| | |
|---|---|
| タイトル | 北海道の地方政治におけるウェブ政治情報システム (栃内香次教授退職記念号) |
| 著者 | 渋木，英潔；木村，泰知；高丸，圭一 |
| 引用 | 北海学園大学経営論集，7(3)：65-85 |
| 発行日 | 2009-12-25 |

# 北海道の地方政治における ウェブ政治情報システム

渋 木 英 潔・木 村 泰 知・高 丸 圭 一

## 1．はじめに

地方自治体の経営では，議会は地域の諸課題を解決するための計画をし，行政はそれを実行する。一方，地域住民は，そういった行政サービスを享受する立場にあり，より快適なサービスを享受するために適切な議員や首長を選ぶ権利をもっている。したがって，自治体が円滑に運営され，住民が適切なサービスを享受するためには，自治体経営における計画者である議員がそれぞれどのような課題に取り組んでいるのかを住民が知る必要がある。しかしながら，TV や新聞のように時間や紙面に限りがあるメディアで取り上げられる政治情報は，国政に関する内容が中心的であり，これに比べて地方政治に関する情報は少ない。議員活動についても同様で，地方議会議員は国会議員と同様に住民による選挙によって選ばれ，かつ，国政よりも身近な存在であるべきであるにもかかわらず，その活動に関する認知度は国会議員よりも低い。住民に提供される地方政治の情報，特に地方議会議員に関する情報量の不足を解決するための方法の一つとして，ウェブ上の情報を有効に利用することが考えられる。

ウェブ上には，新聞社等がニュースサイトで提供するニュースや企業が自社のサイトで提供するプレスリリースなど情報発信者がある程度特定できる情報と，近年増加している，ブログ，SNS，ウィキペディアなど，一般市民が容易に発信できるサイトの情報が，それぞれ大量に存在しており，ウェブは玉石混淆の膨大な情報源とみなすことができる。勿論，政治に関する情報もウェブ上に多く存在しており，上記の発信者という観点からは，住民が発信する政治に対する意見などの情報（住民側の情報）と，議員や政党が発信する情報，または，議員に関する情報（議員側の情報）に分けて考えることができる。

住民側の情報には，個人のホームページ，ブログ，SNS，掲示板，チャットなどによるものがある。この中でもブログは近年，一般市民に急速に普及した情報発信手段であり，2008 年までにその開設数は 1,690 万件を超えているといわれている[1]。一般のブロガー（ブログの執筆者）は日々の生活で感じたことをありのままに書いていることが多く，ブログ記事の中には，地方政治に対する積極的な意見は少ないにしても，日常生活における不満や要望が含まれていると考えることができる。例えば，特に政治的課題と意識することなく，日記として「今日，海水浴に行ったらゴミだらけで不愉快だった。もう行きたくない。」といった内容をブログに書くことは十分に考えられる。そこから，ゴミの収集方法や環境美化，観光客の誘致といった政治的課題を明らかにできたならば，住民と政治家の両方に有益であると考えられる。したがって，こういったブログ記事を分析することで，政治的課題として取り上げるべき住民の意見

を捉えることが重要である。

また，住民側から政治に関する情報が積極的に発信されるサイトは少ないものの，「Yahoo！みんなの政治」のような住民側の意見を投稿するサイトが幾つか存在する[2)]。このようなサイトでは，TV や新聞で取り上げられた国政の話題に関する後追いの議論が多く，地方政治に関する内容はそれほど多くない。それゆえ，地方政治を中心とした住民側の情報発信・集積サイトの重要性は高いと考えられる。

議員側の情報には，議員や政党のホームページ，ニュースサイトの政治ニュース，議員のブログ，マニフェスト，議会の会議録などがある。このうち会議録には，議員からの一方的な情報発信ではなく，議論や反対意見などのやりとりが含まれ，公の場における各議員の活動や考え方を知ることができる。国会の場合，国立国会図書館により会議録サイトが整備されており，第 1 回国会（昭和 22 年）以降のすべての会議録がテキストデータとして公開され，検索システムによって検索を行うことができる[3)]。しかしながら，地方議会会議録については，未だウェブ公開自体がなされていない自治体も多い。ウェブ公開されている会議録も自治体により公開方法が異なっており，国会会議録のように整備されているものはほとんどない。会議録は，定例会のものだけでも膨大な量となる。例えば，北海道小樽市の市議会会議録の場合，定例会 1 回分の会議録は A 4 版で 200 ページを超えている。このような大量のテキストデータを単純にウェブ公開しただけでは，能動的にアクセスしてこれを読もうと考える住民はほとんど存在しないことが予想されるため，公開方法や情報提供形態を工夫し，地方議会会議録を有効に利用することが望ましい。

以上の背景から，我々の研究プロジェクトでは，ウェブ上に存在する政治情報を利用して，メディアで取り上げられる機会の少ない地方議会議員の政治の情報を提供する方法について研究を進めている。住民は日常生活における不満や要望に政治的問題が含まれているとは捉えていない場合が多く，また，住民の関心の対象はそれ自体が多様である。住民の関心に合う情報を探すためには，まず，住民の潜在的な関心を明確化して，それぞれの住民にマッチした情報を抽出・整理して提示するシステムが必要であると考えられる。このため，本プロジェクトは，ウェブ上の情報から住民の関心にあわせた地方議会議員の情報を提示するシステムの開発を目指している。これを「住民本位型政治情報システム」と呼ぶこととする[4)]。

ウェブ上の膨大なテキストから必要な情報を抽出・整理する処理をすべて手作業で行うことは，労力的に不可能である。そこで，これらの作業には「自然言語処理」と呼ばれる電子テキストを自動処理する工学的アプローチを用いることが有効であると考えられる。自然言語処理の分野においても，近年，ウェブ情報に関する応用研究が盛んであり，例えば，ブログを利用した研究として，Inui et al.（2008）は「経験マイニング」と呼ばれる，個人の行動，成功体験，トラブル，興味，感想といった個人の経験の収集を 1 億 5 千万のブログ記事から行い，「みんなの経験」[5)] というブログ検索サービスを提供している。また，池田ら（2008）は，ホテルや旅行，催し物，電化製品など様々な商品やサービスに関する個人の意見や体験をブログから抽出することで，評判情報としてマーケティングや商品開発，企業のリスク分析，商品購入の検討などに役立てようとしている。こういったウェブからの意見抽出に関する自然言語処理の研究動向は乾ら（2006）によるサーベイ論文などに詳しい。本プロジェクトにおいても，自然言語処理技術を応用して住民本位型政治情報システムの開発を行っている。

本稿の構成は以下の通りである。2 では，

本プロジェクトの概要として，政治分野における自然言語処理の従来研究や我々が提案するシステムの概要などを述べる。3では，北海道における地方自治体の会議録公開状況に関する調査結果について述べる。4では，ウェブ上に公開されている情報に基づいて，地方政治に関するカテゴリを体系付けた後，提案システムへの利用に向けた会議録（議員側の情報）の分析，および，住民のブログ（住民側の情報）の分析について述べる。5では，これまでに我々が行ってきた研究成果を紹介する。6に結論と今後の課題を述べる。

## 2．プロジェクトの概要

### 2.1. 政治分野における自然言語処理

政治分野における問題解決の手段として自然言語処理技術を応用するという試みは，本プロジェクト以外にも存在する。例えば，選挙における政策や争点に焦点を当て，有権者の考えに近い候補者の情報を提示することで，選挙時の候補者選びを支援しようとするものとして，岩崎ら（2001）の研究が存在する。我々は，選挙時の政策に加えて，会議録等に含まれる平時の活動を併せて提示することにより，さらに有益な情報を提示できると考えている。

我々の会議録に含まれる有益な情報を提示するという目的と同様の目的をもつ研究としては，国政を対象としたものであるが，川端ら（2007）や山本ら（2005）の研究が存在する。川端ら（2007）や山本ら（2005）は，特徴的な表層表現を手がかりとして国会議事録を自動的に要約した文章を提示することで，膨大な議事録の内容を理解しやすくする研究を行っている。

我々は，既に存在する会議録を利用して有益な情報を提示しようとしているが，有益な情報を提示しやすいように会議録自体を改善しようとする試みもある。友部ら（2005）や本村ら（2005）の研究はディスカッションマイニングと呼ばれるプロジェクトの研究であり，人間同士の知識交換の場であるミーティングにおける活動を記録して，構造化された議事録データを半自動的に生成し，そこから再利用可能な知識を抽出する技術の確立を目指している。

また，我々の目的とは異なるが，音声認識を用いて会議録作成の労力を軽減しようとする研究も存在する。例えば，秋田ら（2008）は，国会会議録の作成支援に向けた音声認識システムの導入を考えており，NECでは，愛知県議会，美唄市議会において音声認識を利用した会議録作成支援システムの導入を試みている。

