SOCIA: Linked Open Data of Context behind Local Concerns for Supporting Public Participation

Shun Shiramatsu*, Tadachika Ozono* and Toramatsu Shintani*
*Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology,
Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan

***Abstract*—To address public concerns that threat the sustainability of local societies, supporting public participation by sharing the background context behind these concerns is essentially important. We designed a *SOCIA ontology*, which was a linked data model, for sharing context behind local concerns with two approaches: (1) structuring Web news articles and microblogs about local concerns on the basis of geographical regions and events that were referred to by content, and (2) structuring public issues and their solutions as public goals. We moreover built a *SOCIA dataset*, which was a linked open dataset, on the basis of the SOCIA ontology. Web news articles and microblogs related to local concerns were semi-automatically gathered and structured. Public issues and goals were manually extracted from Web content related to revitalization from the Great East Japan Earthquake. Towards more accurate extraction of public concerns, we investigated feature expressions for extracting public concerns from microblogs written in Japanese. To address a technical issue about sample selection bias in our microblog corpus, we formulated a metric in mining feature expressions, i.e., bias-penalized information gain (BPIG). Furthermore, we developed a prototype of a public debate support system that utilized the SOCIA dataset and formulated the similarity between public goals for a goal matching service to facilitate collaboration.**

***Keywords*—Semantic Web; social computing; natural language processing; linked open data; e-Participation**

I. Introduction

Japanese regional societies currently face complicated and ongoing social issues or concerns, e.g., dwindling birth rates, an aging population, public finance problems, disaster risks, dilapidated infrastructures, and radiation pollution that threaten the sustainability of societies. The coverage of government services is expected to decrease along with an escalation in these concerns. Some Japanese researchers regard such troubling situations as “a front-runner of emerging issues”[1]. To address these concerns, supporting public participation by sharing background context behind these concerns is essentially important.

We have aimed to develop a Web platform to support public participation, which provides a function for sharing background context behind local concerns [2], [3], [4]. Since citizens who have beneficial awareness or knowledge are not always experts on relevant social concerns, background context needs to be shared to reduce barriers to public participation. It is difficult to participate in addressing concerns without background context. Linked open data (LOD)[5], which are semantically connected data on the basis of universal resource identifiers (URIs) and the resource description framework (RDF), play an important role in fostering open government [6]. To increase transparency and participation in regional communities, it is important for citizens, government officials, and experts to share public concerns. Background context should be structured and open to facilitate the assessment and sharing of public concerns. The LOD framework is suitable for structuring such background contexts and concerns. The structure of public concerns is an important context when building consensus. We have called the process of structuring public concerns “concern assessment”.

We designed a linked data model and built an LOD dataset, which were called *Social Opinions and Concerns for Ideal Argumentation (SOCIA)*, to share the context behind local concerns. The data model of *SOCIA ontology* was designed with two approaches. The first was attained by structuring Web news articles and microblogs about local concerns on the basis of geographical regions and events that were referred to by the content. The second was attained by structuring public issues and their solutions as public goals. We moreover built a *SOCIA dataset*, which was a linked open dataset (LOD), on the basis of the SOCIA ontology. Japanese local news articles, microblog posts, and minutes of city council meetings are semi-automatically structured on the basis of geographical regions and events. The SOCIA dataset also included public issues and goals that were manually extracted from news articles.

Furthermore, we preliminarily investigated feature expressions to extract public concerns from microblogs written in Japanese. The feature expressions were mined from a corpus consisting of microblogs about public concerns (positive examples) and microblogs about irrelevant to public concerns (negative examples). We addressed a technical issue about the sample selection bias in the positive examples, i.e., there were unsuitable feature expressions that were frequently used by only one specific person.

The rest of the paper is organized as follows. Section II presents conventional works related to e-Participation. The SOCIA ontology is described in Section III. Section IV describes the SOCIA dataset built by semi-automatically structuring Web content related to local concerns and manually structuring public issues and goals extracted from Web content. Section V explains how Japanese feature expressions for extracting public concerns from microblogs were mined with a corpus-based approach. Section VI describes applications of the SOCIA dataset and Section VII concludes the paper.

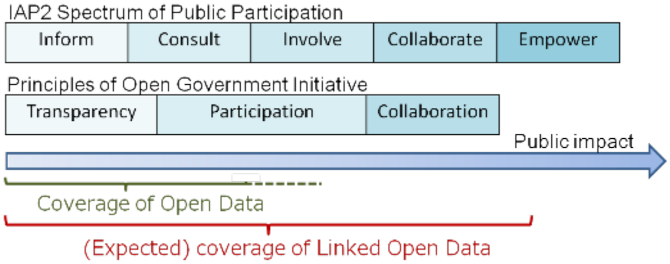

IAP2 Spectrum of Public Participation
Inform
Consult
Involve
Collaborate
Empower
Principles of Open Government Initiative
Transparency
Participation
Collaboration
Public impact
Coverage of Open Data
(Expected) coverage of Linked Open Data

Fig. 1: Expected coverage of Linked Open Data on the spectrum of public participation

II. Related Works

A. Public Participation and Open Data

The International Association for Public Participation (IAP2) and the Obama administration's Open Government Initiative (OGI) have presented similar stages for public participation, i.e., the Spectrum of Public Participation[7] and the Principles of Open Government[8] shown in Figure 1. The gradation in the figure represents the public impact of each stage. The figure also indicates the expected coverage of the use of LOD. Open data generally contributes to transparency, i.e., to the first stage. However, non-linked open data (e.g., CSV table data) generally lack interoperability. LOD is expected to be able to also contribute to the higher/collaborative stages because semantic links compliant with RDF increase the interoperability of data and help us to reuse data for inter-organizational collaboration. Contextual information provided by the semantic links provides the potential for developing social Web services to facilitate public collaboration.

Over 40 countries currently provide open data portals.[1] The number of open data portals has been increasing since 2009. An open data portal by the Japanese government, data.go.jp, was also launched in 2014. One hundred local governments (14 prefectures and 86 municipalities) in Japan also provide their open government data as of Feb. 2015[2].

B. Modeling Public Debate and Participation

Providing background information related to public debate is important in order to support concern assessment. In view of this, argument visualization is an effective approach for supporting eParticipation [9]. Jeong et al. visualized the difference in cognition for several topics among participants in public debates using the co-occurrence of terms [10]. Visualizing an overview of public debate is also effective for grasping the background. Several argument visualization tools currently exist [11]: Compendium [12], Cohere [13], MIT Deliberatorium [14], Araucaria [15], Discourse Semantic Authoring [16], [17], etc. Typically, these tools produce "box and arrow" diagrams in which premises and conclusions are formulated as statements [18].

