3C2-OS-13b-11

住民参画 Web プラットフォーム O_2 における関連情報構造化システム

Web Contents Structuring System on an e-Participation Web Platform O_2

平田紀史[*1] 佐野博之[*1] Robin M. E. Swezey[*1] 白松俊[*1] 大圃忠親[*1]
Norifumi Hirata Hiroyuki Sano Robin M. E. Swezey Shun Shiramatsu Tadachika Ozono

新谷虎松[*1]
Toramatsu Shintani

[*1]名古屋工業大学 大学院 工学研究科 情報工学専攻
Department of Computer Science and Engineering Graduate School of Engineering Nagoya Institute of Technology

We developed an e-Participation web platform O_2 (Open Opinion) for regional communities. Our platform aims at supporting citizen participation in ongoing debates by gathering and openly publishing news and opinions. This paper introduces a LOD dataset and the structuring system. The dataset has news, opinions on Twitter, conference minutes of local assembly and their relations. Users can input opinions for regional concerns and the opinions are saved as LOD. Structured related information support users understanding concerns and opinions.

1. はじめに

本研究では，地域社会での住民参画を促進するための Web プラットフォーム O_2 を開発中である．そのために，地域の社会問題に関連する情報を Linked Open Data (LOD) として蓄積し，構造化するシステムについて述べる．住民参画とは，住民の意見を集約し意思決定に反映させる取り組みであり [Macintosh 09]，様々な問題に関する議論を進めるためには，住民の問題意識，懸案事項などをまとめる構造化と，その共有が必要となる．関連情報の共有のために，ニュース記事やマイクロブログなどの Web コンテンツを LOD として自動的に構造化するシステムを提案する．O_2 では，住民による議題の設定や意見入力が可能であり，これらの情報も LOD として蓄積する．地域住民であるシステムのユーザは，イベントや，個々のニュース記事やツイートを基に議題を作成し，それに対する意見を入力して議論を進めていく．

地域情報の共有に関連するシステムとして，特定の地域に関するニュースに関する共有掲示板 n0tice[*1] や住民の投稿によってニュース記事が作成される Blottr[*2] などがある．これらのシステムは，ユーザ自身がニュース記事を作成し，関連情報や意見を加えていくという特徴がある．他にも，地域に関する情報を提供する Web サービスは数多く存在する[*3]．提案システムでは，ニュース記事自体は既存のニュースサイトによる記事を利用し，イベントとして関連付けを行う．ニュースサイトに掲載されない細かなニュースの場合は，議題のみを設定することになるため，議題の設定作業は n0tice や Blottr におけるニュース記事の作成に類似する．提案システムでは，既にニュースサイトに地域問題として話題となるようなニュース記事が存在する場合は，これらを参照して議題の設定が可能である．議題として明示的に問題提起することで，ニュースの閲覧よりも，問題に対する意見入力や議論の促進に重点をおいたシステムである．

連絡先: 平田紀史，名古屋工業大学大学院 工学研究科 情報工学専攻，〒 466-8555 名古屋市昭和区御器所町，052-733-6550，nori@toralab.org

∗1 n0tice, http://n0tice.com
∗2 Blottr, http://www.blottr.com
∗3 openlylocal (http://openlylocal.com) では，イギリスとアイルランドに関する 560 の地域関連サイトが登録されている．

2. 住民参画 Web プラットフォーム O_2

住民参画 Web プラットフォーム O_2 は大きく 3 構造に分類される．1) 議論の基となる情報の収集，2) 収集した情報の構造化，3) 蓄積された情報の利用である．1) では，ニュース記事や，地方議会が公開している議事録，ツイートを収集する．2) では，収集した情報を構造化し，LOD データセット SOCIA として蓄積する．構造化する際には，ニュース記事やツイートの対象地域の推定やイベントの抽出などを行い，イベントを中心に収集した情報を関連付ける．3) では，構築した LOD を利用して，議論支援システムやコンサーン・アセスメントの支援のための意見分析システム，対面での会議支援システムなどを開発している．本論文では 1) と 2) に関して，Web 上のニュース記事や意見の収集と構造化について述べる．

ニュース記事や Twitter など，Web 上の情報を利用することで，地域に関する社会問題が自動で抽出できる．ニュース記事になるような地域に関する問題は，Twitter などでも話題になることが多くなると考えられる．そのため，O_2 のユーザによる意見入力がない状態でも，議論の基となる情報の提供は可能となる．また，Web 上の情報を自動で関連付けするため，最新の情報を関連情報として提示することも可能となる．

