２Ｅ０９

分野別の自由記述から科学技術政策上意味ある意見を自動抽出する試み

○ 奥和田久美 （文科省科学技術政策研究所）
白井康之、小関悠 （(株)三菱総合研究所）

1、研究の目的

科学技術政策の策定・推進においては、従来から、有識者会議などを召集して意見を求め、それらの意見を取りまとめることが一般的に行なわれてきた。しかし最近では、より多くの科学者・技術者の意見を求めようと、アンケートやパブリックコメントなども重視される傾向にある。質問に選択肢回答や段階評価などなんらかの定量的評価基準を持たせることができる場合は、結果の傾向や推移を掴みやすいため、指標として用いることもできる。しかし、多くのアンケートやパブリックコメントは、自由記述の回答スタイルを採らざるを得ないのが現状である。

一般に、アンケートやパブリックコメントは、有識者会議などで出される意見よりもはるかにサンプル数が多く、また多様な分野の回答者から多様な意見が寄せられる。自由記述から重要なキーワードを抽出し、総合的な意見あるいは平均的な意見をまとめるという作業は、実際には非常に困難な作業である。とりまとめ作業者が、実際に重要なあるいは特徴的なキーワードを見逃していないか、総合的あるいは平均的意見をまとめられているかなどは、実際のところかなり疑問である。回答の一部に有意義な特徴的意見が含まれていたとしても、特徴的な意見と平均的な意見の違いを見極めることは、回答数が多くなるほど難しくなる。また、とりまとめ作業にあたる者には、恣意的でないデータの読み方とともに、大所高所から結果を論ずる心構えが求められるが、そのような心構えを持ってしても結果的には、とりまとめ作業者の「読み手」としての個性や固定観念が入る可能性はかなり高い。作業が正しく実行出来ているのか、そうでないかという検証さえも難しい。

一方、科学技術分野のアンケート回答者やパブリックコメント提供者は、概して、あるレベルの科学技術的素養を身に付けており、比較的ロジカルな議論が可能な参加者である。また、分野の違いはあっても、現在の科学技術の直面している状況を共有しているという点で、文化的には互いを理解できる範囲の参加者である。したがって、一般の社会問題や経済問題などを扱う場合に比べれば、参加者の共感や一定以上の記述レベルは期待しうる。つまり、「テキストデータとしての質」という観点においては、科学技術分野のアンケートやパブリックコメントは質の高いデータであり、機械的な処理が容易な部類に属するはずである。

上記を勘案し、今回は汎用的なテキストマイニングの手法を用い、分野別に行われた同一アンケートの自由記述の回答サンプルから重要キーワードを抽出し、科学技術政策上意味のある平均的意見や特異な意見を、恣意的でない形で自動抽出する方法をいくつか検討した。以下には今回の手法と結果の一部を紹介する。

2、分析方法

2－1、分析の方針

① サンプルとしたアンケート調査の自由回答部分から、各設問単位・分野別単位での重要キーワードを抽出し、「必要」「不足」などの内容に応じた分類を行なうことで、「何が重要キーワードであり、そのキーワードに対してどのような回答が行われているのか」を自動抽出できるようにした。

② キーワード間の関係を図式化することで、各設問における自由回答の傾向の可視化を試みた。

③ クラスタリングにより、平均的な意見に近い代表的な回答と、クラスターから外れた特徴的な回答もそれぞれ抽出しようとした。

④ 分析にはフリーウェア・汎用の辞書等を主として用い、著作権等に問題が生じないようにした。

2－2、分析サンプル

分析サンプルとして、科学技術の状況を探る目的で行なわれた分野別の意識調査(アンケート形式)の自由記述欄への回答を用いた。この調査は全部で37設問から成り、対象分野は第3期科学技術基本計画における分野別推進戦略の8分野。回答者は1分野につき 100 名程度。選択肢や段階評価回答に自由記述欄が付随した設問が多く、段階評価回答必須だが、自由記述欄への回答は任意。結果的に自由記述欄への記述は全分野合計で約 6500 回答あり、これら全てを分析対象データとして用いている。