このように，自然言語処理技術を政治分野に応用しようとする研究は比較的少数であるが存在している。政治分野においてウェブを活用するという流れは今後さらに加速するものと考えられ，ウェブ上の情報を処理する上で自然言語処理分野との連携もさらに活発なものになると考えられる。

### 2.2. プロジェクトの対象地域

本プロジェクトは，有権者である住民の地方政治への関心を高めるために，地方議会議員の情報を地域住民に提示することを目的としている。それゆえ，国内の市町村全てを網羅することを目指しているが，研究開発の初期から網羅的に進めていくことは困難であるため，暫定的に特定の地域を対象として研究を進める必要がある。本稿では，以下の3つの理由から，本プロジェクトの必要性や重要性が高いと思われる北海道内の市町村を最初のターゲットとしている。

(1) 地方議員の多さ

総務省の「地方公共団体の議会の議員及び長の所属党派別人員調等（平成18年12月31日現在）」によると，図2-2-1に示すよう

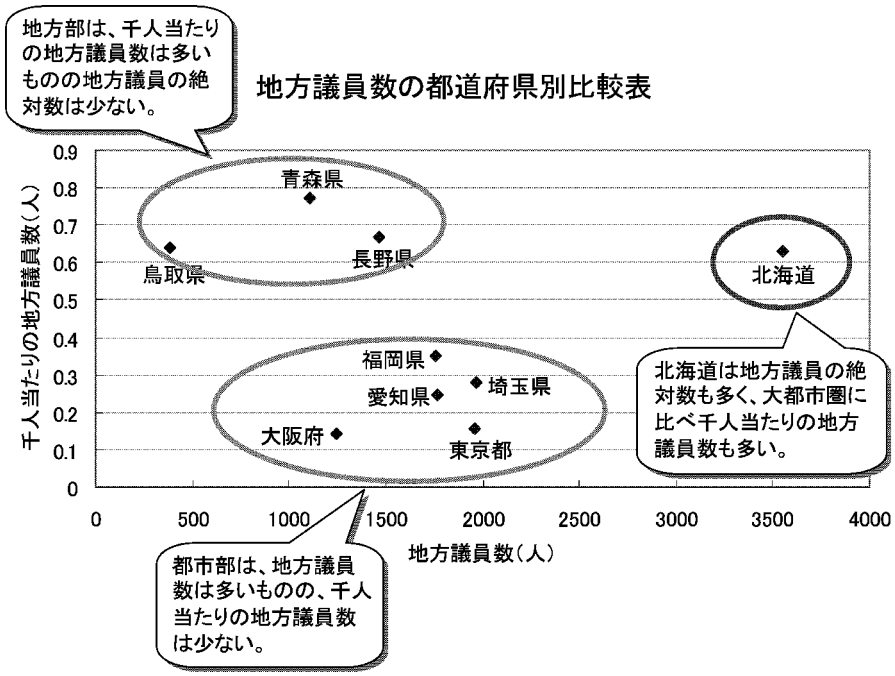


図 2-2-1　北海道における地方議員の数

に，北海道における地方議員の数は 3,549 人であり，次いで多い埼玉県の 1,965 人，東京の 1,957 人，大阪の 1,248 人などと比較して 2 倍から 3 倍もの議員が活動している。また，人口千人当たりの地方議員数での比較においても，全国平均の 0.36 人に対して，北海道は 0.63 人と 1.75 倍の値となっており，青森県や鳥取県のような人口が少ない県並みの高さとなっている。

地方議員が多い理由としては，北海道が広大な面積を有することや，冬季の除雪や知床に代表される自然保護など地域特有の課題を抱えていることなどが考えられるが，他県と比較して地方議員の活動が新聞やテレビなどで紹介されやすいということはない。したがって，ウェブを活用して政治情報を提供する本プロジェクトのモデル地域として適していると考えられる。

⑵　公共投資の多さと地方財政の悪さ

統計局による「社会生活統計指標」によると，平成 16 年度の北海道の普通建設事業費は 6,931 億円と全国で最も多く，全国平均である 1,977 億円の約 3.5 倍の公費が投入されている。公共投資が多いことには，広大な面積を有する北海道の道路やダムなど社会基盤の整備に多額の予算が使われていること，知床などに代表されるように自然環境の保護に取り組んでいること，開発局・道庁・市町村と他地域とは異なり行政機関やその出先機関なども多いこと，などの理由が考えられるが，共通して言えることは，地方議会における活動が道民の生活に大きく影響しているということであり，地方議員の活動の重要性が高いことを示している。それゆえ，地方議会における活動を地域住民に報知することが重要な課題であるにも関わらず，夕張市など財政的に逼迫している自治体が多く，広報活動に多くの予算を割けないのが現状となっている。

総務省の「市町村主要財政指標の都道府県別平均」によると，図 2-2-2 に示すように，平成 17 年度の市町村平均の財政力指数は，北海道が 0.27 と高知県の 0.26 に続き全国ワースト 2 位（島根県と同率）で，全国平均の 0.52 に比べて約半分となっている。このように，北海道の地方財政は危機的な状況にあり，公共予算の使い道や地方財政の状況など，他地域と比較して住民への政治情報の提供や意見の反映が一層求められる地域であり，本プロジェクトの重要性が高いと考えられる。

(3) インターネット利用率の高さ

総務省統計局の「平成 18 年社会生活基本調査」によると，インターネットの利用に関して，ホームページ，ブログの開設・更新を行う 1 年あたりの平均行動日数が，全国平均の 122.4 日に対して北海道は 135.0 日と高く，掲示板・チャットにおいても 127.2 日と全国平均（122.3 日）を上回っている。したがって，北海道の住民がインターネットを通して意見や要望を表明している可能性は比較的高く，また，地方議会議員に関する情報を提供する媒体としても，ウェブの利用が適していると思われることから，本プロジェクトの有効性を検証しやすいと考えられる。

また，本プロジェクトは，総務省戦略的情報通信研究開発推進制度（SCOPE）の平成 20 年度地域 ICT 振興型研究開発課題（北海道総合通信局管轄）に採択されており，同省が定めた地域の活性化などに貢献して豊かなユビキタスネット社会を築くための戦略的な重点研究開発目標を実現する目的も含まれている。

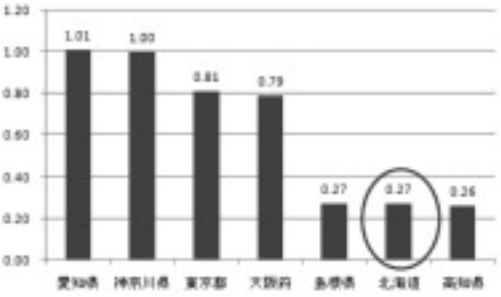


図 2-2-2 市町村主要財政指標の都道府県別平均

### 2.3. 住民本位型政治情報システムの概要

図 2-3-1 と図 2-3-2 は本プロジェクトが提案する住民本位型政治情報システムの全体構成である。図 2-3-1 はインターフェイスなどの外部設計を示しており，幾つかのモジュールの集合として定義されている。モジュール間はネットワーク的に相互連携しているため，直線的な連結と異なり，ユーザは自らの関心に応じたモジュールのみを利用することができ，利便性が高まると考えられる。

このようなシステムを実現するためには，図 2-3-2 に示す内部機構が必要であり，内部機構は大きく 4 つの要素技術に分割できる。第 1 の技術は，議事録等から議員の意見や活動に関する情報を抽出する技術であり，第 2 の技術は，ブログ等から住民の政治的意見や関心を抽出する技術である。第 3 の技術は，抽出された住民の関心と議員の活動を適切に対応付ける技術である。最後に，これらの結果を分かりやすく提示するための技術が第 4 の技術であり，これは図 2-3-1 に示す外部設計と密接に関連する。

以上を踏まえて，3 では議員活動の情報源となる議事録の公開状況を報告し，4 では議事録とブログに含まれる政治的課題の分析を行う。また，5 で紹介する我々の研究は上記の全体構成に従って行われたものである。

