

## 15. 大規模パブリックコメントの意見構造把握のための基礎的研究

岩見麻子

### 1. はじめに

環境政策の分野では住民参加が不可欠であり、広く採り入れられている手法としてパブリックコメント(以下、PC)がある。しかしPCは、提出された意見がどのように政策に反映されたかが不透明である点や、意見に基づく議論の展開など発展的手法が見られず形骸化している点など課題も指摘されている(山田、2011)。また、近年では数万件を超える大規模なものも現れている。たとえば、内閣官房国家戦略室が2012年に実施した「エネルギー・環境に関する選択肢に対する御意見の募集」には、88,634件の意見が提出された。これら膨大な意見の全文は国家戦略室のウェブサイト上においてPDF形式で公開されている(内閣府、2012)。また、各意見をその内容によって、事前に有識者から提起された4つの視点と細かな25の論点に分け、論点ごとに意見数を集計した結果も公表されていた(表1参照)。ただし、各意見をそれぞれの論点に分類した基準や具体的な集計方法などは公表されていない。このように、PCに提出された意見を客観的に把握あるいは集約し、活用するかは解決すべき課題である。

一方、大量のテキストからその文書の話題を表すトピックを把握する手法としてLDA(Latent Dirichlet Allocation)(Blei、2003)が注目を集めている。LDAとは、文書に出現する語とその出現回数から各文書に潜在的に存在するトピックを推定することができるトピックモデルの一つである。LDAは、各文書は確率的に複数のトピックから成り各トピックからは確率的に単語が発生することを表現するものであり(数藤ほか、2013)、クラスター分析などの多変量解析で大きく影響を与える頻度の差に依存することなくトピックを抽出することができる。そこで本研究では、前述した大規模PCに提出された意見集合を対象に、専門家が提起した4つの視点と25の論点を、LDAによって機械的に再現できるかを検証することを目的とする。ここで、LDAで生成されるトピック集合の解釈について、各トピックを説明する単語数が多くなりすぎるとトピックの意味の可読性が低下して論点集合との対応の評価が困難になるため、本稿ではまず、LDAに入力する単語集合 $W$ を最小化した上で、各パブリックコメントがより良く単一トピックに分類されるための方法を考察する。

表1 国家戦略室が設定した4つの視点と25の論点

視点	論点	各論点に分類された意見数	視点ごとの意見数
「原子力の安全確保と将来リスクの低減」	① 原子力安全に不安, 事故原因・影響も不明, 健康被害もある	47,901	118,490
	② 核廃棄物は将来世代に負担を残す	22,691	
	③ 原子力開発は倫理的に適切ではない	33,276	
	④ 今脱原発か推進かを決められない, 決めるべきでない	636	
	⑤ 安全対策を強化することで, リスクを最小化できる	2,172	
	⑥ 時間とコストがかかる廃炉を着実に進めることが重要	8,070	
	⑦ 安全を担う人材と技術が必要である	1,838	
	⑧ 原子力平和利用国としての責務を果たすべき	257	
	⑨ 国家安全保障のため核関連技術を保有すべき	1,169	
	⑩ 原発の不良債権化や立地地域への影響を懸念	480	
「エネルギー安全保障の強化」	⑪ 再生可能エネルギーや新エネ開発こそ急ぐべき	35,063	45,161
	⑫ 国際エネルギー情勢を注視しいずれにも偏らず多様化を進めるべき	904	
	⑬ 化石燃料の, 調達源の多様化, 戦略的活用が重要	1,516	
	⑭ 非化石電源である原子力発電が重要	1,073	
	⑮ 電力の安定供給のためには原子力発電が必要	3,331	
	⑯ 今でも電気は足りている	3,274	
「地球温暖化問題解決への貢献」	⑰ 温暖化対策にもっと積極的に取り組むべき	3,528	4,047
	⑱ 温暖化対策は他国の動向を見極めつつ推進すべき	53	
	⑲ 温暖化対策は国外での実施に貢献すべき	24	
	⑳ 温暖化対策は重視する必要はない	268	
	㉑ 温暖化はしていない	174	
「コストの抑制, 空洞化防止」	㉒ 新産業や雇用創出の好機である	3,756	8,603
	㉓ 経済への影響を見極めながら, エネルギーシフトすべき	842	
	㉔ コストがあがり, 経済に影響が出て, 雇用が失われる	3,609	
	㉕ エネルギー多消費産業の構造転換が必要となる	396	

