Linked Open Dataを用いた コンサーン・アセスメント支援機構の開発

白松 俊[1]・佐野 博之[1]・平田 紀史[1]・Robin SWEZEY[1]・大囿 忠親[1]・新谷 虎松[1]

[1]非会員　名古屋工業大学　工学研究科（〒466-8555 名古屋市昭和区御器所町）
E-mail: siramatu@nitech.ac.jp

本研究では，オープンガバメント3原則（透明性，参画，協働）のうち，特に地域での透明性と参画に焦点を当て，住民参画Webプラットフォームを開発中である．まず透明性を確保するため，データを意味的にリンクさせて蓄積・公開するLinked Open Data (LOD) を用い，コンサーン・アセスメントの結果を公開・共有するための情報共有基盤を構築した．具体的には，コンサーン・アセスメントに特化したLODデータセットSOCIAを設計し，地域ごとのニュース記事，出来事，関連するTwitterのツイート，および地方議会議事録の発言を相互に関連付けるテキストマイニングシステムを開発した．また，SOCIAを用いたコンサーン・アセスメント支援機構を有する議論支援システムcitispe@kを開発した．さらに，コンサーン・コーパスの構築に向け，コンサーンを含むツイートの特徴抽出実験とその分析結果を報告する．

Key Words : *Linked Open Data, text mining, corpus, concern assessment, public involvement, Twitter*

1. はじめに

本研究では，地域社会をターゲットとした住民参画WebプラットフォームO_2 (http://open-opinion.org/) を開発中である．近年，地域社会が備えるべき問題は多様化しており，自然災害，放射能汚染，電力不足，高齢化，経済問題など多岐に渡る．地域社会がこのような多様な問題やリスクに備え，対処するためには，討議を通じて住民の意見を公的な意思決定に反映させるプロセスである住民参画 (public involvement) [1] が非常に重要である．これは，オバマ政権の「透明性とオープンガバメント」覚書[2] で示されたオープンガバメント三原則，すなわち透明性 (Transparency), 参画 (Participation), 協働 (Collaboration) の実現が，日本の地域社会においても急務であることを意味する．本研究では，それら三原則のうち特に透明性と参画に焦点を当て，その基盤としての情報共有インフラの実現を目指す．

地域社会が備えるべき問題や懸案事項，すなわちコンサーンは多岐に渡るため，全ての問題に通じている住民は少ない．多様なコンサーンに対処すべく議論を深め，議の論内容を意思決定に活用するためには，増加していく意見や資料をコンサーンと関連付けて整理し，その結果を共有可能にする仕組みが重要となる．我々は，このようにコンサーンとその背景を整理・構造化する作業が，すなわちコンサーン・アセスメントであると捉えている．

我々は，住民-行政-専門家の間で共有すべきコンサーン，すなわちコンサーン・アセスメントで構造化すべきコンサーンは，以下の8項目に細分化できると考える．

(A) 問題 (Issue): 地域社会にどのような社会問題があるか
(B) 背景 (Background): 各問題にはどのような背景情報があるか
(C) 選択肢 (Option): 各問題にはどのような解決策の選択肢があるか
(D) 評価基準 (Criterion): それら選択肢を検討する上でどのような評価基準を考慮する必要があるか
(E) 評価極性 (Polarity): 各選択肢をある評価基準から検討した場合，どのようなメリット・デメリットがあるか
(F) 利害関係者 (Stakeholder): その問題に係わる意思決定で影響を受けるのは誰か
(G) 決定事項 (Decision): 意思決定が済んだ事項はどれで，未決定事項はどれか
(H) 根拠 (Evidence): 意思決定に使われた根拠情報はどれで，使われなかった根拠情報はどれか

これらの項目は実際には並列ではなく，項目間の依存関係や包含関係を考慮した構造化が必要である．本稿では特に(A)〜(E)の5項目に焦点を当て，その構造化のためのLinked Open Data (LOD)[3] を設計・構築する．

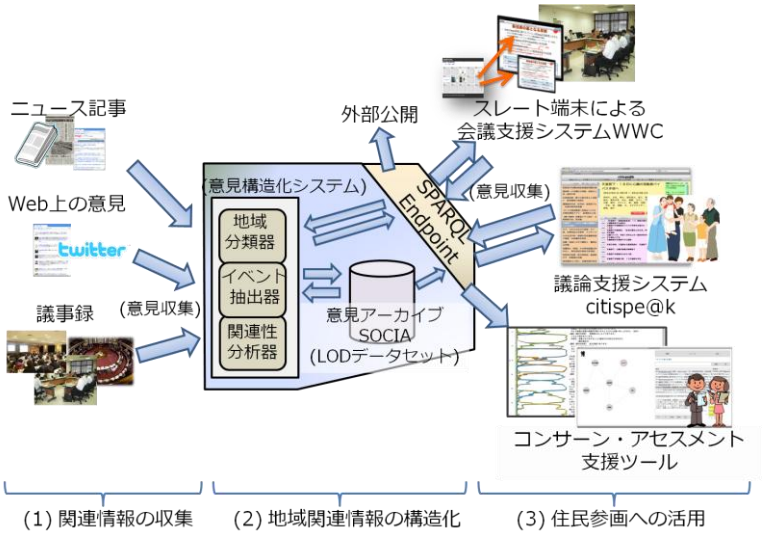

図-1　住民参画 Web プラットフォーム O_2 の概要

LODとは，事象間意味的な関係のリンクを明示的に記述した上で公開されたデータ集合であり，特にオープンガバメント三原則のうち透明性を確保する上で重要な役割を担う[4]．住民参画WebプラットフォームO_2におけるコンサーン・アセスメントへの活用を目指して設計したLODは，SOCIA (Social Opinions and Concerns for Ideal Argumentation) と名付け，http://data.open-opinion.org/ にて公開している．

まず項目(A), (B) の情報共有を可能にするためには，地域に関するコンテンツを相互に関連付けて構造化する必要がある．地域関連コンテンツとして，タウン情報誌やコミュニティ誌など様々な媒体が存在するが，まずは収集・分析が容易なWeb 上のコンテンツを利用する．具体的には，図1(1) に示すように，Web 上のニュース記事，マイクロブログ，および自治体の議会議事録を議論の「種」として収集する．収集したWeb コンテンツは，
図1(2) で地域毎に分類して相互に関連付け，そのメタデータをLOD データセットSOCIA に蓄積する[5,6]．
また項目(C), (D), (E) の情報共有を可能にするためには，発散していく議論を構造化する作業の支援が必要となる．図1(3) では，SOCIA で構造化された関連コンテンツを活用した意見入力や，評価基準タグ等の付与を可能にする議論支援システムcitispe@kを提供する[7]．