Within the context of LOD and the semantic Web, the Talk of Europe project proposed a linked data model to structure public debate [19]. Their data model focuses on transcripts of the plenary meetings of the Talk of Europe. Within a broader context, Porwol et al. designed an e-Participation ontology, which was a semantic model of e-Participation [20]. The ontology contained classes of `epart:Project`, `epart:Platform`, and `epart:DemocraticProcess`.

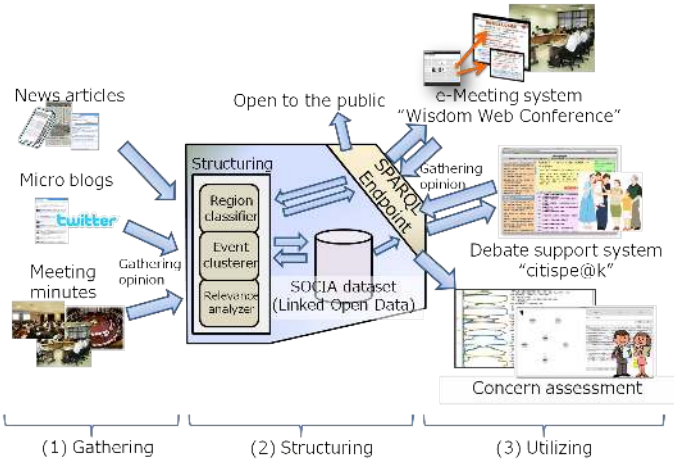


Fig. 2: Outline of O_2, e-Participation Web Platform

III. Designing SOCIA Ontology

This section describes the design of the SOCIA ontology to structure Web news articles and microblogs about local concerns on the basis of geographical regions and events that are referred to by content, and to structure public issues and their solutions as public goals.

A. Structuring Web Content about Local Concerns

To design a data model for sharing background context behind local concerns, we consider applications of the dataset. O_2, an abbreviation for Open Opinion, is our Web platform for citizen participation in debates about regional issues. As shown in Fig. 2, the O_2 platform has three stages. In stage (1), the mining and pre-processing system crawls the Web and gathers information from news articles, microblogs, and meeting minutes that can be used for debates. In stage (2),

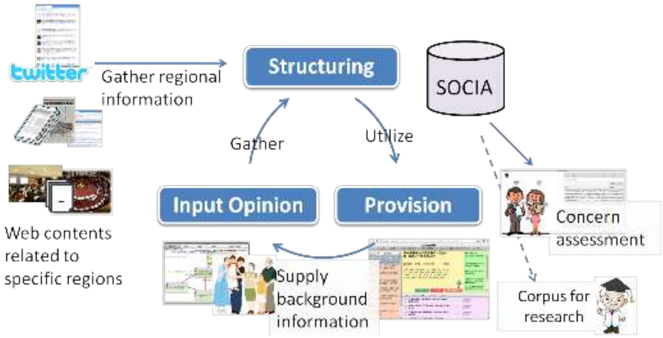

twitter
Gather regional information
Structuring
SOCIA
Gather
Utilize
Input Opinion
Provision
Web contents related to specific regions
Supply background information
Concern assessment
Corpus for research

Fig. 3: Cycle of utilizing regional information for e-Participation

[1] http://www.data.gov/opendatasites

[2] http://fukuno.jig.jp/2013/opendatamap (in Japanese)

the system geographically classifies the gathered contents and clusters them by event. Relevant information is then structured and stored in the SOCIA dataset in accordance with the SOCIA ontology as openly published Linked Open Data. In stage (3), the structured information is used for public participation, i.e., debate support, concern assessment, etc.

The cycle of utilizing regional information in SOCIA for eParticipation is illustrated in Fig. 3. To help citizens understand public concerns and express 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 the SOCIA dataset. 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 the URLs of Web contents and then link them with regional issues.

The SOCIA dataset is openly published on the Web using the SOCIA ontology,[3] designed using Web Ontology Language (OWL) as shown in Fig. 4. Through this process, eParticipative data becomes re-usable and transparent.

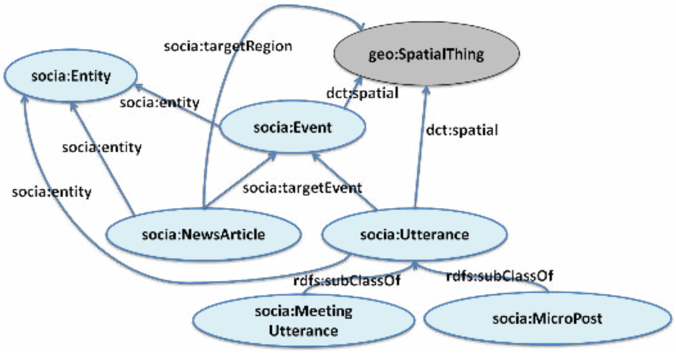


Fig. 4: Core classes for structuring regional information in SOCIA ontology

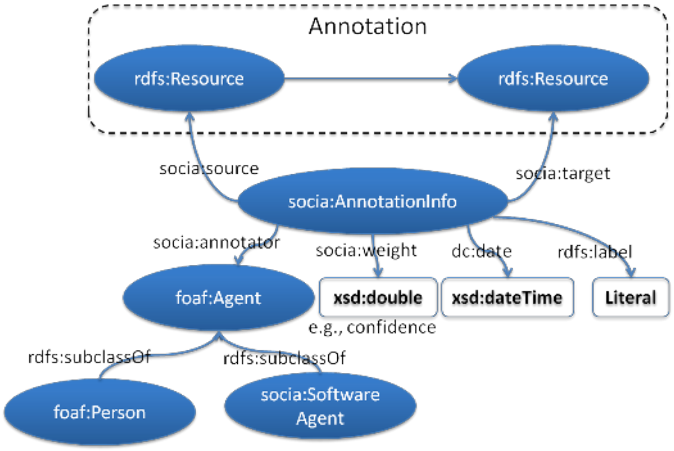


Fig. 5: AnnotationInfo: meta-context information related to property annotation

Text mined from the Web is structured in the form of events by region, which are then used as discussion seeds to further build the SOCIA dataset. 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. 5. Such meta-context information is necessary when the data set is used as a corpus for research on natural language processing.

B. Structuring Public Issues and Goals

Public collaboration and consensus building between stakeholders are essential to enable revitalization from disasters, e.g., the Great East Japan Earthquake. Collaboration between multiple agents generally requires the following conditions:

- Similarity of the agents' goals or objectives
- Complementarity of the agents' skills, abilities, or resources

As the first step, this study focuses on the similarity of the goals. Sharing a data set of public goals can help citizens, who have similar goals, build consensus and collaborate with one another.

