

地方議会会議録の自動収集に向けた公開パタンの分析

乙武 北斗^{*1} 高丸 圭一^{*2} 渋谷 英潔^{*3}
Hokuto OTOTAKE Keiichi TAKAMARU Hideyuki SHIBUKI
木村 泰知^{*4} 荒木 健治^{*1}
Yasutomo KIMURA Kenji ARAKI

^{*1} 北海道大学大学院情報科学研究科 ^{*2} 宇都宮共和大学シティライフ学部
^{*3} 横浜国立大学大学院環境情報研究院 ^{*4} 小樽商科大学商学部社会情報学科

{hokuto,araki}@media.eng.hokudai.ac.jp
takamaru@kyowa-u.ac.jp, shib@forest.eis.ynu.ac.jp, kimura@res.otaru-uc.jp

1. はじめに

今日のインターネット環境の発達に伴って、地方自治体の議会会議録が Web 上に公開されるケースが増加しており、政治的関心と情報収集の知識を持つ住民であれば自由に調べることができる。しかしながら、積極的・能動的に調べることはない住民も多く存在するのが現状である。議会会議録は相手が能動的に読んでくれることを期待するものであるため、消極的な相手には効果が薄いと考えられる。そこで我々は、積極的に情報を提示するようなシステムを開発することで、そのような問題を解決することを目指す¹。

しかしながら、システムを使うユーザを不愉快にさせないためには、ユーザの関心に沿った情報のみを提示する必要がある。したがって、システムを実装する際には、事前に各地方自治体の政治情報に関心毎に整理しておくことが求められる。そのためには、Web 上で公開されている各地方自治体の議会会議録のテキストデータを自動収集して最新の政治活動を把握しておく必要がある。

議会会議録の公開方法は各自治体によって様々である。[1]では小樽市、[2]ではそれに加えて帯広市の議会会議録を用いた研究を行ってきたが、これら2つの自治体においても議会会議録の公開方法には明確な違い

が存在した。それゆえ、各自治体の膨大な議会会議録を自動収集するためには多くの公開方法の違いを考慮する必要がある。本稿では北海道内の各市町村を対象に、どのような形式で議会会議録が公開されているかについての分析を行う。

2. 議会会議録の公開方法

公開されている議会会議録は多くの場合、自治体の議会に関する Web ページにリンクが配置されている。そのため、`wget` 等の既存のファイル取得ツールを用いることで、会議録のデータそのものは取得可能であると考えられる。しかしながら、我々の目指すシステムにおいては、会議録の年度、開催日、議会の種別、段落情報といったメタデータは非常に重要である。これは、年度や議会種別によって議題内容の傾向が変わることがあるためである。

会議録のメタデータは会議録へのリンク本文、またはその付近にあることが多い。しかしながら、会議録やメタデータの配置方法は各自治体で様々であるのが現状である。したがって、会議録とメタデータを共に自動収集する際には、そのような公開方法の違いを考慮する必要がある。本章では自動収集の際に考慮すべき公開方法の違いを4つの項目別に分析する。

¹ 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/>

2.1 対象とする議会会議録

本稿において分析の対象とする議会会議録は、北海道内の全180市町村のうちWeb上に議会会議録を公開している63市町村とした。

2.2 各市町村の公開方法

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。本稿では、公開方法の違いを以下の項目毎に分析した。

- ① 検索機能の有無
- ② 階層構造の有無
- ③ 公開データ形式
- ④ 公開単位

①の検索機能の有無というのは、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の階層構造の有無は、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の公開データ形式は、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の公開単位は、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。

以上の4つの項目に関して、対象市町村の会議録公開方法を分類した。採用件数の多い公開方法順にソートした結果を表1に示す。表1より、最も共通して採用されている公開方法は、全体のおよそ24%を占めていることがわかる。以下より個別の項目毎に分析する。

2.2.1 検索機能の有無

今回対象とした全ての市町村において議会会議録データへのリンクが用意されていたため、検索機能の有無が直接会議録の自動収集に影響することはないと考えられる。しかしながら、検索機能の提供のために既存の市販システムを用いる自治体が多く存在した。こ

のような場合、使用されている既存システムに合わせた自動収集パターンを用意することで、同様の既存システムを使用している自治体の会議録収集に対応することができる。

議会会議録の検索機能を提供している自治体は対象全体のおよそ33%となった。そのうちの約7割にあたる15市（帯広市、留萌市、釧路市、伊達市、北見市、北広島市、稚内市、室蘭市、苫小牧市、登別市、函館市、旭川市、根室市、芦別市、千歳市）が、NTTアドバンステクノロジー株式会社が販売する自治体向け会議録検索システム「DiscussNet」を用いていることが確認できた。これは表1の1行目の部分を表している。DiscussNetは平成20年9月の段階で全国約400自治体に採用されている[3]。

また、音更町では株式会社大和速記情報センターが販売している会議録検索システムDB-Searchを用いている。DB-Searchは全国約120の自治体で採用されている[4]。

