.g., Web documents, Web query logs, and e-shopping logs) in the Web world, and investigated how correlated Web-sensed spatiotemporal data are with physically-sensed spatiotemporal data (e.g., rainfall, snowfall, and earthquake statistics of JMA (Japan Meteorological Agency, 2016)) as shown in Figure 1.

Document-based Web Sensors with "Temporal Shift" (Hattori, 2011a, 2013d) showed that

1. The optimized 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 Web 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 Web leads the physical

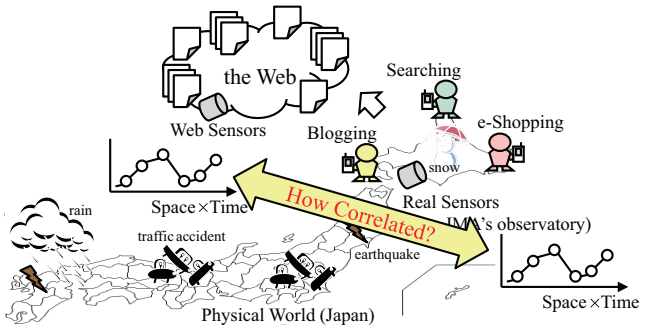

Figure 1. Can Web Sensors sense the physical world sensitively?

world) for snowfall, and Shifted-to-Past Web Sensor whose temporal shift parameter δ is positive gives the highest correlation coefficient (i.e., the Web follows the physical world) for earthquake,

2. The optimized 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 optimized temporal shift parameter δ for snowfall and earthquake varies more widely, and

3. More shaken geographical spaces and time periods are given higher correlation coefficient between Web-sensed spatiotemporal data and physically-sensed spatiotemporal data by the Great East Japan Earthquake (3.11).

Query-based Web Sensors using Web search query logs (Hattori, 2013c) are superior to Document-based Web Sensors using Web documents such as blogs for snowfall and earthquake, while Query-based Web Sensors are inferior to Document-based Web Sensors 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.

This paper introduces a novel method to interpolate the loss of physically-sensed spatiotemporal data about a targeted physical phenomenon (e.g., Japan Meteorological Agency's rainfall, snowfall, and earthquake statistics) by regression analysis between physically-sensed spatiotemporal data and Web-sensed spatiotemporal data about the targeted physical phenomenon, as an industrial application of variously defined "Web Sensors" with Temporal Shift, Temporal Propagation, and Geospatial Propagation to sense the Web for knowledge about a targeted physical phenomenon, i.e., to extract its spatiotemporal data sensitively by analyzing big data on the Web (e.g., Web documents, Web queries, and e-shopping logs).

The rest of this paper is organized as follows. Section 2 shows various definitions of Web Sensors, and Section 3 introduces a novel method of interpolating lost spatiotemporal data of physical statistics by Web Sensors and regression analysis. And Section 4 concludes this paper.

2 Web Sensors

This section shows various definitions of Web Sensors with Temporal Shift, Temporal Propagation, and Geospatial Propagation to 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 physical phenomenon (e.g., rainfall, snowfall, and earthquake).

First, the simplest and spatiotemporally-normalized Web Sensor (Hattori and Tanaka, 2008b; Hattori, 2013b) by using only Web documents (not Web search query logs (Hattori, 2013c)) 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 365 days or 52 weeks in 2011 such as January 1st (1st day) or from January 1st to 7th (1st week) and from December 24th to 30th (52nd week), and a linguistic keyword kw representing a targeted physical phenomenon, e.g., "rain," "snow," and "earthquake," is defined as

$$\mathrm{ws}(kw,s,t) := \frac{\mathrm{df}_t(\texttt{["}kw\texttt{" AND "}s\texttt{"]})}{\mathrm{df}_t(\texttt{["}s\texttt{"]})}, \qquad (1)$$

where $\mathrm{df}_t(\texttt{["}s\texttt{"]})$ stands for the Frequency of Web Documents searched from the Web, especially the Weblog, by submitting the search query q with the custom time range t to Google Web Search. Note that the Weblog is superior to the whole Web, Twitter, Facebook, and News as a corpus of Web Sensors (Hattori, 2012c).