## 3．北海道を対象とした会議録のウェブ公開状況と収集方法

### 3.1. 会議録のウェブ公開状況に関する調査

ウェブ上に公開されている会議録等の活用という目的において，現時点でどの程度の市

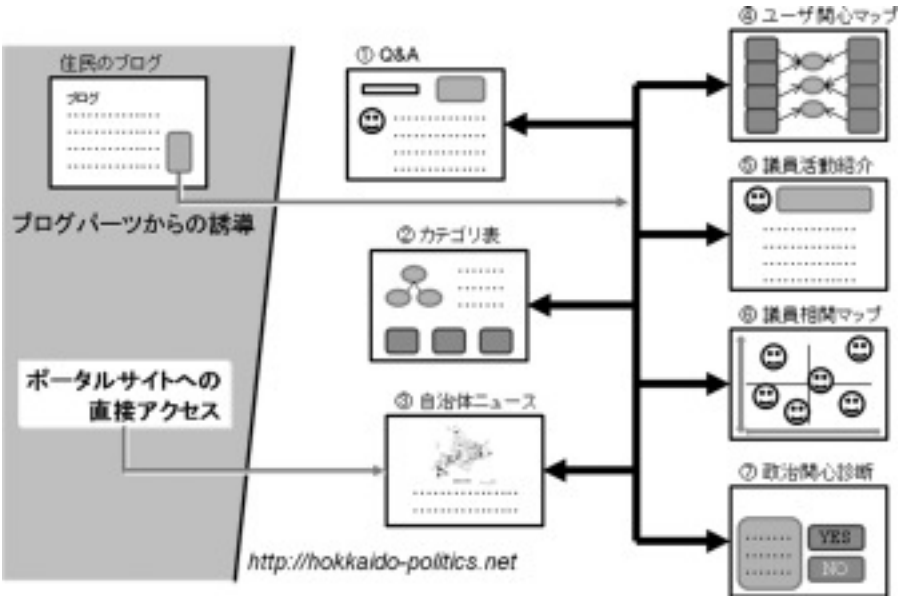


図 2-3-1　住民本位型政治情報システムの外部設計

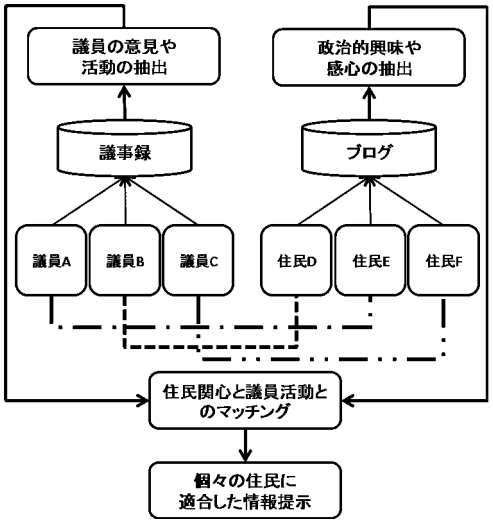


図 2-3-2　住民本位型政治情報システムの内部設計

町村がウェブ上で公開しているか，また，それらの会議録をどのように管理しているか，といった点が明確にされていることが必要であるが，そのような調査はこれまで十分に行われていなかった。そこで，本節では，各自治体に対して会議録のウェブ公開状況をアンケート調査した結果を報告する。また，住民とのマッチングへ向けて，収集した会議録をコンピュータ上で利用できる形式に変換するために，会議録がどのように記述されているかを把握することも必要であるので，議事の進行順序や一問一答といった質問形式についても併せて調べることとした。

本調査の項目は，次の 5 点である。

調査 1　議会会議録のウェブ公開状況
調査 2　会議録の管理方法（アウトソーシングの状況）
調査 3　議事の進行順序
調査 4　議事の進行ルールが明文化されているのか
調査 5　質問形式（一括質疑一括回答か，一問一答か）

アンケート調査は，北海道の 180 市町村を対象として実施した。調査票は，E メールにより送付した。E メールアドレスが公開されていない自治体には郵便で送付した。回答は

表 3-1-1　アンケート回収率

| 区分 | 回答 | 未回答 | 合計 |
|---|---|---|---|
| 市 | 25（71.4%） | 10（28.6%） | 35（100.0%） |
| 町村 | 90（62.1%） | 55（37.9%） | 145（100.0%） |
| 合計 | 115（63.9%） | 65（36.1%） | 180（100.0%） |

表 3-1-2　回収結果に基づいた公開状況

| 区分 | ①掲載している | ②掲載していない | 合計 |
|---|---|---|---|
| 市 | 20（80.0%） | 5（20.0%） | 25（100.0%） |
| 町村 | 21（23.3%） | 69（76.7%） | 90（100.0%） |
| 合計 | 41（36.0%） | 74（64.9%） | 114（100.0%） |

表 3-1-3　会議録の管理方法

| 区分 | ①業者委託 | ②職員管理 | 合計 |
|---|---|---|---|
| 市 | 12（63.2%） | 7（36.8%） | 19（100.0%） |
| 町村 | 0（0.0%） | 20（100.0%） | 20（100.0%） |
| 合計 | 12（30.8%） | 27（69.2%） | 39（100.0%） |

Eメール，FAX，または郵便で受け付けることとした。Eメールによる送付は2008年8月11日に行い。回答の締切日は2008年8月29日とした。

まず，アンケートの回収結果について述べる。アンケート回収率を表3-1-1に表す。180市町村のうち，115の自治体より回答を得ることができた。全体の回収率が63.9%であり，市に関しては7割を超え，高い回収率となった。

まず，**調査1「議会会議録のウェブ公開状況」**の結果を表3-1-2に示す。8割の市が会議録をウェブ公開している一方，町村の8割近くが未公開であり，小さな自治体ほどウェブ公開が進んでいないことが確認できる。

次に，**調査2「会議録の管理方法」**についての結果を述べる。表3-1-3に示すように，市では議事録のウェブ公開に関するシステムを業者委託の形態で管理している自治体が多いが，業者委託を行っている町村は存在せず，職員が管理していることが明らかになった。

表3-1-4は，**調査3「議事の進行順序」**についての結果である。市では「①本会議⇨委員会⇨採決」の議事順序が多く，町村では①②ほぼ同数であった。

表 3-1-4　議事順序

| 区分 | ①本会議⇨委員会⇨採決 | ②委員会⇨本会議⇨採決 | ③その他（本会議⇨採決など） | 合計 |
|---|---|---|---|---|
| 市 | 20（80.0%） | 0（0.0%） | 5（20.0%） | 25（100.0%） |
| 町村 | 36（41.4%） | 38（43.7%） | 13（14.9%） | 87（100.0%） |
| 合計 | 56（50.0%） | 38（33.9%） | 18（16.1%） | 112（100.0%） |

表 3-1-5　進行ルール明文化の有無

| 区分 | ①明文化されている | ②明文化されていない | 合計 |
|---|---|---|---|
| 市 | 13（52.0%） | 12（48.0%） | 25（100.0%） |
| 町村 | 41（47.1%） | 46（52.9%） | 87（100.0%） |
| 合計 | 54（48.2%） | 58（51.8%） | 112（100.0%） |

表 3-1-6　質疑応答の形式

| 区分 | ①一括質疑一括回答 | ②一問一答 | ③その他（選択制など） | 合計 |
|---|---|---|---|---|
| 市 | 15（62.5%） | 5（20.8%） | 4（16.7%） | 24（100.0%） |
| 町村 | 26（29.2%） | 47（52.8%） | 16（18.0%） | 89（100.0%） |
| 合計 | 41（36.3%） | 52（46.0%） | 20（17.7%） | 113（100.0%） |

表3-1-5は，**調査4「議事の進行ルールが明文化されているのか」**についての結果である。明文化されている自治体とされていない自治体の数はほぼ同数であった。なお，明文化された議事進行ルールの入手には情報公開請求が必要な場合があるため，進行ルールの詳細については調査を行っていない。

表3-1-6は，**調査5「質問形式（一括質疑一括回答か，一問一答か）」**の結果である。市では一括質疑一括回答の形式が多く，町村では一問一答の形式が過半数を占めた。

これらの調査結果に加えて，無回答の自治体における会議録のウェブ公開情報を独自に調査した。その結果，北海道内では63の自治体が会議録を公開していることが確認された。図3-1-1は北海道を対象とした会議録の

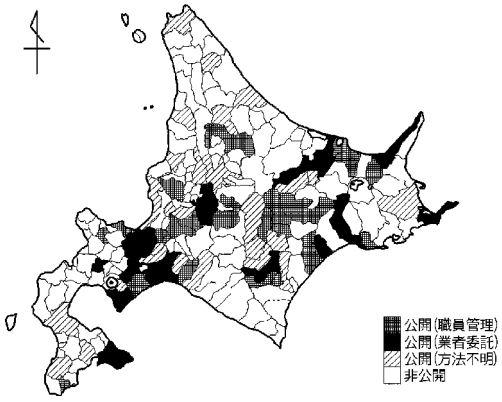


**図3-1-1　北海道を対象とした会議録のウェブ公開状況**

ウェブ公開状況であり，公開（業者委託），公開（職員管理），公開（方法不明），非公開の4つに分類している。この結果から，公開している地域に偏りがあることがわかる。そして，市の周辺は比較的公開している町村があり，周辺の自治体への影響が少なからずあるように見える。

### 3.2．地方議会会議録の自動収集

前節の調査により，北海道内の全180市町村のうち63市町村がウェブ上に会議録を公開していることが明らかになった。年々新しい会議録が追加されていくことを考えると，ウェブ上から自動的に会議録を収集することがコストの面から望ましい。本節では，公開されている地方議会会議録の自動収集方法および収集項目について説明する。