2－3、キーワード抽出方法

形態素解析技術(「茶筅」・「和布蕪」等を使用)により、自由回答から名詞および名詞の連語を抽出した。それぞれの名詞について、各設問での重要度を、設問全体との比較からTF-IDF(Term Frequency-Inverted Document Frequency:索引語頻度-逆文書頻度)により推定した。TF-IDF は右の関係式で、対象テキスト中のある文字列の出現頻度に、その文字列が他のテキストに現れていなければいないほど大きな重みを付けるものである。総数 N は抽出したい対象に応じて変えている。

$$w_{ij} = tf_{i,j} \times \log(\frac{N}{df_i})$$

$w_{i,j}$:名詞iの設問j(分野j)での重要度

$tf_{i,j}$:名詞iの設問j(分野j)での出現回数

N :設問数(37)、または分野数(8)

df_i :自由記述に一度でも名詞iが出現する設問数(分野数)

2－4. 分析項目

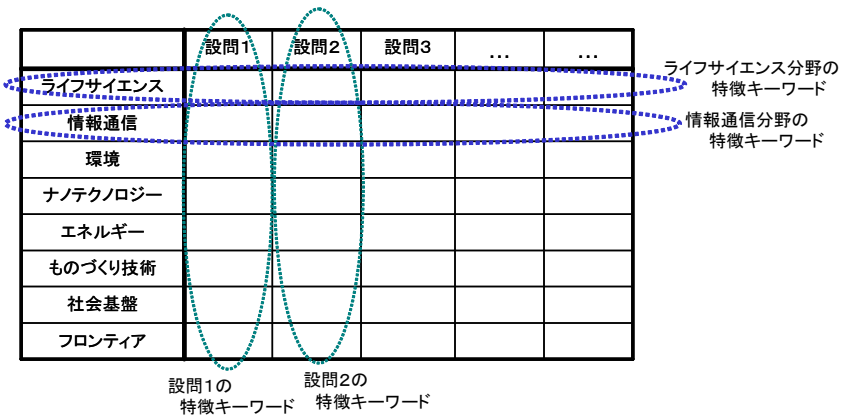
対象サンプルにおける自由回答文を対象に、以下の分析を実施した。

(1) 分野別特徴キーワードの抽出と内容分類

図表1のように、設問を横断して各分野に関する特徴キーワードを抽出し、さらに文の語尾の言い回しによって内容分類(「必要」「不足」「減少」「増加」「困難」「原因」)を機械的に分類してマトリクスを作成した。

(2) 設問別特徴キーワードの抽出と内容分類

図表1のように、分野を横断して各設問に関する特徴キーワードを抽出し、(1)同様の内容分類を行なってマトリクスを作成した。



図表1 分野別および設問別の特徴キーワード抽出のイメージ

(3) 特徴キーワードマップ

一部の設問において、TF-IDF の高い分野別並びに設問別特徴キーワードに関して、バネモデル(Spring Embedder Algorithm)で関連付け、描画ツールとして yed を用いてマップとして示した。

(4) 設問別特徴キーワードの分野別分布

設問別に、(2)で求めた設問別特徴キーワードのうち分野共通なものを拾い出し、分野間で比較した。

(5) 設問ごとの分野別特徴キーワード一覧

設問ごとに、分野別の TF-IDF を計算して特徴キーワードを算出し、高い順に並べた。

(6) 特徴キーワードによる回答文のクラスタリング

各回答の特徴量をベクトル表現し、各回答間の距離に基づいて、代表的な回答、特徴的な回答を抽出した。クラスタリングには、それぞれの特徴キーワードを含むクラスターを初期値とし、反復法によりクラ

スターを構成していくWillettのアルゴリズムを適用した。

3、分析結果

（1） 分野別特徴キーワードの抽出と内容分類

37設問全てについて設問横断で分野別特徴キーワードを抽出し、TF-IDF の高い順から並べた結果を、ライフサイエンス分野を例として図表2に示した。分野を特徴づけるキーワードがうまく抽出されており、内容分類とクロスで見ることにより、例えば「臨床（研究）の不足」といった特徴的な回答を引き出すことができた。図表1には、単語切り出しの問題、または回答件数の問題から、必ずしも適切ではないと思われるキーワード（「言語」「投下」「研修」など）も若干は含まれている。これを改善するには、人手による精査のほか、多年度連続分析などにより分野別特徴キーワードを同定する方法などが考えられる。