3. 関連情報の構造化

3.1 構造化システムの概要

関連情報の構造化のために，Web 上の情報を収集する．収集対象はニュースサイト，Twitter，地方議会の議事録公開ページである．同一対象について言及されたニュース記事や発言を関連付け，LOD として蓄積する．LOD として公開することで，住民参画に必要とされる透明性 (transparency)[Lathrop 10] が確保されるようになる．本論文では，ニュース記事やツイートの言及対象をイベントして表す．ただし，イベントとして表す対象は，ニュース記事や発言の細かな内容ではなく，中心的な内容に関するものとする．

Web 上のニュース記事や発言などは，明示的に URL のリンクなどがない限り，関連付けが行われていない状態である．

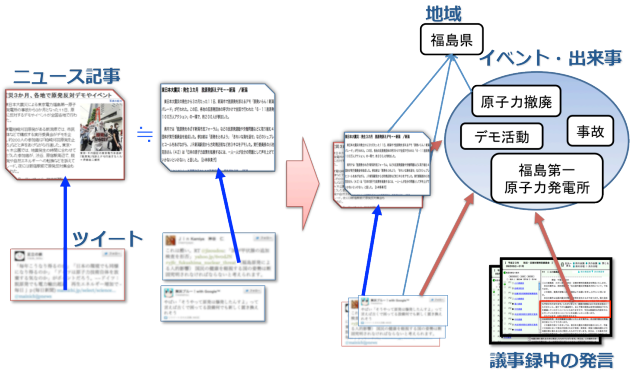

図 1: イベントによる構造化の例

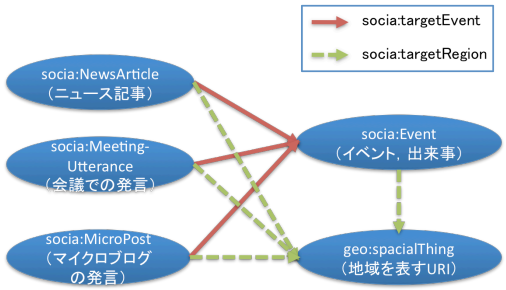

図 2: クラス間の関係

図 1 に示すように，イベントによって関連付けることによって，関連するニュース記事や発言などを提示可能になる．図 1 の例では，原発事故のデモ関連のニュース記事や発言をイベントに関連付けている．このような構造化が行われれば，ユーザは原発事故のデモに関連するニュース記事や発言を取得可能になる．

ニュース記事や発言を表すクラスについて，イベントに関する関係を図 2 に示す．イベントを Event，ニュース記事を NewsArticle，Twitter 上の発言を MicroPost，議事録上での発言を MeetingUtterance として表す．また，プロパティ targetEvent で関連付けられた NewsArticle や MicroPost，MeetingUtterance は，対象の Event に言及している状態を表す．プロパティ targetRegion はニュース記事やイベントなどの各クラスが関連している地域を表す．このような関連付けが行われれば，特定の地域に関するイベントやイベントに関連付けられたニュース記事一覧を取得することが可能となる．

3.2 ニュース記事を利用したイベント作成

ニュース記事を利用したイベントの生成と，ニュース記事や発言の関連付けを自動で行う手法について述べる．表 1 に示すように，イベントは関連する単語や地域，イベントの発生していた期間，イベントを表す短い文字列などで表される．また，socia:namedEntity と socia:term，socia:targetRegion で表すオブジェクトは複数である場合がある．

ニュース記事から得られる情報を利用して，Event の各プロパティに対応する情報を取得する．類似する記事をイベントに追加し，イベントの情報を更新していく．具体的には，新たに収集したニュース記事と既存のイベントとの類似度を計算し，閾値以上であればニュース記事をイベントに関連付け，イベントの情報も更新する．イベントが存在しない，または，閾値より小さければ単体のニュース記事からイベントの生成を行う．類似度計算には tf·idf を利用したコサイン類似度を利用した．ただし，ニュース記事のタイトルに出現する単語は tf·idf の値を 3 倍し，類似度は時間窓関数を乗算した値とした．(1) 式にニュース記事 a とイベント e との類似度を，(2) 式に窓関数を示す．

表 1: Event のプロパティ

プロパティ	オブジェクトの説明
rdf:type	クラス
dc:title	イベントを表す短い文字列
socia:term	イベントを表す単語
socia:namedEntity	イベントを表す固有名詞
socia:targetRegion	関連する地域の URI
socia:start	イベントの開始日時
socia:end	イベントの終了日時
socia:annotator	イベント作成者の URI

$$sim(a,e) = f(a,e) \cdot cos(a,e) \tag{1}$$

$$f(a,e) = \begin{cases} 1.0 & \text{if } start(e) < t_a < end(e), \\ \frac{-t_a + end(e) + T}{T} & \text{if } end(e) < t_a < end(e) + T, \\ \frac{t_a - (start(e) - T)}{T} & \text{if } start(e) - T < t_a < start(e), \\ 0 & \text{otherwize.} \end{cases} \tag{2}$$

