地方議会会議録検索システム「ぎ〜みる v2」の概要

乙武北斗[1]　高丸圭一[2]　内田ゆず[3]　木村泰知[4]
[1] 福岡大学　[2] 宇都宮共和大学 [3] 北海学園大学　[4] 小樽商科大学
ototake@fukuoka-u.ac.jp　takamaru@kyowa-u.ac.jp
yuzu@hgu.jp　kimura@res.otaru-uc.ac.jp

概要

本論文では，多様な地方議会会議録を統一的に扱うデータスキーマを提案し，それを活用することで自治体を横断して議会会議録を検索・視覚化するシステム「ぎ〜みる v2」の概要について述べる．本システムは従来システムと比較して Embedding に基づく発言検索，任意の区域を対象とした視覚化の機能が追加され，さらにシステム運用の簡便化を図ったことで，異なる自治体セットを対象とした複数のシステムを 1 つのサーバで容易に運用可能とした．

1　はじめに

現在，多くの地方自治体が議会活動や財政状況に関する資料をインターネット上で公開しており，議会での発言記録である議会会議録もその一つである．議会会議録は，議員の発言内容や質問内容，議案の審議内容などが記録されており，地方自治体の政策や課題を知る上で重要な情報源である．議会会議録の公開方法やデータ形式は自治体によって多様であり，ベンダーの検索システムを導入している自治体も少なくないが，検索対象は基本的にその自治体の範囲内となる．また，PDF 形式の議会会議録を単に配置している自治体も多い [1].

このような背景から，筆者らは 47 都道府県の議会会議録を収集・整理する手法を確立して「地方議会会議録コーパス」の構築を進めた [2]．さらに，地方議会会議録コーパスのデータを検索・視覚化する Web システム「ぎ〜みる」（以下，ぎ〜みる v1 とする）を開発し，一般公開した [3].

ぎ〜みる v1 は，都道府県議会会議録の横断検索を目的として構築したものであり，市区町村などの自治体にそのまま対応することは難しかった．また，システムの内部構成や依存する外部サービスが複数あって複雑であることから，動作環境を整備することが難しく，1 台のサーバ上で複数のぎ〜みる v1 を運用することも困難であった．

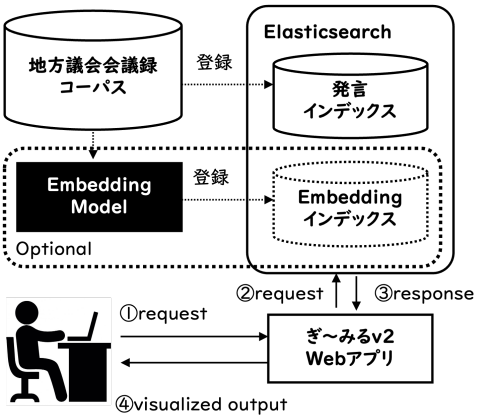

図 1　システムの概要

これらの問題を解決するために，筆者らは都道府県および市区町村の議会会議録を統一的に扱えるデータスキーマを提案し，ぎ〜みる v1 を代替する新たなシステム「ぎ〜みる v2」を開発した．本論文では提案するデータスキーマとぎ〜みる v2 の概要について述べる．

本研究の貢献は，以下の 4 点である．

- 地方議会会議録データを統一的に扱うデータスキーマの提案した
- 都道府県だけでなく市区町村など任意の区域を対象とした検索・視覚化機能を実現した
- 異なる自治体セットを対象とした複数の検索システムを 1 つのサーバで容易に運用可能とした
- Embedding を用いた発言文のあいまい検索の実現した

2　システムの概要

2.1　システムの全体像

本システムは，地方議会会議録を検索・視覚化することを目的とした Web アプリケーションとして実装されている．本システムのアウトラインを図 1

に示す．①ユーザは，一般的な Web ブラウザを通じて本システムにアクセスし，検索クエリを入力すると，②本システムは入力されたクエリに基づき，全文検索エンジンを用いて地方議会会議録データを検索する．③本システムは検索エンジンから検索結果を受け取り，④検索結果をユーザに提示，あるいは視覚化して表示する．