図表2
ライフサイエンス分野の
分野別特徴キーワードの
抽出と内容分類

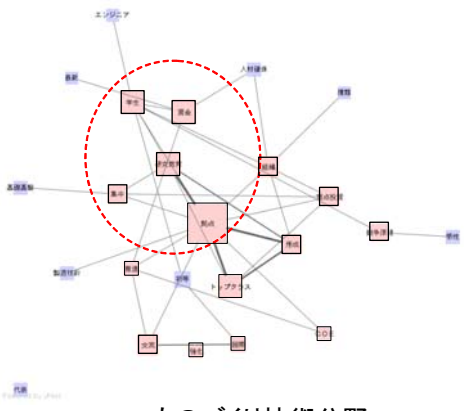
#1 ライフサイエンス	TF-IDF	1.必要	2.不足	3.減少	4.増加	5.困難	6.原因	9.未分類
臨床	30.405	4	9	2	2	2	2	10
言語	9.808		4					6
細胞	9.704						1	6
医学系	8.827	3	2	1			1	2
医師	7.846		4	1			1	2
ライフサイエンス	7.05	7			2	1		5
投下	6.931							5
ゲノム	6.931	1		1				3
再生医療	6.865							7
研修	5.884		2	2			1	1
長続き	5.884	1	2		2		1	
治療	5.884		2			1		3
前述	5.884	1	1	2	1			1
医療機器	5.545		2		1			1
厚労省	5.545	1	1				1	1
発掘	5.545	1	2					5
典型	5.545		3					1
規定	5.545		1					3
ポストドク	4.904		1	1	1			2
批判	4.904		1					4
遺伝子	4.904				2		1	2

（2） 設問別特徴キーワードの抽出と内容分類

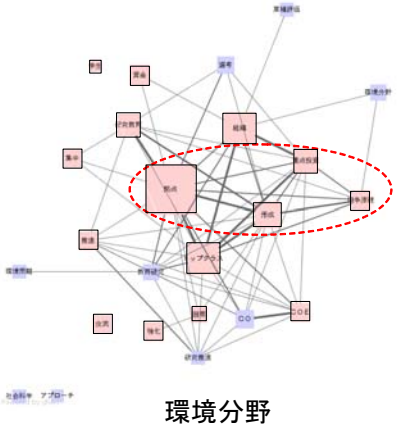
分野横断で、各設問を特徴づけるキーワードを（1）と同じように抽出した。ここでも内容分類とクロスで見ることにより、この設問において特徴的なキーワードとその回答の内容的分布を俯瞰することができた。

（3） 特徴キーワードマップ

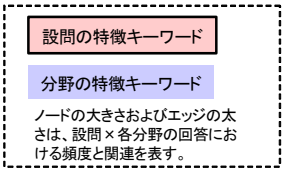
段階評価の無い自由記述のみの設問を選んで、分野別と設問別の特徴的なキーワードの関連のマップ化を試みた。図表3では、設問キーワードを固定した後、分野別キーワードを配置している。例えば、図表3の2つの分野を比べると、環境分野のほうが研究教育拠点と競争原理や重点投資などのキーワードとの繋がりが強いと言える。



ものづくり技術分野



環境分野



図表3
分野別特徴キーワードの
マップ化
（研究教育拠点形成の
必要なものを問う設問）

（4） 設問別特徴キーワードの分野別分布

各設問別に、（2）で求めた設問別特徴キーワードのうち分野共通なものをマトリクスにして示し、分野間で比較した。すなわち（2）と同様の分析を分野間で比較すると、「何がどの分野で不足しているか」などを一覧できる。この分析

は、自由回答のみのサンプルでは必須であるが、今回のサンプルのように選択肢や段階評価が付随している場合には、それらの補足という意味にとどまる。

(5) 設問ごとの分野別の特徴キーワード一覧

(1)では設問横断的に分野別特徴キーワードを求めたが、ここでは各設問に対して分野別特徴キーワードを求め、TF-IDF の大きい順に並べた。過不足などの内容分析部分は省いている。今回のサンプルはこの分析を行なうには設問×分野の回答件数が少ないため、統計的には意味のある結果とは言いがたいが、傾向を見る程度のことはできると考えた。このような分析は、分野別に大きな差が出る設問、あまり差が出ない設問などを見分けることができ、問い方の参考になると考えられるからである。図表4に結果の例を示すが、設問の問い方により、特徴キーワードの出現には差異が見られている。段階評価で最近の傾向を問う上段の一覧では、分野独自のキーワードのみとは言えないが、各分野の問題意識が端的に表れている。また、自由記述のみで重要なものや注目すべきものを問う下段の一覧では、分野別の特徴が非常に良く出ており、設問どおりの回答が得られていると見なすことができる。