$cos(a,e)$ は a と e のコサイン類似度であり，t_a は a の配信日時，$start_e$ と end_e はイベントの開始日時と終了日時を表す．

イベントの期間，すなわち，socia:start と socia:end は，イベントに関連付けられているニュース記事の，最初と最後の日時としている．また，イベントを表すタイトルは，イベントに関連付けられたニュース記事集合中から一つを選択する．これには，記事の本文を含めた単語と各タイトル中の単語とを比較し，評価値の高いタイトルを選択する [平田 12]．

3.3 ニュース記事とマイクロブログの地域判定

特定の地域に関する情報を提示するために，ニュース記事とマイクロブログを地域分類する．分類のためにはナイーブベイズ分類器の改良した Transformed Weight-normalized Complementary Nave Bayes (TWCNB) [Rennie 03, Swezey 11] を利用する．TWCNB では，単純ナイーブベイズで計算したクラス確率 $p(c|d)$ を以下の式で補正したスコアを用い，これが最大となるクラス c に文書 d を分類する．

$$logscore(d,c) = \log p(c|d) - \sum_w tf(w,d) log \left(\frac{1 + \sum_{k=1}^{|C|} tf(w,c_k)}{N + \sum_{k=1}^{|C|} \sum_{x=1}^{N} tf(x,c_k)} \right) \tag{3}$$

C は地域を表すクラスの集合，$tf(w,d)$ は文書 d における語 w の頻度，$tf(w,c)$ はクラス c における語 w の頻度，N は異なり単語の総数である．また，クラス間の偏りに対処するためにオーバーサンプリングした．増加させるサンプルはランダムに決定した．

さらに，信頼度を計算し，信頼度が閾値 θ 以上の値のみを分類結果とする．これには，曖昧な分類結果を除外する狙いが

表 2: 関連情報のクラス毎の数

クラス名	数
Event	40814
NewsArticle	63421
MicroPost	215561
MeetingUtterance	142293

ある．信頼度は (4) 式によって求める．

$$
\begin{aligned}
confidence(d) &= score_1 - score_2 \qquad (4)\\
score_1 &= \max_{c \in C} logscore(d, c)\\
c_1 &= \operatorname*{argmax}_{c \in C} logscore(d, c)\\
score_2 &= \max_{c \in C - \{c_1\}} logscore(d, c)
\end{aligned}
$$

すべてのクラスの集合 C に関して，$logscore(d, c)$ の最大値を $score_1$ とし，2 番目の値を $score_2$ とする．$score_2$ は $score_1$ となるクラス c_1 を除外した場合の値である．すなわち，$logscore(d, c)$ の最大値と 2 番目の値の差が大きいほうが，分類の信頼性が高いと判定する．

4. 構築した関連情報

4.1 関連情報の収集と構築

構造化した関連情報の数を表 2 に示す*4．これらは，2011 年 10 月 1 日からのニュース記事やツイートである．議事録中の発言は，名古屋市会の Web サイトで公開されている情報を利用した．ニュース記事は，ニュースサイトは asahi.com*5，MSN 産経ニュース*6，YOMIURI ONLINE*7，毎日 jp*8から得られた記事である．また，ツイートは，Streaming API*9により取得できたツイートから，地域やイベントと関連付けが可能であったもののみを登録している．平均すると 1 イベントに対して約 1.5 記事が関連付けられている．全国的なニュースの他に地域別のニュースも取得しているため，細かくなる傾向があったものと考えられる．

4.2 イベントに関連付けられたニュース記事

イベントに関連付けられているニュース記事の適合率を求めた．ランダムサンプリングしたニュース記事の内容と，それに関連付けられているイベントのタイトルが表す内容とを比較した．同一イベントであると判定した場合を正解したとき，136 イベント中 119 記事が正解であった．そのため，適合率は 87.5%となる．ここで，136 のイベントは 2 記事以上に関連付けられ，記事タイトルとイベントのタイトルが異なる場合のみを集計した．

表 3 にニュース記事のタイトルとイベントのタイトルの具体例を示す．ニュース記事の内容とイベントのタイトルが表す内容が同一である判定した場合に true としている．記事 1 の場合，タイトル中には共通の単語があるが，異なる内容である．記事 2 の場合も，タイトル中には共通の単語が存在するが，同一イベントと判定した．また，記事 3 の場合，イベントのタイトルは現状の解説と展望についての内容であり，ニュース記事の表す事件自体を表す内容ではない．そのため，意味的な関連性が高いが，イベントとしては別であると判定した．