まず，会議録のウェブ公開方法の違いを以下の項目に分類する。

1．検索機能の有無
　議会会議録専用の検索インターフェイスが提供されているかどうかを表している。
2．階層構造の有無
　一つのインデックスページにすべての会議録へのリンクがあるか，または年度や開催日，発言者毎にリンクが階層構造をとっているかどうかを示している。
3．公開データ形式
　議会会議録がどのようなデータ形式で公開されているかを表している。主にPDF，静的HTML，サーバサイドで生成される動的HTMLがある。
4．公開単位
　議会会議録が一度にどの単位で閲覧できるかを意味する。例えば，議案毎にPDFファイルが用意されている場合，公開単位は議案単位となる。

表3-2-1からわかるように，市議会会議録の公開方法は統一された規則といったものは存在せず，各自治体の意向や技量に任されているのが現状である。これらの結果を踏まえ，自動収集プログラムを作成して収集を試みた結果，63市町村の議会会議録を収集するために，51種類の収集パターンが必要となった。会議録をウェブ上で公開している約94%にあたる59の市町村の議会会議録を自動収集することができた。自動収集プログラムを作る前には，数パターン程度で収集可能と考えていたが，階層構造，表示方法の違いによりパターンが増えた。他にパターンが増えた理由として，収集項目がある。我々は収集するだけではく，都市名，開催年度，委員会名を抽出する必要があるため，6%にあたる4つの自治体からは，Web上で議会会議録が公開されているのにもかかわらず，自動収集することができなかった。その原因としては，何らかのアクセス制限によるもの，インデックスページのリンクから会議録データを追うことができなかった等といったものが挙げられる。

**表 3-2-1　ウェブ公開方法の分類**

発言者単位とは，一度に閲覧できる範囲が一発言者

| | 検索機能 | 階層構造 | データ形式 | 公開単位 | 割合 | 自治体例 |
|---|---|---|---|---|---|---|
| 1 | 有 | 有 | 動的 HTML | ページ・発言者単位 | 23.8% | 帯広市，釧路市など |
| 2 | 無 | 無 | PDF | 日単位 | 12.7% | 石狩市，三笠市など |
| 3 | 無 | 無 | PDF | 議会単位 | 12.7% | 上士幌町，幕別町など |
| 4 | 無 | 有 | PDF | 日単位 | 11.1% | 美唄市，釧路町など |
| 5 | 無 | 有 | PDF | 議会単位 | 7.9% | 小樽市，網走市など |
| 6 | 無 | 有 | 静的 HTML | 議題単位 | 4.8% | 士別市，清水町など |
| 7 | 有 | 有 | 動的 HTML | 発言単位 | 3.2% | 札幌市，音更町 |
| 8 | 無 | 無 | 動的 HTML | 議会単位 | 3.2% | 月形町，清里町 |
| 9 | 有 | 有 | 動的 HTML | 日単位 | 1.6% | 岩見沢市 |
| 10 | 有 | 有 | 静的 HTML | ページ単位 | 1.6% | 江別市 |
| 11 | 有 | 無 | 静的 HTML | 日単位 | 1.6% | 深川市 |
| 12 | 有 | 有 | 静的 HTML | 議題単位 | 1.6% | 八雲町 |
| 13 | 有 | 有 | PDF | 日単位 | 1.6% | 恵庭市 |
| 14 | 無 | 有 | 動的 HTML | 日単位 | 1.6% | 富良野市 |
| 15 | 無 | 有 | 動的 HTML | 議会単位・一般質問のみ | 1.6% | 雄武町 |
| 16 | 無 | 無 | 静的 HTML | 日単位 | 1.6% | 羽幌町 |
| 17 | 無 | 無 | 静的 HTML | 議会単位 | 1.6% | 北斗市 |
| 18 | 無 | 無 | 静的 HTML | 議題単位 | 1.6% | 厚真町 |
| 19 | 無 | 有 | 静的 HTML | 議会単位 | 1.6% | 置戸町 |
| 20 | 無 | 有 | 静的 HTML | 議会単位・一般質問のみ | 1.6% | 上ノ国町 |
| 21 | 無 | 無 | 動的 HTML | 議会単位・行政報告のみ | 1.6% | 壮瞥町 |

## 4．会議録とブログを対象とした政治的意見に対する注釈

本章では，収集された会議録やブログに含まれている政治的な課題や意見・関心にはどのようなものがあるか，また，それらの課題や意見・関心は議員側と住民側に共通のものであるのか，といった点を中心に調査する。最初に両者の共通基盤となる政治的カテゴリ体系の構築を述べた後，会議録とブログに含まれる政治的カテゴリの出現頻度を比較することで両者の特徴を分析する。

### 4.1. 政治的カテゴリ体系の構築

政治的カテゴリは，地方議員と住民を結びつけるためのものであり，地方議員の活動を抽象化するために利用する。既存の政治に関するカテゴリの多くは国政に関する内容が多く，地方議会の内容として相応しくないカテゴリが含まれるため，我々は小樽市，帯広市，函館市，釧路市の 4 市を対象とした予備調査を行った。その結果，議題を区分するために存在する委員会体系が 4 市に共通していることが確認された。そこで，我々は，委員会体系が地方政治における基本となる概念体系であると仮定し，最も細目化されている帯広市の市議会における常任委員会とその所管事項の名称をもとに基本となる概念体系を作成することにした。帯広市の委員会とその所管事項の例を図 4-1 に示す。

これらの委員会名と所管事項から，概念体系を作成する手順を以下に示す。

委員会名：総務文教委員会

●調査事項
●重要政策の企画及び総合計画に関する事項
●財務に関する事項
●広報及び広聴に関する事項
●総合的な行政の推進に関する事項
●職員に関する事項
●財産に関する事項
●…

**図 4-1　帯広市の委員会とその所管事項の例**

1．名称末尾の「委員会」を削除
2．名称末尾の「に関する事項」を削除
3．名称末尾の「に属する事項」を削除
4．それぞれの上位概念に属する概念として「その他」を追加
5．上位概念に並列な概念として「その他」を追加

上記の手順により，政治的カテゴリを作成し，平成 17 年度の小樽市の市議会会議録の第 1 回定例会および第 2 回定例会に対して，含まれていない政治的カテゴリが存在するか調査を行い，政治的カテゴリの修正を行った。その結果，5 つの上位概念からなる 96 の政治的カテゴリを作成した。表 4-1-1 は，作成した政治的カテゴリの例である。

**表 4-1-1　政治的カテゴリの例**

| カテゴリ番号 | 大カテゴリ | 中カテゴリ | 小カテゴリ |
|---|---|---|---|
| 1000 | 総務文教 | | |
| 1010 | | 財務 | |
| 1011 | | | 地方税 |
| 1012 | | | 予算 |
| 1013 | | | 地方債 |
| 1020 | | 総合的な行政の推進 | |
| 1021 | | | 条例 |
| 1022 | | | 高齢化対策 |
| 1023 | | | 少子化対策 |
| 1024 | | | 男女共同参画 |
| 1025 | | | 改革 |
| 1030 | | 職員 | |
| 1040 | | 財産 | |

## 4.2. 会議録における政治的カテゴリの出現頻度

本節の目的は，4.1 で作成した政治的カテゴリが，会議録に含まれる議員の発言にどの程度含まれているかを明らかにし，政治的カテゴリの観点から会議録や議員の特徴を分析することである。

ここで，政治的カテゴリが含まれる発言の単位について説明する。本研究の目的からは，議員の活動または意見単位で政治的カテゴリを判断することが望ましいが，会議録において，活動や意見の単位で記述されているとは限らない。しかしながら，一般に，議題には議員の活動や意見が反映されていると考えられ，会議録ではある一つの議題に関する発言が一段落にまとめられる傾向にある。そこで，本分析では，段落単位で政治的カテゴリを判断することとした。

分析の対象は平成 19 年度の小樽市市議会の会議録とした。小樽市は 1 年に定例会が 4 回あるため，第 1 回から第 4 回の定例会を対象となる。定例会の段落数は約 1,700 段落であり，各定例会に対して大学生 2 名による分析を行った。これは，分析作業が主観的判断になりやすいため，2 名で同一の会議録を分析することで，できるだけ客観的な分析結果となるようにするためである。したがって，4 回の定例会に対して 8 人で分析を行った。

ここでは，平成 19 年度の小樽市の市議会会議録第 1 回から第 4 回までを対象としているが，平成 19 年 4 月に市議会議員選挙が行われたため，第 1 回定例会（3 月）の議員と第 2 ～ 4 回の議員が異なる。そこで，2 期に渡って市議会議員の職に就いている議員を対象とすることとした。また，市長，議員以外の発言者，発言していない議員については対象外とした。その結果，対象議員は 17 名と

なった。

表 4-2-1 は，会議録の分析結果である。項目のカテゴリ名と ID は 4.1 のカテゴリ体系に基づくものであり，段落数は分析者がその政治的カテゴリであると判断した段落の数である。また，割合は段落数が会議録中の全段落に対して占める割合であり，表 4-2-1 は割合が大きい上位 12 位までを表示している。ただし，ここでの段落数は，2 名の分析者がそれぞれ判断した段落数の和集合としている。

