

発言録データに基づく文脈マイニング手法の開発

広島大学大学院工学研究科

学生員 ○難波雄二

広島大学大学院工学研究科

正会員 塚井誠人

広島大学大学院工学研究科

正会員 桑野将司

1. はじめに

社会資本整備や公共交通施策の決定に、複数の利害関係者が参画する場合、互いの意見が適切に伝わらず、コミュニケーションに齟齬が生じることがある。利害関係者間の合意形成を円滑に行うには、発言者の意見やその推移を把握する必要がある。これまで意見の分析は、協議会の傍聴や協議会議事録の整理、または、利害関係者へのヒアリングなどによって行われてきたが、これらの分析には時間と労力がかかる。また、議論の内容を把握する際に、分析者の主観が入りやすく、分析結果が形式知として蓄積されない。膨大な既往事例から議論の進行に有益な知見を見出したり、PI 等のコミュニケーションの場において論点や意見を整理する上では、意見の推移を分析できる客観的な分析手順が必要である。

本研究は、膨大な文書情報の分析に関するデータ整理の手間を省き、また客観性の高い分析結果を得るために、発言録を統計的に分析する手法の開発を目的とする。具体的には、自然言語処理とマイニング処理の手順に従って発言者の意見を抽出し、データベースを作成した上で、ある意見から別の意見への推移確率を算出して、それに関する分析を行う。

2. 分析手法

本研究は、テキストマイニング手法を用いて、発言者の発言から意見を抽出し、意見カテゴリに関する定量的な意見データベースを構築する。さらに、得られた意見間の推移確率を求め、意見の推移を分析する。本研究の分析手順を図-1 に示す。

(1) 意見の抽出

コーパス文書の文章を語幹となる品詞を中心とした「語」(キーワード)に分割し、その出現頻度や共起(同時出現)関係などを分析することによって、発言録データから意見の抽出を行う。①のコーパスデータの作成では、発言録データをテキストマイニングにより分

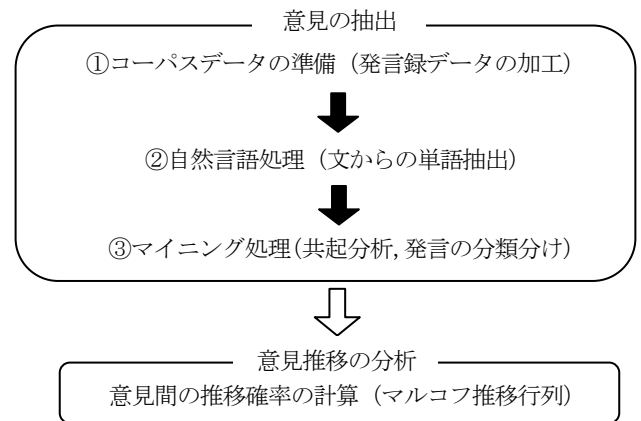


図-1 分析の流れ

析できるように加工する。なお、後段での分析を考慮して、コーパスデータ作成段階のデータ単位を句点「。」までの文単位とする。次に自然言語処理では、文を語に区切り、品詞タグを付けることによって、コーパス文書中の各文を統計的に分析可能な状態に加工する。②の自然言語処理では、文書中の各文を名詞、動詞、形容詞、形容動詞、地名、人名、組織名、その他の8種類に分類する。その際助詞、助動詞などの単独では意味を成さない付属語は、名詞や動詞などの語幹となる独立語に付随する語として、単独では抽出しないこととした。

③のマイニング処理では、自然言語処理により抽出されたキーワードの共起関係を分析し、発言の分類を行う。共起分析の指標は、一方の語が出現するときの、他方の語の共起のしやすさを表すMI値¹⁾を用いる。MI値は、値が大きい二つのキーワードは独立ではなく、一方のキーワードが出現していれば他方のキーワードが共起しやすいことを示す。以下同一発言者の意見に出現するキーワードの共起状態を分析する。そこで共起とみなす範囲(スパン)を、通常共起分析で設定される対象語の前後数語の範囲とせず、発言者が同一の範囲まで拡張してMI値を算出した。

以上の手順から明らかなように、自然言語処理で抽

出したキーワードから、意見を抽出し、各発言がどの意見カテゴリを含んでいるかを把握するまでの手順が、本研究におけるマイニング処理である。

(2) 意見推移の分析

(1)の手順により抽出された意見カテゴリは、発言ごとの発言者、発言順序、委員会の回次情報などに関するタグを付したデータベースとして整理される。その意見推移データベースを用いて、各意見カテゴリ間の推移確率に基づく分析を行う。意見カテゴリ間の推移確率は、マルコフ過程を仮定した推移確率行列によって算出する。マルコフ推移確率 p_{st} を(1)式に表す。

$$p_{st} = \Pr(X_n = t | X_{n-1} = s, \quad s, t = 1, 2, \dots, S)$$

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1t} \\ p_{21} & p_{22} & \cdots & p_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ p_{s1} & p_{s2} & \cdots & p_{ss} \end{bmatrix} \quad (1)$$