We focus on the following three problems related to public collaboration.

1) Citizens cannot easily find somebody whose goals are similar to their ones.
2) Stakeholders who have similar goals occasionally conflict with one another when building consensus because subgoals are sometimes difficult to be agreed on even if the final goal is generally agreed on.
3) A too abstract and general goal is hard to be contributed collaboratively.

We presume that the hierarchies of goals and subgoals play important roles to address these problems. First, the hierarchical structure can make methods of calculating the similarity between public goals more sophisticated. The hierarchy provides rich context to improve retrieval of similar goals. If the data set of public goals had only short textual descriptions without hierarchical structures, calculating the similarity between goals would be difficult and the recall ratio in retrieving similar goals would be lower. Second, visualizing the hierarchies is expected to support people in conflict to attain compromises. Third, dividing goals into fine-grained subgoals reduces barriers to participation and collaboration because small contributions to fine-grained subgoals are more easily provided.

Fig. 6 shows an extention of the SOCIA ontology to represent public issues and goals. The classes `socia:Issue` and `socia:Goal` are connected with the `socia:solution` property. These classes are linked with `foaf:Agent` corresponding to participants or stakeholders and with

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

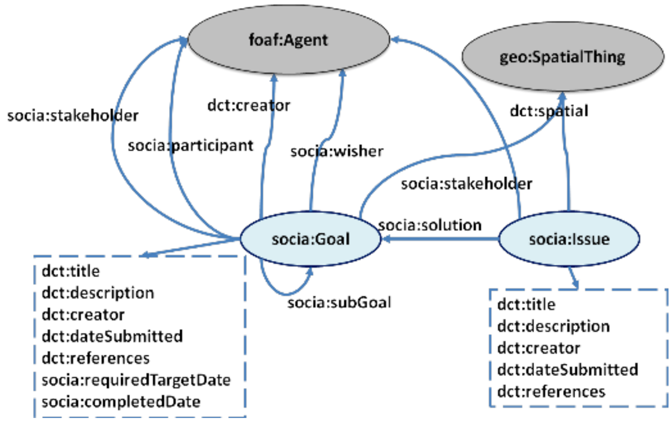


Fig. 6: Core classes for structuring public issues and goals in SOCIA ontology

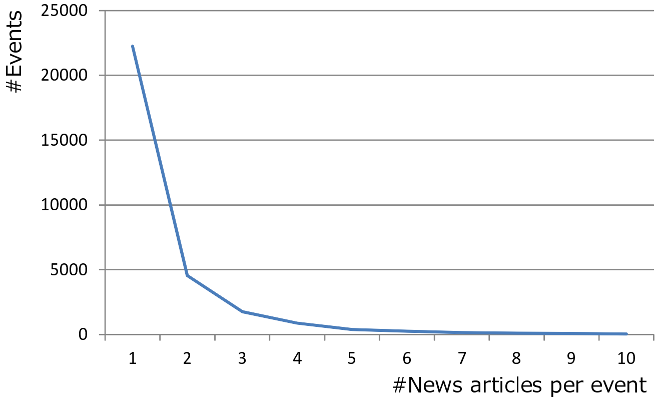


Fig. 7: Distribution of news article counts per event

`geo:SpatialThing` corresponding to geographical regions.

IV. Building SOCIA Dataset

This section describes semi-automatic structuring of Web content on local concerns and manual structuring of public issues and goals.

A. Gathering Web Content about Local Concerns

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).

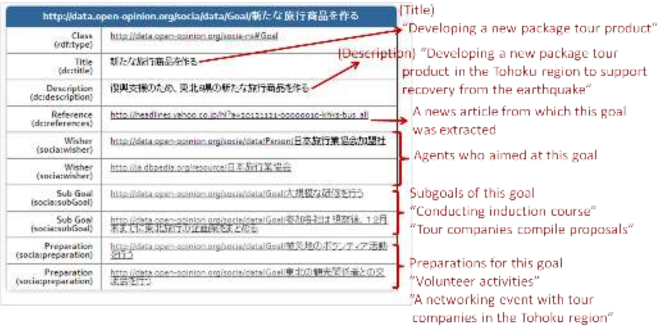
1) Classification by Geographic Region: After the mining, the gathered news articles and tweets are classified geographically (by the 47 prefectures of Japan). To this end, we use Transformed Weight-normalized Complementary Naive Bayes (TWCNB) algorithm [21]. In the classification, the feature vectors 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[4] 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 [22], [23].

2) Clustering by Events: The SOCIA dataset stored 54,854 news articles, with about 13,000 ones classified as related to a prefictures.[5] The events are extracted as clusters of similar news articles [23]. 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. 7, about 35,000 events were extracted through the clustering of these articles.

B. Manual Extraction of Public Goals from Web News Articles

We built an LOD set[6] by manually extracting public goals from news articles and related documents. The 657 public



http://data.open-opinion.org/socia/data/Goal/新たな旅行商品を作る	
Class (rdf:type)	http://data.open-opinion.org/socia-ns#Goal
Title (dc:title)	新たな旅行商品を作る
Description (dc:description)	復興支援のため、東北6県の新たな旅行商品を作る
Reference (dc:references)	http://headlines.yahoo.co.jp/hl?a=20121121-00000010-khks-bus_all
Wisher (socia:wisher)	http://data.open-opinion.org/socia/data/Person/日本旅行業協会加盟社
Wisher (socia:wisher)	http://ja.dbpedia.org/resource/日本旅行業協会
Sub Goal (socia:subGoal)	http://data.open-opinion.org/socia/data/Goal/大規模な研修を行う
Sub Goal (socia:subGoal)	http://data.open-opinion.org/socia/data/Goal/参加各社は研修後、12月末までに東北旅行の企画案をまとめる
Preparation (socia:preparation)	http://data.open-opinion.org/socia/data/Goal/被災地のボランティア活動を行う
Preparation (socia:preparation)	http://data.open-opinion.org/socia/data/Goal/東北の観光関係者との交流会を行う

(Title) "Developing a new package tour product"

(Description) "Developing a new package tour product in the Tohoku region to support recovery from the earthquake"

A news article from which this goal was extracted

Agents who aimed at this goal

Subgoals of this goal
"Conducting induction course"
"Tour companies compile proposals"

Preparations for this goal
"Volunteer activities"
"A networking event with tour companies in the Tohoku region"

Fig. 8: Instance of public goal: "Developing new package tour product"