**表 4-2-1　会議録に含まれる政治的カテゴリの割合**

| 順位 | 政治的カテゴリ名 | ID | 段落数 | 割合 |
|---|---|---|---|---|
| 1 | 財務 | 1010 | 615 | 11.05% |
| 2 | 病院事業 | 1101 | 273 | 4.91% |
| 3 | 教育 | 1120 | 235 | 4.22% |
| 4 | 学校 | 1121 | 209 | 3.76% |
| 5 | 医療 | 1100 | 204 | 3.67% |
| 6 | 総合的な行政の推進 | 1020 | 201 | 3.61% |
| 7 | 施設 | 1160 | 186 | 3.34% |
| 8 | 予算 | 1012 | 153 | 2.75% |
| 9 | 職員 | 1030 | 146 | 2.62% |
| 10 | 住民活動 | 1061 | 144 | 2.59% |
| 11 | 観光 | 3040 | 132 | 2.37% |
| 12 | 建築 | 4000 | 126 | 2.26% |

表 4-2-1 の内容をみると，「財務」に関する内容が第 1 位となっており，全体の 11.05%を占める結果となった。「財務」に関する内容は地域に関係なく議論されていると考えられるが，2 位の「病院事業」に関しては，地域特有の議題と考えられる。なぜなら，小樽市立病院に関する議題が多く，病院よりも，小樽市立病院に限定した内容となっていたためである。

次に，市議会議員の発言について考察する。まず，各議員の発言数の違いを調べる。表 4-2-2 は，政治的カテゴリと議員のクロス表であり，全体を通して発言の多い上位 12 議員を示す。最も発言数が多かった議員 A は 829 回と最下位の議員 L の 143 回と比較して

**表 4-2-2　政治的カテゴリと議員のクロス表**

| 発言数順位 | カテゴリ名 | 各議員議員の発言数（左から総発言数の多い議員を順番に並べている。） | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A | B | C | D | E | F | G | H | I | J | K | L |
| 1 | 財務 | 94 | 122 | 44 | 73 | 40 | 43 | 30 | 17 | 19 | 20 | 32 | 10 |
| 2 | 病院事業 | 26 | 63 | 25 | 59 | 19 | 21 | 3 | 13 | 2 | 8 | 7 | 1 |
| 3 | 教育 | 39 | 2 | 22 | 19 | 18 | 22 | 4 | | 23 | 9 | 4 | 1 |
| 4 | 学校 | 53 | 14 | 20 | 19 | 12 | 19 | 5 | | 19 | 5 | 4 | |
| 5 | 医療 | 24 | 43 | 2 | 38 | 21 | 17 | 2 | 13 | 7 | 9 | 5 | |
| 6 | 総合的な行政の推進 | 23 | 20 | 24 | 20 | 19 | 24 | 22 | 9 | 7 | 2 | | 3 |
| 7 | 施設 | 25 | 27 | 22 | 6 | 24 | 26 | 10 | 7 | 3 | 6 | 7 | 17 |
| 8 | 予算 | 22 | 33 | 20 | 12 | 5 | 9 | 12 | 10 | 12 | 2 | | |
| 9 | 職員 | 10 | 18 | 17 | 14 | 8 | 18 | 20 | 2 | 10 | 6 | | |
| 10 | 住民活動 | 23 | 15 | 13 | 11 | 20 | 2 | 12 | 5 | 12 | 9 | 2 | 2 |
| 11 | 観光 | 14 | 1 | 10 | 8 | 1 | 15 | 19 | | 3 | 1 | 6 | 31 |
| 12 | 建築 | 21 | 24 | 16 | 14 | 8 | 16 | 3 | 6 | | 2 | 1 | 4 |
| ～途中省略～ | | | | | | | | | | | | | |
| | 合計 | 829 | 803 | 605 | 541 | 527 | 464 | 326 | 213 | 182 | 175 | 159 | 143 |

5 倍以上の差があった。勿論，所属する会派の影響などもあるため，発言数をもって単純に優劣を論じることはできない。しかしながら，カテゴリ単位で見た場合，議員ごとに発言内容の偏りが見られる。例えば，議員Ａは「教育」や「学校」,「住民活動」についての発言が比較的多く，他の議員よりも力を入れていると考えられる。また，発言数が比較的少ない議員Ｇの「職員」や議員Ｌの「観光」のように，カテゴリ単位で見た場合に多く発言していることを考慮することで，議員が重要視する政治的カテゴリを特徴づけることができると考えられる。このような特徴を活用することで,「観光客が来なくて困っている」住民に対しては，議員Ａよりも議員Ｌ，Ｇ，Ｆに関する意見や活動情報を提示するといったことが可能になると考えられる。

### 4.3. 住民ブログにおける政治的カテゴリの出現頻度

近年，CMS（Contents Management System）が発達したことにより，CGM（Consumer Generated Media）と呼ばれる一般市民（消費者）側からの情報提供が注目されている。そして，住民の政治的意見が内在するメディアとしては，ブログ，SNS，掲示板，チャット，動画配信，個人の HP などが挙げられる。これらの中でも，ブログは最も普及しているメディアとなっており，テキストを対象に個人の意見および関心を抽出する観点から，本研究ではブログを対象とした分析を行うこととした。

本研究では，前節で述べた会議録の分析対象と合わせるために，小樽住民のブログを対象にする。ブログ収集に関する予備調査の結果，ブログの内容から地域を特定することは単純にキーワードを利用する程度では，収集精度が低いことが確認されたため，人手によりブログの内容を判断し，URL を特定した後に，自動収集をすることとした。その結果，小樽に関するブログを 40 件見つけることができ，それらの各ブログに含まれる記事を全て自動で収集した。ブログ間のバランスを保つために，各ブログから最大 100 記事を抽出し，合計で 2,581 件を分析対象とした。

次に，分析方法について説明する。分析方法は会議録と同様の方法で進める。ここでは，会議録の分析方法と異なる個所だけ説明する。分析単位は，会議録のように段落単位ではなく，ブログの記事（１日の投稿内容）とした。また，ブログ記事の内容は政治的課題と関連が薄い傾向にあることが予備調査から確認されていたため，ごく僅かでも関係性があると思われた政治的カテゴリを判断するように指示した。分析者は会議録と同じく大学生 4 名で，２名ずつ同一のブログ記事を判断するようにしたため，１名の分析者は約 1,300 記事を担当することになった。

上記の方法で分析した結果，ブログに含まれる政治に関する政治的カテゴリの割合を図 4-3-1 および表 4-3-1 に示す。この分析時間については，１名の分析者は約 1,300 記事を 50 時間程度で終了し，会議録よりは短い時間で終了することができた。

図 4-3-1，表 4-3-1 の結果に基づいて，ブログ記事で判断された政治的カテゴリの割合の上位 20 件と地方議会会議録で判断された

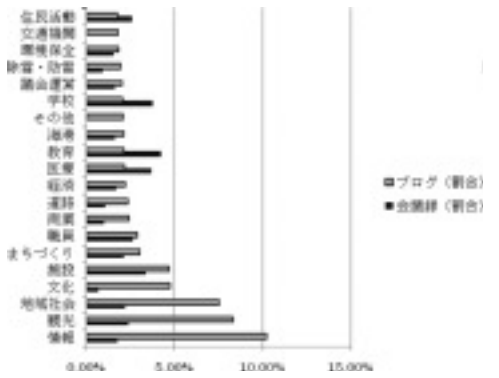


**図 4-3-1　会議録とブログ記事で判断された政治的カテゴリの比較**

表 4-3-1　ブログに含まれる政治的カテゴリの割合

| | カテゴリ名 | ID | ブログ（個数） | ブログ（割合） |
|---|---|---|---|---|
| 1 | 情報 | 1050 | 917 | 10.21% |
| 2 | 観光 | 3040 | 746 | 8.30% |
| 3 | 地域社会 | 1060 | 676 | 7.52% |
| 4 | 文化 | 1161 | 425 | 4.73% |
| 5 | 施設 | 1160 | 423 | 4.71% |
| 6 | まちづくり | 1062 | 271 | 3.02% |
| 7 | 職員 | 1030 | 257 | 2.86% |
| 8 | 商業 | 3011 | 216 | 2.40% |
| 9 | 道路 | 4020 | 209 | 2.33% |
| 10 | 経済 | 3020 | 196 | 2.18% |

政治的カテゴリの割合を比較している。この結果から，ブログ記事と会議録において，判断される政治的カテゴリの頻度分布は異なっていた。しかしながら，ブログ記事と会議録の両方とも比較的幅広い政治的カテゴリが含まれており，量的な問題はあるが相互に関連付けられる内容が存在していた。このような結果から，住民の興味と政治家の活動を結びつける可能性を示すことができたと考えられる。ただし，「交通機関」のようにブログ側にしか存在しない政治的カテゴリも存在しており，このようなギャップを埋める必要があると考えられる。