## 2. 分析の枠組み

LDAを用いる場合、抽出するトピック数を分析者が任意に設定する必要があり、その決定方法について検討した研究も見られる（藤野、2014）。対象語数についても、語数が多すぎる場合、結果が複雑になりトピックを解釈することが困難であったり、比重が発散したりするのに対して、少なすぎる場合、文書全体のトピックを抽出するには不十分であることが考えられるため、これらのバランスを考慮した対象語数を慎重に検討する必要がある。

本稿では、まず50~1,000語まで、出現頻度順に $N_i = 50 \cdot i$ 個の単語を選定した出現単語の部分単語集合 $W_i$  ( $i = 1 \sim 20$ )を作成し、 $W_i$ でコメント集合 $C$ をBag of Words表現にしてLDAを適用した。このとき、トピック数は国家戦略室が設定した論点の数と同じ25とした。次に、各 $W_i$ を用いて生成したトピックモデル $l$ について、各コメント $c_j$  ( $j = 1 \sim 88,634$ )がトピック $k$  ( $k = 1 \sim 25$ )に所属する事後確率分布を求め、各コメント $j$ について情報エントロピー $H_j = -\sum p \log 2p$ を算出、全コメントの情報エントロピーの合計値を算出し、 $\max H_i(W_i)$ を最適な部分単語集合 $W_i$ と決定した。その後、部分単語集合 $\arg \max H_i(W_i)$ を構成する単語集合 $w_i$ は、事後確率の最大値 $\max(p_{k|w_i})$ により所属するトピック $k$ を決定した。その上で、 $H_i$ の値から $W_i$ を選び、各語について、それぞれのトピックに所属する事後確率分布を求め、その最大値 $\max(p_{k|w_j})$ のトピックに所属するものとして対象語を分類してみる。

なお、形態素解析にはttmを、LDAにはRのtopicmodelsパッケージを用いる。

## 3. 結果と考察

本稿ではttmの品詞体系の名詞のうち、一般と固有名詞、サ変接続、形容動詞語幹、ナイ形容詞語幹、副詞可能、複合名詞に分類された語を用い、代名詞と英数字のみの1文字語は分析から除外した。なお、対象品詞について

出現語と出現回数を把握した結果、異なり語数は136,238語、のべ出現回数は2,654,039回であった。

まず、 $W_i$ に対してLDAをそれぞれ適用し、 $\arg \max H_i(W_i)$ を把握した結果を図1に示す。図には $W_i$ ののべ出現回数が全出現回数に占める割合を併せて示している。図に示すように、 $\arg \max H_i(W_i)$ は、語数が増加するにつれてその値が低くなる傾向が見られ、特に450語の時点において平均値が低下していた。

