Extraction of Context Information from Web Content Using Entity Linking

Norifumi Hirata†, Shun Shiramatsu†, Tadachika Ozono†, and Toramatsu Shintani†

†Dept. of Computer Science and Engineering, Graduate School of Engineering Nagoya Institute of Technology Gokiso-cho, Showa-ku, Nagoya, Aichi, 466-8555 Japan

Summary
We developed a system for extracting context information from microblogs, such as Twitter, using entity linking. A feature of Twitter is real-time. Our proposed system uses news articles to generate entity links since tweets are posted in real-time. It is difficult to extract context information from Twitter because the maximum length of a tweet is 140 characters. Therefore, we use both news articles and microblogs for entity linking. Entity links contain context entities. When our system extracts regional context, it uses entity links about geographical regions. The experimental results suggest that our proposed system can extract context entities based on Twitter users and news articles.

Key words:
Microblogs, News Articles, Context Information, Entity Linking

1. Introduction

We developed a system for extracting context information from microblogs, such as Twitter, using entity linking. Twitter[1] is one of microblogging services. The maximum length of a tweet is 140 characters; therefore it is difficult to identify the content of a tweet compared to news articles and blog posts. The aim with our system is to extract context information using entity linking based on news articles and tweets. In this paper, we define an entity is a term in news articles and tweets.

Real time is a feature of Twitter. Wikipedia and WordNet are used to generate entity linking [1] and contain comparatively static information. Our proposed system extracts novel information. The system uses news articles and tweets because they contain such information. Furthermore, the content of a tweet depends on each user. The contexts of tweets are not same even if the texts of tweets are same.

There are many types of context information. The proposed system generates entity links with specific context information. If we would like to extract context information about geographical regions, the system generates entity links on those regions. It extracts context information from tweets using generated entity links.

[1] http://twitter.com

Manuscript received February 5, 2013
Manuscript revised February 20, 2013

2. Related Works

It is difficult to extract context information from Twitter because the maximum length of a tweet is 140 characters. A context extraction system must make up for the lack of context information from other information sources. Context information represents features of a tweet such as personal name, geographical name, and time. Our proposed system uses news articles for context extraction. News articles often have information regarding the five Ws – the 'Who, What, Where, When, and Why'. They are written about topics. Therefore, they are appropriate for context extraction from tweets.

We have developed e-Participation Web platform [2][3] based on Linked Open Data [4] that targets regional communities in Japan. The aim is to utilize web contents related to target regions for sharing public concerns among citizens, government officials, and experts. It is important to show Web content classified based on geographical regions for supporting discussion.

Methods for extracting keywords related to an individual have been proposed [4][5]. The method [4] extracts keywords using co-occurrence terms on the Web. The method uses context words such as "research", "sport", and "economy" for word sense disambiguation. For extracting context extraction from a tweet, it is necessary to restrict genres because the content of each tweet is not the same.

The method [5] generates résumés from the Web. The method classifies extracted information by the type of personal history, i.e., academic, career family, and award. Classification is based on a support vector machine (SVM). This study showed that SVM is a sufficient method for extracting specific attribute information.

Named entity extraction has been extensively long time. The sixth in a series of Message Understanding Conferences (MUC-6) [7][8] defined seven categories such as person, location, and organization. Seven categories are sufficient because the aim of MUC-6 was the structuring activities about business and country from newspapers. However, necessary categories depend on the extraction target.