LODの枠組を採用することにより，他組織が公開したデータやシステムを相互に活用し合うことが可能となるが，そのためにはデータそのものやリンクの意味定義を含むオントロジーも公開する必要がある．本稿でも，上記(A)～(E)の5項目を対象としたコンサーン・アセスメントのためのドメインオントロジーとして，SOCIA オントロジーの概要を示す．ドメインオントロジーとは，ある特定の分野（ドメイン）の事象間の関係を記述するための意味定義であり，その分野のクラス（概念）やプロパティ（概念／事象間の関係）の定義を含む辞書的な記述である．セマンティックWeb研究においては，まず事象はURI (Uniform Resource Identifier) が割り当てられたリソースとして表現され，事象間の関係はRDF (Resource Description Framework) で記述される．また事象間の関係だけでなく，関係の意味定義のためのオントロジーも，RDFに基づくOWL (Web Ontology Language) で記述される．本稿で概要を示すSOCIAオントロジーも，OWLで記述し，http://data.open-opinon.org/socia-ns にて公開している．

2.　関連研究

2.1　住民参画のためのLinked Open Data

住民参画やオープンガバメントの促進にあたり，特に透明性を確保する上でLOD は重要な役割を担う[4]．現在，オープンガバメントを志向したLODプラットフォームとして，各国でData.gov, Data.gov.uk, Data.gov.au, data.gouv.fr, India.gov.in, などが公開されている．中でも2009 年から米国政府が運営するData.gov は，国民全体にアイディアを募り(Brainstorming)，議論を深め(Discussion)，案を作る(Drafting) という三段階の参加型ダイアログへのデータ活用を先駆的に試みており[8]，この取り組みはソーシャルメディアを用いた住民参画の代表例と言える．またData.gov とIndia.gov.in は，共同でデータ管理システム等をオープンソース化する作業を進めている[9]．

Joinup[10] は，ヨーロッパ各国の行政機関の連携を促進するための協働プラットフォームである．Joinup では，公的データ資源共有のためのオントロジーADMS (Asset

Description Metadata Schema) を提案している．

日本でも経産省によるオープンガバメントラボ[11]がデータ公開や意見収集のためのWebアプリケーションを試験運用しており，東日本大震災後の復興への活用を目指したコンサーン・アセスメントが 試みられている[12]．

また，地域情報の構造化のために必要となる地名や地域に関しても，GeoNames[13] 等のLOD データセットが構築されている．日本国内の地域に特化した事例としては，LODAC プロジェクトが地名・施設名のデータを公開しており[14]，本研究でもLODAC の地名データを利用している．

以上のように，公開されたデータを住民参画に活用しようとする試みが盛んになってきているが，本研究で目指しているような問題意識の共有基盤を実現するためには，データを活用して行なわれた討議をどのように分析・構造化し，LOD として蓄積すれば良いかが重要な課題である．この課題に関する統一的な解は未だ明らかでない．

2.2 議論の構造化と可視化

公的討議から問題意識を抽出して共有可能にするコンサーン・アセスメントのためには，討議の背景情報を構造化して共有する必要がある．この観点から，議論の構造を可視化するアプローチが効果的である[15]．Jeong ら[16] は，議事録に含まれる発言毎の語の共起頻度を用い，参加者間の認識のさを可視化した．また，議論の全体像を把握可能にするためには，議論構造の可視化が効果的である．多くの議論支援システムは，発言をノードとし，発言間の関係をリンクとして議論構造を可視化するアプローチをとる[17]．そのようなシステムとして，Cohere[18]，Deliberatorium[19]，Discourse Semantic Authoring[20]，等が挙げられる．Cohere は，Question（問題），Option（解決策の選択肢），Criteria（評価基準）という3 種類のノードに基づくQOCモデル[21] によって議論を構造化する．Deliberatoriumは，Issue（問題），Idea（解決策のアイディア），Argument（論拠）という3 種類のノードにより議論を構造化するが，これはIBIS (Issue-Based Information System) モデル[22] を参考にしている．また，Discourse Semantic Authoringでは，RST (Rhetorical Structure Theory) [23] に基づく談話関係によって議論を構造化する．

本研究の議論支援システムcitispe@kもQOCモデルやIBISモデルを参考にした議論構造化を行う[7] が，地方議会の書き起こし議事録については，図2の右側に示すような議題遷移図をユーザに提示する．このような可視化も併用することで，長い議事録全体の議題の流れを把握しやすくする効果がある[24]．

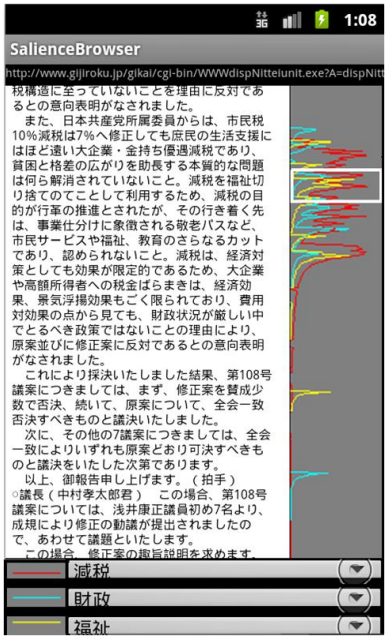

図-2 地方議会の書き起こし議事録の議題遷移図

3. SOCIA: コンサーン・アセスメントのための Linked Open Data

地域に関連するWeb コンテンツをデータセットSOCIA に蓄積・構造化し，住民参画に活用する運用サイクルを図3 に示す．SOCIA 中に地域関連コンテンツを構造化することで，議論の「種」として共有可能にする．まず，第1. 節で述べた項目(A)，すなわち地域にどのような問題があるのかを提示する．また，意見の入力を補助するため，(B)問題の背景情報となる関連コンテンツを活用する．入力された意見もSOCIA 上に蓄積し，(A), (B) の情報共有に活用する．さらに，関連コンテンツや入力された意見を効果的に整理するため，(C)解決のための選択肢，(D)コンテンツが言及している評価基準（経済的側面，環境面など），(E)各評価基準からのメリット・デメリットをタグ付け可能にする．さらに，自動解析の信頼度やアルゴリズムの精度，手動タグ付けの作業者など，アノテーションされた状況をも併せて管理することで，SOCIA 中のデータをコーパスとして利用できる可