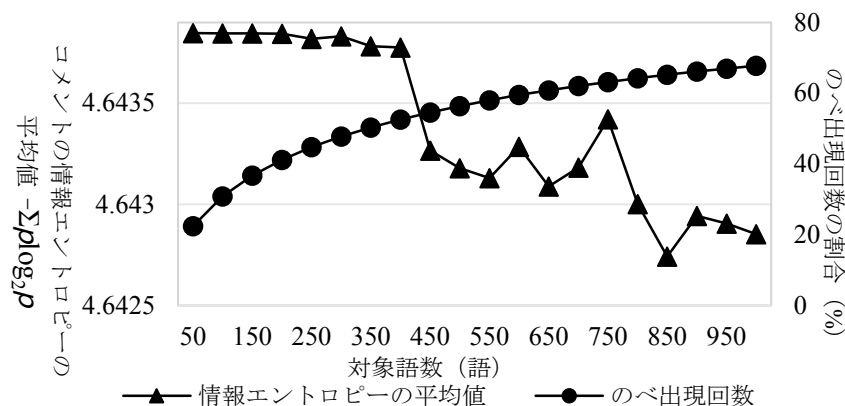


図1 情報エントロピーの平均値とのべ出現回数の割合

次に、各コメント $c_j$ がトピック $k$ に所属する事後確率分布の最大値である $H_i$ に基づき語を分類した。ここでは図1に示した結果から、450語を用いて分類を試みた。その結果を表2に示す。なお、同450語（全出現語の約0.33%）、のべ出現回数は1,334,136回（同約54.6%）であった。表に示すように、各トピックに分類された語数は、2～42語（最少：トピック3と21、最多：トピック4と23）であった。各トピックに含まれる語を見ていくと、論点を表していると考えられるトピックも見られる。たとえばトピック7には「経済」や「企業」「電力不足」などの語が分類され、経済への影響を懸念したトピックであると考えられる。同様に、トピック8の「温室効果ガス」や「気候変動」「削減目標」から地球温暖化、トピック11の「地震」「安全性」「地震大国日本」から災害時の安全性、トピック17の「事故」「影響」「放射能汚染」などから原発のリスク、トピック23の「エネルギー政策」「安全対策」「クリーンエネルギー」などからエネルギー政策に関するトピックであることがそれぞれ推察される。これらは、表1に示した国家戦略室が設定した4つの視点を網羅するものであり、25の論点のようなより具体的なトピックは抽出することは困難であるが、大きな意見のカテゴリー（視点）は抽出することができたものと考えられる。

一方で、雑多な語を含み、あるいは語数が極端に少なく論点の判断が困難なトピックも多く見られる。この原因として、対象語の選定に関する課題が考えられる。本稿では名詞のうち代名詞と英数字のみの1文字語を対象から除外したのみで、単純な出現頻度を用いて対象語を決定した。対象語数と併せて、TFIDFなど、分析に意味のある語を客観的に選定する手法も用いて検討する必要がある。また、対象語数についても、多くなれば各語がそれぞれのトピックを意味する比重は減少し、トピック内の論点を推定することが困難になるため、慎重に検討する必要がある。

表2 語の分類結果 (450語)

1	選択, すべて, 現実, 化石燃料, 活用		
2	生活, 開発, 場合, 維持, 安全神話, 普及, 利権, 電気代, 不足, 報道	15	放射能, 技術, 力, 不安, 利用, 核廃棄物, 大飯原発, 全て, 危険性, 太陽光発電, 東日本大震災, 破壊, 一刻, 安定, 共存, 技術開発, 時期, 核燃料サイクル, 移行, 安価, 実行, 幸せ, 協力, 面
3	管理, 本当		
4	選択肢, 反対, 絶対, 福島原発事故, 議論, 福島原発, 負担, 声, 研究, 原因, 世界中, 情報, 故郷, いま, 月, 無視, 教訓, 無責任, 検討, 考え, 災害, 話, 様々, 拡大, 使用済み燃料, 燃料, 被爆, 保障, 確実, 方針, 自体, 電力供給, 現時点, 放出, 構築, 専門家, 最後, 存続, 見直し, 確認, 風力発電, 長期的	16	自然エネルギー, 処理, 今回, 現状, 使用済み核燃料, 政策, 風力, 使用, 時間, 前提, 重要, 子供達, 早急, 期待, 次, CO <sub>2</sub> , 税金, 発電方法, 大量, 莫大, 水力, 認識, 導入, 設置, 考慮, 家族, 他国, 観点, 官僚, 目標, 提示, 不便, 課題, 排出, 輸出, 省エネルギー, 増加, 国策, 悪影響, 結論
5	原子力, 核, 自分, 目, 火力発電, コントロール, 経験, 明確, 心配, 言葉, 自分たち, 将来的, 比率	17	事故, 影響, 原子力発電所, 汚染, 国土, 子どもたち, 処理方法, 費用, 事実, 放射能汚染, 国内, 決定, 昨年, 甚大, 私達, 子ども, 一つ, 主張
6	政府, 廃炉, 現在, 推進, 人類, 社会, ゴミ, お金, 不可能, 方々, 一部, 利益, 判断, 火力, 豊か, 証明, 疑問, システム, 国家, 現実的, 知恵, エネルギー源, 住民, 健康被害, 家庭	18	シナリオ, 危険, 原子力発電, 電力, リスク, 電力会社, 稼働, 発電, 確立, 理解, 省エネ, 活断層, 暮らし, 原爆, 我が国, 保証, 継続, 多大, 保管, アメリカ, 中心
7	国, 経済, 状況, 地球, 節電, 企業, 状態, 十分, 人たち, 生命, 制御, 雇用, 電力不足, 海外, 時代, 広島, 長崎, 被曝, 信頼, 原発推進, 発電所, 段階	19	今, 命, 大切, 土地, 遺産, 犠牲, 地震大国, 大事, 反省, 運転, 便利, 孫, 姿勢, 被爆国, 気持ち, 金
8	再生可能エネルギー, 私たち, 人々, 意見, 依存, 世代, 支持, 前, 優先, 原子炉, 積極的, 具体的, 天然ガス, 崩壊, 原子力エネルギー, 提案, 生産, 発電分離, 石炭, 耳, 真摯, 再生, 加速, 誇り, 温室効果ガス, 要求, 核エネルギー, エネルギー効率, きれい, 約束, エネルギー計画, 気候変動, 削減目標, 国際公約, エネルギー部門	20	今後, 可能性, 確保, 廃止, 対策, ドイツ, 電気料金, 説明, 核燃料, 避難, 同様, 経済成長, 指摘, 一番
9	福島, 国民, 問題, 人間, 責任, 転換, 発生, 政治家, 東電, 信用, 夏, 唯一, 行動, 立場, 中国, 明白	21	未来, 自然
10	稼働, 多く, 放射性物質, 先, 太陽光, シフト, 事態, 無理, 大変, 規模, クリーン, 原発依存度, 破綻, 施設	22	方法, 放射性廃棄物, 存在, 子供たち, 海, 無駄, 過去, 目先, 真剣, 放射線, 効率, 仕方
11	日本, 膨大, 実施, 形, 外国, 事故後, 想像	23	世界, 原発事故, 将来, 電気, 被害, 明らか, エネルギー政策, コスト, 他, 可能, 対応, 手, 停止, 津波, 収束, 新た, 負, 発展, 経済的, 地震国, 地熱, 想定外, 技術力, 決断, 仕事, 困難, 供給, 建設, 震災, 納得, 人災, 安全対策, 自然災害, 恐怖, 次世代, 深刻, 即刻, 間違い, 投資, 東京電力, クリーンエネルギー, 持続可能
12	地震, 安全性, 努力, 代替エネルギー, 想定, 取り返し, 処分, 速やか, 地震国日本, 即時	24	原発, エネルギー, 理由, 安心, 環境, お願い, 希望, 健康, 意味, 道, 方向, 結果, 子孫, たくさん, 政治, 人達, 心, 後世, 覚悟, 根拠, 日本経済
13	安全, 人, 解決, 子供, 思い, 賛成, チャンス, あと	25	必要, 地域, 廃棄物, 産業, 場所, 完全, 非常, 資源, 水, 予算, チェルノブイリ, 削減, もんじゅ, 石油, ウラン, 地熱発電, 再生エネルギー, 家, レベル, 地球温暖化, 早期, 地球上, 視点, 機会, フクシマ
14	日本人, 実現, 原発依存, 間, 日本国民, 歴史		