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

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

[6] http://data.open-opinion.org/socia/data/Goal?rdf:type=socia:Goal&limit=700 (in Japanese)

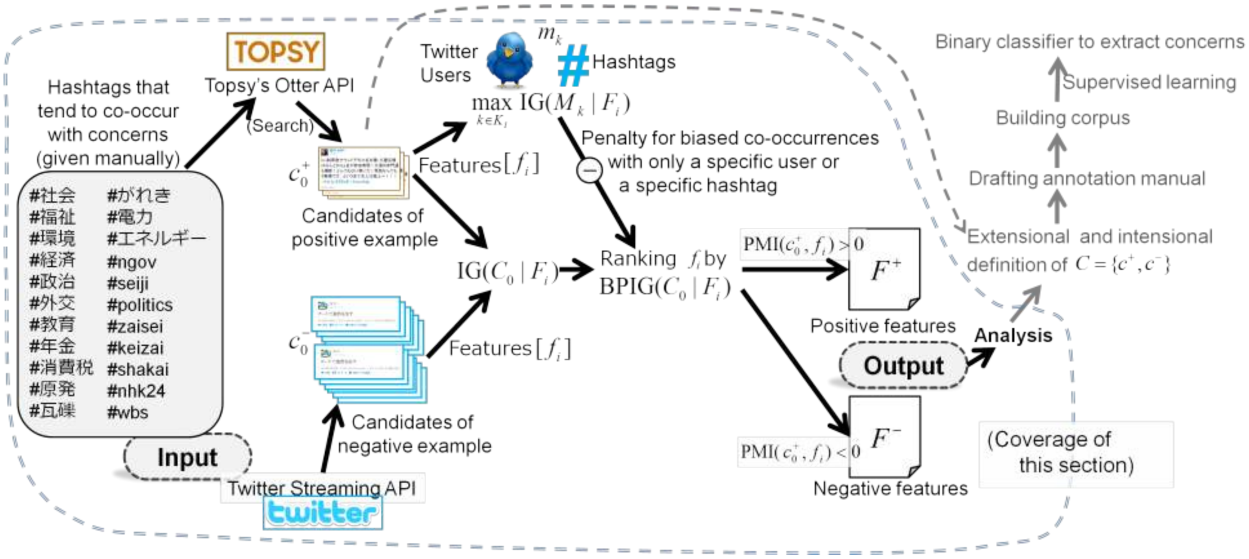

Hashtags that tend to co-occur with concerns (given manually)
TOPSY
Topsy's Otter API
(Search)
#社会
#福祉
#環境
#経済
#政治
#外交
#教育
#年金
#消費税
#原発
#瓦礫
#がれき
#電力
#エネルギー
#ngov
#seiji
#politics
#zaisei
#keizai
#shakai
#nhk24
#wbs
Input
Twitter Streaming API
twitter
c_0^+
Candidates of positive example
c_0^-
Candidates of negative example
Features$[f_i]$
Features$[f_i]$
Twitter Users
m_k
Hashtags
$\max_{k \in K_i} \mathrm{IG}(M_k \mid F_i)$
Penalty for biased co-occurrences with only a specific user or a specific hashtag
$\mathrm{IG}(C_0 \mid F_i)$
Ranking f_i by $\mathrm{BPIG}(C_0 \mid F_i)$
$\mathrm{PMI}(c_0^+, f_i) > 0$
F^+
Positive features
$\mathrm{PMI}(c_0^+, f_i) < 0$
F^-
Negative features
Output
Analysis
Binary classifier to extract concerns
Supervised learning
Building corpus
Drafting annotation manual
Extensional and intensional definition of $C = \{c^+, c^-\}$
(Coverage of this section)

Fig. 9: Processing flow for mining features to extract public concerns

goals and 4349 RDF triples were manually extracted from 96 news articles and two related documents by one human annotator. The most abstract goal that is the root node of the goal-subgoal hierarchy is "revitalization from the earthquake".[7] The subgoals are linked from this goal with the `socia:subgoal` property.

The manually built LOD set can be used for developing a method of calculating the similarities between public goals. It can also be used as example seed data when citizen users input their own goals for revitalization. Fig. 8 shows an instance of a public goal to revitalize the Tohoku region from the Great East Japan Earthquake. This goal of "developing a new package tour product", has a title in Japanese, a description in Japanese, and two subgoal data resources.

This dataset about public goals for revitalization won the 2nd Prize of Dataset Track of the Linked Open Data Challenge Japan 2013[8].

V. Mining Feature Expression to Extract Concerns

Automatic structuring needs to become more accurate with a filter for noisy text to support concern assessment because consumer-generated Web content (e.g., microblogs) frequently contains noise information on the target regions. We aimed to construct a binary classifier between tweets including public concerns and others. To define the boundary between the positive class c^+ (corresponding to public concerns) and the negative class c^- (corresponding to tweets other than public concerns), we investigate approximative examples collected through hashtag search. Figure 9 represents the processing flow for investigating the approximate examples. Firstly, we manually prepare the list of hashtags that may frequently co-occur with public concerns in Japanese tweets: #政治 (politics), #社会 (society), #環境 (environment), and so on. The tweets collected through searching by these hashtags from Topsy's Otter API[9] are regarded as candidates of positive examples. These examples are labeled as class c_0^+, an approximative positive class. However, note that the c_0^+ examples also include noise tweets that are not suitable for concern assessment. Secondly, we gather general tweets from Twitter Streaming API[10]. The ratio of public concern in this set is much less than that in the c_0^+ set. Therefore, these general tweets are regarded as candidates of negative examples and labeled as class c_0^-, an approximative negative class. In this section, we empirically analyze features for classifying tweets into $C_0 = \{c_0^+, c_0^-\}$ towards building a corpus annotated with $C = \{c^+, c^-\}$, a more sophisticated concern definition.

