Structuring Japanese Regional Information Gathered from the Web as Linked Open Data for Use in Concern Assessment

Shun Shiramatsu, Robin M. E. Swezey, Hiroyuki Sano, Norifumi Hirata, Tadachika Ozono, and Toramatsu Shintani

Graduate School of Engineering, Nagoya Institute of Technology

Abstract. We are developing an eParticipation web platform based on Linked Open Data that targets regional communities in Japan. To increase transparency and public participation, we aim to utilize web contents related to target regions for sharing public concerns among citizens, government officials, and experts. We have designed a Linked Open Data set called SOCIA (Social Opinions and Concerns for Ideal Argumentation) to structure regional web contents (e.g. regional news articles, microblog posts, and minutes of city council meetings) and utilize them for eParticipation and concern assessment. The web contents are semi-automatically structured by our text mining system, Sophia, on the basis of regions and events extracted from news articles on the web. Minutes of city council meetings stored in SOCIA are annotated with discourse salience in order to visualize topic transitions in a meeting transcript. We also developed a prototype debate support system called citispe@k that uses SOCIA to help citizens share their concerns. Users can submit agendas, ideas, questions, and answers by referencing the structured regional information in SOCIA. Moreover, they can annotate SOCIA data with tags representing criteria for assessing concerns or utterance intentions.

Keywords: Linked Open Data, concern assessment, information structuring, public involvement

1 Introduction

In addition to having endured huge earthquakes and a nuclear catastrophe, Japanese regional communities face ongoing social issues. Besides, the issues and risks have become diversified: radiation pollution, climate changes, financial problems, aging population, welfare problems, etc. Public involvement, i.e. citizen participation in deciding public policy, has thus become more important, especially in regional communities. Although public involvement is characterized as an interactive communication process among stakeholders[1], stating opinions is not easy for Japanese citizens because they tend to be reticent, and are not experts about the diversified social issues. To facilitate public involvement, we

are developing an eParticipation web platform, O_2,[1] based on Linked Open Data (LOD).

LOD, semantically connected data with universal resource identifiers (URIs) and the resource description framework (RDF) can be used for supporting citizens' deliberation because the LOD mechanism enables sophisticated information provision with semantic links [2]. Our platform aims to increase transparency, participation, and collaboration in Japanese regional communities. We focus on transparency and participation by using web contents as background information related to regional concerns.

To increase transparency and participation in regional communities, it is important to share public concerns among citizens, government officials, and experts. Background information should be structured and open in order to facilitate assessing and sharing public concerns. We have developed an LOD data set called SOCIA (Social Opinions and Concerns for Ideal Argumentation) that consists of Japanese regional news articles, microblog posts, and minutes of city council meetings. SOCIA is designed to be used for supporting concern assessment. It is semi-automatically structured by our text mining system, Sophia, on the basis of regions and events extracted from news articles on the web. Minutes of city council meetings stored in SOCIA are annotated with discourse salience in order to visualize topic transitions in a meeting transcript.

Sophia is a mining and intelligent pre-processing platform that classifies and clusters news articles and tweets. SOCIA is a data set desined for structuring public debate and regional information. Citispe@k is a web application that supports public debate related to regional issues identified by Sophia using SOCIA.

2 Literature Review

2.1 Linked Open Data for Open Government

LOD plays an important role in fostering open government [3]. There are over 20 international open data platforms in the open government community: Data.gov, Data.gov.uk, Data.gov.au, data.gouv.fr, India.gov.in, etc.[2] Data.gov and India.gov.in are in progress on making their open data platform Data.gov open source [4]. Joinup, [3] a collaborative platform in Europe, proposes Asset Description Metadata Schema (ADMS), which describes semantic assets, that is, a collection of highly reusable metadata and reference data [5]. In Japan, Ministry of Economy, Trade and Industry (METI) operates a web site called "Open Government Laboratory"[4] as an experimental web site towards realizing eParticipation and eGovernment. It also launched the "Apps for Japan" project, which utilizes various type of data for tackling the unprecedented damage wrought by the Great East Japan Earthquake [6].