Secondly, the temporally-shifted Web Sensor (Hattori, 2011a, 2013d) 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

$$\text{ws-ts}_\delta(kw,s,t) := \mathrm{ws}(kw,s,t+\delta). \qquad (2)$$

As shown in Figure 2, 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 (Hattori, 2011a) with a "Temporal Propagation" parameter σ_t^2, a geographic space s, a time period t, and

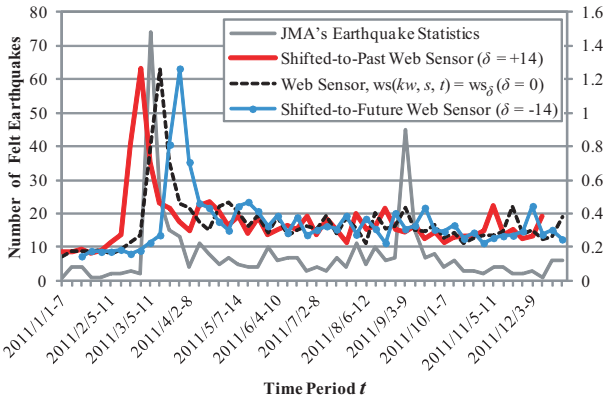

Figure 2. Temporally-shifted Web Sensors for earthquake and JMA's weekly earthquake statistics in Hokkaido prefecture, 2011.

a linguistic keyword kw representing a physical phenomenon is defined by integrating the surrounding time periods as

$$\text{ws-tp}^{\sigma_t^2}(kw,s,t) := \sum_{\forall\delta} \text{ws-ts}_\delta(kw,s,t)\cdot p^{\sigma_t^2}(\delta) \tag{3}$$

$$p^{\sigma_t^2}(\delta) := \frac{1}{\sqrt{2\pi\sigma_t^2}}\cdot\exp\Big(-\frac{\delta^2}{2\sigma_t^2}\Big) \tag{4}$$

where $p^{\sigma_t^2}(\delta)$ stands for a Normal Distribution $N(0,\sigma_t^2,\delta)$ with a mean 0 and a variance σ_t^2. In this paper, $\forall\delta$ is restricted to $[-30,30]$.

Next, the geospatially-propagated Web Sensor (Hattori, 2014, 2015) 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

$$\text{ws-sp}^{\sigma_s^2}(kw,s,t) := \sum_{\forall s_i} \text{ws}(kw,s_i,t)\cdot p^{\sigma_s^2}(\text{distance}(s,s_i)) \tag{5}$$

$$p^{\sigma_s^2}(d) := \frac{1}{\sqrt{2\pi\sigma_s^2}}\cdot\exp\Big(-\frac{d^2}{2\sigma_s^2}\Big) \tag{6}$$

where $\text{distance}(s,s_i)$ stands for the geographic distance [km] between geographic spaces s and s_i and is calculated based on their latitude and longitude. In this paper, $\forall s_i$ is restricted to 47 prefectures in Japan, and the latitude and longitude of its prefectural capital are used for calculating $\text{distance}(s,s_i)$ by using the Survey Calculation API of GSI (GeoSpatial Information Authority of Japan, 2016). 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].

3 Data Interpolation

As an industrial application of variously above-defined "Web Sensors" with Temporal Shift, Temporal Propagation, and Geospatial Propagation to the loss or error of physically-sensed spatiotemporal data due to some sort of troubles (e.g., temporary faults of Japan Meteorological Agency's observatories), this section proposes a novel method to interpolate lost spatiotemporal data about a targeted physical phenomenon (e.g., Japan Meteorological Agency's rainfall, snowfall, and earthquake statistics).

For a lost spatiotemporal numerical value $\text{ps}(s,t,kw)$ about a targeted physical phenomenon (which is represented by a linguistic keyword kw, e.g., "rain," "snow," and "earthquake") in a geographic space s, e.g., one of 47 prefectures in Japan such as "Hokkaido" over a time period t, e.g., one of 365 days or 52 weeks in 2011 such as January 1st (1st day) or from January 1st to 7th (1st week) and from December 24th to 30th (52nd week), the proposed method interpolates it by regression analysis with its surrounding N physically-sensed spatiotemporal data, their corresponding N Web-sensed spatiotemporal data, and its corresponding Web-sensed spatiotemporal data $\text{ws}(s,t,kw)$ or $\text{ws-XX}(s,t,kw)$ (where $\text{XX}\in\{$"ts," "tp," "sp"$\}$). In this paper, N is restricted to [1,30]. The variety of N physically-sensed spatiotemporal data surrounding a lost physically-sensed spatiotemporal numerical value $\text{ps}(s,t,kw)$ has:

1. N physically-sensed spatiotemporal data followed by it (i.e., only N past data),
$$\text{ps}(s,t-N,kw),\cdots,\text{ps}(s,t-1,kw),$$
2. N physically-sensed spatiotemporal data following it (i.e., only N future data),
$$\text{ps}(s,t+1,kw),\cdots,\text{ps}(s,t+N,kw),$$
3. $\lfloor N/2\rfloor$ physically-sensed spatiotemporal data followed by it and $\lceil N/2\rceil$ physically-sensed spatiotemporal data following it (i.e., both $\lfloor N/2\rfloor$ past data and $\lceil N/2\rceil$ future data, future-preferred when N is odd-numbered),
4. $\lceil N/2\rceil$ physically-sensed spatiotemporal data followed by it and $\lfloor N/2\rfloor$ physically-sensed spatiotemporal data following it (i.e., both $\lceil N/2\rceil$ past data and $\lfloor N/2\rfloor$ future data, past-preferred when N is odd-numbered).