Here, we denote a feature vector of a tweet by $[f_i]_i$. Let $F_i = \{f_i^+, f_i^-\}$ where f_i^+ denotes a label representing that the feature f_i appears in a tweet, and f_i^- denotes a label representing that f_i does not. A feature f_i's significance for extracting c_0^+ tweets can be estimated by the information gain:

$$\mathrm{IG}(C_0|F_i) = \mathrm{H}(C_0) - \mathrm{H}(C_0|F_i), \tag{1}$$

with

$$\mathrm{H}(C_0) = -p(c_0^+)\log p(c_0^+) - p(c_0^-)\log p(c_0^-), \tag{2}$$

$$\begin{aligned}&\mathrm{H}(C_0|F_i) = \\ &-p(c_0^+|f_i^+)\log p(c_0^+|f_i^+) - p(c_0^-|f_i^+)\log p(c_0^-|f_i^+) \\ &-p(c_0^+|f_i^-)\log p(c_0^+|f_i^-) - p(c_0^-|f_i^-)\log p(c_0^-|f_i^-).\end{aligned} \tag{3}$$

The features f_i extracted from c_0^+ tweets with the information gain, however, are biased due to sample selection

[7] `http://data.open-opinion.org/socia/data/Goal/%E9%9C%87%E7%81%BD%E5%BE%A9%E8%88%88 (in Japanese)`

[8] http://lod.sfc.keio.ac.jp/blog/?p=2074 (in Japanese)

[9] http://otter.topsy.com/

[10] https://dev.twitter.com/docs/streaming-api

bias dependent on the input hashtags. To address the sample selection bias, we formulate bias-penalized information gain (BPIG) with considering a penalty for biased occurrence of feature f_i as follows:

$$\text{BPIG}(C_0|F_i) = \text{IG}(C_0|F_i) - \alpha \max_{k \in K_i} \text{IG}(M_k|F_i) \tag{4}$$

with

$$K_i = \{k \mid \text{PMI}(m_k, f_i|c_0^+) > 0\} \tag{5}$$

$$\text{PMI}(m_k, f_i|c_0^+) = \log \frac{p(m_k, f_i|c_0^+)}{p(m_k|c_0^+)p(f_i|c_0^+)} \tag{6}$$

$$M_k = \{m_k^+, m_k^-\}, \tag{7}$$

where let m_k^+ be a label representing that m_k, a hashtag or a user, appears in a tweet or is the author of the tweet, m_k^- be a label representing that m_k does not, and $\alpha \in [0, 1]$ be a weight of the penalty term. Here, $\max_{k \in K_i} \text{IG}(M_k|F_i)$ can be regarded as a penalty for f_i that co-occurs only with a particular hashtag or user m_k.

Table I shows the hashtags for gathering c_0^+ tweets from Topsy's otter API. We specified Japanese as the language of gathered tweets in query URLs for the API. Temporal distribution of the 32,844 tweets collected as c_0^+ is shown in Figure 10. The c_0^+ tweets consist mostly of the tweets in the latest months due to the characteristics of time window of the Topsy search.

TABLE I: Hashtags for gathering c_0^+ tweets

Hashtags	#Tweets
#社会 (society)	1,981
#福祉 (welfare)	1,629
#環境 (environment)	1,380
#経済 (economy)	1,985
#政治 (politics)	3,131
#外交 (diplomacy)	986
#教育 (education)	1,865
#年金 (pension)	940
#消費税 (consumption tax)	1,592
#原発 (nuclear plant)	3,129
#瓦礫 (rubble)	2,367
#がれき (rubble)	1,949

Hashtags	#Tweets
#電力 (electricity)	1,020
#エネルギー (energy)	797
#ngov	1,040
#seiji (politics)	4,796
#politics	1,775
#zaisei (finance)	1,014
#keizai (economy)	2,406
#shakai (society)	1,018
#nhk24	1,844
#wbs	289
Total	38,933
Total without duplication	32,844

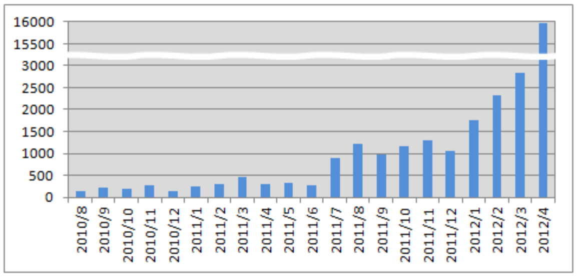


Fig. 10: Temporal distribution of c_0^+ tweets gathered from Topsy Otter's API

TABLE II: Temporal distribution of c_0^- tweets gathered from Twitter streaming API

Duration (JST)	#Tweets
2011-10-16 21:44:25〜23:55:31	49,998
2012-02-20 11:19:25〜15:25:04	49,994
2012-04-14 00:59:15〜07:57:55	49,992
Total	149,984

The c_0^- tweets are gathered from Twitter Streaming API. The ratio of public concerns in c_0^- is predicted to be much less than that in c_0^+. Temporal distribution of the 149,984 tweets collected as c_0^- is shown in Table II. Since we presume that the ratio of c^- is greater than that of c^+, the ratio of c_0^- is also set as greater than that of c_0^+. We conducted an experiment for feature extraction using these 182,828 tweets consisting of c_0^+ and c_0^-. Features representing c_0^+ and c_0^- are extracted with the following procedure:

1) Rank features f_i by $\text{IG}(C_0|F_i)$ and $\text{BPIG}(C_0|F_i)$, respectively.
2) As features for c_0^+, extract high-ranked features f_i, such that $\text{PMI}(c_0^+, f_i) = \log \frac{p(c_0^+, f_i)}{p(c_0^+)p(f_i)} > 0$.
3) As features for c_0^-, extract high-ranked features f_i, such that $\text{PMI}(c_0^+, f_i) < 0$.

In this experiment, we regard morpheme N-grams as features of each tweet. Table III and IV represent the results of feature extraction where let $N = 3$, i.e., in case of morpheme tri-grams. There are some pre-processings before extracting morpheme N-grams; URL strings and user names (starting with @) in tweets are replaced by "[URL]" and "[USER]". Hashtags in tweets are omitted. "[B]" and "[E]" are inserted into beginning and end of a tweet, respectively.

The features for the positive example, c_0^+, are shown in Table III. The features extracted by information gain, which are ranked in the left side of the table, are greatly biased due to the input hashtags. For example, both of "NEWS WEB 24" (a name of TV news program) and "番組で紹介" (introducing it in our program) are dependent on the hashtag #nhk24. In contrast, the features extracted by BPIG in the right sides of the tables are not specific to a particular hashtag of a user. These N-gram features are commonly used for describing public concerns, e.g., expressions for stating fact or question. Table V represents features f_i which have higher penalties for bias, that is, higher $\max_{k \in K_i} \text{IG}(M_k|F_i)$. The result shows that BPIG can appropriately filter out features that co-occurs only with a particular hashtag or user.

The features for the negative example, c_0^-, are shown in Table IV. Both the c_0^+'s features and the c_0^-'s features are needed for classifying the positive examples and the negative ones. The c_0^-'s features can be used for filtering the negative examples as noise tweets. Although in both cases of information gain and BPIG, expressions for greeting or communication are higher ranked, features with higher $p(c_0^+|f_i)$, such as " !! [E] " and " !!! ", are lower-ranked in BPIG than in information gain.