[1] http://open-opinion.org/ (in Japanese)

[2] http://www.data.gov/, http://data.gov.uk/, http://data.gov.au/, http://www.data.gouv.fr/, http://india.gov.in/, etc.

[3] https://joinup.ec.europa.eu/

[4] http://openlabs.go.jp/ (in Japanese)

There are several competitive challenges designed to promote the use of LOD: Challenge.gov and the New York City Challenge in the U.S., the Open Data Challenge in Europe, and the LOD Challenge in Japan. SOCIA and citispe@k, which we developed, received the ChallengeDay Award at LOD Challenge Japan 2011.[5]

2.2 Supporting Analysis of Public Debate

Providing background information related to public debate is impotant in order to support concern assessment. In view of this, argument visualization is an effective approach for supporting eParticipation [7]. Jeong et al. visualized the difference in cognition for several topics among participants in public debates using the co-occurrence of terms [8]. Visualizing an overview of public debate is also effective for grasping the background. Several argument visualization tools currently exist [9]: Compendium [10], Cohere [11], MIT Deliberatorium [12], Araucaria [13], Discourse Semantic Authoring [14, 15], etc. Typically, these tools produce "box and arrow" diagrams in which premises and conclusions are formulated as statements [16]. We have developed a method for visualizing the transitions of a topic [17] because understanding discussion flow requires overviewing the whole transition of a long meeting rather than local diagramming.

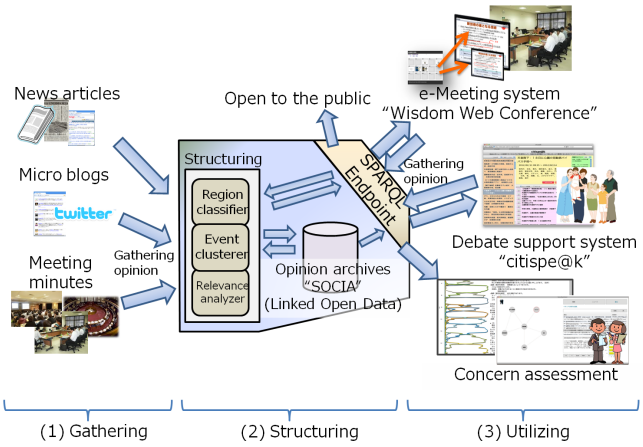

Fig. 1. Outline of O_2, eParticipation Web Platform

[5] http://lod.sfc.keio.ac.jp/challenge2011/result2011.html (in Japanese)

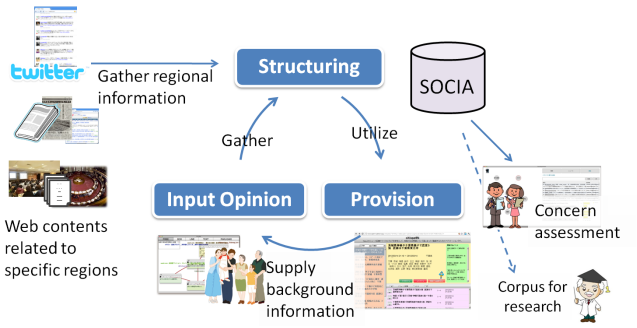

Fig. 2. Cycle of utilizing regional information for eParticipation

3 Designing Platform and Ontology

3.1 O_2: eParticipation Web Platform

O_2, an abbreviation for Open Opinion, is a web platform for citizen participation in debates about regional issues. As shown in Fig. 1, the O_2 platform has three stages. In stage (1), the mining and pre-processing system, Sophia, crawls the web and gathers information from news articles, microblogs, and meeting minutes that can be used for debates. In stage (2), the system geographically classifies the gathered contents and clusters them by event. Relevant information is then structured and stored in the SOCIA data set in accordance with the SOCIA ontology as openly published Linked Open Data. In stage (3), the structured information is used for debate support, e-Meeting, and concern assessment. In this paper, we focus on the eParticipation system for supporting debates using web contents related to specific regions structured in SOCIA.