全文検索エンジンにはぎ〜みる v1 から引き続き，Elasticsearch を用いた．本システムを運用する際に，あらかじめ地方議会会議録データを全文検索エンジンにインデックスしておく必要がある．また，本システムは Elasticsearch の k-nearest neighbor (kNN) search を活用し，発言文の Embedding を用いた「あいまい検索」機能を実現している．図 1 の Embedding Model 部分は，外部の Embedding 取得サービスや Ruri[4] などの公開されているモデルによって Embedding を取得する部分である．ただし Embedding を取得するためには費用や計算リソースが必要であることから，図 1 に示すように Embedding インデックス関連の処理はオプショナルとした．

2.2 ぎ〜みる v1 からの改善点

従来システムであるぎ〜みる v1 と比較して，ぎ〜みる v2 には以下の改善・機能拡張がある．

- Embedding に基づくあいまい検索機能の追加
- 都道府県だけでなく市区町村など任意の区域を対象とした視覚化機能の追加
- システム運用の簡便化

システム運用の簡便化は，本システムを Docker によるコンテナ群として構成することで実現した．従来システムは Web ブラウザ表示を担うフロントエンド部分と内部処理を担うバックエンド部分が完全に分離しており，さらに全文検索を担う Elasticsearch や検索履歴を記録するデータベースなどの外部サービスに依存し，それらを運用するための環境構築が複雑であった．本システムはフロントエンド・バックエンドを統合して開発できる Remix フレームワーク[1)]を採用し，Docker を用いて本システムの全てのコンポーネントをコンテナ化することで，システムの構築・運用を容易にした．その結果，検索対象とする自治体セットを複数用意し，それぞれの自治体セットに対して本システムを独立して運用することも容易となった．

それ以外の改善点については，**3.** で詳細に述べる．

1） https://remix.run/

表 1 地方議会会議録データのスキーマ

1.	**識別子**（発言文に固有の ID）
2.	**自治体名**（東京都，栃木県宇都宮市，…）
3.	**回**（「第 335 回」など）
4.	**号**（各会議の会期における号数（日数））
5.	**年**（開催年（西暦））
6.	**月**（開催月）
7.	**日**（開催日）
8.	**会議種別**（定例／臨時など）
9.	**会議名** （例）平成 27 年第 341 回臨時会（第 1 号 5 月 13 日）
10.	**発言者フルテキスト**（「99 番乙武北斗君」など）
11.	**発言者 ID**（発言者リストを参照する外部キー）
12.	**発言者名**（「乙武北斗」など）
13.	**発言者の役職**（「議長」「知事」「議員」など）
14.	**発言文**（例：「次に節電対策について伺います。」）
15.	**発言以外の記録文**（「(拍手)」など）
16.	**原本 URL**

2.3 データスキーマ

本システムが対象とするデータは都道府県や市区町村といった地方自治体が公開している議会会議録である．しかしながら，自治体ごとにデータの形式や公開方法が異なるため，本システムが対応するためにはデータのスキーマを統一する必要がある．従来のぎ〜みる v1 は都道府県議会の会議録データを対象としていたが，本システムでは全国の自治体の会議録データを対象とするために，データスキーマを一部改善した．本システムが対象とするデータのスキーマを表 1 に示す．

「10. 発言者フルテキスト」は，議会会議録中に記載されている発言者部分の抜き出し文字列であり，通常は発言者の名前に加えて役職や議員番号などが含まれる．同一人物であっても会議が異なると役職が異なったり，氏名の表記も一部ひらがなになるなどの揺れがあったりすることから，この項目は同一人物で同一文字列になる保証がない．そのため，名寄せ処理の結果を記録する項目「11. 発言者 ID」および「12. 発言者名」を設けている．発言者名は発言者フルテキストから人物名以外を取り除いて表記揺れを解消したものであり，同一発言者には同一の名前を付与することが望ましい．筆者らが過去に整備した都道府県議会会議録コーパスにおいては，人手で名寄せを行い，発言者 ID と発言者名を付与している [5]．しかしながら，名寄せは時間や労力のコストがかかるため，本システムは発言者 ID や発