ここで、 s, t は意見カテゴリ、 n は時点、 S は状態の数を表す。1次マルコフにおいて P は時不変のため、時点 n の状態 X_n は(2)式で表すことができる。

$$X_n = X_{n-1} P \quad (2)$$

式(2)は、特に $n \rightarrow \infty$ のとき、

$$X^* = X_\infty = X_0 \prod_{n=1}^{\infty} P = X_0 P^* \quad (3)$$

となる。ここで、 X^* は定常状態である。

$$X^* = S^* P^* \quad (4)$$

が成り立つので、

$$X^* - X^* P = X^* (I - P)^* = 0, X^* \neq 0 \quad (5)$$

ならば、 X^* は P^* の固有値を要素に持つベクトルとなる。

マルコフ連鎖において、ある状態が時間の経過によって別の状態に推移しても、元の状態へ必ず戻ってくる性質を再帰性と呼ぶ。また、全ての状態間の推移が可能な性質を既約と呼ぶ。状態 s を出発したマルコフ連鎖が初めて状態 t を訪れるまでの時点数の期待値 $E[T_{st}]$ を、初到達時間という。すなわち、既約なマルコフ連鎖において $E[T_{st}] < \infty$ である。 T_{st} 初到達時間（時点数）とすると $E[T_{st}]$ は(6)式で表される。

$$E[T_{st}] = \inf\{n \geq 1 \mid X_0 = s, X_m \neq t (m = 1, \dots, n-1), X_n = t\} \quad (6)$$

また、推移確率行列 P を用いると、 $E[T_{st}]$ は以下の手順で計算できる。まず、初期状態ベクトルを X_s ($X_{s0} = 1, X_{t0} = 0, \forall t \neq s$) とする。このとき、1 時点後に状態 t に至る初到達時間の期待値時間の $E[T_{st,1}]$ は、

$$E[T_{st,1}] = 1 \times X_{t,1} \quad (7)$$

$$X_{t,1} = X_s P \quad (8)$$

$T_{st,2}$ は、 $n=1$ の状態ベクトル X_1 のうち、 $X_{t,1}=0$ としたベクトル $X_1^{(t\phi)}$ を用いて表される。

$$E[T_{st,2}] = 2 \times X_{t,2} \quad (8)$$

$$X_{t,2} = X_{t,1}^{(t\phi)} P \quad (9)$$

と表される。ここで、ベクトル $E[T_t] = (E[T_{t1}], \dots, E[T_{ts}])$ 、大きさが $S \times S$ で要素 $u_{ss}=1, u_{st}=0, s \neq t$ の行列を U_t とすると、

$$E[T_t] = 1 \times X_{t,1} + \sum_{n=1}^{\infty} (n+1) (X_{t,n} - X_{t,n-1} \cdot U_t) P \quad (10)$$

$$X_{t,n} = X_{t,n-1} \cdot P \quad (11)$$

によって求められる。また、発言の推移確率における時点は発言順であるため、時点差は発言間隔となる。

以上の手順で、マルコフ推移確率行列から、意見の出現確率、出現頻度、定常状態における確率、および各意見の発言間隔を求める。

3. 使用データ

本研究は、「横浜市市営交通事業あり方検討委員会」（以下、委員会）²⁾の発言録を用いる。この委員会では、第1回から第7回までは市営地下鉄事業、第8回から第12回までは市営バス事業について議論されている。本研究では、計12回の委員会のうち、第1回から第7回まで開催された市営地下鉄事業に関する委員会の発言録を用いて分析する。

4. 意見抽出結果

作成したコーパスデータに、自然言語処理を適用してキーワードを抽出した。品詞別では、名詞が最も頻度が高く、以下動詞、形容動詞、形容詞の順となった。

名詞の抽出結果から、グループ化を行った。主要な話題と関連の深いキーワードを含むと考えられる、出現頻度が8回以上の名詞を抽出すると、310種類の名詞が抽出された。これは、出現する全名詞の8.56%を捕捉する。次に、指示語など話題を抽出する上で、有用でない名詞とその名詞単独で意味を成さない名詞を不要語として除き、主要話題を構成すると考えられる189種類の名詞を抽出した。