## 5．自然言語処理による取り組み

本章では，これまでに我々が行ってきた自然言語処理による幾つかの取り組みを紹介する。紙面の都合により，各研究の概要を述べるに留め詳細は各参考文献に譲る。

### 5.1. 定型表現による会議録からの議員活動抽出

本節では，渋木ら（2007）で行った最初期の研究を紹介する。この研究では，北海道小樽市の市議会を対象とし，市議会会議録を入力して，そこに含まれる議員名とその議員が取り組んでいる活動のリストを出力することを目的とした。ただし，リストはその議員の重点を置いていると思われる順にランク付けし，類似する活動は同じクラスタにまとめて出力するものとした。

我々は，この課題を解決するにあたり，会議録の特徴的な言い回しを最大限利用することとした。小樽市市議会の会議録は図 5-1-1 に示すような形式で記述されており，ある議員の質問に対してまとめて回答するという形式をとっている。そこで，まず，議員が関心をもって取り組んでいる活動内容を表す語句を重要フレーズと定義し，会議録の中である問題に対してある議員が質問を行っているならば，その議員はその問題に関心があるという仮定に基づいて重要フレーズの抽出を行った。このように定義することで，会議録から質問した議員の名前と質問内容を抽出するというタスクに捉えなおすことができる。

また，議会における発言形式は比較的定まっており，図 5-1-1 に示すように，質疑の回答は，最初に「○○議員の御質問にお答えいたします」という形式で質問者の名前を述べた後，「最初に，△△についてですが」と質問内容に言及することが多い。したがって，このような定型表現を質問者と質問内容を抽出するためのテンプレートとして利用した。質問者の抽出テンプレートには，「○○議員の御質問にお答え」の 1 パターンを，関心がある問題の抽出テンプレートとしては，「まず，△△について」，「初めに，△△について」，「最初は，△△について」，「また，△△について」，「また，△△につきまして」，「次に，△△について」，「最後に，△△について」の 7 パターンを用いた。ただし，このようなテンプレートにより抽出されたフレーズの中には，質問内容として相応しくないフレーズがあるため，それらのフレーズをあらかじめ登録しておいた不要フレーズと比較し

○市長（山田勝麿）秋元議員の御質問にお答えいたします。
初めに，小樽市の防災に関連しての御質問でありますけれども，まず要援護者支援班の設置につきましては，小樽市では平成 19 年 9 月から，これまで総務部防災担当，福祉部，生活環境部，消防本部，小樽市社会福祉協議会の担当者によるプロジェクトチームで議論を進めてまいりました。現在，このチームを支援班と位置づけまして，この中で災害時要援護者避難支援プランの作成に向け，情報の共有，連携などについて協議しているところであります。
次に，災害時要援護者名簿の作成の進ちょく状況でありますけれども，総務部防災担当が主体となり，民生・児童委員の方々の協力をいただきまして，基礎調査を実施して，災害時要援護者のデータの集約をしているところであります。次の段階として，要援護者個々の避難支援プランを作成することとなりますが，実際の災害時における要援護者の避難をサポートする支援者の方々を選定することが最大の課題でありまして，今後，多大な労力を要するものと考えております。このため，町会や民生委員などの協力を最大限にいただきまして，できる限り早い時期に避難支援プランをまとめた台帳を作成してまいりたいと考えております。
次に，行政外の関係機関などとの情報の共有も含めた連携の問題でありますけれども，先ほどもお答えしましたとおり，昨年実施しました基礎調査や今年の個々の避難支援プラン作成の段階においても，民生委員などの協力を得ながら作業を進めております。また，今後個々の避難支援プラン登録台帳が完成した段階では，町会，自主防災組織や福祉関係機関などと情報の共有をし，災害時に対応すべく連携を進めていかなければならないものと考えております。

**図 5-1-1　小樽市議会会議録の例**

てフィルタリングを行った。この不要フレーズは，「ただいま決定いたしました以外の各案件」の 1 フレーズとした。

フィルタリングされたフレーズのランク付けは，フレーズの重要度をフレーズ中の名詞の重要度の平均で近似することで行った。名詞の重要度には，TF-IDF 値と呼ばれる，文章中での名詞の出現頻度（TF，Term Frequency）とその名詞が出現する文書の逆出現頻度（IDF，Inverse Document Frequency）の積を用いており，特定の文書にしか頻出しない名詞の重要度を上げることを意味している。

ランク付けされたフレーズに対して，ランク上位のフレーズから順に，類似した下位フレーズが存在するか判断し，類似度の高い下位フレーズを上位フレーズと同じクラスタにまとめることを行った。フレーズ間の類似度は，フレーズ中に含まれる全ての名詞を要素としたベクトル空間の中に各フレーズを象徴するベクトルを配置し，2 つのベクトルが成す角度の余弦を類似度とすることで計算した。この類似度が閾値 0.8 を超えたフレーズを同一のクラスタと判断している。

実験は，平成 12 年から 18 年までの小樽市議会の会議録を用いて行った。入力された会議録の文字数は 7,821,573 文字であり，会議録から抽出された議員数は 35 人であった。表 5-1-1 は出力結果の例であり，2 人の議員を対象に上位 5 位までの重要フレーズをまとめたリストである。最初のフレーズがその順位で抽出された重要フレーズであり，括弧内は同一クラスタと判断された類似フレーズを表している。

表 5-1-1 で例示されるように，関心が高い問題を表すために特徴的なフレーズがリストされており，全体的に良好な結果が得られている。しかしながら，議員Ａの「このたび示された三位一体の改革」において「このたび示された」の部分は不要であると考えられるため，抽出されたフレーズから不要部分を除去するための処理が必要である。また，議員Ｂの 2 位と 4 位のフレーズ「懲戒処分」と「分限処分」は同一のクラスタにまとめるなど，クラスタリング処理に関しても検討が必要であると考えられる。これらが今後の課題として残されている。

表 5-1-1　出力された議員活動の例

| 順位 | 議員A［発言 50 回］ | 議員B［発言 35 回］ |
|---|---|---|
| 1 | 海洋開発（海洋エネルギーの利用，海洋開発の推進） | ホームレス（ホームレス対策） |
| 2 | 乳がん，子宮けいがん検診（乳がん，子宮がん検診） | 懲戒処分（懲戒処分と分限処分） |
| 3 | 高齢者の就労機会 | 除雪（除雪費補助） |
| 4 | 福祉医療助成（医療助成制度の見直し，老人医療・福祉医療助成制度） | 分限処分 |
| 5 | このたび示された三位一体の改革 | 18 年度一般会計予算 |

## 5.2. 議員活動とブログ記事との対応付け

本節では，木村ら（2008）で行った研究を紹介する。この研究は，小樽市市議会議員へのアンケート調査により得られた議員活動を特徴づけるフレーズと，小樽市の住民が書いたブログ記事中の政治的関連性の高い記述との対応付けを目的としたものである。

この対応付けは，利用者にとっての理解の容易さという点から，出力を「記事中の『○○○○』という記述は，△△議員の『××××』という活動に近い」という，ブログ記述と議員活動が一対一で対応するものとし，複数のブログ記述や議員活動が混在するような出力はしないこととした。また，対応付けという目的から，ブログ記述と議員活動の両方が抽出されることが前提であり，どちらか一方からしか抽出できない情報は用をなさない。そこで，議員活動を特徴付けるフレーズを手がかりとして，関連度の高いブログ中の記述を抽出することで，抽出と対応付けを同時に行い処理の効率化を行った。

議員活動とブログ記述の関連度は，5.1 で述べた類似度と同様に，議員活動のフレーズに含まれる全ての名詞を要素とするベクトル空間を生成し，議員活動とブログ記述を象徴するベクトル間の余弦を計算することで求めた。このとき，議員活動のフレーズには一般に硬い表現が用いられるのに対し，ブログは口語に近い表現が用いられる傾向があるため，概念的に近いベクトル間であってもベクトル要素となる名詞の表現が一致するとは限らないという問題が生じる。それゆえ，分類語彙表を用いて名詞の拡張を行った。

分類語彙表は表 5-2-1 に示すように，15 項目から構成されており，ある単語の概念は 5 桁の分類番号の下，段落番号 2 桁，小段落番号 2 桁，語番号 2 桁の合計 11 桁の数字で階層的に表現されており，概念的に近い単語がグループ化されている。これを利用して，小段落番号までの上位 9 桁，または，段落番号までの上位 7 桁が一致する名詞を，表現が一致しなくとも同じ概念の単語と判断した。ただし，表層が一致しない場合には関連度を計算する際に 0.5 の重みを乗じることで表現が一致する名詞の方を重視することとした。

上記の手法を実装し，分類語彙表を用いた名詞の拡張による対応付けの件数への影響と，対応付けられた議員活動とブログ記述との妥当性を調査するための実験を行った。実験で用いた議員活動のフレーズは 150 個であり，小樽市議会議員 15 名を対象に自分の行っている政治活動を 10 個ずつ回答してもらった結果である。また，ブログ記事は，2 年間で「小樽」という単語を含んだ記事を 3 回以上発信したことがあるドメインから収集した 6 万件である。