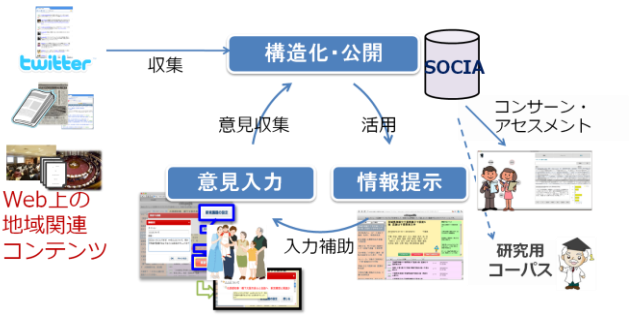

図-3 Web 上の情報を活用した住民参画のサイクル

図-4　地域情報構造化の中核となるクラス群

能性があると考える．

3.1　地域関連情報の構造化

上述したような地域関連コンテンツの活用を可能にするためには，同一地域に言及したコンテンツ群や，同一事象に言及したコンテンツ群を相互に関連付けておく必要がある．具体的には，まずコンテンツが言及する地域へそのコンテンツを分類し，次にニュース記事をクラスタリングして得られるイベントへの関連付けを行う．データセットSOCIA には，そのようなWeb コンテンツ間の関係を構造化して蓄積する．図4 に，そのための中核となるクラス群を示す．これらのクラス群はSOCIA オントロジーにて定義されている．図5 は，SOCIA 中で同一のイベントに言及した複数のコンテンツが，SOCIA オントロジーに則って構造化された例を示している．この例では，「愛知県が震災がれきの受け入れ方針を固めた」という同一イベントに言及している複数のニュース記事が，socia:targetEventプロパティで同イベントに紐付けされている．また，これらは愛知県に言及したニュース記事・イベントであるため，socia:targetRegionプロパティでLODACの愛知県に紐付けされている．この例ではニュース記事のみを示しているが，実際にはマイクロブログの発言や，議事録中の発言，および議論支援システムcitispea@k から入力された意見についても同様の構造化を行う．

さらに，評価基準と評価極性を考慮した構造化のために，経済＋，経済－，治安＋，治安－，環境＋，環境－などの評価基準タグを定義した．評価基準タグは議論支援システムcitispea@k上から人手で付与可能であり，SOCIA上では図6のような形で定義されている．評価基準タグはsocia:polar プロパティを持ち，評価極性が正（メリット）の場合は+1, 負（デメリット）の場合は-1の値をとる．

3.2　アノテーションされた状況に関する構造化

SOCIA 中のデータはまず自動的に収集・構造化されるが，自動解析には必ず解析誤りを伴う．そこで，データの精度を向上させるためには，自動解析の確信度や精度，人手の修正といったアノテーション付与時の状況を管理する必要がある．SOCIA オントロジーでは，その

