１Ｇ１２
科学技術予測のテキストにおける意味あるメッセージの自動抽出

○奥和田久美、横尾淑子　（科学技術政策研究所）
小関悠、鵜戸口志郎　（（株）三菱総合研究所）

1、研究の目的

近年は、インターネットやブログ等の普及により、ウェブ上にテキスト形式の情報が爆発的に増大しており、意見収集やアンケート自体も容易になってきている。それらの膨大なテキストから、意味のあるメッセージやキーワードを自動抽出する方法も数多く提案され、市場調査等ではこのような手法が盛んに試みられている。多くの情報から、まだ弱いが、しかし意味のあるシグナルを見出す方法として、例えば 1997 年に Weak Signal Research が提案されている[1)]。

科学技術政策の決定・推進においてもイビデンスの明示が重要視されるようになり、パブリックコメントやアンケートなどと通じてより多くの意見を求めるケースは増えている。しかし、これらのパブリックコメントやアンケートにおいては、議論の対象や方向性が明確に定まっている場合を除くと、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ケースは少なく、多くの場合は自由記述の回答スタイルを採らざるを得ない。にもかかわらず、これらをテキストとして扱いメッセージを自動抽出する方法などは、まだほとんど試みられていない。

そこで発表者らは、パブリックコメントやアンケートなどで得られる記述（テキスト）を、できるだけ恣意的でない手段により、科学技術政策上意味のあるメッセージ性のある表現で総合的にまとめる方法や、問題点を浮かび上がらせていくような方法を検討している。これまでの試みから、科学技術政策に関して寄せられる意見は、総じてテキストデータとしての質が高く、このような自由記述のテキストデータから、個人の既成概念にとらわれず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ている[2)]。今回は、このようなテキストからの自動抽出を、科学技術予測において得られたテキスト情報に対して試みた。

科学技術予測は世界各地で行なわれており、各国の政策立案合議による方針決定、企業や業界団体による戦略策定などに役立てられ、予測手法としても様々なものが提案されている。日本などアジアの国々ではデルファイ法を用いた調査などが中心的に行なわれてきた経緯があるが、欧州などでは主に会議・ワークショップ・ウェブアンケートなどを通じて収集した種々の意見のなかから、合議等を通じて方向性を見出そうとする予測手法が中心的である。今後のメガトレンドなどを、多くの書誌や意見等から自動的に導き出そうとする試みも盛んに行なわれている。手法によらず、こうした予測活動では、特定の一人の見解による戦略策定よりも、むしろ「集合知(Wisdom of Crows)」[3)]を重要視することが基本にされている。したがって、多くのテキストからメッセージを自動的に引き出す方法は、今後の予測活動でも重要なツールのひとつになっていく可能性もある。実際に、欧州の予測プログラムのなかには、前述の Weak Signal Research の考え方を採り入れたものが出てきている[4)]。さらに将来的には、自動抽出されたキーワードなどを用いて、将来の方向性を示すシナリオとしてのテキストを再構築する方向に進展するものと考えられる。

今回は、2010 年6月に公表された「将来社会を支える科学技術の予測調査」[5)]において、デルファイ調査で出された科学技術のトピックや専門家グループによって書かれた将来へのシナリオのテキストを対象とし、潜在意味分析を用いて、恣意性を排除した形で科学技術全体の方向性を見出すとともに、社会の課題に貢献度の高い科学技術の抽出や科学技術の成果がより社会に貢献するための学際性などに関して可視化を試みた結果を報告する。

2、分析対象および分析方法

分析対象としたテキストは、「将来社会を支える科学技術の予測調査」[5)]の報告書のなかの以下の2種類である。

①「第9回デルファイ調査」報告書に記載された、12の学際的分科会から出されたトピック（総数832）と、いくつかのトピックの集まりである区分名（総数94）

②「将来を支える科学技術の予測調査」報告書に記載された、専門家グループによって書かれたシナリオ（総数12）における本文

各分析は、以下の(1)～(3)を基本として進めた。

(1)キーワード抽出

テキストからキーワードを抽出し、その出現度合を分析し、両テキスト間の対応状態を調べた。恣意性や固定概念を排除するために意味論に立ち入らず、高速かつ確実により多くのキーワードを抽出するため、テキストの意味的な検討は行なわず、漢字の連続・アルファベット・カタカナなどのパターンにより抽出した。キーワードとして抽出される例としては、3 文字以上のアルファベット、3 文字以上のカタカナ語、漢字の繰り返しとひらがなの組み合わせ(ひらがなが 2 文字続くものは除く)などである。

(2)キーワードの重み付け

このような潜在意味分析において、一般的に重み付けの手法として用いられるTF－IDF(索引語頻度―逆文書頻度)を用い、出現頻度(TF)と出現する文書数(DF)の関係からキーワードの重要度を導き出した。多くの文書に登場する語が一般的と見なされて重みが小さくな特定の文書のみに登場する語は重みが大きくなるため、テーマや領域で専門用語が異なる自然科学系の科学技術の分析には適していると考えられる。