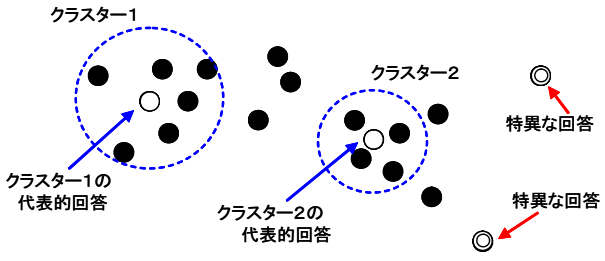
問6	ライフサイエンス	情報通信	環境	ナノテクノロジー	エネルギー	ものづくり技術	社会基盤	フロンティア
1	技官	敬遠	コンピュータ	博士後期課程	官僚体質	能力不足	敬遠	理科系
2	一途	啓蒙	情報交流	多忙化	悪しき	質向上	大学院重点	処遇改善
3	この後	ボリューム	植物	後期課程学生	一時期	ドクターコース	大学院後期	否定
4	大学院後期	一優先	臨床医	ものごと	思考力	システム研究	基礎基本	エンジニアリング会社
5	医師研修制度	若人	遺伝子解析	量産	工学等	現物	ため基礎	理数教育
6	スポーツ界	大学教育	医学領域	一流企業	平衡	研究現場	推察	へい害
7	米国等	根底	質向上	根拠	法人研究機関	机上	淘汰	ESA
8	憂慮	起因	操作	育成プログラム	前半	伝承	オリジナリティ	現代
9	程度継続	根幹	二極	基幹材料	エネ	きらい	任期付研究員	ISAS
10	定員削減等	横ばい	好奇心	環境改善	後半	理工系	派閥	メリハリ

問12	ライフサイエンス	情報通信	環境	ナノテクノロジー	エネルギー	ものづくり技術	社会基盤	フロンティア
1	細胞	医工連携	リスク評価	ナノテクノロジー	燃料	加工	災害	惑星
2	分子	Web	生態	ナノマテリアル	水素	マイクロ	防災	探査
3	生物	融合領域	物質	分子	電池	レーザ	土地	宇宙輸送
4	神経	コンピュータ	シナリオ	金属	エネルギーシステム	元素	地震	有人
5	再生医療	検索	緩化	電池	石炭	細胞	減災	無人
6	ゲノム	ロボット	環境問題	融合領域	発電	分析技術	物流	食料
7	生理学	情報工学	数値	素材	太陽	デバイス	都市	掘削
8	生体	ユビキタス	環境影響	複合	電力	MEMS	生態	宇宙環境
9	生命	通信技術	毒性	材料技術	石油	生体	ロジスティクス	環境利用
10	医学	セキュリティ	リモートセンシング	線形	変換	計測技術	交通システム	海洋

図表4 設問ごとの分野別の特徴キーワード一覧の例
上段:段階評価で最近の傾向を問う設問(緑色カラムは分野の独自性がありそうなキーワード)
下段:自由記述のみで重要なものや注目すべきものを問う設問

(6) 特徴キーワードによる回答文のクラスタリング

設問ごとに、各分野に含まれる回答を、設問別キーワードと分野別キーワードで特徴ベクトル化し、各ベクトル要素に対して TF-IDF により重み付けを行ない、特徴的なキーワードの出現を重視したクラスタリングを行なった。設問により、使用キーワードは変えている。このような分析により、似通った回答のなかで標準的と言えるような「代表的な意見」(そのクラスターの中心的意見)と、他とは変わった特異な意見(クラスターを形成しない意見)をそれぞれ選択することができた。自由回答文が非常に多い場合のとりまとめ作業では、これらの回答を優先的に読めばよい。



図表5 回答文のクラスタリング

4、まとめ

科学技術分野のアンケート回答やパブリックコメントのような自由記述のテキストデータから、自動的に効率よく、恣意的でないとりまとめが可能になる感触を得た。複数の方法で、人の技量に頼らない自動的な分野別とりまとめ作業が可能であり、これらは少なくとも検証としては有効に機能すると考えられる。今回のサンプルは選択肢や段階評価を併用した自由記述であったことから設問に対して回答がぶれにくく、データとしての質が高かった。手法としてまだ検討の余地があり、改善のためにも今後はタイプの違う自由記述を対象とした試みを行なっていく必要があるだろう。