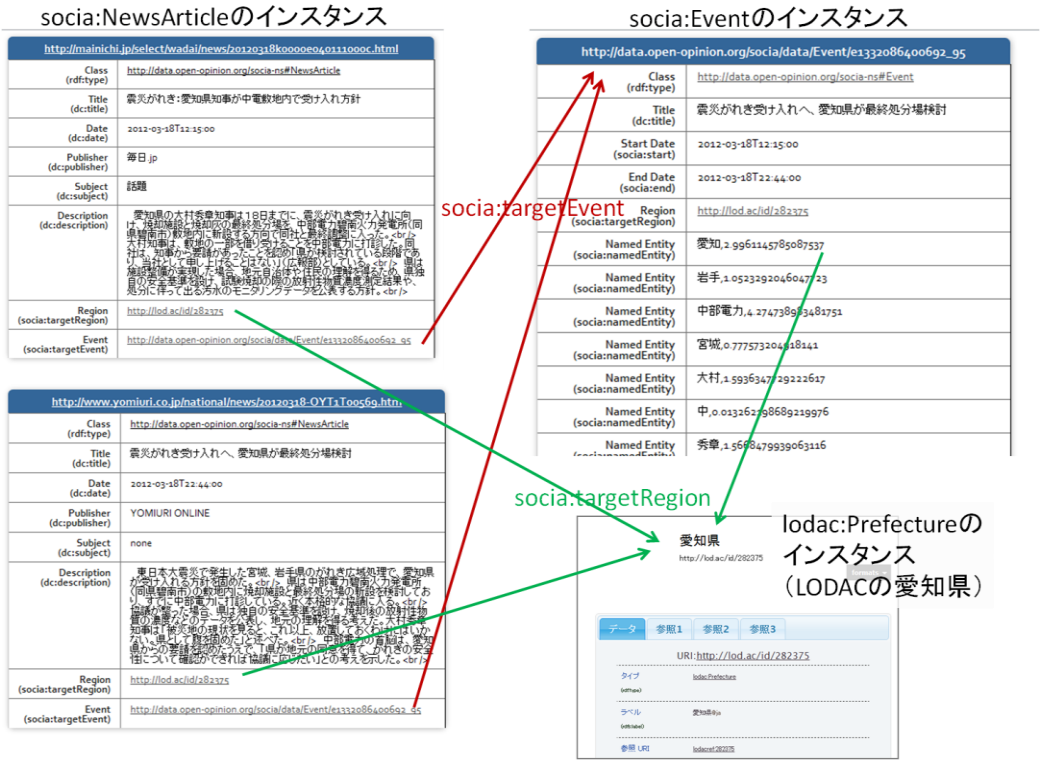

図-5　イベントと地域を基点にしたコンテンツ構造化の実例

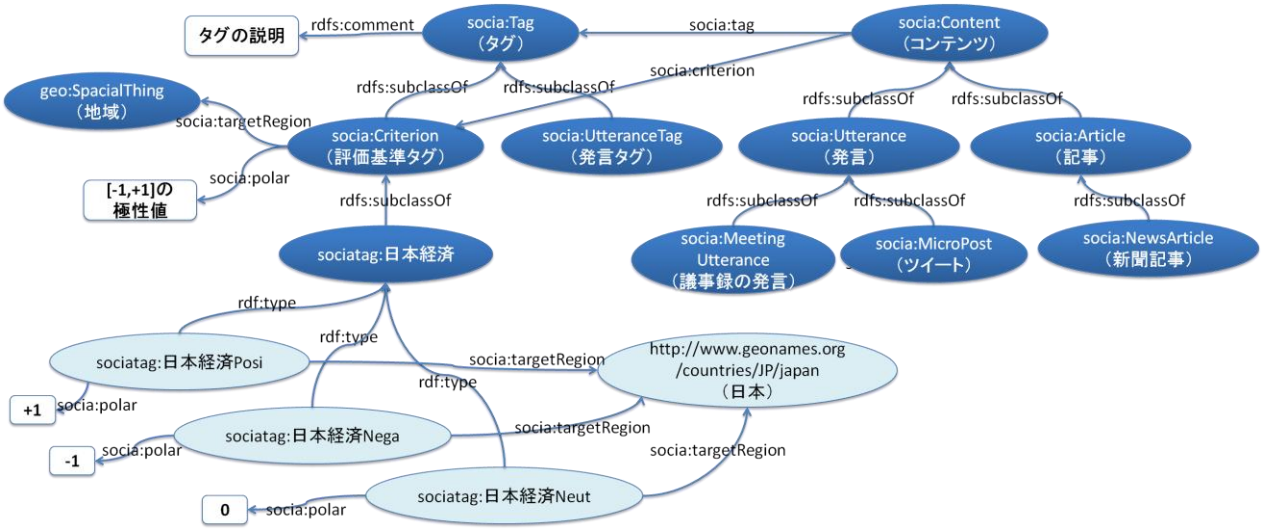

図-6　評価基準・評価極性を表すタグの定義例

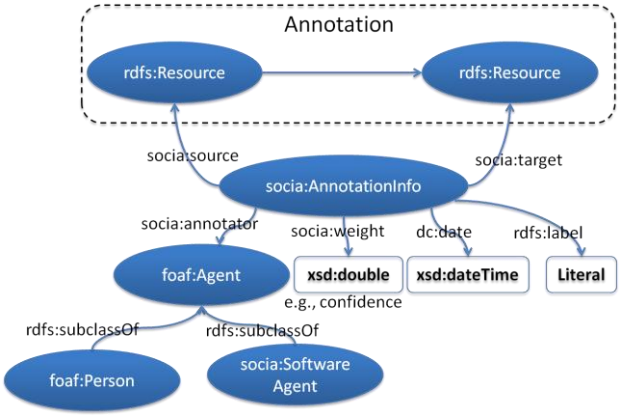

図-7　AnnotationInfo: プロパティがアノテーションされた状況

ようなアノテーション付与時の状況を記述するためのクラスとして，図7に示すsocia:AnnotationInfoクラスを定義している．これは，プロパティの付与が自動か手動か，自動ならどの解析アルゴリズムで確信度はどの程度か，手動なら誰が作業者で何を修正したのか，といった状況の管理を可能するためのクラスである．これにより，SOCIA で蓄積したメタデータの品質を向上させ，研究用のコーパスとして利用することが可能になると考える．

3.3　SOCIA の公開

データセットSOCIA の公開には，Apache Jena プロジェクトによるオープンソースのSPARQL サーバであるFusekiを拡張して用いた．SPARQL エンドポイントの公開にはFusekiをそのまま用いることができたが，RDF グラフの公開には拡張が必要であった．Fuseki 0.2.0 ではRDF グラフ全体の公開はサポートしているが，この仕様では大きなグラフに対するスケーラビリティが無く，またリソースURL からの簡便なデータ取得にも未対応であったため，以下の拡張を行なった．

まず，グラフURL に対し，1 回の表示エントリ数やプロパティをGET パラメタで指定可能にした．ニュース記事から抽出したイベントのグラフURL は http://data.open-opinion.org/socia/data/Event であるが，例えば1 回の表示エントリ数limit が50, 対象地域socia:targetRegion が愛知県であるイベントを以下のようなGETパラメタを含むURLで指定できるよう拡張した．

http://data.openopinion.org/socia/data/Event?limit=50
&socia:targetRegion=http://lod.ac/id/282375

また，Fuseki では未対応であったリソースURL からのデータ取得をサポートするよう拡張した．前ページ図5 に示した各インスタンスは，ブラウザからリソースURL にアクセスした場合に表示される表である．RDF をブラウザで閲覧した場合は，XSLTスタイルシートによってHTML テーブルに変換される．

4.　citispe@k：SOCIAを用いた議論支援システム

SOCIAを活用したコンサーン・アセスメントのために議論支援システムcitispe@kを試作した．これは，Web 上のニュース記事やTwitterを参考にして地域の社会問題について議論し，コンサーンに関するコンテンツを共同編集するための議論支援システムである．citispe@k はWeb アプリケーションとして実装されており，PC, タブレット端末を問わず，Web ブラウザから利用可能である．citispe@k は大きく分けて，関連情報の提示，議論の構造化という2 つの役割を持つ．

4.1　関連情報の提示

Web ブラウザを用いてcitispe@k のURL にアクセスすると，図2 に示すように，イベントや関連情報が提示され

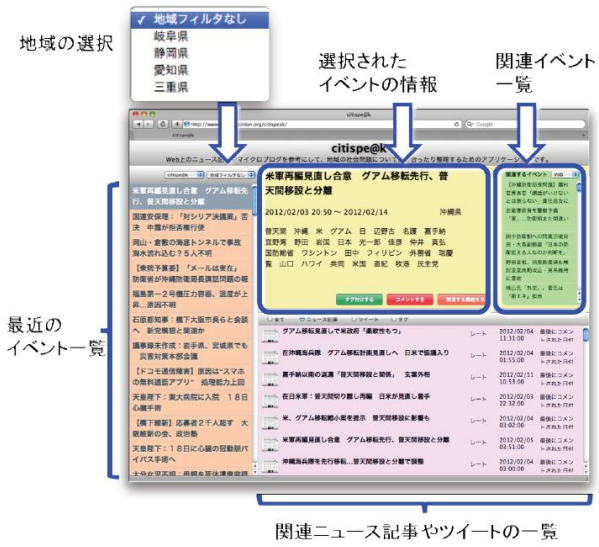

図-8　イベント一覧と関連情報の提示例

る．左側には，最近のイベントの一覧が表示され，地域のフィルタを選択することで，特定の地域に関するイベントのみの提示が可能となる．そして，特定のイベントを選択すると，右側にイベントの情報が提示される．それぞれは，イベント自体の情報，イベントとの関連ニュース記事やツイート，他の関連イベントへのリンクである．