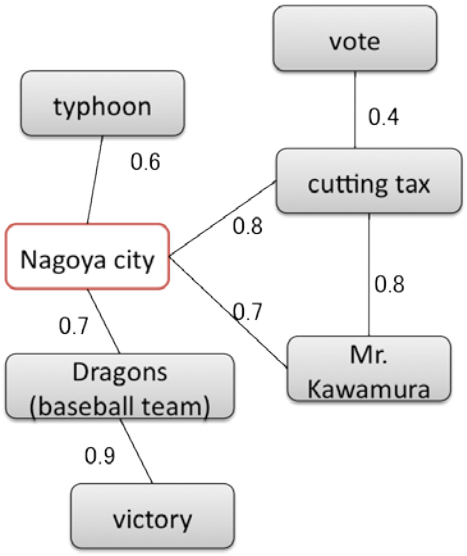

Fig.1 Example of a entity link about Nagoya city

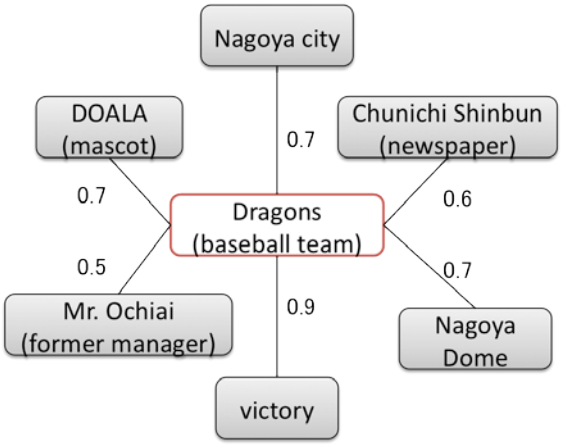

Fig.2 Example of a entity link about a baseball team

3. Structuring Entity Link

3.1 Structure of Entity Link

We define an entity is a term in news articles and tweets in this paper. Two entities are related using a link. A link value represents the relation between two entities. The proposed system generates entity links on each context entity that represents context information.

Fig.1 shows an example of an entity link about "Nagoya city". The entity link contains related entities such as "typhoon", "vote", and "Dragons (the local professional baseball team)". The number on the link presents the relation between the entities. For example, the value of a relation between the entities "typhoon" and "Nagoya city" is 0.6.

The system does not require entities that are not related to context entities because the aim of this study was to extract context information. Fig.2 is another example of an entity link. If our system extracts context entities on baseball teams, an entity link in Fig.1 is not appropriate. Our system requires the entity link in Fig.2.

It is important to extract novel information to generate entity links for extracting context from a tweet. Wikipedia, WordNet, and dictionaries are static and general information sources, and each tweet is a dynamic and personal information source. News articles are a dynamic and general information sources. Our proposed system uses news articles to generate entity links because tweets are posted in real-time.

3.2 Method for Structuring Entity Links

If a relation value exceeds a threshold, the proposed system links between entities and evaluates the relation between entity e_1 and entity e_2 as follows:

$$s_t(e_1,e_2)=\frac{\left|f_{t1,t2}(e_1)\cap f_{t1,t2}(e_2)\right|}{\left|f_{t1,t2}(e_1)\cup f_{t1,t2}(e_2)\right|} \quad (1)$$

$$t_1=t-T_1$$

$$t_2=t-T_2$$

A relation is evaluated using the co-occurrence frequency [9] of entities and time restrictions T_1 and T_1. In Eq. 1, $f_{t1,t2}(e)$ denotes a set of news articles containing entity e. $f_{t1,t2}(e_1) \cap f_{t1,t2}(e_2)$ denotes a set of news articles that contains both entities e_1 and e_2, and $t1$ and $t2$ are time restrictions. If $f_{t1,t2}(e)$ is zero, $s_t(e_1, e_2)$ is zero.

The system uses Japanese language morphological analysis software called MeCab [10] for evaluating context. The system obtains attribute information of terms such as personal names, organization names and geographical names by using MeCab.

The proposed system generates entity links about each context entity that represents context information. Equation (2) expresses the relation value between context entities e_0 and e_N. Entity e_N has N links between it and e_0.

$$relation_t(e_0,e_N)=\prod_{i=0}^{N-1}s_t(e_i,e_{i+1}) \quad (2)$$

where t is the post time of a tweet from which to extract contexts. If e_0 equals e_N, $relation_t$ is 1. The system selects a minimum $relation_t$ path from e_0 to e_1. Therefore, the system extracts entity links from context entities.

4. Context Extraction Using News Articles

4.1 Method for Context Extraction from Tweets

Our proposed system extracts context information from tweets using entity links. We define context information extraction as context entity extraction. The system compares entity links to terms in a tweet text.

The system evaluates whether a tweet m at time t is related to e_0 as follows:

$$context(e_0, m) = \sum_{e_j \in m.entities} relation_{t,u}(e_0, e_j) \quad (3)$$

where *m.entities* is a set of entities in a tweet *m*. The system evaluates a term as an entity. If $context(e_0, m)$ exceeds a threshold, *m* is assumed to e_0. Therefore, the system can obtain some context entities.