記事 1 も記事 3 も，“ビール” や “イランと制裁” という広い観点から見れば，類似するイベントである．そのため，イベントの範囲の広さや観点の違いなどを考慮してイベントを抽出する手法も考えられる．地域でのイベントの閲覧が O_2 の目的であれば，ユーザごとに異なる観点を考慮して提示することは有用である．しかし，O_2 における議論支援システムでは，議論となるような話題提示によって，ユーザが話題に対する議題を設定し議論を進めるという状況を考えている．そのために，議論の前段階として話題の共有が必要であり，議論のためには話題を共有するユーザが一定数以上必要である．話題の共有という点においては，ユーザごとに異なるイベントを提示するより，同一イベントを提示した方が，議題当たりの参加ユーザ数は増加することが期待できる．ユーザごとに異なるイベントを提示して，それぞれのユーザがそれぞれの観点で議題を設定すれば，議論に参加するユーザが分散すると考えられるためである．しかし，システムの利用ユーザが十分に多ければ，提示するイベントが異なっても，観点が類似するユーザも多くなると考えられる．この場合は，ユーザごとにイベントを提示しても，他ユーザとの話題の共有や十分なユーザ数での議論は可能である．Web 上での議論において，議論のための適切なユーザ数や発言数を仮定するなら，議論への参加ユーザ数や発言数によって，提示するイベントの範囲や観点を調整するといった手法も検討する必要がある．また，類似する議題があっても，議題ごとに分断されているという問題がある．複数の類似する議題があった場合に，内容や議論の進行速度などを考慮して，議題の統合や関連議題の提示の仕組みが必要となる．

記事 4 では，本文やタイトル中での共通の単語は多いが，異なるイベントと判定した．記事のタイトルとイベントのタイトルが表す内容は，異なる日付の内容であった．このように日付に強く依存するイベントもあれば，依存が弱いイベントもあり，固定された窓関数では，正しく関連付けられる類似度を求めることが困難である．例えば “外為” や “為替” といった特定の単語が多く出現する場合は，窓関数の範囲を狭くしたり，記事本文中に特定の日付が出現する場合は，その日付の前後で窓関数が表す値を大きくしたりするといった工夫が必要である．

4.3 地域推定の精度

ニュース記事とツイートに対して，都道府県レベルでの地域分類実験を行った．47 都道府県と地域との関連性が低い “NO_REGION” という 48 クラスへ分類する．ここでは，訓練データとテストデータが等しい closed test を行った．適合率は，クラスに分類されたコンテンツ中での正解数の総数を，クラスとして分類されたコンテンツ数の総数で割った値とした．再現率は，クラスに分類されたコンテンツ中での正解数の総数を，クラスごとの本来の正解のコンテンツ数の総数で除算した値とした．

ニュース記事の分類の場合は，訓練データは，Yahoo ニュースから得られた地域ごとに分類されたニュース 8,811 と，地域とは無関係のニュース 1,133 とした．2011 年 6 月 13 日から 2011 年 7 月 12 日に配信されたニュース記事を利用した．(4) 式に関して閾値 θ を変化させた場合の精度と再現率を図 3 に示す．適合率と再現率の調和平均である F 値は，98.1%が最大であった．F 値が最大となる閾値では，適合率が 98.4%，再現率が 97.8%という値であった．

ツイートの場合は，都道府県別にハッシュタグを用意し，そ

*4 2012 年 4 月上旬時点での情報
*5 http://www.asahi.com/
*6 http://sankei.jp.msn.com/
*7 http://www.yomiuri.co.jp/
*8 http://mainichi.jp/
*9 https://dev.twitter.com/docs/streaming-api

表 3: ニュース記事のタイトルと，ニュース記事に関連付けられたイベントのタイトルの具体例

	ニュース記事のタイトル	イベントのタイトル	同一判定
記事 1	出っ腹、ビールのせいじゃない	ビール類出荷 最低を更新	false
記事 2	文科省調査：精神疾患休職の教諭　１８年ぶり減	心の病で休職の教員、半数が在校２年未満　公立校調査	true
記事 3	在イラン英国大使館に数百人侵入　追加制裁に怒り	ＩＡＥＡ報告書：イランと欧米　対立先鋭化…制裁強化に道	false
記事 4	外為：東京＝１７時　１ドル＝７７円７１-７２銭	１ドル７８円をはさんで小動き　東京外国為替市場	false