3.2 SOCIA: Linked Open Data set for eParticipation

The cycle of utilizing regional information in SOCIA for eParticipation is illustrated in Fig. 2. To help citizens understand public concerns and exporess their opinions, background information needs to be provided because most citizens are not experts about diversified public concerns. The opinions expressed can also be utilized as background information after being structured in SOCIA. For web contents (e.g. news articles, blogs, and tweets) to be used as background information, they need to be classified by region and then presented to citizens in an understandable way. Our platform and ontology can be used to structure news and opinions and then link them with regional issues. The data is openly published on the web using the SOCIA ontology,[6] designed using tWeb Ontology Language (OWL) as shown in Fig. 3. Through this process, eParticipative data becomes re-usable and transparent.

[6] http://data.open-opinion.org/socia-ns

Fig. 3. Core classes for structuring regional information in SOCIA ontology

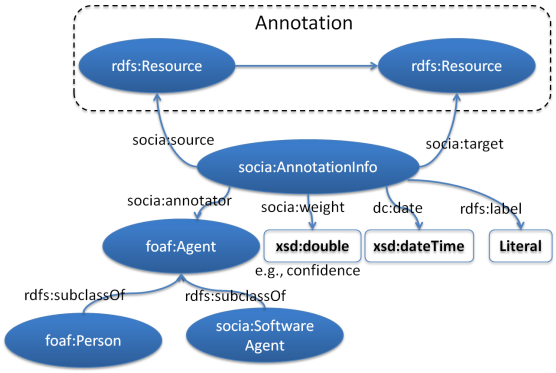

Fig. 4. AnnotationInfo: meta-context information related to property annotation

Text mined from the web is structured in the form of events by region, which are then used as discussion seeds to further build SOCIA. Citizens then create discussion topics out of each seed, e.g., a cluster of news articles related to the same event, and input their opinions by using the system, among other functionalities.

To improve the structuring accuracy, the history of how the LOD properties were annotated (e.g., which algorithm, which parameter, by whom is needed) because the automatic structuring by Sophia has an inherent error of a few percent. To maintain the annotation history, we defined the AnnotationInfo class, as shown in Fig. 4. Such meta-context information is necessary when the data set is used as a corpus for research on natural language processing.

4 Structuring Regional Information on the Web

The system first collects news articles, microblog posts (in this work, tweets), and minutes of city council meeting from the web along with necessary metadata (dates, emission sources, etc). It then classifies this crawled web contents by region and filters out contents unrelated to the interests of regional communities or to current events. Next, the system extracts target events from the news articles and microblogs, and links them using the ontology.

Citizens can then add further links to events, news articles, and microblogs, by creating relevant topics and can debate them by inputting their opinions, polling, or sharing further resources. Those resources and new links are also incorporated in the data set, as are the opinions and the discussion. This creates a virtuous cycle in which the intelligent platform, by creating understandable and relevant discussion seeds, involves citizens in eParticipation. The citizens add further data to the data set, making it grow over time, and this data can be used as input again (e.g. for training better learning models and developing better ontologies).

4.1 Classification by Region

After the mining, the gathered news articles and tweets are classified geographically (by the 47 prefectures of Japan). To this end, we use Transformed Weight-normalized Complementary Naive Bayes (TWCNB) algorithm [18]. In the classification, the feature vectors for each document consist of the TF*IDF value of morpheme bi-grams. To decide whether contents should be filtered out or not, we use a confidence threshold where the confidence value is defined as the difference between log scores of the highest-ranked class and that of second-ranked class.

We conducted a classification experiment through varying threshold of confidence value, using 8,811 news articles related to Japanese prefectures crawled from Yahoo! Japan News[7] during Jun. 13 to Jul. 12, 2011, and 1,133 ones that do not related to any prefectures. The experimental result showed that the precision is 98.2% and the recall is 98.0% for the optimal threshold [19, 20].

4.2 Clustering by Events

SOCIA stored 54,854 news articles, with about 13,000 ones classified as related to a prefictures.[8] The events are extracted as clusters of similar news articles[20]. The similarity between news articles are calculated as a cosine similarity which is weighted by a window function determined by for considering dates/times the news articles were published. As shown in Fig. 5, about 35,000 events were extracted through the clustering of these articles.