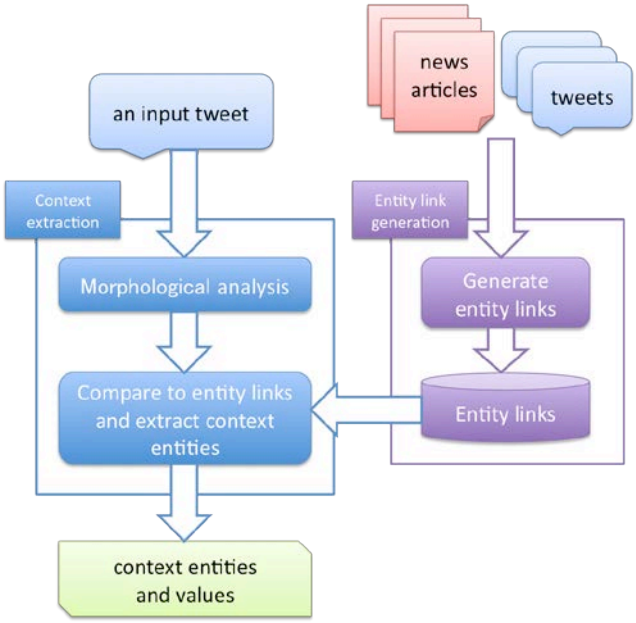

Fig.2 A system structure for context extraction

If the system obtains a tweet m_a states, "I cannot vote because of the typhoon.", It calculates $context("Nagoya city", m_a)$ is 1.02 from the entity links in Fig.1, calculates the relation value between "typhoon" and "Nagoya city" as 0.6 from equation (2), and calculates the relation value between "vote" and "Nagoya city" as 0.32 (=0.4*0.8). Therefore, $context("Nagoya city", m_a)$ is 0.92 (=0.6+0.32). The system evaluates 0.92 as the relation value of "Nagoya city" and m_a.

In equation (3), $context(e_0, m)$ represents the sum of $relation_t(e_i, e_j)$, and $context(e_0, m)$ is not an average of $relation_t(e_i, e_j)$. If *m* contains an *e*, the system evaluates the relation regardless of other entities.

4.2 System Structure

Fig. 3 shows the structure of our proposed system. A system input is a tweet and the system outputs are context entities and values. The proposed system has two modules, i.e., context evaluation, and entity link generation.

The entity link generation module generates entity links from news articles and tweets. The task of this module is to save entity links for the context evaluation module. The context evaluation module extracts context entities using entity links. First, the context evaluation module analyzes morphemes of an input tweet. Then, it compares morphemes to entity links. The task of the context evaluation module is to extract context entities from an input tweet.

Table 1 News articles for generating entity links

	published date of news articles	number of news articles	number of extracted context entities about region names
Entity links A	2011/08/ 26-28	222	328
Entity links B	2011/09/ 26-28	352	462

Table 2 Extracted contexts from entity links A

region name	relation value
North Carolina	0.917
New York	0.569
Virgina	0.458
NY	0.388
America	0.362

Table 3 Extracted contexts from entity links A

region name	relation value
Kinka	0.250
Wakayama	0.199
Amami	0.182
Tatsugo	0.182
Kagoshima	0.182

4.3 Examples of Context Extraction from a Tweet

The system extracted context information using equation (2). We compared two entity links generated at different times. The system used news articles published on MSN Sankei News[2] to generate entity links. Table 1 summarizes the information of the news articles. The text of an input tweet was, "We had been ordered to evacuate". The system extracted context information about geographical region names.

Tables 2 and 3 list the results. "North Carolina" and "New York" were extracted from entity links A. On August 27th 2011, hurricane Irene made landfall in Cape Lookout. "North Carolina" and "New York" were extracted because news articles about the hurricane were published. "Kagoshima" and "Amami" were extracted from entity links A. They are regions in Japan. On September 26th, it rained heavily in the Amami Islands in Kagoshima Prefecture. "Kinka" and "Wakayama" are areas geographically distant from Amami Islands. Some news articles about past floods in these areas were published. We confirmed that the system can change extracted context information by entity links.

[2] http://sankei.jp.msn.com/

5. Context Extraction Based on Twitter Users

5.1 Entity Links and Twitter Users

The system extracted context information using entity links that are generated from news articles. The extracted information was general. If the same tweets are posted, the system extracts the same context entities even if the twitter users are different. Twitter users have each context. Therefore, the system should extract context entities based on Twitter users.