リンクをクリックすると，関連イベントの場合は，そのイベントが選択された状態（図8）となる．ニュース記事やツイッターの場合は，そのURL のWebページを表示する．

ヘッダー部分にボタンが追加された状態でWeb ページが表示され，ヘッダーの[コメントする] ボタンを押すことで，そのニュース記事に対する意見を入力できる．ここで入力された意見はLOD サーバに登録され，Twitter にも投稿される．現在は特定のアカウントでの投稿となっているが，今後はユーザ自身のアカウントによる投稿にも対応する予定である．

本システムを用いて議論を行う場合，ユーザはまず，議題を作成する必要がある．図9 に議題作成のインターフェースを示す．

イベントを選択し，「関連する議題を見る」ボタンを押すと，ユーザが作成した議題の一覧と，「新規議題の設定」ボタンが表示される．ここで「新規議題の設定」ボタンを押すと，議題のタイトルと詳細を入力するためのビューが表示される．タイトルと詳細を入力した後に「OK」ボタンを押すと，閲覧中のイベントと関連付けられた状態で新たな議題が作成される．

4.2　タグ付けによる議論の構造化

citispe@k における議論はグラフ構造として表現され

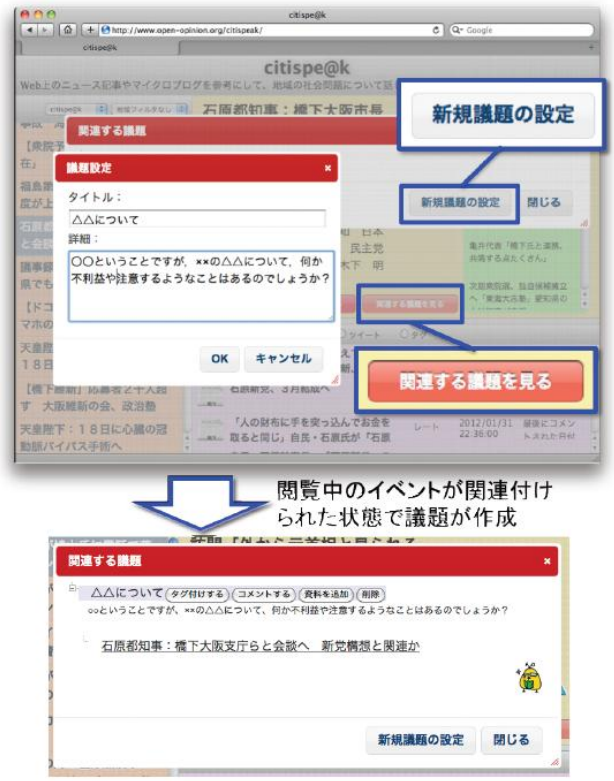

図-9　議題の設定と，コンテンツ・意見に対するタグ付け

る．このグラフ構造のノードに相当するのは，イベント，SOCIA 上に蓄積されたWeb コンテンツ，citispe@k 上で作成された議題，およびcitispe@k から入力された意見である．各コンテンツに対してタグを付与し，コンテンツ同士をリンクで接続していくことによって議論の構造化を目指す．本システムでは，イベント，SOCIA 上に蓄積されたWeb コンテンツ，作成した議題に対して，(1) タグ付けを行う，(2) コメントする，(3) 資料を追加，の3 つのアクションが可能である．タグとしては，3.1 節で示した評価基準タグ（経済＋／－，治安＋／－，環境＋／－，教育＋／－など）の他にも，発言の意図（質問，アイディア，ツッコミ，非難，ファシリテーション）を表す発言タグが定義されている．

5.　コンサーンを述べたツイートの特徴抽出

コンサーン・アセスメント支援のために，コンサーンを様々な媒体・コンテンツから自動抽出する技術の確立が望まれる．本節では特にTwitterからのコンサーン自動抽出技術の開発に向け，コンサーンを述べたツイートの特徴抽出実験を行う．