図 2　発言検索「子育て支援」の例

言者名が未付与のデータであっても利用可能としている．

3　機能の詳細

3.1　発言検索

発言検索機能は，ユーザが入力した検索文字列を含む発言文を検索する機能である．図 2 に都道府県議会会議録を対象とした発言検索の例を示す．図 2 では検索文字列として「子育て支援」が入力されており，検索対象の自治体を北海道・福岡県・東京都に，会議の開催年を 2016〜2017 年に絞り込んでいる．検索条件として発言者を加えることも可能であるが，47 都道府県のように自治体数が多い場合は，発言者を指定することは一般的ではないため，本システムでは発言者の条件入力欄の表示・非表示の切り替えを可能としている．

検索結果の発言文は，検索文字列部分がハイライトされて表示される．また，前後 3 発言を取得するボタンを各発言に設けており，このボタンを押下するたびに 3 発言ずつ表示範囲を広げることができ，ユーザが発言文の文脈を把握しやすいようにしている．

図 1 で示した Embedding インデックス関連の処理が有効な場合，「あいまい検索」機能が有効になるほか，発言検索結果には「類似発言を検索」ボタンが表示される．あいまい検索の例を図 3 に示す．図

図 3　あいまい検索「害獣対策」の例

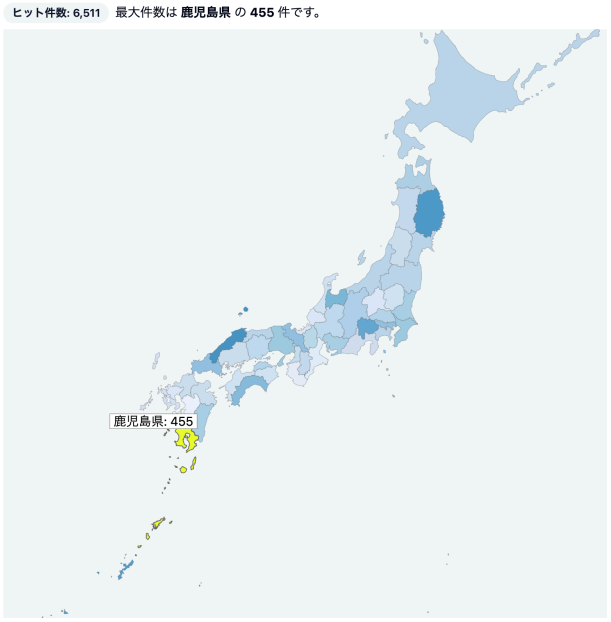

図 4　都道府県マップ検索「子育て支援」の例

3 では検索文字列として「害獣対策」が入力されているが，検索結果には「野生鳥獣対策」や「野生鳥獣による被害の防止策」など，検索文字列と類似した発言文が表示されている．政策や課題の表現が多様である場合に，あいまい検索機能はユーザが意図する検索結果を得るために有用である．

検索結果の並び順は，発言検索の場合は会議開催日の降順（新しい順），あいまい検索の場合は類似度の降順（高い順）を既定としているが，ユーザが基準を変更することも可能である．