The generalization of the above-mentioned examples is m ($\in[0,N]$) physically-sensed spatiotemporal data followed by it and $N-m$ physically-sensed spatiotemporal data following it (i.e., m past data and $N-m$ future data) as shown in Figure 3. Meanwhile, Figure 4 shows a simple method to interpolate a lost physically-sensed datum by average function using only physically-sensed data.

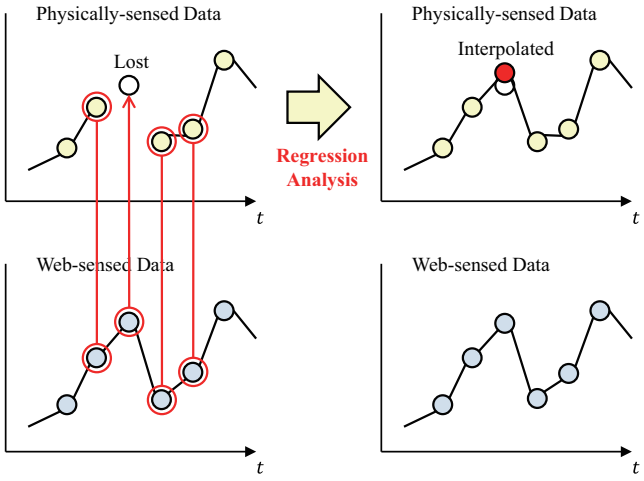

Figure 3. Interpolating a lost physically-sensed datum by Web Sensors and regression analysis using not only physically-sensed data but also Web-sensed data (when $N = 3$ and $m = 1$).

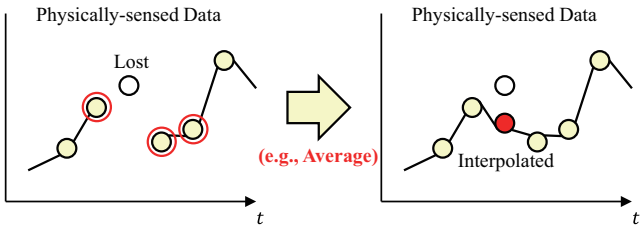

Figure 4. Interpolating a lost physically-sensed datum by average function using only physically-sensed data (adopted as a baseline in the experiment).

4 Conclusions

This paper has introduced a novel method to interpolate the loss of physically-sensed spatiotemporal data about a targeted physical phenomenon (e.g., Japan Meteorological Agency's rainfall, snowfall, and earthquake statistics) by regression analysis between physically-sensed spatiotemporal data and Web-sensed spatiotemporal data about the targeted physical phenomenon, as an industrial application of variously defined "Web Sensors" with Temporal Shift, Temporal Propagation, and Geospatial Propagation to sense the Web for knowledge about a targeted physical phenomenon, i.e., to extract its spatiotemporal data sensitively by analyzing big data on the Web (e.g., Web documents, Web queries, and e-shopping logs).

The future work has to perform experiments to validate the introduced method of interpolating lost spatiotemporal data of physical statistics by Web Sensors and regression analysis, and also will try to apply the other kinds of physical phenomena to the proposed interpolation. In addition, Web Sensors will be able to forecast future data about a targeted physical phenomenon and to alert falsified data of real statistics.

Acknowledgment

This research project was partially supported by a grant-in-aid for scientific research from the Japan Society for Promotion of Science (15K00329).

References

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: Detecting influenza epidemics using Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'11)*, pages 1568–1576, 2011.

Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the Peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International World Wide Conference (WWW'03)*, pages 519–528, 2003.

Shigeru Fujimura, Masashi Toyoda, and Masaru Kitsuregawa. A reputation extraction method considering structure of sentence. In *Proceedings of the 16th IEICE Data Engineering Workshop (DEWS'05)*, 6C-i8, 2005.

GeoSpatial Information Authority of Japan. Sokuchi survey calculation API No.2, 2016. http://vldb.gsi.go.jp/sokuchi/surveycalc/main.html.

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014, 2009.

Shun Hattori. Object-oriented semantic and sensory knowledge extraction from the Web. In *Web Intelligence and Intelligent Agents*, chapter 18, pages 365–390. In-Tech, 2010.