表 5-2-2 に，分類語彙表を用いて拡張した場合の拡張語数と対応付けの件数を示す。拡

表 5-2-1　分類語彙表の例

| 整理番号 | 項目 | 値 |
|---|---|---|
| 1 | レコード ID 番号 | 30548 |
| 2 | 見出し番号 | 29140 |
| 3 | レコード種別 | A |
| 4 | 類 | 体 |
| 5 | 部門 | 活動 |
| 6 | 中項目 | 言語 |
| 7 | 分類項目 | 言語 |
| 8 | 分類番号 | 1.3101 |
| 9 | 段落番号 | 03 |
| 10 | 小段落番号 | 01 |
| 11 | 語番号 | 01 |
| 12 | 見出し | 国語 |
| 13 | 見出し本体 | 国語 |
| 14 | 読み | こくご |
| 15 | 逆読み | ごくこ |

表 5-2-2　素性拡張によるマッチング件数の変化

| | 素性単語数 | 拡張語数 | マッチング件数 |
|---|---|---|---|
| 上位 7 桁 | 295 | 5,745 | 1,177 |
| 上位 9 桁 | 295 | 1,162 | 535 |
| 拡張なし | 295 | 0 | 369 |

張する前の名詞数は 295 語であり，小段落番号までの上位 9 桁に拡張した場合では 1,162 語に，段落番号までの上位 7 桁では 5,745 語に拡張され，それに応じて対応付けの件数も増加している。しかしながら，対応付け件数の増加数は拡張語数の増加数と比べて緩やかなものとなっており，爆発的に増える状態にはなっていない。その意味では，まだ改善の余地があると考えられる。

対応付けの評価は，著者 2 名による「正解」,「準正解」,「不正解」,「評価不能」の 4 段階評価で行った。まず，ブログ記述が短すぎるなどの理由により，正誤の判断が困難なものを「評価不能」とした。残りの記述の中で，議員活動のフレーズと結びつける解釈が

表 5-2-3　上位 15 件の正解数

| | 正解 | 準正解 | 不正解 | 評価不能 | 合計 |
|---|---|---|---|---|---|
| 上位 7 桁 | 2 | 9 | 3 | 1 | 15 |
| 上位 9 桁 | 10 | 0 | 1 | 4 | 15 |
| 拡張なし | 9 | 0 | 1 | 5 | 15 |

表 5-2-4　上位 30 件の正解数

| | 正解 | 準正解 | 不正解 | 評価不能 | 合計 |
|---|---|---|---|---|---|
| 上位 7 桁 | 6 | 13 | 6 | 5 | 30 |
| 上位 9 桁 | 21 | 0 | 1 | 8 | 30 |
| 拡張なし | 21 | 0 | 2 | 7 | 30 |

非常に困難な記述を「不正解」，単純な仮定を介することで解釈が可能となる記述を「準正解」とし，以上の基準に当てはまらない記述を「正解」とした。

正解と判断された記述として以下の例がある。

・U 1 さんの「小樽市中心部にある『小樽都通り・サンモール一番街・花園銀座商店街』の 3 つの商店街を会場に様々なイベントが催されます」という記事は A 議員の「中心商店街」という考えに近い。

この例では，「商店」と「中心」という名詞に基づいて正しく判断できていることが分かる。

表 5-2-3 と表 5-2-4 は，関連度が高い上位 15 位と 30 位までの結果における正解数である。正解数を比較すると，15 件と 30 件という小規模での実験ではあるが，上位 9 桁に拡張した場合に，最も良い結果となることを確認した。このことから無制限に拡張を行うのではなく，適切な概念体系に基づいて拡張することが重要であると考えられる。4.1 で述べた概念体系の構築は，上記の結果を受けて行われた部分がある。

### 5.3. 定型表現による議員活動抽出手法の改善

本節では，長谷川ら（2008）で行った研究

表 5-3-1　分析対象データ

| | 段落数 | 段落当りの文字数 |
|---|---|---|
| 平成 14 年度小樽市 | 598 | 107 |
| 平成 19 年度小樽市 | 695 | 173 |
| 平成 14 年度帯広市 | 1,246 | 61 |
| 平成 19 年度帯広市 | 1,314 | 73 |

を紹介する。この研究は，5.1 で述べた定型表現を用いた会議録からの抽出手法が，小樽市以外の会議録においても有効であるか調査することを目的としたものである。また，小樽市以外の都市と比較することで，政治問題に地域による差異が存在するか，といった点の調査も併せて行った。

地域差を調査するために，都市の規模が近い小樽市と帯広市の会議録を比較することとした。小樽市の人口は平成 19 年の時点で 137,456 人，帯広市は 169,156 人であり，小樽市は海に面している一方で，帯広市は内陸に位置しているため，地域性による政治問題の相違が現れやすいと考えられる。さらに，年代による相違も考察するため，両市の平成 19 年度と平成 14 年度の定例会会議録を選択した。分析に使用したデータの詳細を表 5-3-1 に示す。

4.1 で述べた概念体系を用いて，政治的概念を想起させるキーワードを会議録に注釈付けする作業を行った。この作業は大学院生 2 名により行われ，両名がキーワードとして注釈付けした記述の内，共にキーワードと注釈付けした記述の割合は 68.4%であった。表 5-3-2 と表 5-3-3 は，帯広市と小樽市の会議録からキーワードとして注釈付けした記述とその頻度を示したものである。

特定の地域に固有の政治問題の例として，表 5-3-2 の「ばんえい競馬」がある。「ばんえい競馬」は帯広市が主催する地方競馬であり，小樽市には存在しない。また，年度による差異としては，平成 14 年度の帯広市では乳幼児医療など福祉制度の問題が主に議論さ

表 5-3-2　帯広市会議録から抽出されたキーワード

| 平成 14 年度 | | 平成 19 年度 | |
|---|---|---|---|
| キーワード | 頻度 | キーワード | 頻度 |
| 乳幼児医療 | 4 | ばんえい競馬 | 16 |
| 介護保険 | 4 | 後期高齢者医療制度 | 2 |
| 児童扶養手当 | 3 | 後期高齢者 | 2 |
| 学童保育 | 3 | 北海道市営競馬組合 | 2 |
| 予算 | 3 | 事故 | 2 |

表 5-3-3　小樽市会議録から抽出されたキーワード

| 平成 14 年度 | | 平成 19 年度 | |
|---|---|---|---|
| キーワード | 頻度 | キーワード | 頻度 |
| 加配 | 12 | 財政再建 | 8 |
| TT | 11 | 財政 | 8 |
| 生徒指導 | 8 | 予算 | 7 |
| TT 加配 | 7 | 協働 | 6 |
| ﾃｨｰﾑ･ﾃｨｰﾁﾝｸﾞ | 7 | 病院 | 4 |

れている一方で，平成 19 年度では「ばんえい競馬」の存続をめぐる議論が多数を占めていることが分かる。小樽市においても同様に，平成 14 年度では，学校におけるティーム・ティーチングの制度に関する不正が問題となっていたことから，ティーム・ティーチングに関わるキーワードが多く含まれているが，平成 19 年度には自治体の財政赤字が問題となっていることから財政に関するキーワードが多く含まれていることを確認できた。

5.1 で述べた定型表現による抽出手法を用いて，これらのキーワードを自動的に抽出することができるか調査を行った。ただし，特定市町村の記述スタイルに依存せず自動的に抽出することを目指すため，定型表現を自動的に設定するよう以下の改善を行っている。まず，上記調査において注釈付けしたキーワードの前後 n 単語から成る部分文字列を抽出し，定型表現として設定する。抽出された部分文字列には定型表現として不適格なものも存在するため，定型表現としての尤度を，その部分文字列が会議録中に出現する頻度に

基づいて付加する。これは，適格な文字列は不適格な文字列よりも会議録中に多く出現するであろうという多数決の原理に基づいている。

自動的に設定された定型表現を用いて，どの程度キーワードを抽出できるか実験を行った。前後 3 単語で設定した定型表現を用いた場合，キーワードの抽出精度は 88.4%，前後 4 単語で設定した場合には 91.3%となることを確認した。したがって，注釈付けを行うことにより，自動的に定型表現を設定し，適切にキーワードを抽出できることが示された。残された課題としては，両市の会議録を用いて設定された定型表現の間に共通する表現が少ないことから，汎用性を高めるために適切な一般化を行うことなどがあげられる。

### 5.4. ツールによる注釈支援

今日の自然言語処理において，コーパスの整備は非常に重要な課題である。コーパスとは，テキストや音声などの言語データを，コンピュータ上で処理できるように大量に集めたものである。その用途は，4 で行ったような対象分野の分析のためだけではなく，システムの言語モデルを構築するための機械学習のデータとしての利用や，システムの性能評価を行う上で参照する正解としての利用など多岐にわたっている。それゆえ，大量の言語データを収集することに加え，システムの研究開発に必要な情報をそれらのデータに対して人手で注釈付けすることが求められている。