Morpheme N-grams ($N = 2, 3, 4, 5$) extracted as features for c_0^+ can be classified by modality types as shown in Table

TABLE III: Morpheme tri-grams extracted as features representing c_0^+

Ranking by IG			Ranking by BPIG		
Tri-gram f_i	$\mathrm{IG}(C_0\|F_i)$	$p(c_0^+\|f_i)$	Tri-gram f_i	$\mathrm{BPIG}(C_0\|F_i)$	$p(c_0^+\|f_i)$
）[URL][E]	8.38×10^{-3}	0.804	」[URL][E]	4.14×10^{-3}	0.758
』[URL][E]	8.25×10^{-3}	0.843	ている。	1.21×10^{-3}	0.602
:[URL][E]	6.79×10^{-3}	0.845	ているの	9.23×10^{-4}	0.560
...[URL][E]	5.96×10^{-3}	0.751	ています	9.14×10^{-4}	0.539
」[URL][E]	5.51×10^{-3}	0.758	している	9.13×10^{-4}	0.602
。[URL][E]	4.13×10^{-3}	0.465	。RT [USER]	7.66×10^{-4}	0.563
NEWS WEB 24	3.70×10^{-3}	1.00	された	6.21×10^{-4}	0.492
。』[URL]	3.68×10^{-3}	0.954	れている	6.17×10^{-4}	0.589
している	3.64×10^{-3}	0.602	してい	5.83×10^{-4}	0.499
のベストセラー→	3.36×10^{-3}	0.984	ではない	3.39×10^{-4}	0.480
番組で紹介	3.22×10^{-3}	0.997	０万円	2.77×10^{-4}	0.83
WEB 24 です	3.21×10^{-3}	1.00	のエネルギー政策	2.26×10^{-4}	0.97
24 です。	3.21×10^{-3}	1.00	、２０	2.25×10^{-4}	0.90
ツイートには	3.18×10^{-3}	1.00	yes or no	2.16×10^{-4}	1.0
で紹介し	3.15×10^{-3}	0.986	・‘) yes or	1.97×10^{-4}	1.0
してよい	3.14×10^{-3}	0.997	ω ・‘) yes	1.97×10^{-4}	1.0
よいツイートに	3.11×10^{-3}	1.00	? [URL] 拡散	1.97×10^{-4}	1.0
てよいツイート	3.11×10^{-3}	1.00	or no?	1.97×10^{-4}	1.0
編集部)	3.10×10^{-3}	1.00	no? [URL]	1.97×10^{-4}	1.0
SankeiBiz 編集部	3.10×10^{-3}	1.00	、日本の	1.93×10^{-4}	0.82

TABLE IV: Morpheme tri-grams extracted as features representing c_0^-

Ranking by IG			Ranking by BPIG		
Tri-gram f_i	$\mathrm{IG}(C_0\|F_i)$	$p(c_0^+\|f_i)$	Tri-gram f_i	$\mathrm{BPIG}(C_0\|F_i)$	$p(c_0^+\|f_i)$
(笑)	1.84×10^{-3}	0.009	[B][USER] お	1.83×10^{-3}	0.001
[B][USER] お	1.83×10^{-3}	0.001	(笑)	1.60×10^{-3}	0.009
笑)[E]	1.43×10^{-3}	0.005	＼(ˆo	1.11×10^{-3}	0.006
!! [E]	1.39×10^{-3}	0.044	笑)[E]	1.09×10^{-3}	0.005
＼(ˆo	1.25×10^{-3}	0.006	[B][USER] おはよう	9.66×10^{-4}	0.002
[B][USER] おはよう	9.66×10^{-4}	0.002	[B][USER] おやすみ	8.98×10^{-4}	0.000
・ω・	9.08×10^{-4}	0.028	[B][USER] そう	8.67×10^{-4}	0.000
[B][USER] おやすみ	8.98×10^{-4}	0.000	´▽‘	6.02×10^{-4}	0.002
[B][USER] そう	8.67×10^{-4}	0.000	[USER] おは	5.98×10^{-4}	0.002
•• [E]	8.24×10^{-4}	0.030	▽‘)	5.89×10^{-4}	0.002
!!!	7.00×10^{-4}	0.061	[B][USER] え	5.76×10^{-4}	0.000

TABLE V: N-grams that frequently co-occur only with a specific hashtag or user in c_0^+ (excerpted)

N-gram f_i	Hashtag or user $\arg\max_{m_{k \in K_i}} \mathrm{IG}(M_k\|F_i)$	Penalty for f_i $\max_{k \in K_i} \mathrm{IG}(M_k\|F_i)$
』[URL][E]	#介護	9.05×10^{-2}
のベストセラー→	#本	4.19×10^{-2}
受験のベストセラー→	#学参	3.99×10^{-2}
SankeiBiz 編集部	#news	3.84×10^{-2}
...[URL][E]	#newsJP	3.60×10^{-2}
NEWS WEB 24	#nhk24	3.48×10^{-2}
番組で紹介し	#nhk24	3.10×10^{-2}
:[URL][E]	@snn007	9.29×10^{-2}
産経新聞）[URL][E]	@selection_news	4.76×10^{-2}
）[URL][E]	@selection_news	3.75×10^{-2}
ヨミドクター）[URL][E]	@yomidr	3.24×10^{-2}

VI. Suggestions, questions, and fact statements with some references (quotation) can be extracted as public concerns from Japanese tweets, according to this analysis result. We suppose that these analyses can be used to define the boundary between positive example c^+ and negative example c^- towards drafting annotation manual and building a concern corpus.

VI. Application

A. Public Debate Using SOCIA Dataset

Citispe@k (pronounced "citi-speak") is a prototype Web application that supports public debate by utilizing the SOCIA dataset. 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 the SOCIA dataset. Figure 7 shows a screenshot of citispe@k. The screenshot has lists of event 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

TABLE VI: Modality types of morpheme N-grams extracted as features representing c_0^+ (excerpted)