用言は、自然言語処理より抽出したキーワードに手作業による修正を加えた結果、624種類の動詞、70種類の形容詞、131種類の形容動詞を抽出した。用言は、

意味が同じ、もしくは類似する場合が多いので、抽出されたそれぞれの語の意味や文書における役割を考慮して、40 グループに分類した。

名詞、用言の抽出後の意見の抽出は、次の手順で行った。まず、名詞の MI 値を求め、共起関係をもつ名詞の組み合わせ（話題）を抽出した。次に、名詞の組み合わせと用言グループの MI 値を求め、その組み合わせを抽出した。話題と用言の組み合わせを主要発言と呼ぶ。抽出された主要発言は 107 種類であった。次に、抽出された 107 種類の主要発言を、内容の類似したグループごとに意見にまとめる。意見は 13 種類定義した。抽出した 13 グループの意見を表-1 に示す。意見 1~12 は、主要発言に基づいて定義した意見である。また、意見 1~12 のどれにも該当しない意見は、ID13 の「その他」と定義する。

抽出された意見を意味で分類すると、「1.修繕費縮減」と「2.ワンマン化」、「4.借入金」はコスト削減という観点から見て、同系統の意見、一方、「3.税金」は公金収入によるサービスの推移という観点から見て、1、2、4 に対する対立意見と考えられる。なお、その他の意見に含まれる個別意見に意味上の相関は見られず、それぞれが独立した意見と考えられる。

5. 意見推移分析結果

4 章で定義した意見の推移確率を求め、定常状態における議題と関連の深い意見に初めて到達するまでの発言間隔の期待値を求める。ここで本研究では、委員会の議題と意味の類似する意見を議題と深く関連する意見と定義した。各回次の委員会の出現頻度、出現確率、定常状態における確率、最初の発言と議題に関連する意見までの発言間隔の期待値を算出した。第 1 回委員会の算出結果を表-2、発言間隔を図-2 に示し、第 5 回委員会の意見の抽出頻度、出現確率、定常状態を表-3、発言間隔を図-3 に示す。なお、図中の◎は議論と深く関連する意見を示す。

第 1 回委員会の議題に深く関連する意見は、「11.財政状況」である。表-2 より、意見の出現確率と定常分布は比較的一致しており、1 次マルコフ性を仮定した推移確率モデルはほぼ妥当である。また、図-2 より「9.自由度」の発言が議題に深く関連する「11.財政状況」に対して、発言間隔が最も小さいことを示し、「9.自由度を持たせたやり方により利便性を高くできる」か

表-1 抽出された意見

意見 ID	意見	略称
1	3 号線の車両や施設保守のための修繕費の削減を目指した業務の改善を目指す	1.修繕費縮減
2	民営化、乗務員のワンマン化によるコスト削減を目指す	2.ワンマン化
3	税収入がピークのため、赤字の部分は税金を上げて補うしかない	3.税金
4	財政が厳しく、借入金は運賃収入で賄う	4.借入金
5	高齢化が進んでいる	5.高齢化
6	都市への通勤のマイカー利用が変化した	6.マイカー利用
7	都市計画局による沿線の調整区域の市街化を目指す	7.沿線市街化
8	4 号線をリニア方式にするかを検討する	8.4 号線リニア
9	自由度を持たせたやり方により利便性を高くできる	9.自由度
10	公営企業の事業体は責任を持たなければならない	10.事業体責任
11	一般会計の財政状況を計算した上での市営地下鉄の課題について	11.財政状況
12	協議会の進行に関する話題	12.協議進行
13	議題にあまり関係のない話題	13.その他

表-2 意見の出現頻度、出現確率、定常状態（第 1 回）

意見	出現頻度	出現確率	定常分布
1.修繕費縮減	2	1.40%	1.28%
2.ワンマン化	8	5.59%	3.07%
3.税金	2	1.40%	1.49%
4. 借入金	3	2.10%	2.32%
5.高齢化	11	7.69%	8.72%
6.マイカー利用	0	0.00%	0.00%
7.沿線市街化	5	3.50%	3.81%
8.4 号線リニア	2	1.40%	1.49%
9.自由度	4	2.80%	2.16%
10.事業体責任	10	6.99%	4.55%
11.財政状況	3	2.10%	2.03%
12.協議進行	2	1.40%	1.49%
13.その他	91	63.64%	67.60%

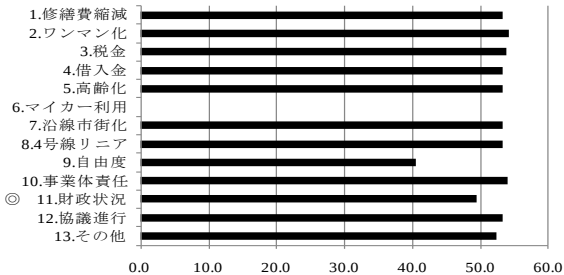


図-2 議題に関連する意見までの発言間隔（第 1 回）

ら、「11.一般会計の財政状況を計算した上での市営地下鉄の課題について」への推移がおこりやすいことがわかる。また、他の意見の発言間隔は、約 52~54 回で大きな差がない。なお第 1 回は、後の回次と比較して、発言間隔が大きいことが特徴である。以上より第 1 回委員会では、特定の話題が連続することなく、意見が検討されたと考えられる。