[7] http://headlines.yahoo.co.jp/hl?c=loc

[8] The number of news articles stored in SOCIA was counted on Mar. 16, 2012. It has been constantly increasing.

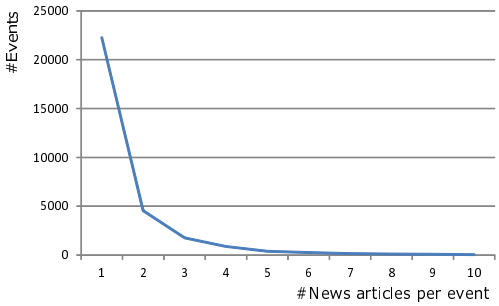

Fig. 5. Distribution of news article counts per event

4.3 Analyzing Topic Transition in Transcript of City Council

For promoting the participation of young citizens in public debates, the use of smartphones is important tool. Browsing the long transcript of a public debate with a smartphone requires a lot of time and effort because the semantics essentially depends on the preceding context. The system automatically analyzes the topic transitions in the minutes of city council meetings stored in SOCIA. The analysis is based on our proposed metric for discourse salience, that is, reference probability [21].

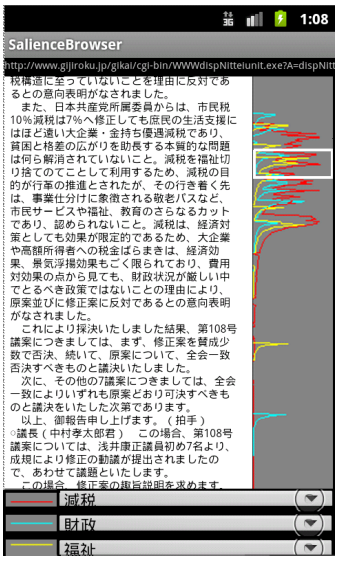

Fig. 6. Visualization of topic transition in a transcript of city council meeting

The use case we consider is browsing past meeting minutes or related documents while attending a public meeting. In public meetings, participants sometimes reference a statement in a past meeting. Trying to find the referenced statement from the long meeting minute in order to confirm it on site would increase the cognitive load due to the need to grasp the topic transitions in a long discourse.

The topic transitions in a transcript of Nagoya city council[9] on a smartphone can be visualized as shown in Fig. 6. The user can grasp the topic transitions from "SalienceGraph" shown on the right of the user interface. The horizontal axis corresponds to discourse salience, that is, the degree of focus on each term or latent topic. The vertical axis represents each sentence in the target discourse.

The GUI for the visualizer was designed in accordance with Shneiderman's *Visual Information-Seeking Mantra*, that is "overview first, zoom and filter, then details-on-demand" [22]. After getting the overview of the topics, users may either browse the salience dynamics of a particular latent topic, or inspect the discourse at a particular point by consulting the record.

5 Citispe@k: Debate Support System Using SOCIA

Citispe@k is a prototype web application that supports public debate by utilizing SOCIA. It provides mobility and reach by supporting web browsers running on smart phones and tablets. The term citispe@k is based on the idea that citizens speak about social issues and current events of the regions in which they live. Users can discuss and sort out regional issues by referencing news articles, tweets, or other relevant resources on the web by using citispe@k. By creating discussion topics or inputting opinions into the system, those topics and opinions are also stored as Linked Open Data in SOCIA.

Fig. 7 shows a screenshot of citispe@k. The screenshot has lists of events or related information. Events recently updated are listed on the left of the screenshot. The system initially shows all events. The user can then limit the list to show only events related to a region. When the user selects an event from the list, information about the event is shown on the right side of the screenshot. Information consists of news articles, tweets, and events related to the event. Those resources can be easily shown and visualized in an iFrame without leaving the system. Users can append comments, e.g. ideas, questions, and answers, by selecting specific content provided by citispe@k. A comment can also be posted to Twitter (via @citispeak for now) to further its reach and be stored in SOCIA.