本研究においても，注釈付きコーパスの存在が必要不可欠であるが，このような注釈付きコーパスは，労力の観点から高価なものであると同時に，研究対象となる分野ごとに注釈情報が限定されてしまうものが多い。我々が目的とする地方政治の分野においても同様であり，国会議事録ではなく地方議会会議録となると文書収集の段階から始める必要があった。文書の収集に関しては，3.3 や 4.2 で詳しく述べているため，本節では，収集した会議録やブログなどの文書に対して，我々がどのように注釈付けを行ったか，注釈支援ツールの紹介を通して説明する。

本研究において付与した注釈情報は，議員名などの発言者情報，記述に対応する政治的カテゴリ，政治的カテゴリを判断するために重要と思われるキーワードおよびキーフレーズである。ここで，キーワードとは，名詞あるいは，名詞連続と定義している。例えば，「環境問題」などはキーワードとなる。また，キーフレーズとはキーワードよりも長く，助詞等を含む意味のまとまっている範囲としており，原則 1 文以内から構成されると定義している。例えば，「環境の問題」，「環境に非常に問題がある」などである。なお，4 における分析はキーワードに基づいて行われている。

これらの注釈情報は，コンピュータ上で処理しやすいように，図 5-4-1 に示されるような XML 形式で付与されている。図 5-4-1 の例では，〈Paragraph〉というタグで囲まれた部分が一つの段落を示しており，Member という属性に発言した議員名を保持している。キーワードに関しても，山田市長の最初の発言における「市政運営」のように〈Keyword〉タグで囲み，発言議員と政治的カテゴリの情報を Member と Category という属性に保持している。〈Keyphrase〉タグを用いることでキーフレーズも同様に注釈付けされている。

XML 形式による注釈付けは，コンピュータと人間の両方の可読性が比較的高い表現であるが，テキストエディタなどで図 5-4-1 に示すような注釈情報を直接入力することは作業者にとって極めて労力が高いものである。また，直接入力にはスペルミスなどのヒューマンエラーによる問題が多い。このような問題を完全に無くすことはできないが，可能な限り軽減するために支援ツールの利用といっ

図 5-4-1　XML 形式による注釈情報

図 5-4-2　注釈支援ツール画面

たことが考えられる。我々は図 5-4-2 に示すような注釈支援ツールを作成し，全ての作業者がこのツールを用いてコーパスの注釈付けを行うことで，作業者の労力軽減を図ると同時に，スペルミスが無いなどの最低限の品質が保証された注釈付きコーパスの整備を行っている。

我々のツールでは，キーワードまたはキー

フレーズとなる記述の範囲をマウスでドラッグすることにより指定し，図 5-4-2 の左側にあるようなリストから議員名や政治的カテゴリを選択させることで作業者の労力を軽減している。また，文書中に全ての政治的カテゴリに関する記述が均等に現れているわけではなく，ある程度の範囲で偏って現れることが多い。それゆえ，最近選択した政治的カテゴリのリストを別に表示することでカテゴリ選択に関する労力を更に軽減させている。

人手で注釈付けを行う際に問題となるのは，スペルミスなどの誰が見ても明らかに分かるような過誤だけではない。大量のデータに注釈付けを行うことで作業者本人も気付かない内に，以前の判断結果との間に揺れが生じることがある。当然，大きな揺れが生じないように予め設定したガイドラインに従って注釈付けを行っているが，小さい揺れが生じることは避けられない。このような揺れを作業者に自覚させることが重要であり，我々のツールでは以下の機能によりこの問題に対処している。

キーワードの抽出揺れに関しては，ある時点で作業者により抽出されたキーワードと同一表記である文書中の文字列全てに対して，色を変えて表示することにより作業者にキーワードの可能性がないか注意を喚起している。また，そのような文字列へボタン一つで直接ジャンプできる機能も備えている。政治的カテゴリの選択揺れに関しては，あるキーワードに対して作業者が選択した政治的カテゴリの情報を，他の同一キーワード候補（上記のキーワード抽出における同一表記文字列）に対して暫定的に付与している。これらの揺れに対する処理はあくまでも暫定的なものであり，作業者が確定させるまでは注釈情報として付与されることはない。これにより，人間の繊細な判断を要する事例に対して作業者の意識を特別に向けさせることができる。

## 6．ま と め

本論文では，北海道を対象とした住民本位型政治情報システムの構築を目指していることを述べ，2 では本プロジェクトの概要について説明した。また，3 では北海道を対象とした会議録のウェブ公開に関する調査により，ウェブによる公開は 3 割程度であることが明らかになった。4 では，議員活動および住民の関心の抽出へ向けた分析を行い，問題点を明らかにした。そして，5 では，本システム構築に向けた今までの取り組みについて述べた。

今後は，システムを完成させ，ウェブ上で公開する予定である。また，全国の地方議会を対象に進めていく予定である。

**謝辞**

本研究の一部は総務省 SCOPE 補助金（No.082301004）の支援により行われた。

## 参 考 文 献

K. Inui, S. Abe, H. Morita, M. Eguchi, A. Sumida, C. Sao, K. Hara, K. Murakami, and S. Matsuyoshi 2008 “Experience Mining: Building a Large-Scale Database of Personal Experiences and Opinions from Web Documents” 2008 IEEE/WIC/ACM International Conference on Web Intelligence, pp.314-321.

S. Sekine, C. Nobata 2004 “Definition, Dictionary and Tagger for Extended Named Entities” Forth International Conference on Language Resources and Evaluation.

秋田祐哉・三村正人・河原達也．2008．「会議録作成支援のための国会審議の音声認識システム」電子情報通信学会技術研究報告，SP 2008-99，NLC 2008-44（SLP-74-21）．

池田佳代・田邊勝義・奥田英範・奥　雅博．2008．「Blog からの体験情報抽出」，情報処理学会論文誌，Vol.49，No.2，838-847 頁．

乾　孝司・奥村　学．2006．〝テキストを対象とした評価情報の分析に関する研究動向〟，自然言語処理，Vol.13，No.3．

岩橋雄一郎・佐藤哲也・坂野達郎．2001．「争点態度投票理論に基づいた投票エージェントの制作・評価」第八回社会情報システム学シンポジウム，55-60 頁．

川端正法・山本和英．2007．「話題の継続に着目した国会会議録要約」言語処理学会第 13 回年次大会，696-699 頁．

喜連川優．2008．「情報学を創る ── 科研プロジェクトがめざしたもの：特定領域研究「情報爆発（Info-plosion）」」情報処理学会誌　Vol.48，No. 8，917-919 頁．

木村泰知・渋木英潔．2008．「ブログに潜在する政治的意見と議員活動とのマッチング手法」，電子情報通信学会言語理解とコミュニケーション（NLC）研究会，19-23 頁．

渋木英潔・木村泰知・山崎記敬．2007．「議員発言録からの重要単語抽出システムの提案」FIT 2007 情報科学技術フォーラム　一般講演論文集　第 2 分冊，275-276 頁．

友部博教・長尾　確．2005．「ディスカッションマイニング：議事録集合からの知識発見」，情報処理学会第 67 回全国大会．

長谷川大・乙武北斗・木村奉知・渋木英潔・高丸圭一・荒木健治．2008．〝市議会会議録を対象とした概念体系構築へ向けた分析〟，情報処理学会研究報告（2008-NL-187），23-28 頁．

藤井　敦．2008．「OpinionReader：意思決定支援を目的とした主観情報の集約・可視化システム」電子情報通信学会論文誌　D，Vol. J 91-D，No. 2，459-470 頁．

本村可奈子・友部博教・長尾　確．2005「ディスカッションマイニングシステムにおける会議活性化支援」，情報処理学会第 67 回全国大会．

諸岡　心・福本淳一．2005．「国会議事録の質疑・応答システム」，電子情報通信学会　第二種研究会資料　Web インテリジェンスとインタラクション，電子情報通信学会，35-40 頁．

山本和英・安達康昭．2005．「国会会議録を対象とする話し言葉要約」，自然言語処理，Vol.12，No.1，51-78 頁．

渡辺一郎・桝井文人・福本淳一．2004．「固有表現抽出ツール NExT の精緻化とユーザビリティの向上」第 10 回言語処理学会年次大会発表論文集 413-415 頁．

## 注

1）総務省情報通信政策研究所の 2008 年 7 月の発表では，2008 年 1 月現在，インターネット上で公開されている国内のブログの総数は約 1,690 万とされている。http://www.soumu.go.jp/iicp/chousakenkyu/seika/houkoku.html#2008I02

2）Yahoo！Japan「みんなの政治」http://seiji.yahoo.co.jp/

3）国立国会図書館　http://www.ndl.go.jp/

4）北海道における地方議員と住民間の協働支援システム　http://hokkaido-politics.net/

5）「みんなの経験」サイト　http://minna.naist.jp/