5.1　コンサーンの定義に向けたアプローチ

Twitter からコンサーンを述べたツイートを自動的に抽出するためには，コンサーンを述べたツイートの特徴

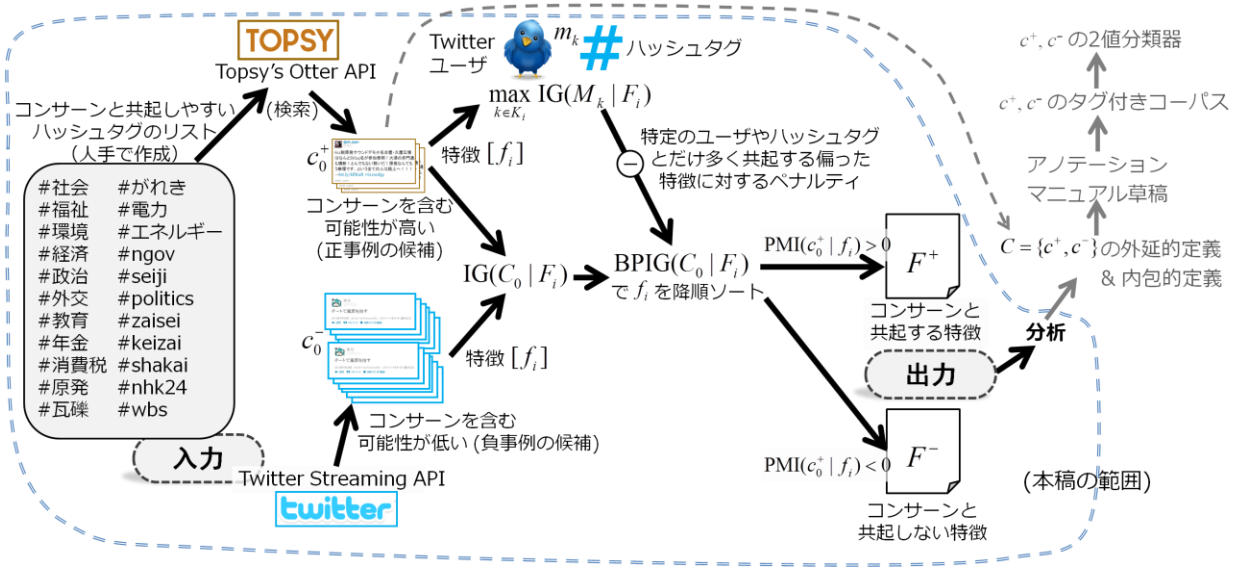

図-10　Twitter からのコンサーン抽出のための特徴分析

を何らかの方法で定め，コンサーンとそれ以外を分類する必要がある．しかし，住民-行政-専門家の間で共有すべきコンサーンを含むツイートの言語的特徴をトップダウンに定義することは難しい．自然言語処理の分野では，内包的な特徴を明確に定義できない分類タスクの場合，人手で分類した正解コーパスを用いた教師付き学習により，2値分類器の分類モデルを訓練するアプローチを取ることが多い．

正解コーパスを構築するにも，やはり複数のアノテーション作業者にとって解釈のぶれが少なく了解性の高い分類基準を定義し，複数作業者の分類結果がなるべく一致するようなアノテーションマニュアルを用意する必要がある．そのためには，先験的・内包的な考察のみに頼るのではなく，実際のデータの分析を通じた経験的・外延的な分類基準が必要となる．例えば，コンサーンを述べているか否かの分類基準を，1節で述べた8項目（問題，背景，選択肢，評価基準，評価極性，利害関係者，決定事項，根拠）を含むか否かによって内包的に定義したとする．このような定性的・内包的な分類基準だけで実データの分類作業を試みた場合，作業者ごとに解釈が異なる事例が多くなり，自動分類モデルの訓練に適さない正解コーパスになる恐れがある．これを防止するために，実際のツイート分類事例による外延的な定義の併用が有効である．

また，Twitter中に含まれるコンサーンの絶対数は少なくないが，膨大なツイート全体からすれば割合が小さい．コーパス構築時に，アノテーション作業者が膨大な日本語ツイートを全てチェックするのは効率が悪く，非現実的である．よって，コンサーンを表す特徴を定めておき，コンサーンの可能性がある候補ツイートを作業者に自動推薦する機構が必要である．

本節では，実際のTwitterでつぶやかれているコンサーンの言語的特徴を抽出する実験を行い，コンサーンの外延的定義とコンサーン候補自動抽出の実現に向けた分析を行う．

5.2　バイアス罰則付き情報利得による特徴抽出

コンサーンについて述べた実際のツイートの特徴を分析するため，図 10 のような手順を踏んだ特徴抽出実験を行う．まず，コンサーン・アセスメントで扱うべきコンサーンを含むツイートのクラスをc^+，含まないツイートのクラスをc^-とする．最終的にはc^+とc^-を分類したコンサーン・コーパスを構築するのが目的だが，本稿ではまず近似的に Twitter のハッシュタグを使った検索により，コンサーンを含む可能性が高い正例候補c_0^+を収集し，これにより近似的な分析を行う．ハッシュタグとはシャープ記号#で始まる文字列であり，ツイートが言及するトピック等を表すために用いられている．ハッシュタグの意味は，Twitter ユーザやコミュニティが自然発生的に定めており，その運用は自由であり厳格でない．

コンサーンを含む可能性が高い正例候補c_0^+としては，表 1 に示した 22 種のハッシュタグをクエリとし，Topsy

表-1 c_0^+ の収集に用いたハッシュタグとツイート数内訳

ハッシュタグ	ツイート数	ハッシュタグ	ツイート数
#社会	1,981	#電力	1,020
#福祉	1,629	#エネルギー	797
#環境	1,380	#ngov	1,040
#経済	1,985	#seiji	4,796
#政治	3,131	#politics	1,775
#外交	986	#zaisei	1,014
#教育	1,865	#keizai	2,406
#年金	940	#shakai	1,018
#消費税	1,592	#nhk24	1,844
#原発	3,129	#wbs	289
#瓦礫	2,367	合計（延べ数）	38,933
#がれき	1,949	合計（異なり数）	32,844

Otter's API でツイート検索を行って得られた日本語の 32,844 ツイートを用いる．表 1 にその内訳を示す． また，コンサーンを含まない可能性が高い負例候補c_0^-としては，Twitter Streaming API から得られた日本語の 149,984 ツイートを用いる．

このように，ハッシュタグで検索した正例候補c_0^+をそのまま機械学習の訓練データとして用いた場合，入力ハッシュタグ集合に依存した標本選択バイアスがあるため，偏った特徴が抽出されてしまう恐れがある．従来，特徴が分類精度に寄与する度合を評価する尺度としては，以下の(1) 式が表す情報利得 (Information Gain) がよく用いられていた．

$$\mathrm{IG}(C_0|T_i) = \mathrm{H}(C_0) - \mathrm{H}(C_0|T_i) \quad (1)$$

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

$$\mathrm{H}(C_0|F_i) = -p(c_0^+|f_i^+)\log p(c_0^+|f_i^+) - p(c_0^-|f_i^+)\log p(c_0^-|f_i^+) - p(c_0^-|f_i^+)\log p(c_0^-|f_i^+) - p(c_0^-|f_i^-)\log p(c_0^-|f_i^-) \quad (3)$$

ただし，f_iは特徴，f_i^+はf_iを含むツイートのクラス，f_i^-はf_iを含まないツイートのクラスを表し，$C_0 = \{c_0^+, c_0^-\}$, $F_i = \{f_i^+, f_i^-\}$である．この，従来の情報利得では標本選択バイアスに対処できない．これに対処するため，我々は以下のようなバイアス罰則付き情報利得 (BPIG; Bias-Penalized Information Gain) を定める．

$$\mathrm{BPIG}(C_0|F_i) = \mathrm{IG}(C_0|F_i) - \max_{k \in K_i} \mathrm{IG}(M_k|F_i) \quad (4)$$