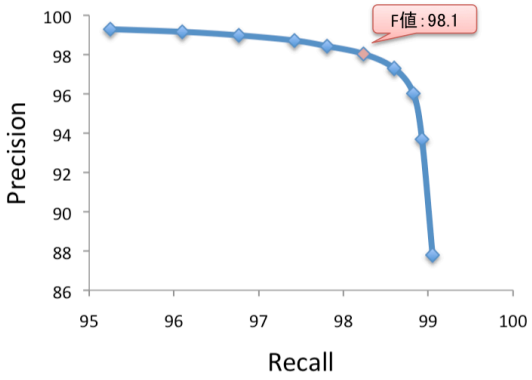

図 3: ニュース記事の地域分類の適合率と再現率

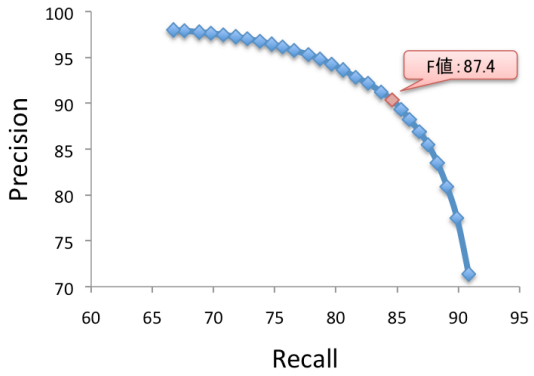

図 4: マイクロブログの地域分類の適合率と再現率

のハッシュタグが付けられたツイートをその地域に関するツイートであるとした．訓練データは，Topsy Otter API[*10]から得られた地域に対応するハッシュタグのついた 20,000 ツイートと，ハッシュタグのないツイートを “NO_REGION” とした．例えば，愛知県なら “#aichi” のように地名をそのまま利用したハッシュタグを利用した．ニュース記事の場合と同様に，(4) 式に関して閾値 θ を変化させた場合の精度と再現率を図 4 に示す．最大となる F 値は 87.4%であり，この場合，適合率 90.4%，再現率が 84.6%という値であった．

確信度を利用しない場合の結果は，図 3 と図 4 での最も右下のプロットであり，再現率は高いが，精度が低い結果となっている．確信度を利用することで，F 値の高い結果を得ることが可能となった．しかし，実際のデータと本実験で利用したデータでは，各都道府県や “NO_REGION” と判定されるツイートやニュース記事の割合が異なる．そのため，実際のデータの分布に基づいて，高い適合率と再現率を示す閾値 θ を検討する必要がある．

5. おわりに

本論文では，地域社会での住民参画を促進するための Web プラットフォーム O_2 について述べた．特に，住民参画のための情報を共有するために，Web コンテンツ中から地域の社会問題に関連する情報を抽出し，LOD として蓄積するシステムについて述べた．関連情報の構造化において，ニュース記事のイベントへの精度は 87.5%という結果が得られた．また，地域分類については，確信度を用いることで，ニュース記事の場合では F 値が 98.1%，マイクロブログの場合では F 値が 87.4%という結果が得られた．Web コンテンツが対象とするイベントや地域に関して構造化することで，地域ごとに議論の種となる情報を提示可能であり，住民参画を行うためのシステムへの利用が可能となった．

*10　http://code.google.com/p/otterapi/

謝辞

本研究の一部は，総務省戦略的情報通信研究開発推進制度 (SCOPE) の支援を受けた．

参考文献

[Macintosh 09] Ann Macintosh, Thomas F. Gordon, and Alastair Renton, “Providing argument support for e-participation”, Journal of Information Technology & Politics, Vol. 6, No. 1, pp. 43–59, 2009.

[Lathrop 10] Daniel Lathrop and Laurel Ruma, “Open Government - Collaboration, Transparency, and Participation in Practice ”, O ’ Reilly Media, 2010.

[平田 12] 平田紀史，白松俊，大囿忠親，新谷虎松，“同一イベントを対象としたニュース記事集合の表現について”，2012 年電子情報通信学会 総合大会，2012.

[Rennie 03] Jason D.M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger, “Tackling the poor assumptions of naive bayes text classifiers”, In Proceedings of the 20th International Conference on Machine Learning, pp. 616–623, 2003.

[Swezey 11] Robin Swezey, Shun Shiramatsu, Tadachika Ozono, and Toramatsu Shintani, “Intelligent page recommender agents: Real-time content delivery for articles and pages related to similar topics”, In Modern Approaches in Applied Intelligence, Proceedings of IEA/AIE 2011, Part II, pp. 173–182, 2011.