If the system obtains a tweet stating, “victory!!”, the system calculate the relation value between “victory” and “Nagoya city” as 0.63 as shown in Fig. 1. If the system uses the entity links in Fig. 4, it calculates the relation value between “victory” and “Fukuoka city” as 0.64. Our system based on the method in Section 4.1 selects “Fukuoka city” as the context entity regarding the input tweets. If a user who posted the tweet lived in Nagoya city, it is not appropriate to extract “Fukuoka city”. We propose a method for extracting context information based on past tweets of a Twitter user.

5.2 Weighting Based on Context Entities of Twitter Users

The system based on this method extracts context entities from past tweets of a Twitter user. It then weights equation (2) by using these context entities.

The system calculates the relation value using $w1_t(u,e_0)$, which denotes the weight of e_0 of a user u. $w1_t(u,e_0)$ is expressed in equations (4) and (5).

$$s_{t,u}(e_0,e_1) = w1_t(u,e_0)\cdot(1-s_t(e_0,e_1)) + s_t(e_0,e_1) \quad \text{(4)}$$

$$w1_t(u,e_0) = \frac{\left|f_{t1,t2}(u,e_0)\right|}{\sum_{e_c \in context_u}\left|f_{t1,t2}(u,e_c)\right|} \quad \text{(5)}$$

where e_1 is an entity linked to a context entity e_0. $w1_t(u,e_0)$ is the rate of a user's context, and $f_{t1,t2}(u, e)$ is a set of entities contained in tweets. The Tweets are posted by a user u between $t1$ and $t2$. The term $context_u$ denotes a set of user's context entities. If a user u frequently posts a context entity, the weight of the context entity is evaluated highly.

Fig. 5 is an example of a entity link weighted by a user's context entity “Nagoya city”. If $w1_t(u, \text{“Nagoya city”})$is 0.5, the relation values are weighted, as shown in Fig. 5. A weighed entity link provides a new value. The relation value between “Nagoya city” and “victory” is 0.765 (=0.9*0.85). If a twitter user posts the context entity “Nagoya city”, the system evaluates determines that the weight of “Nagoya city” is higher than the weight of “Fukuoka city”.

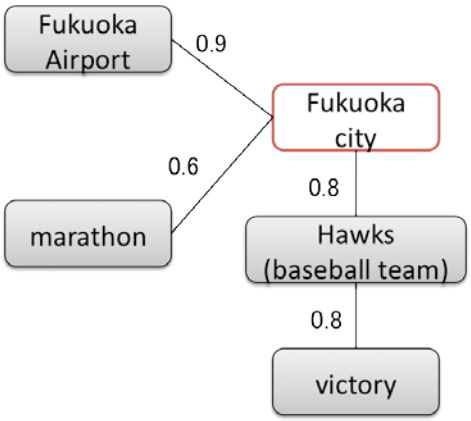

Fig. 4 Example of a entity link about Fukuoka city

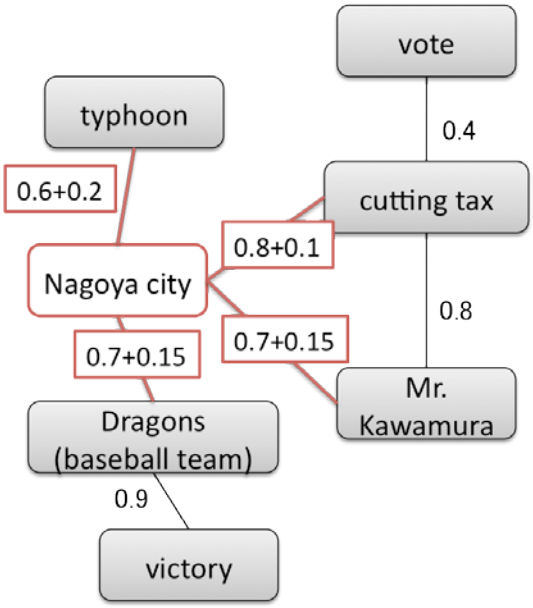

Fig. 5 Example of a entity link weighted by a user's context entity

5.3 Example of Context Extraction based on Context Entities of Twitter Users

The system using the method proposed in Section 5.2 extracted context entities based on information of a Twitter user. The system used Entity links A mentioned in Section 4.3. A text of an input tweet was “We had been ordered to evacuate”. The input tweet was the same as that mentioned in Section 4.3. The system used the following past tweets:

- “Today, does the Fukushima nuclear power plant have any problems?”
- “It is too late to order evacuation in Fukushima in eastern Japan”

Table 4 Extracted contexts from entity links using user's context entities

region name	relation value
North Carolina	0.917
Fukushima	0.633
eastern Japan	0.576
New York	0.569
Virgina	0.458
NY	0.388

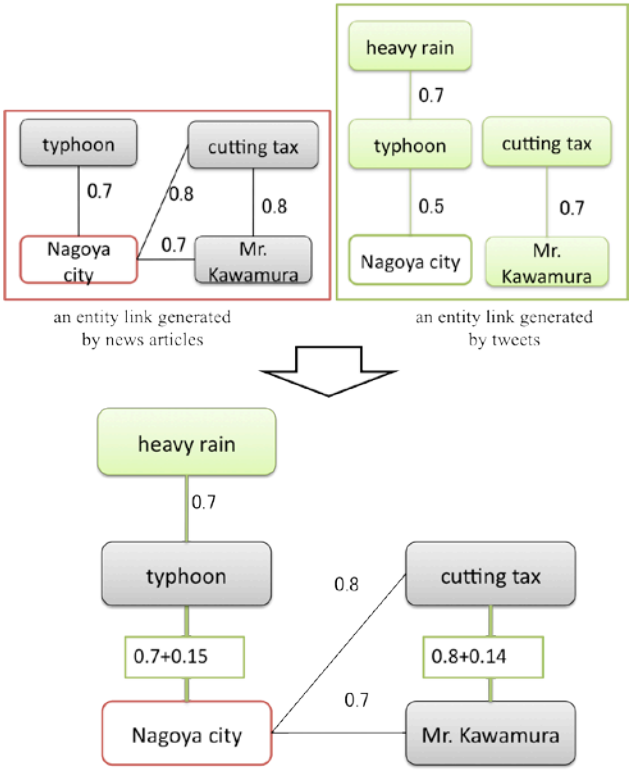

Fig. 6 Example of a entity link weighted by a user's entity links

- "Do you get a leave order after the Tohoku earthquake?"

The context entities about regions were "Fukushima" and "eastern Japan".

Table 4 lists the results of context extraction using user's context entities. The weight of "Fukushima" and "eastern Japan" was higher than the results in table 2. The results suggest that the system can extract context entities based on Twitter users and news articles. However, the maximum value of a relation is "North Carolina". The results show that the effectiveness of this method is not sufficient.

5.4 Weighting Based on Entity Links of a Twitter User

We propose another method for extracting contexts based on past tweets of users. The system generates entity links from past tweets of a Twitter user by combining equation (2) using user's context entities.

Table 5 Extracted contexts from entity links using past tweets

region name	relation value
eastern japan	1.076
North Carolina	0.917
Fukushima	0.895
New York	0.569
Virgina	0.458
NY	0.388

Fig. 6 shows an example of an entity link weighted by a user's entity links. An entity link generated by news articles is combined with an entity link generated by tweets. Because the proposed system uses all entities in past tweets, this method is more effective than that method proposed in Section 5.2.

The system calculates the relation value using $w2_t(u,e_i,,e_i)$, which denotes a user's relation value of a relation between entity e_i and entity e_j. Details of $w2_t(u,e_i,,e_i)$ are expressed in equations (6) and (7).

$$s_t(u,e_0,e_1) = \prod w2_t(u,e_i,e_j)\cdot(1-s_t(e_0,e_1)) + s_t(e_0,e_1) \quad (6)$$

$$w2_t(u,e_i,e_j) = \frac{\left|f_{t1,t2}(u,e_i)\cap f_{t1,t2}(u,e_j)\right|}{\left|f_{t1,t2}(u,e_i)\cup f_{t1,t2}(u,e_j)\right|} \quad (7)$$

$$t1 = t - T_1$$

$$t2 = t - T_2$$

where $f_{t1,\,t2}(u,e)$ is a set of entity e that is contained in a user's past tweets. The Tweets are posted by a user u between time $t1$ and $t2$.

5.5 Example of Context Extraction based on Context Entities of a Twitter User

The system, using the method proposed in Section 5.2, extracted using entity links of a Twitter user. The system used entity links A mentioned in Section 4.3. The input tweets were the same as those mentioned in Section 5.2.

Table 5 lists the results of context extraction using entity links of a Twitter user. The region name that had the maximum relation value was "eastern Japan". "Fukushima" and "North Carolina" have similar values. The results show that effectiveness of the method is sufficient.