Shun Hattori. Secure spaces and spatio-temporal Weblog sensors with temporal shift and propagation. In *Proceedings of the 1st IRAST International Conference on Data Engineering and Internet Technology (DEIT'11)*, LNEE vol.157, pages 343–349, 2011a.

Shun Hattori. Linearly-combined Web sensors for spatio-temporal data extraction from the Web. In *Proceedings of the 6th International Workshop on Spatial and Spatiotemporal Data Mining (SSTDM'11)*, pages 897–904, 2011b.

Shun Hattori. Hyponym extraction from the Web based on property inheritance of text and image features. In *Proceedings of the 6th International Conference on Advances in Semantic Processing (SEMAPRO'12)*, pages 109–114, 2012a.

Shun Hattori. Peculiar image retrieval by cross-language Web-extracted appearance descriptions. *International Journal of Computer Information Systems and Industrial Management (IJCISIM)*, 4:486–495, 2012b.

Shun Hattori. Spatio-temporal Web sensors by social network analysis. In *Proceedings of the 3rd International Workshop on Business Applications of Social Network Analysis (BASNA'12)*, pages 1020–1027, 2012c.

Shun Hattori. Hyponymy-based peculiar image retrieval. *International Journal of Computer Information Systems and Industrial Management (IJCISIM)*, 5:79–88, 2013a.

Shun Hattori. Granularity analysis for spatio-temporal Web sensors. In *Proceedings of the WASET International Conference on Knowledge Management (ICKM'13)*, pages 192–200, 2013b.

Shun Hattori. Spatio-temporal Web sensors using Web queries vs. documents. *Journal of Automation and Control Engineering (JOACE)*, 1(3):192–197, 2013c.

Shun Hattori. Spatio-temporal dependency analysis for temporally-shifted Web sensors. In *Proceedings of the 2nd SDIWC International Conference on Informatics & Applications (ICIA'13)*, pages 30–35, 2013d.

Shun Hattori. Spatio-temporal propagation for Web sensors. In *Proceedings of the SDIWC International Conference on Computer Science, Computer Engineering, and Social Media (CSCESM'14)*, pages 69–76, 2014.

Shun Hattori. Deflection analysis for spatially propagated Web sensors. In *Proceedings of the SDIWC International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC'15)*, pages 20–28, 2015.

Shun Hattori and Katsumi Tanaka. Extracting concept hierarchy knowledge from the Web based on property inheritance and aggregation. In *Proceedings of the 7th IEEE/WIC/ACM International Conference on Web Intelligence (WI'08)*, pages 432–437, 2008a.

Shun Hattori and Katsumi Tanaka. Mining the Web for access decision-making in secure spaces. In *Proceedings 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, pages 370–375, 2008b.

Shun Hattori and Katsumi Tanaka. Object-name search by visual appearance and spatio-temporal descriptions. In *Proceedings of the 3rd International Conference on Ubiquitous Information Management and Communication (ICUIMC'09)*, pages 63–70, 2009.

Shun Hattori, Taro Tezuka, and Katsumi Tanaka. Mining the Web for appearance description. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA'07)*, LNCS vol.4653, pages 790–800, 2007.

Shun Hattori, Hiroaki Ohshima, Satoshi Oyama, and Katsumi Tanaka. Mining the Web for hyponymy relations based on property inheritance. In *Proceedings of the 10th Asia-Pacific Web Conference (APWeb'08)*, LNCS vol.4976, pages 99–110, 2008.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING'92)*, volume 2, pages 539–545, 1992.

Kentaro Inui, Shuya Abe, Kazuo Hara, Hiraku Morita, Chitose Sao, Megumi Eguchi, Asuka Sumida, Koji Murakami, and Suguru Matsuyoshi. Experience mining: building a large-scale database of personal experiences and opinions from Web documents. In *Proceedings of the 7th IEEE/WIC/ACM International Conference on Web Intelligence (WI'08)*, pages 314–321, 2008.

Japan Meteorological Agency. Weather, climate & earthquake information, 2016. http://www.jma.go.jp/jma/en/menu.html.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. Automatising the learning of lexical patterns: an application to the enrichment of WordNet by extracting semantic relationships from Wikipedia. *Data & Knowledge Engineering*, 61 (3):484–499, 2007.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International World Wide Web Conference (WWW'10)*, pages 851–860, 2010.

Taro Tezuka and Katsumi Tanaka. Visual description conversion for enhancing search engines and navigational systems. In *Proceedings of the 8th Asia-Pacific Web Conference (APWeb'06)*, LNCS vol.3841, pages 955–960, 2006.

Taro Tezuka, Takeshi Kurashima, and Katsumi Tanaka. Toward tighter integration of Web search with a geographic information system. In *Proceedings of the 15th International World Wide Web Conference (WWW'06)*, pages 277–286, 2006.