Modality	N-gram f_i	BPIG$(C_0\|F_i)$	$p(c_0^+\|f_i)$
Quotation Retweet	」[URL][E]	4.14×10^{-3}	0.758
	ニュース [URL]	1.11×10^{-3}	0.825
	RT [USER]:	7.66×10^{-4}	0.563
	：日本経済新聞	1.46×10^{-4}	0.82
	読売新聞）[URL][E]	1.00×10^{-4}	0.78
	-MSN 産経ニュース	9.19×10^{-5}	0.884
Suggestion Assertion	べき。	2.20×10^{-4}	0.76
	すべき	1.66×10^{-4}	0.725
	べきだ	1.48×10^{-4}	0.61
	するべき	6.03×10^{-5}	0.60
	したらどうだろう	3.92×10^{-5}	1.0
Averment Fact	ている。	1.21×10^{-3}	0.602
	ています	9.14×10^{-4}	0.539
	している	9.14×10^{-4}	0.602
	されている	2.26×10^{-4}	0.629
Question Doubt	・‘) yes or no?	1.97×10^{-4}	1.0
	では？	1.18×10^{-4}	0.63
	ているのか	7.49×10^{-5}	0.55
	のでしょうか	4.19×10^{-5}	0.53
Content	日本の	5.32×10^{-3}	0.824
	億円	1.14×10^{-3}	0.900
	委員会	1.04×10^{-3}	0.884
	兆円	8.79×10^{-4}	0.940
	政策を	3.60×10^{-4}	0.97
	研究機関	2.62×10^{-4}	0.98

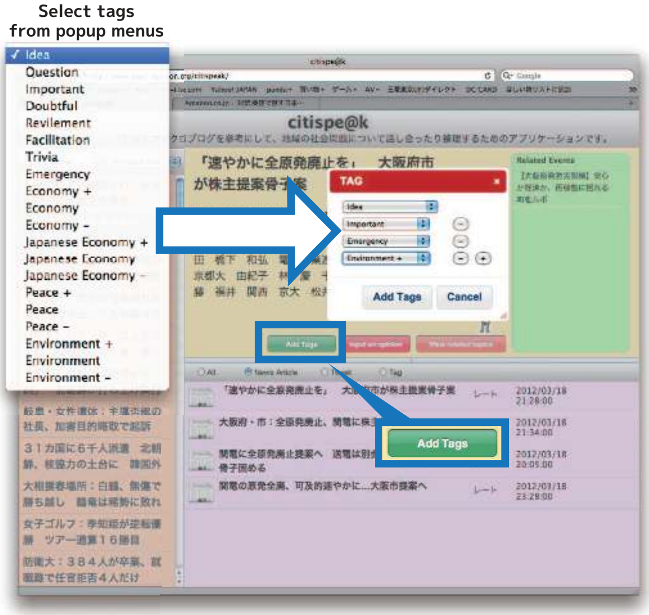


Fig. 12: Annotating selected event with tags representing criteria

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 account) 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 (Figure 11) 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

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 City analyzes road construction.[11] 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. 12 shows an example of tagging an event. We designed the tags by referencing the QOC model [24] and the Deliberatorium [14] for supporting concern assessment through public debates using citispe@k and the contents in SOCIA.

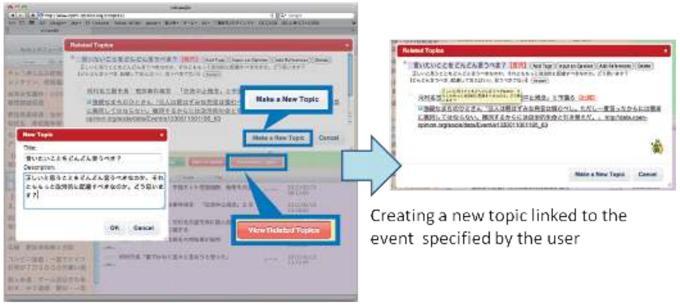


Fig. 11: Creating a new discussion topic

B. Goal Matching Service Using SOCIA Dataset

We are planning to develop a Web service to match citizens and agents who are aiming at similar goals to facilitate collaboration. Toward this end, we expanded the SOCIA ontology to describe the public goals in Fig. 6. The property `socia:subgoal` enables us to describe the hierarchical structure of goals and subgoals. The public goal matching service that we aim to develop requires high-recall retrieval of similar goals to facilitate inter-domain, inter-area, and inter-organizational collaboration.

Pairs of similar goals are connected by the `schema:isSimilarTo` property[12]. The similarity between public goals can be calculated on the basis of a recursive definition of a bag-of-features vector as:

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

[12] http://schema.org/isSimilarTo

$$\mathrm{sim}(g_i, g_j) = \frac{\mathrm{bof}(g_i) \cdot \mathrm{bof}(g_j)}{\|\mathrm{bof}(g_i)\| \|\mathrm{bof}(g_j)\|} \tag{8}$$

$$\begin{aligned}\mathrm{bof}(g) = {} & \frac{\alpha}{\|\mathrm{tfidf}(g)\|}\mathrm{tfidf}(g) + \frac{\beta}{\|\mathrm{lda}(g)\|}\mathrm{lda}(g) \\ & + \frac{\gamma}{|\mathrm{sub}(g)|} \sum_{sg \in \mathrm{sub}(g)} \frac{\mathrm{bof}(sg)}{\|\mathrm{bof}(sg)\|}\end{aligned} \tag{9}$$

$$\mathrm{tfidf}(g) = \begin{pmatrix} \mathrm{tfidf}(w_1, g) \\ \vdots \\ \mathrm{tfidf}(w_{|W|}, g) \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{|W|+|Z|}, \tag{10}$$

$$\mathrm{lda}(g) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \mathrm{p}(z_1|g) \\ \vdots \\ \mathrm{p}(z_{|Z|}|g) \end{pmatrix} \in \mathbb{R}^{|W|+|Z|}, \tag{11}$$

where g denotes a public goal, $\mathrm{bof}(g)$ denotes a bag-of-features vector of g, and $\mathrm{sub}(g)$ denotes a set of subgoals of g. Here, $w \in W$ denotes a term, $z \in Z$ denotes a latent topic derived by a latent topic model [25], and $\mathrm{tfidf}(w, g)$ denotes the TF-IDF, i.e., the product of term frequency and inverse document frequency, of w in a title and a description of g. The $\mathrm{p}(z|g)$ denotes the probability of z given g, $0 \leq \alpha, \beta, \gamma \leq 1$, and $\alpha + \beta + \gamma = 1$. The reason this definition incorporates a latent topic model is to enable short descriptions of goals to be dealt with because TF-IDF is insufficient for calculating similarities in short texts. The parameters α, β, and γ are empirically determined on the basis of actual data.

This prototyped method of calculating similarities should be tested, verified, and refined though experiments in future work using the LOD set of public goals that we present.

VII. Conclusion

We designed the SOCIA ontology, which is a linked data model to share context behind local concerns with two approaches: (1) structuring Web news articles and microblogs about local concerns on the basis of geographical regions and events that were referred to by content, and (2) structuring public issues and their solutions as public goals. We moreover built the SOCIA dataset, which was a linked open dataset, on the basis of the SOCIA ontology. Web news articles and microblogs related to local concerns were semi-automatically gathered and structured. 54,854 news articles were stored to the SOCIA dataset and they were automatically linked with prefectures and events. Moreover, 657 public goals were manually extracted from Web content related to revitalization from the Great East Japan Earthquake.