6. Conclusion

We proposed a system for extracting context entities from entity links, generated from news articles and tweets. The experimental results suggest that the system can extract context entities based on Twitter users and news articles. The system can extract context entities that appear frequently in news articles.

The proposed system uses news articles to generate entity links because tweets are posted in real-time. However, context information from news articles is limited since they contain static information. Each tweet is a dynamic and personal information. Therefore, it is necessary to combine the information sources to extract context information.

Acknowledgement

This work was supported in part by Strategic Information and Communications R&D Promotion Programme (SCOPE) from the Japan's Ministry of Internal Affairs and Communications, Japan and Program for Revitalization Promotion, JST A-STEP.

References

[1] Xianpei Han and Le Sun and Jun Zhao, "Collective Entity Linking in Web Tex, A Graph-Based Method," Proceedings of the 34th international ACM SIGIR conference on Research and development in Information, pp.765–774, 2011.
[2] Shun Shiramatsu, Robin M. E. Swezey, Hiroyuki Sano, Norifumi Hirata, Tadachika Ozono and Toramatsu Shintani, "Structuring Japanese Regional Information Gathered from the Web as Linked Open Data for Use in Concern Assessment", The 4th International Conference on eParticipation (ePart 2012), 2012.
[3] Robin M. E. Swezey, Shun Shiramatsu, Tadachika Ozono, and Toramatsu Shintani, "An Improvement for Naive Bayes Text Classification Applied to Online Imbalanced Crowdsourced Corpora", In Modern Advances in Intelligent Systems and Tools, Studies in Computational Intelligence, Vol. 431, pp. 147–152, Springer, (2012). 2012.
[4] Christian Bizer, Tom Heath and Tim Berners-Lee, "Linked Data - The Story So Far", International Journal on Semantic Web and Information Systems, pp.1—22, 2009.
[5] Mori Junichiro, Yutaka Matsuo, and Mitsuru Ishizuka "Personal Keyword Extraction from the Web", Journal of the Japanese Society for Artificial Intelligence, Vol.20, No.5, pp.337–345, 2005.
[6] Hiroshi Ueda, Harumi Murakami, and Shoji Tatsumi, "Creating Curriculum Vitae for Understanding People on the Web", Journal of the Japanese Society for Artificial Intelligence, No.25, Vol.1, pp.144–156, 2010.
[7] MUC-6, the sixth in a series of Message Understanding Con- ferences, http://cs.nyu.edu/faculty/grishman/muc6.html
[8] Ralph Grishman, Beth Sundheim, "Message Understanding Conference-6: a brief history", In Proceedings of the 16th conference on Computational linguistics (COLING '96), Vol. 1, 1996.
[9] Anish Das Sarma, Alpa Jain and Cong Yu, "Dyanmic Relationship and Event Discovery", Proceedings of the 4th ACM International Conference on Web Search and Data Mining, pp. 207–216, 2011.
[10] MeCab: Yet Another Part-of-Speech and Morphological An- alyzer, http://mecab.sourceforge.net

Norifumi Hirata is currently a PhD candidate at Nagoya Institute of Technology, Japan. He received his MS and BS degrees in computer science from Nagoya Institute of Technology of Nagoya City, Japan, in 2010 and 2008. He is currently (2013) a student at Nagoya Institute of Technology, Japan. His research interests include document classification, system-user interaction and text mining.

Shun Shiramatsu received his PhD degree in information science from Kyoto University, Japan, in 2009 and his MS degree in information science from Tokyo University of Science in 2003. He is currently (2013) an Assistant Professor of Computer Science. His research interests include discussion support and conversation modeling.

Tadachika Ozono received his MS and PhD degrees in computer science from Nagoya Institute of Technology of Nagoya City, Japan, in 1998 and 2000. He is currently an Associate Professor of Computer Science at Nagoya Institute of Technology of Nagoya City, Japan. Currently (2013) his main research interest is Web intelligence.

Toramatsu Shintani received his MS degree in industrial engineering and his PhD degree in computer science from Tokyo University of Science in 1982 and 1993, respectively. He was a research staff member at Fujitsu Limited from 1982 to 1993. He is currently (2013) a Professor of Computer Science at Nagoya Institute of Technology of Nagoya City, Japan. His current research interests include decision support systems and Web intelligence.