ただし

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

表-2 c_0^+ の特徴として抽出された形態素 trigram

情報利得の上位		バイアス罰則付き情報利得の上位	
形態素 trigram f_i	$\mathrm{IG}(C_0\|F_i)$	形態素 trigram f_i	$\mathrm{BPIG}(C_0\|F_i)$
）[URL][E]	8.38×10^{-3}	」[URL][E]	4.14×10^{-3}
』[URL][E]	8.25×10^{-3}	ている。	1.21×10^{-3}
:[URL][E]	6.79×10^{-3}	ているの	9.23×10^{-4}
...[URL][E]	5.96×10^{-3}	ています	9.14×10^{-4}
」[URL][E]	5.51×10^{-3}	している	9.13×10^{-4}
。[URL][E]	4.13×10^{-3}	。RT [USER]	7.66×10^{-4}
NEWS WEB 24	3.70×10^{-3}	された	6.21×10^{-4}
。』[URL]	3.68×10^{-3}	れている	6.17×10^{-4}
している	3.64×10^{-3}	してい	5.83×10^{-4}
のベストセラー→	3.36×10^{-3}	ではない	3.39×10^{-4}
番組で紹介	3.22×10^{-3}	０万円	2.77×10^{-4}
WEB 24 です	3.21×10^{-3}	のエネルギー政策	2.26×10^{-4}
24 です。	3.21×10^{-3}	、２０	2.25×10^{-4}
ツイートには	3.18×10^{-3}	yes or no	2.16×10^{-4}
で紹介し	3.15×10^{-3}	•‘) yes or	1.97×10^{-4}
してよい	3.14×10^{-3}	ω•‘) yes	1.97×10^{-4}
よいツイートに	3.11×10^{-3}	? [URL] 拡散	1.97×10^{-4}
てよいツイート	3.11×10^{-3}	or no?	1.97×10^{-4}
編集部）	3.10×10^{-3}	no? [URL]	1.97×10^{-4}
SankeiBiz 編集部	3.10×10^{-3}	、日本の	1.93×10^{-4}
にはを	3.10×10^{-3}	研究機関の	1.56×10^{-4}

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

であり，m_kはハッシュタグやユーザ，$M_k = \{m_k^+, m_k^-\}$はm_kを含むツイートのクラスと含まないクラスである．

従来の情報利得に対し，特定のハッシュタグやユーザと共起する度合をペナルティとして引くことで，標本選択バイアスに対処可能であると考える．

前述した方法で収集したc_0^+, c_0^-を合わせた 182,828 ツイートを用い，以下の手順で特徴抽出を行った．

1. ツイートの特徴f_iを，情報利得$\mathrm{IG}(C_0|F_i)$，バイアス付き情報利得$\mathrm{BPIG}(C_0|F_i)$でそれぞれランク付けする

2. ランク上位のf_iのうち，$\mathrm{PMI}(c_0^+, f_i) = \log \frac{p(c_0^+, f_i)}{p(c_0^+)p(f_i)} > 0$である$f_i$を$c_0^+$のツイートの特徴として抽出する

3. ランク上位のf_iのうち，$\mathrm{PMI}(c_0^+, f_i) < 0$である$f_i$を$c_0^-$のツイートの特徴として抽出する

表2は，形態素trigramを特徴とし，上記の手順で特徴抽出した結果である．従来の情報利得で抽出された特徴はハッシュタグ#news, #nhk24やユーザ@selection_newsなどに特有の表現であり，広く使われているものではない．

表-3 c_0^+ の特徴として抽出された形態素 *N*-gram のモダリティ分類

モダリティなど	形態素 N-gram f_i	$\mathrm{BPIG}(C_0\|F_i)$
引用／リツイート	」[URL][E]	4.14×10^{-3}
	ニュース [URL]	1.11×10^{-3}
	RT [USER]:	7.66×10^{-4}
	：日本経済新聞	1.46×10^{-4}
	読売新聞）[URL][E]	1.00×10^{-4}
	-MSN 産経ニュース	9.19×10^{-5}
主張／提案	べき。	2.20×10^{-4}
	すべき	1.66×10^{-4}
	べきだ	1.48×10^{-4}
	するべき	6.03×10^{-5}
	したらどうだろう	3.92×10^{-5}
事実	ている。	1.21×10^{-3}
	ています	9.14×10^{-4}
	している	9.14×10^{-4}
	されている	2.26×10^{-4}
質問／疑問	•‘) yes or no?	1.97×10^{-4}
	では？	1.18×10^{-4}
	ているのか	7.49×10^{-5}
	のでしょうか	4.19×10^{-5}
内容語	日本の	5.32×10^{-3}
	億円	1.14×10^{-3}
	委員会	1.04×10^{-3}
	兆円	8.79×10^{-4}
	政策を	3.60×10^{-4}
	研究機関	2.62×10^{-4}
	のエネルギー政策	2.26×10^{-4}
	ガスシフト	1.31×10^{-4}
	沈静化	1.10×10^{-4}

しかし，バイアス罰則付き情報利得により抽出された特徴からは，特定のハッシュタグやユーザに固有の表現が適切に除外されるという結果を得た．

さらに，形態素trigramだけでなく，N = 2, 3, 4, 5についてもバイアス罰則付き情報利得で形態素N-gramを抽出したところ，引用，主張，事実，質問，社会問題を表す内容語など，コンサーンを述べる際に一般的に用いられると考えられる表現を抽出することができた（表3）．

これらの結果をさらに詳しく分析することで，誤解の余地の少ないコンサーン定義や，コンサーン・コーパスの構築，ひいてはコンサーンの自動抽出器の開発にも繋がると考える．

6.まとめ

本稿では，LOD を用いた住民参画Web プラットフォームO2 の概要を示し，コンサーン・アセスメントのために構築したLinked Open DataであるSOCIA を構築した．SOCIA オントロジーの設計にあたり，コンサーンに関連する情報の構造化手法と，アノテーションが付与された状況を管理するための構造化手法の概要を示した．また，SOCIAを用いた議論支援システムcitispe@kを試作し，評価基準タグや発言タグの付与による構造化手法を示した．さらに，Twitterからコンサーンを表すツイートの候補を取得し，特徴抽出実験とその実験結果の分析を行った．バイアス罰則付き情報利得を用いることにより，特定のハッシュタグやユーザに偏って現れる形態素Ngramを除外し，コンサーンを述べる際に使用されるモダリティ（引用，事実，主張，質問）を表す形態素Ngramを抽出できた．

