E-057

議員発言録からの重要単語抽出システムの提案
A Proposal of a System for Opinion Word Extraction from Minutes

渋木 英潔 † 木村 泰知 †† 山崎 記敬 ‡
Hideyuki Shibuki Yasutomo Kimura Noriyuki Yamazaki

1. まえがき

近年，有権者の政治離れなどが問題となっている一方で，インターネットなどを通して政治への関心を高めようとする運動が行われている [1]．また，個々の議員がブログやメールなどを利用して有権者と直接コミュニケーションを図ろうとする試みも行われており，これらの活動は非常に有意義なものであると考えられる．しかしながら，国会議員と比較して地方議会議員はマスメディア等への露出も少なく，どのような考えをもつ議員がいるのかさえ有権者に理解されていないことが多い．そのような状況では，政治家に対する質問内容も限られたものとなり，直接対話できる場があっても有効に活用されにくいと考えられる．それゆえ，どの政治家がどのような問題に関心をもっているかを知ることが第一であるが，多くの有権者にとって政治家の意見を知る手段は限られている．

このような観点から議事録を見たとき，議会における質疑応答は，質問した議員がどのような問題に関心をもっているかを知るための指標の一つとなりうる．しかしながら，議事録の分量は極めて多く全てに目を通すことは困難である．それゆえ，有権者の関心がある箇所を議事録から抽出し整理して提示することは，有権者が政治家の意見を知る手助けとなり，政治への関心を高めることにもつながると考えられる．以上の背景から，住民本位型政治情報システムの構築を目指して，議事録から重要単語を抽出するシステムの提案を行う．

2. 処理の概要

本手法では，抽出する重要フレーズを議員が関心をもっている問題を表す語句と定義し，重要フレーズの抽出は，議会においてある問題に対して質問を行っているならば，その問題に関心があるという仮定に基づいて行われる．入力は一般に公開されている議事録であり，出力は議員ごとに関心がある問題をまとめたリストである．ただし，リストはその議員の関心が高い順にランク付けし，類似する問題は同じクラスタにまとめて出力するものとする．

本手法の流れを図 1，発言録の例を図 2 にそれぞれ示す．まず入力された議事録から，質問した議員の名前と質問内容を，定型表現を手がかりとして抽出する．議会における発言形式は比較的定まっており，図 2 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する．本稿では，質問者の抽出テンプレートとして「○○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「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[1]．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てフィルタリングを行う．本稿では不要フレーズを「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．

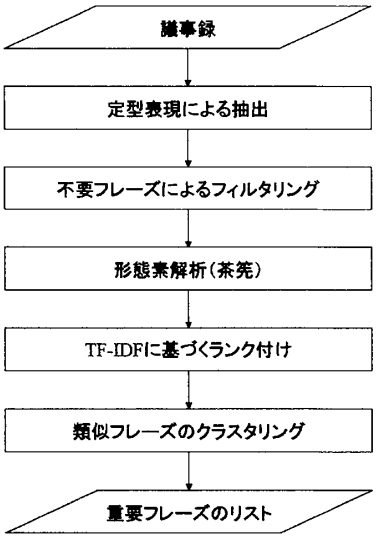

図 1: 全体の流れ

フィルタリングされたフレーズは，以下の手順で尤度を計算しランク付けされる．まずフレーズ中の名詞を同定するために ChaSen[3] を用いて形態素解析を行った後，式 (1) にしたがって TF-IDF 値を計算し議員 m における名詞 n の尤度 $P(m,n)$ を求める[2]．

$$P(m,n) = tf(n,m) \times log\frac{N_m}{df(n)} \quad (1)$$

$tf(n,m)$ は議員 m がフレーズ中で名詞 n を発言した回数であり，$df(n)$ は名詞 n を発言した議員数である．ま

†北海学園大学ハイテク・リサーチ・センター, High-Tech Research Center, Hokkai-Gakuen University
††小樽商科大学商学部社会情報学科, Graduate School of Information Science and Technology, Hokkaido University
‡中小企業診断協会北海道支部, Japan Small and Medium Enterprise Management Consultants Association