We investigated feature expressions to extract public concerns from microblogs written in Japanese towards more accurate extraction of public concerns. To address a technical issue about sample selection bias in our microblog corpus, we formulated a metric for mining feature expressions, i.e., bias-penalized information gain (BPIG). We conducted an experiment for extracting features representing positive examples and negative examples. The experimental results showed that BPIG is more suitable for dealing with training data with hashtag-dependent sample selection bias than the conventional information gain.

Furthermore, we presented applications of the SOCIA dataset, i.e., a public debate support system and a goal matching service. These applications utilize the SOCIA dataset to share context behind local concerns. We are planning to sophisticate the SOCIA ontology and dataset towards facilitating public collaboration in the real world.

Acknowledgment

This work was supported by a Grant-in-Aid for Young Scientists (B) (No. 25870321) from the Japan Society for the Promotion of Science (JSPS) and SCOPE from the Ministry of Internal Affairs and Communications, Japan.

References

[1] H. Komiyama, "Vision 2050 and the role of Japan toward the sustainable society," in *Proceedings of the 4th International Symposium on Environmentally Conscious Design and Inverse Manufacturing*, 2005, pp. 2–4.

[2] S. Shiramatsu, R. Swezey, H. Sano, N. Hirata, T. Ozono, and T. Shintani, "Structuring Japanese Regional Information Gathered from the Web as Linked Open Data for Use in Concern Assessment," in *Electronic Participation - Proceedings of the 4th IFIP WG 8.5 International Conference, ePart 2012*, ser. Lecture Notes in Computer Science, vol. 7444. Springer, 2012, pp. 73–84.

[3] S. Shiramatsu, N. Hirata, R. Swezey, H. Sano, , T. Ozono, and T. Shintani, "Gathering Public Concerns from Web towards Building Corpus of Japanese Regional Concerns," in *Proceedings of the 2012 IIAI International Conference on Advanced Applied Informatics*, 2012, pp. 248–253.

[4] S. Shiramatsu, T. Ozono, and T. Shintani, "Approaches to Assessing Public Concerns: Building Linked Data for Public Goals and Criteria Extracted from Textual Content," in *Electronic Participation - Proceedings of the 5th IFIP WG 8.5 International Conference, ePart 2013*, ser. Lecture Notes in Computer Science, vol. 8075. Springer, 2013, pp. 109–121.

[5] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 2011.

[6] J. Hochtl and P. Reichstadter, "Linked open data - a means for public sector information management," in *Proceedings of the 2nd International Conference on Electronic Government and the Information Systems Perspective*, ser. Lecture Notes in Computer Science, vol. 6866. Springer, 2011, pp. 330–343.

[7] International Association for Public Participation, "IAP2 Spectrum of Public Participation," http://www.iap2.org/associations/4748/files/IAP2%20Spectrum_vertical.pdf, 2007.

[8] White House, "Open government initiative," http://www.whitehouse.gov/open, 2009.

[9] N. Benn and A. Macintosh, "Argument visualization for eparticipation: towards a research agenda and prototype tool," in *Electronic Participation - Proceedings of the 3rd IFIP WG 8.5 international conference, ePart 2011*, ser. Lecture Notes in Computer Science, vol. 6847. Springer, 2011, pp. 60–73.

[10] H. Jeong, S. Shiramatsu, T. Hatori, and K. Kobayashi, "Discourse analysis of public debates using corpus linguistic methodologies," *Journal of Computers*, vol. 3, no. 8, pp. 58–68, 2008.

[11] P. Kirschner, S. Shum, and C. Carr, *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*. Springer, 2003.

[12] A. Selvin and S. Shum, "Hypermedia as a productivity tool for doctoral research," *New Review of Hypermedia and Multimedia, Special Issue on Scholarly Hypermedia*, vol. 11, no. 1, pp. 91–101.

[13] A. D. Liddo and S. B. Shum, "Cohere: A prototype for contested collective intelligence," in *Workshop on Collective Intelligence in Organizations: Toward a Research Agenda, ACM Computer Supported Cooperative Work*, 2010.

[14] L. Iandoli, M. Klein, and G. Zolla, "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*, vol. 1, no. 1, pp. 69–92, 2009.

[15] C. Reed and G. Rowe, "Araucaria: Software for argument analysis, diagramming and representation," *International Journal of AI Tools*, vol. 13, no. 4, pp. 961–980, 2004.

[16] N. Kamimaeda, N. Izumi, and K. Hasida, "Evaluation of Participants' Contributions in Knowledge Creation Based on Semantic Authoring," *The Learning Organization*, vol. 14, no. 3, pp. 263–280, 2007.

[17] K. Hasida, "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*, ser. Lecture Notes in Computer Science, vol. 3609. Springer, 2007, pp. 137–149.

[18] S. W. van den Braak, H. van Oostendorp, H. Prakken, and G. A. W. Vreeswijk, "A critical review of argument visualization tools: Do users become better reasoners?" in *Workshop Notes of the ECAI-2006 Workshop on CMNA*, 2006, pp. 67–75.

[19] A. van Aggelen, "Modelling the european debates," http://www.talkofeurope.eu/2014/05/modelling-the-european-debates/, 2014.

[20] L. Porwol, A. Ojo, and J. Breslin, "A semantic model for e-participation: detailed conceptualization and ontology," in *Proceedings of the 15th Annual International Conference on Digital Government Research*. ACM, 2014, pp. 263–272.

[21] J. Rennie, L. Shih, J. Teevan, and D. Karger, "Tackling the poor assumptions of naive bayes text classifiers," in *Proceedings of the 20th International Conference on Machine Learning*, 2003, pp. 616–623.

[22] R. Swezey, H. Sano, S. Shiramatsu, T. Ozono, and T. Shintani, "Automatic detection of news articles of interest to regional communities," *International Journal of Computer Science and Network Security*, vol. 12, no. 6, pp. 99–106, 2012.

[23] R. Swezey, H. Sano, N. Hirata, S. Shiramatsu, T. Ozono, and T. Shintani, "An e-participation support system for regional communities based on linked open data, classification and clustering," in *Proceedings of the 11th IEEE International Conference on Cognitive Informatics & Cognitive Computing*, 2012, pp. 211–218.

[24] A. MacLean, R. Young, V. Bellotti, and T. Moran, "Questions, options, and criteria: elements of design space analysis," *Human Compututer Interaction*, vol. 6, no. 3, pp. 201–250, 1991.

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.