Users can create discussion topics related to events, news articles and tweets. The "View related topics" button (Fig. 8) is used to see topics related to the event being viewed. Users can create a new discussion topic about the event by clicking the "Make a new topic" button. The cycle of the discussions in citispe@k is that users browse events, get topics related to an event, and add their opinion

[9] http://www.gijiroku.jp/gikai/c_nagoya/index.html

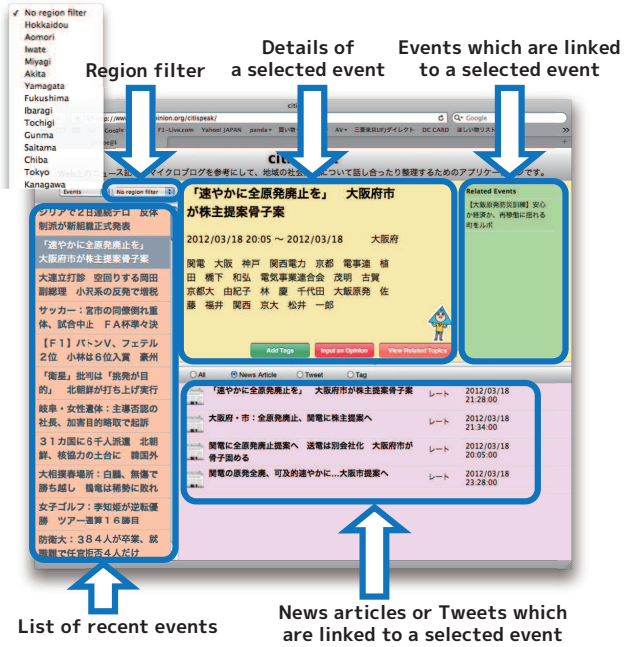

Fig. 7. Screenshot of citispe@k

Fig. 8. Creating a new discussion topic

to a topic of interest. The system supports adding web contents to topics as information sources for discussion as well as adding opinions to topics.

Citispe@k also has a function supporting concern assessment. The system aim to support the analysis of the trends in citizens' awareness, its background information, and the anxiety about social issues. For example, a committee for scientific verification of road construction in Aioiyama-Ryokuchi Park in Nagoya

Fig. 9. Annotating selected event with tags representing criteria

City analyzes road construction.[10] A report on their analysis was made based on several criteria: "economic chance", "life, educational or cultural chance", "safety, security", etc. Thus, classifing opinions on the basis of criteria is effective for concern adjustment. Citispe@k provides tags for such criteria. Users can add tags composed of criteria and polarity, such as "Environment +" or "Environment -". Citispe@k also provides tags that can be used to express the intention of an utterance, like "Question", "Idea", and "Refutation". If events or news articles have many such tags, the tags can be used to support the analysis of concerns. Fig. 9 shows an example of tagging an event. We designed the tags by referencing the QOC model [23] and the Deliberatorium [12] for supporting concern assessment through public debates using citispe@k and the contents in SOCIA.

6 Conclusion

We are developing an eParticipation web platform called O_2 with the aim of increasing transparency and participation in Japanese regional communities. Our Sophia text mining system automatically structures news articles, microblog posts, and transcripts of city council meetings and stores them in SOCIA, an LOD dataset designed for supporting concern assessment. Our Citispe@k web

[10] http://www.city.nagoya.jp/shisei/category/53-3-7-4-0-0-0-0-0-0.html (in Japanese)

application helps citizens debate issues by utilizing regional information structured in SOCIA. It also enables assessment of public concerns through manual annotation of criteria tags. As the next step, personal data will be incorporated into SOCIA to facilitate collaboration among citizens.

Acknowledgments. This work was supported by SCOPE, Ministry of Internal Affairs and Communications, Japan.