[1]本稿で対象とした発言録は小樽市議会 [2] のものであり，これに応じた抽出テンプレートを用いている．議会の違いによる表現の差が存在すると考えられるが，この調査に関しては今後の課題である．

[2]TF を名詞の尤度とした場合と比較した結果，著者らが TF-IDF を用いた方が良好な結果であると判断した．

○教育長（菊　譲）　斉藤陽一良議員の御質問にお答えいたします。

最初に、「大すきおたる」の発刊状況についてでありますが、子供と保護者に向けてのイベント情報誌として、平成13年12月に創刊しましたが、現在年３回、各9,000部を市内の小中学校や幼稚園などに配布しており、今月下旬には第17号を発刊する予定であります。

次に、おたる子どもプラン協議会についてでありますが、子供たちにさまざまな体験活動の場の情報提供や活動機会の拡大を主な任務として、教育や福祉、青少年育成団体などの関係者14名から成る委員により構成されております。また、会議は年間３回開催することとし、今年度は蘭島川水辺の楽校や地域子ども教室の運営、さらには「大すきおたる」の発刊などについて協議してきたところであります。

次に、本市におけるスポーツ・文化芸術活動など、さまざまな活動に取り組んでいる市民の数についてでありますが、内閣府の平成15年から平成18年までの世論調査報告書に基づいて小樽市のものを推計

図 2: 発言録の例

表 1: 出力リストの例

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

た，N_m は発言した議員の総数である．次に，式 (2) にしたがってフレーズ p の尤度 $P(m,p)$ を計算する．

$$P(m,p) = \frac{\sum_{i=1}^{N_n} P(m,n_i)}{N_n} \tag{2}$$

N_n はフレーズ p 中に含まれる名詞の数であり，n_i はフレーズ中で i 番目に出現する名詞である．フレーズの尤度は，含まれる名詞の数に影響されないよう平均を求めることとした．

ランク付けされたフレーズは，以下の手順で類似したフレーズごとにクラスタリングされる．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，類似性の高いフレーズは上位フレーズと同じクラスタとする．類似性の判断には，フレーズ中に含まれる名詞を次元としたベクトル空間を利用し，2 つのベクトルが成す角度の余弦を類似度とする．2 つのフレーズ p_1 と p_2 の間の類似度 $S(p_1,p_2)$ は式 (3) にしたがって計算される．

$$S(p_1,p_2) = \frac{\boldsymbol{p_1}\cdot\boldsymbol{p_2}}{|\boldsymbol{p_1}||\boldsymbol{p_2}|} \tag{3}$$

$\boldsymbol{p_1}$と$\boldsymbol{p_2}$は，p_1 と p_2 に含まれる全ての名詞を次元とするベクトルであり，その要素は $P(m,n)$ で重み付けした値である．同一のクラスタと判断する閾値は 0.8 とした[3]．

3. 実験

インターネット上で一般に公開されている平成 12 年から 18 年までの小樽市議会 [2] の議事録を用いて実験を行った．入力された議事録の文字数は 7,821,573 文字であり，議事録から抽出された議員数は 35 人であった．

[3]閾値を 0.6, 0.7, 0.8 の 3 段階で調査した結果，著者らが 0.8 が精度的に妥当であると判断した．

表 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である．表 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いることから，今後，定量的な調査が必要ではあるが，全体的に良好な結果が得られていると考えられる．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「懲戒処分」と「分限処分」は同一のクラスタにまとめた方がよいなど，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．これらは今後の課題である．

4. 今後の予定

今後，実際に議員の関心が高い問題が何かを調査するため議員本人に直接聴取を行い，抽出された重要フレーズとどの程度一致しているかで定量的評価を行いたいと考えている．

参考文献

[1] 岩橋雄一郎, 佐藤哲也, 坂野達郎: 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 (2001).

[2] 小樽市議会会議録: http://www.city.otaru.hokkaido.jp/gikai/gijiroku.htm

[3] ChaSen/茶筅: 奈良先端科学技術大学大松本研究室, http://chasen.naist.jp/hiki/ChaSen/