第 5 回委員会の議題に深く関連する意見は、「1.修繕費削減」である。表-3 より、「2.ワンマン化」、「3.税金」は出現確率よりも定常分布の値が低い。一方、「1.修繕費削減」、「11.財政状況」は、出現確率より定常分布の値が高い。よって、「2.ワンマン化」、「3.税金」の発言回数は期待値よりも多い一方、「1.修繕費削減」と「11.財政状況」は期待値よりも過少である。これは、前回の発言にのみ依存して、次の発言の出現頻度が決定されると仮定した 1 次マルコフモデルの当てはまりはあまり良くないことを表している。

また、図-3 より、「11.財政状況」の議題に深く関連する「1.修繕費削減」に対しては、発言間隔が最も小さいことを表し、「11.一般会計の財政状況を計算した上での市営地下鉄の課題について」から、「1.3 号線の車両や施設保守のための修繕費の削減を目指した業務の改善を目指す」への推移がおこりやすいことがわかる。また、他の意見の発言間隔は約 13~21 回であり、以上より第 5 回委員会は、それぞれの意見に頻繁に言及しつつ、議題に関連する意見につながる議論がされていたと考えられる。

「13.その他」の発言間隔に着目すると、第 1 回では「13.その他」は議題に関連する意見◎との発言間隔が他の意見に比べて小さい。すなわち、頻繁に「13.その他」が発言され、議論が散漫になっていた可能性がある。一方第 5 回は、「13.その他」は議論に関連する意見との発言間隔は、1~11 の意見と比較して相対的に大きくなり出現確率も小さい。これは回を重ねることによって、参加している委員が「13.その他」を発言せず、議論がかみ合っている傾向を表していると考えられる。

また、「12.協議進行」に着目すると、第 1 回では、議論に関連する意見◎に到達するまでの発言間隔が 1~11 の意見と比較すると相対的に多く、第 5 回は最多となっている。これは、第 1 回と比較して第 5 回は、「12.協議進行」を発言することなくスムーズに議論が進んだ傾向を表している。

表-3 意見の出現頻度, 出現確率, 定常状態(第 5 回)

意見	出現頻度	出現確率	定常分布
1.修繕費削減	4	3.15%	6.58%
2.ワンマン化	10	7.87%	5.52%
3.税金	10	7.87%	5.58%
4.借入金	3	2.36%	0.79%
5.高齢化	1	0.79%	0.79%
6.マイカー利用	4	3.15%	1.58%
7.沿線市街化	1	0.79%	5.79%
8.4号線リニア	6	4.72%	2.92%
9.自由度	9	7.09%	4.79%
10.事業体責任	4	3.15%	3.15%
11.財政状況	6	4.72%	8.15%
12.協議進行	2	1.57%	1.58%
13.その他	67	52.76%	52.80%

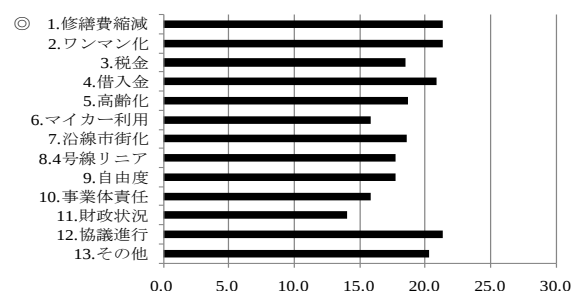


図-3 議題に関連する意見までの発言間隔(第 5 回)

6. おわりに

本研究では、発言録データにテキストマイニング手法を適用し、意見データベースとして抽出し、意見の出現頻度、出現確率、定常状態、各委員会の議題と関連の深い意見までの発言間隔を算出した。その結果、議題と関連の深い意見と、関連性が強い意見や、関連性が低い意見を発言間隔の期待値に基づいて、定量的に分類することが可能となった。しかし意見の抽出段階では、発言に対する定義した意見の捕捉率が十分ではなかった可能性がある。今後は、意見の定義をより幅広い内容をカバーできるように拡張して、捕捉率を向上させる必要がある。

参考文献

- 1) 石田基弘：R によるテキストマイニング入門，森北出版，2008
- 2) 横浜市市営交通事業あり方検討委員会：
http://www.city.yokohama.jp/me/keiei/seisaku/koutsu_arikata/，2010