References

1. Jeong, H., Hatori, T., Kobayashi, K.: Discourse analysis of public debates: A corpus-based approach. In: Proceedings of 2007 IEEE International Conference on Systems, Man and Cybernetics. (2007) 1782–1793
2. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool (2011)
3. Hochtl, J., Reichstadter, P.: Linked open data - a means for public sector information management. Proceedings of the 2nd international conference on Electronic government and the information systems perspective, Lecture Notes in Computer Science **6866** (2011) 330–343
4. Howard, A.: White house to open source data.gov as open government data platform. http://radar.oreilly.com/2011/12/data-gov-open-source.html (2011)
5. ISA: Improving semantic interoperability in european egovernment systems. http://ec.europa.eu/isa/documents/isa_action1-1.pdf (2011)
6. METI Japan: Open government laboratory to launch the "apps for japan" project. http://www.openlabs.go.jp/apps4japan (2011)
7. Benn, N., Macintosh, A.: Argument visualization for eparticipation: towards a research agenda and prototype tool. In: Proceedings of the Third IFIP WG 8.5 international conference on Electronic participation. 60–73
8. Jeong, H., Shiramatsu, S., Hatori, T., Kobayashi, K.: Discourse analysis of public debates using corpus linguistic methodologies. Journal of Computers **3**(8) (2008) 58–68
9. Kirschner, P., Shum, S., Carr, C.: Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making. Springer (2003)
10. Selvin, A., Shum, S.: Hypermedia as a productivity tool for doctoral research. New Review of Hypermedia and Multimedia, Special Issue on Scholarly Hypermedia **11**(1) 91–101
11. Liddo, A.D., Shum, S.B.: Cohere: A prototype for contested collective intelligence. In: Workshop on Collective Intelligence in Organizations: Toward a Research Agenda, ACM Computer Supported Cooperative Work. (2010)
12. Iandoli, L., Klein, M., Zolla, G.: Enabling online deliberation and collective decision making through large-scale argumentation: A new approach to the design of an internet-based mass collaboration platform. International Journal of Decision Support System Technology **1**(1) (2009) 69–92
13. Reed, C., Rowe, G.: Araucaria: Software for argument analysis, diagramming and representation. International Journal of AI Tools **13**(4) (2004) 961–980
14. Kamimaeda, N., Izumi, N., Hasida, K.: Evaluation of Participants' Contributions in Knowledge Creation Based on Semantic Authoring. The Learning Organization **14**(3) (2007) 263–280

15. Hasida, K.: Semantic Authoring and Semantic Computing. In: New Frontiers in Artificial Intelligence: Joint Proceeding of the 17th and 18th Annual Conferences of the Japanese Society for Artificial Intelligence. Volume 3609 of Lecture Notes in Computer Science., Springer (2007) 137–149
16. van den Braak, S.W., van Oostendorp, H., Prakken, H., Vreeswijk, G.A.W.: A critical review of argument visualization tools: Do users become better reasoners? In: Workshop Notes of the ECAI-2006 Workshop on CMNA. (2006) 67–75
17. Shiramatsu, S., Komatani, K., Ogata, T., Okuno, H.G.: SalienceGraph: Visualizing Salience Dynamics of Written Discourse by Using Reference Probability and PLSA. In: Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence, Springer (2008) 890–902
18. Rennie, J., Shih, L., Teevan, J., Karger, D.: Tackling the poor assumptions of naive bayes text classifiers. In: Proceedings of the 20th International Conference on Machine Learning. (2003) 616–623
19. Swezey, R., Sano, H., Shiramatsu, S., Ozono, T., Shintani, T.: Automatic detection of news articles of interest to regional communities. to appear in: International Journal of Computer Science and Network Security **12**(6) (2012)
20. Swezey, R., Sano, H., Hirata, N., Shiramatsu, S., Ozono, T., Shintani, T.: An e-participation support system for regional communities based on linked open data, classification and clustering. In: to appear in: Proceedings of the 11th IEEE International Conference on Cognitive Informatics & Cognitive Computing. (2012)
21. Shiramatsu, S., Komatani, K., Hasida, K., Ogata, T., Okuno, H.G.: A Game-Theoretic Model of Referential Coherence and Its Empirical Verification Using Large Japanese and English Corpora. ACM Transactions on Speech and Language Processing **5**(3) (2008) Article 6
22. Shneiderman, B.: Designing the User Interface: Strategies for Effective Human-Computer Interaction (3rd Edition). Pearson Addison Wesley (1998)
23. MacLean, A., Young, R., Bellotti, V., Moran, T.: Questions, options, and criteria: elements of design space analysis. Human Compututer Interaction **6**(3) (1991) 201–250