#### 4. おわりに

本稿では、大規模PCに提出された意見集合を対象に、専門家が提起した25の論点を、LDAによって機械的に再現できるか検証を試みた。その結果、次のような課題が明らかになった。まず、国家戦略室が設定した4つの視点のような、大きな意見のカテゴリーは抽出することができたと考えられた。また、トピックを的確に推定するためには対象語数を慎重に検討する必要があることが確認できた。本稿では単純な出現回数の上位語を用いてLDAを実施したが、品詞を含め対象とする語の選定方法も併せて検討する必要がある。さらに、本稿で対象とした大規模PCは、国家戦略室によって25の論点が設定されており、それとの比較によって分析結果の妥当性を検討することが可能であるが、その対応関係を比較する方法を検討することも課題として残されている。加えて今後は、語の分類結果を比較しよりの確に論点を表すトピックを評価する方法や、各意見の比重を用いたトピックの推定についても検討する必要があると考えられる。

#### 参考文献

Brei, D.M, Ng, A.Y and Jordan, M.I.: Latent Dirichlet Allocation, Journal of Learning Research, Vol.3, pp.993-1022, 2003.

藤野巖, 星野祐子: LDA法におけるトピック数の決定法およびトピックの評価法について: Twitterストリーミングデータを応用例として, 電子情報通信学会技術研究報告. DE, データ工学, Vol.114-101, pp.67-72, 2014.

内閣官房国家戦略室 エネルギー・環境会議: 一政策一エネルギー・環境会議 パブリックコメント, <http://www.cas.go.jp/jp/seisaku/npu/policy09/archive11.html> (最終閲覧日: 2017年3月27日)

数藤京子, 村崎和彦, 島村潤, 谷口行信: レシピのテキスト及び画像特徴の学習による画像からの素材・調理法の推定, 電子情報通信学会技術研究報告, Vol.113-196, pp.195-200, 2013.

山田久美子, 柳下正治: 我が国の気候変動政策における意思決定プロセスへの市民関与の発展, 環境科学会誌, Vol24-5, pp.422-439, 2011.