今後は，議論支援システムcitispe@kを洗練化し，実際の地域社会でのコンサーン・アセスメントに活用すべく実証実験を行う予定である．さらに，1節で述べた8項目（問題，背景，選択肢，評価基準，評価極性，利害関係者，決定事項，根拠）を全てカバーするようにSOCIAオントロジーを拡張した上で，特徴抽出実験の結果を踏まえたコンサーン・コーパスの構築に臨む予定である．

謝辞：この研究の一部は，総務省 戦略的情報通信研究開発推進制度 (SCOPE) の支援を受けたものです．

参考文献

1) Jeong, H., Hatori, T., and Kobayashi, K.: Discourse Analysis of Public Debates: A Corpus-based Approach, *Proceedings of 2007 IEEE International Conference on Systems, Man and Cybernetics*, pp. 1782-1793, 2007.
2) Obama, B.: Transparency and Open Government, http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment/, 2009.
3) T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*, Morgan & Claypool, 2011.
4) Hochtl, J. and 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*, Vol. 6866, pp. 330-343, 2011.
5) Swezey, R., Sano, H., Hirata, N., Shiramatsu, S., Ozono, T., and Shintani, T.: An e-Participation Support System for Regional Communities Based on Linked Open Data, Classi_cation and Clustering, *Proceedings of the 11th IEEE International Conference on Cognitive Informatics & Cognitive Computing*, 2012 (to appear).
6) 平田紀史, 佐野博之, Swezey, R., 白松俊, 大囿忠親, 新谷虎松：住民参画Web プラットフォームO2 における関連情報構造化システム, 第26 回人工知能学会全国大会論文集, 3C2-OS-13b-11, 2012.
7) 佐野博之, 平田紀史, Swezey, R., 白松俊, 大囿忠親, 新谷虎松：住民参画Web プラットフォームO2 における関連情報を用いた議論支援システム, 第26 回人工知能学会全国大会論文集, 3C2-OS-13b-12, 2012.
8) 奥村裕一：オバマのオープンガバメントの意味するもの～今後も続く完成への長い道のり～, 三菱UFJ リサーチ&コンサルティング季刊政策・経営研究2010, Vol. 4, pp. 51-79, 2010.
9) 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
10) ISA: Joinup, https://joinup.ec.europa.eu/, 2011.
11) オープンガバメントラボ事務局: オープンガバメントラボ, http://www.openlabs.go.jp/, 2010.
12) オープンガバメントラボ事務局: 行政機関における情報分析ツール活用ガイドなどを公開, http://www.openlabs.go.jp/home/news/xingzhengjiguanniokeruqingbaofenxitsuruhuoyonggaidonadowogongkai, 2012.
13) GeoNames Project: GeoNames, http://www.geonames.org/, 2007.
14) 嘉村哲郎, 加藤文彦, 大向一輝, 武田英明, 高橋徹, 上田洋: Linked Data による芸術・文化情報統合の試み, 第24回セマンティックウェブとオントロジー研究会, No.SIG-SWO-A1101-04, 2011.
15) Benn, N. and Macintosh, A.: Argument Visualization for eParticipation: Towards a Research Agenda and Prototype Tool, *Proceedings of the Third IFIP WG 8.5 International Conference on ElectronicParticipation*, pp. 60-73, 2011.

16) Jeong, H., Shiramatsu, S., Hatori, T., and Kobayashi, K.: Discourse Analysis of Public Debates Using Corpus Linguistic Methodologies, *Journal of Computers*, Vol. 3, No. 8, pp. 58-68, 2008.
17) Braak, van den S. W., Oostendorp, van H., Prakken, H., and Vreeswijk, G. A. W.: A critical review of argument visualization tools: Do users become better reasoners?, *Workshop Notes of the ECAI-2006 Workshop on Computational Models of Natural Argument*, pp. 67-75, 2006.
18) Liddo, A. D. and Shum, S. B.: Cohere: A prototype for contested collective intelligence, *Workshop on Collective Intelligence in Organizations: Toward a Research Agenda, ACM Computer Supported Cooperative Work.*, 2010.
19) Iandoli, L., Klein, M., and 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*, Vol. 1, No. 1, pp. 69-92, 2009.
20) Kamimaeda, N., Izumi, N., and Hasida, K.: Evaluation of Participants' Contributions in Knowledge Creation Based on Semantic Authoring, *The Learning Organization*, Vol. 14, No. 3, pp. 263-280, 2007.
21) A. Maclean, RM. Young, VME. Bellotti and T. Moran: Questions, Options, and Criteria: Elements of Design Space Analysis, *Journal of Human-Computer Interaction*, Vol.6, pp.201-250 1991.
22) Conklin, J. and Begeman, M.L., gIBIS: A hypertext tool for team design deliberation, Proceedings of the ACM conference on Hypertext, pp. 247-251, 1987.
23) Mann, W. and Thompson, S: Rhetorical Structure Theory: Toward a Functional Theory of text Organization, *Text*, Vol. 8, No.3 , pp. 243-281, 1988.
24) Shiramatsu, S., Komatani, K., Ogata, T., and Okuno, H. G.: SalienceGraph: Visualizing Salience Dynamics ofWritten Discourse by Using Reference Probability and PLSA, *Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence*, pp. 890-902, Springer, 2008.

Developing a Mechanism to Support Concern Assessment Using Linked Open Data

Shun SHIRAMATSU, Hiroyuki SANO, Norifumi HIRATA, Robin M. E. Swezey, Tadachika OZONO, and Toramatsu SHINTANI

We have developed an eParticipation Web Platform based on Linked Open Data, which targets regional communities in Japan. To increase transparency and public participation, it is important to share public concerns among citizens, government officers, and experts. In this paper, we present a Linked Open Data set, SOCIA, which consists of regional news articles, microposts in Twitter, and minutes of city council. SOCIA is designed to be utilized for supporting concern assessment. It is semi-automatically structured by our text mining system on the basis of regions and events extracted from news articles on the web. We report development of functions for supporting concern assessment in our discussion support system based on SOCIA Furthermore, we conducted an experiment of feature extraction towards automatic concern extraction from Twitter.