その他の検索機能を提供している自治体では、独自のシステムを用いていると考えられる。そのような自治体の会議録を収集するためには、以下に説明するような階層構造やデータ形式を考慮した個別の収集パターンを用意する必要がある。

2.2.2 階層構造の有無

議会会議録の公開インターフェイスで階層構造をとっている自治体は全体の64%となった。また、検索機能を提供する自治体の中では、95%が階層構造をとっていることも確認できた。階層構造を持たない場合、単独のインデックスページからすべての会議録へのリンクを取得することができるため、自動収集のためのロジックが比較的単純になると考えられる。逆に、階層構造を持つ場合、自動収集の際には階層を辿っていく処理を追加する必要がある。階層の分け方は各自治体によって異なり、階層やディレクトリの名前も異なるため、会議録とメタデータを同時に収集するためには個別に収集パターンを記述する必要がある。

表1 対象市町村の会議録公開種別状況

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%	美瑛市, 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

2.2.3 データ形式

データ形式については、PDF、静的HTML、および動的HTMLの3種類が見られた。PDF形式で公開している自治体は全体の46%で、約半分という結果になった。しかしながら、PDF形式の場合、議会会議録データに標準外のフォントや文字コードが含まれていると、テキストデータ変換時に問題が発生することがある。

2.2.4 公開単位

公開単位に関しては、発言者単位、日単位、議会単位がそれぞれ約3割を占める形になった。

の発言に納まっている公開単位である。したがって、議会会議録を自動収集した際に議会単位のデータとして保存したい場合は、各発言者のHTMLテキストを結合する処理が必要である。市販の会議録検索システムを採用する自治体や札幌市など、比較的豊富な検索機能を提供する自治体では、発言者単位で公開する傾向が強い。

日単位とは、議会の開催日単位で会議録を公開しているケースである。例えば定例会は何日間に渡って開催されることが多いが、その場合、日単位の公開方法をとっている自治体では、1つの定例会の会議録を表すデータは複数となる。

議会単位とは、一つの定例会や臨時会を単位として公開することを意味する。議会単位で公開されている会議録の約7割はPDF形式を採用している。

3. 議会会議録の自動収集

我々は本稿で対象とする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を用いた自動収集プログラムを作成して収集を試みた。その結果、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。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。

2.2.1 で述べた DiscussNet を用いた自治体が全体の26%あったため、DiscussNet 用の収集パターン1つで24%の議会会議録収集を処理することができるものの、残り8割の議会会議録を収集するためには、自治体の数だけ個別に収集パターンを用意しなければならない。自治体独自で議会会議録公開インターフェイスを構築している場合、各自治体によって階層構造やデータ形式に差があるのが現状である。しかしながら、分析の結果、表1で示したように、議会会議録の公開方法の違いはおおよそ4種類の項目に整理できることがわかった。したがって、少なくとも現状の道内の市町村においては、これらの項目をパラメータとして設定することで自動収集が可能となった。

本稿では対象自治体を北海道内に限定したが、対象を全国の自治体に拡大した場合、DiscussNet や DB-Search などの市販システムを使用している地方自治体の数も増加するが、その数は全国の地方自治体の総数と比較して少ないと考えられる。したがって、現状では自動収集の際に各自治体の公開方法を考慮してパラメータ設定をする必要があると考えられる。

また、今回対象とした北海道内の市町村の中で、現在 Web 上で議会会議録を公開していない自治体も存在する。今後 Web 上で公開される議会会議録が増加すると考えられるが、その際に共通した公開方法が用意

されれば、議会会議録の自動収集が容易になると考えられる。今後、会議録を Web 公開していない市町村を含めた北海道内の市町村にアンケート調査を実施し、会議録電子化の状況や Web 以外の方法も含めた会議録公開の状況について更なる分析を進める予定である。

4. まとめ

本稿では、北海道内の各市町村を対象に地方議会会議録の自動収集に向けた公開パターンを分析した。その結果、約2割の自治体は共通する議会会議録検索システムを用いており、その他の自治体は各自異なるシステムを用いていることが確認できたが、議会会議録の公開方法は4種類の項目におおよそ分類できることが明らかとなった。また、我々が作成した51種類の収集パターンによる議会会議録自動収集プログラムを用いたところ、約94%の自治体から会議録を収集することができた。

今後は、収集対象の議会会議録を全国に拡大し、住民本位型政治情報システムの開発を目指す。また、収集したデータはコーパスとして公開したいと考えている。

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。また、会議録自動収集プログラムの構築にあたり、(株)テクノラボ米谷信昭氏にご協力頂いた。

参考文献

- [1] 木村泰知, 渋木英潔, “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, pp.19-23, 2008
- [2] 長谷川大, 乙武北斗, 木村泰知, 渋木英潔, 高丸圭一, 荒木健治, “市議会会議録を対象とした概念体系構築へ向けた分析”, 第187回自然言語処理研究会, pp.23-28, 2008
- [3] (株)会議録研究所, <http://www.kaigiroku.co.jp/>
- [4] (株)大和速記情報センター, <http://www.yamatosokki.co.jp/>