3.2　マップ検索

マップ検索機能は，検索結果の件数を集計し，地図上の塗りつぶしの濃度によって視覚化する機能である．図 4 に都道府県議会会議録を対象とした，検

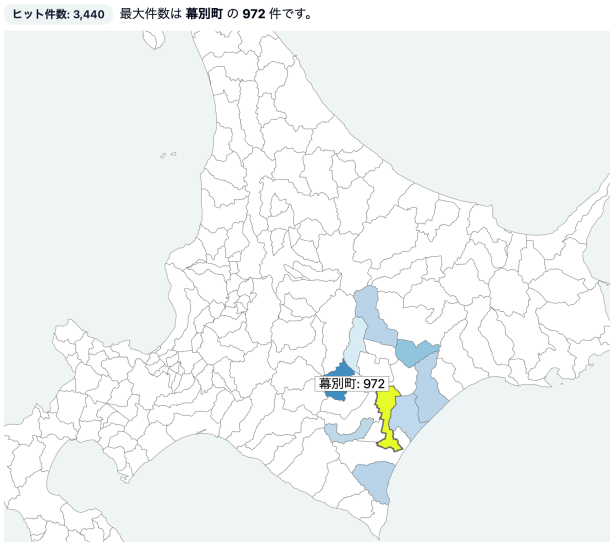

図 5　十勝管内マップ検索「子育て支援」の例

索文字列「子育て支援」のマップ検索の例を示す．地図上の都道府県単位で塗られた色が濃いほど検索ヒット件数が多いことを示しており，区域をクリックすることで当該地域の発言検索結果を閲覧することが可能である．

従来のぎ〜みる v1 からの改善点として，本システムでは汎用的な GeoJSON および TopoJSON 形式の地理空間データを用いて地図を描画しているため，都道府県だけでなく，市区町村など任意の区域を対象とした視覚化を実現した．図 5 に北海道十勝管内の市町村議会を対象とした「子育て支援」のマップ検索の例を示す．

マップ検索においても，Embedding インデックス関連の処理が有効な場合は，「あいまい検索」機能が利用可能である．しかしながら，あいまい検索は通常の検索と異なり，検索文字列に類似する順で発言文が列挙されることから，集計の際には類似度が高い発言も低い発言も同様にカウントされる．あいまい検索では類似度閾値の設定が可能であるため，ユーザが閾値を調整することで，検索結果の集計精度を調整することができる．

3.3　クロス表検索

クロス表検索機能は，行・列要素にそれぞれ項目名を指定することで，検索結果の件数のクロス集計結果を表形式で表示する．図 6 に都道府県議会会議録の長崎県を対象とした，検索文字列「子育て支援」における発言者・開催年のクロス表検索の例を示す．

ヒット件数: 102

	発言者	2015	2016	2017	2018	2019
1	中村法道	5	3	6	6	6
2	園田俊輔	0	0	0	1	3
3	中村和弥	0	0	0	0	1
4	山本由夫	0	0	0	0	1
5	永松和人	7	6	1	0	0
6	大久保潔重	4	0	0	0	0
7	小林克敏	3	0	0	0	0
8	橋村松太郎	2	3	0	0	0
9	山田朋子	1	0	1	4	0
10	山口経正	1	1	0	0	0

図 6　長崎県におけるクロス表検索「子育て支援」の例

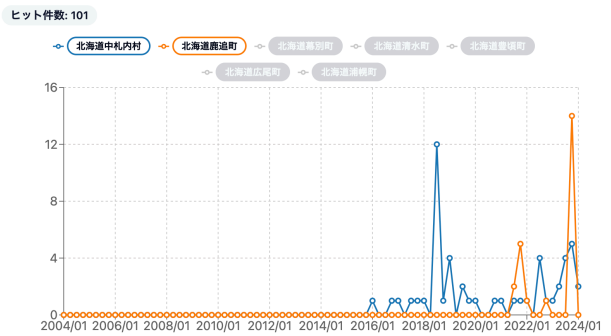

図 7　十勝管内における時系列検索「ヒグマ」の例

3.4　時系列検索

時系列検索機能は，検索結果の件数を時系列で集計し，折れ線グラフで表示する機能である．図 7 に北海道十勝管内の市町村を対象とした，検索文字列「ヒグマ」の時系列検索の例を示す．検索結果は四半期ごとに集計され，各自治体で色分けされた折れ線グラフが表示される．ユーザは凡例の自治体名部分をクリックすることで，当該自治体のグラフの表示・非表示を切り替えることができる．図 7 から，表示されている 2 つの自治体において近年ヒグマに関する議論が活発になってきているが，そのタイミングに差があることが見受けられる．

4　おわりに

本論文では，全国の地方自治体の議会会議録を統一的に扱うデータスキーマの提案と，それを活用した議会会議録検索システム「ぎ〜みる v2」の概要について述べた．現在，既に整備済みのデータを用いて本システムの試験運用を行っており，公開に向けて最終調整を進めている．試験運用中のシステムの内訳は付録を参照されたい．システムは筆者らのプロジェクトの Web サイト[2)]で公開予定である．

2)　https://local-politics.jp/

謝辞

本研究は JSPS 科研費 JP22K12740, JP21H03769, JP22K00582 の助成を受けたものです.

参考文献

[1] 乙武北斗, 内田ゆず, 高丸圭一, 木村泰知. 構造化データ作成を目的とした pdf 地方議会資料のテキスト抽出に関する分析. 第 37 回ファジィシステムシンポジウム講演論文集, pp. 431–436, 2021.

[2] Yasutomo Kimura, Keiichi Takamaru, Takuma Tanaka, Akio Kobayashi, Hiroki Sakaji, Yuzu Uchida, Hokuto Ototake, and Shigeru Masuyama. Creating japanese political corpus from local assembly minutes of 47 prefectures. In **Proceedings of the 12th Workshop on Asian Language Resources**, pp. 78–85, 2016.

[3] Hokuto Ototake, Hiroki Sakaji, Keiichi Takamaru, Akio Kobayashi, Yuzu Uchida, and Yasutomo Kimura. Web-based system for japanese local political documents. **International Journal of Web Information Systems**, Vol. 14, No. 3, pp. 357–371, 2018.

[4] Hayato Tsukagoshi and Ryohei Sasano. Ruri: Japanese general text embeddings, 2024. https://arxiv.org/abs/2409.07737.

[5] 高丸圭一, 内田ゆず, 木村泰知. 地方政治コーパスにおける都道府県議会会議録パネルデータの基礎分析. 宇都宮共和大学シティライフ学論叢, No. 18, pp. 136–155, 2017.

A　試験運用中のシステム

本付録では試験運用中のぎ〜みる v2 システムの詳細を述べる．

(i) **都道府県議会会議録 2011-2014:** 2011 年 4 月から 2015 年 3 月までの 4 年分の全国 47 都道府県議会の本会議会議録データを対象としたシステム．人手による発言者の名寄せ処理済みであり，総発言数は約 165 万件．

(ii) **都道府県議会会議録 2015-2018:** 2015 年 4 月から 2019 年 3 月までの 4 年分の全国 47 都道府県議会の本会議会議録データを対象としたシステム．人手による発言者の名寄せ処理済みであり，総発言数は約 225 万件．

(iii) **東京 23 区議会 2011-2018:** 2011 年から 2018 年までの東京都特別区議会の会議録データを対象としたシステム．総発言数は約 258 万件．

(iv) **十勝管内 9 自治体:** 北海道十勝管内の 9 町村（広尾町，本別町，上士幌町，幕別町，中札内村，鹿追町，清水町，豊頃町，浦幌町）の会議録データを対象としたシステム．総発言数は現在のところ約 10 万件であるが，データの整備を進めている．

(v) **小樽市:** 北海道小樽市の会議録データを対象としたシステム．総発言数は現在のところ約 20 万件であるが，データの整備を進めている．小樽市では公式の議会会議録の検索システムは提供されておらず，PDF ファイルによる会議録のみが公開されている．

(ii) と (iv) については，すべての発言文の Embedding を取得した上で，Embedding インデックス関連の処理を有効に設定している．Embedding の取得には `cl-nagoya/ruri-small` モデル [4] を用いた．

都道府県議会会議録データは統一地方選挙の年を基準に 4 年単位で収集・整備を行っている．**都道府県議会会議録 2019-2022**（2019 年 4 月から 2023 年 3 月までの 4 年分）については現在データ整備中であり，間もなく公開する予定である．