TF－IDF(索引語頻度―逆文書頻度)

$$W_R(i) = tf_R(i) \cdot \log\left(\frac{n}{df_R(i)} + 1\right)$$

$W_R(i)$：キーワードの重要度

$tf_R(i)$：キーワードの出現頻度

$df_R(i)$：キーワードの出現する文書数

(3)テキスト間の比較および関連付け

まず各シナリオで抽出したキーワードの重みを用いて、テキスト間の比較を行なった。また、デルファイ調査のトピックとのマッチング(類似度分析)を行なった。デルファイ調査のトピックのいくつかの集合は「区分」とされ、区分名が付与されており、この区分名も含める形でマッチングを行なっている。これらを生かして、シナリオ間の類似度分析およびトピック間の類似度分析を行ない、コレスポンデンス分析によるマップ化も行なった。

3、分析結果

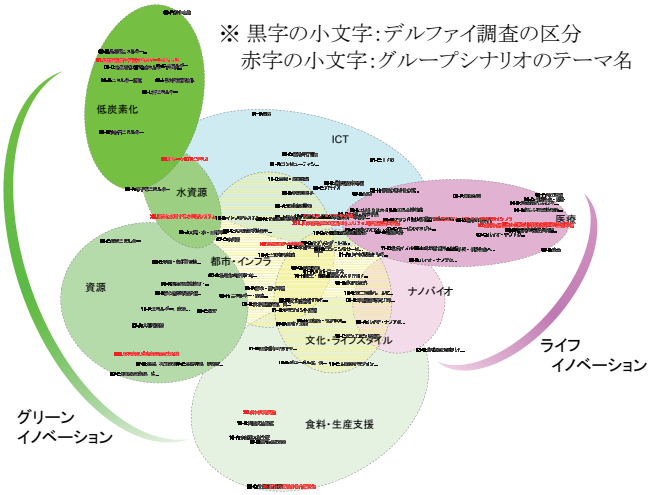
3－1. テキスト全体のマップ化

デルファイ調査の各トピックとグループシナリオの特徴キーワードとの関連性を、相対的位置としてニ次元にマッピングした。

この図は、テキスト全体で扱っている言葉の相関図であり、テキスト全体を表す「雲」のようなものである。小さい雲が集合してより大きな雲を形成し、それらが集まって、全体が1つの雲のようになっている。

この全体像からは、グリーンイノベーションとライフイノベーション関係が特に注目を集めていることがわかる。今後のイノベーション創出の柱として、これらを2つの方向性を推進していくことは妥当であろう。また、ICTを基盤技術とし、インフラやライフスタイルなどの共通的な議論のもとで展開されていくことが望ましいということもわかる。

右図は全体像を構成する各括り(小さい雲)のなかに出現するキーワードと、その中に含まれるデルファイ調査のトピックの例を示す。

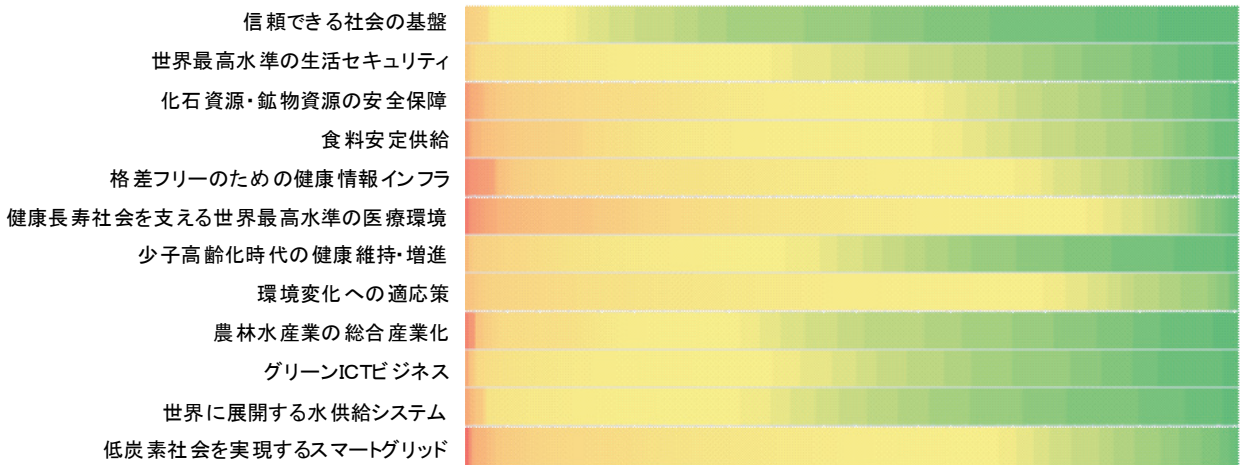


マップ中の括り	出現キーワードの例*	デルファイ調査トピックの例
低炭素化	エネルギー、電池、電力、発電、ネットワーク、太陽、変換、供給、材料、再生可能	◆全体の需給バランスをICTを活用し最適に運用し、低コスト、安定供給、低炭素化電力供給が可能な送配電ネットワーク技術 ◆エネルギー変換効率60%以上の水素電池
水資源	下水、資源、水資源、循環、地下水、水道、観測、処理、水管理、汚染	◆経済的・実用的海水淡水化技術、汚染水浄化再利用技術などを活用し、水の偏在化に対応した地域固有の水循環利用システムの構築 ◆水利用・水質汚濁実態の地球規模観測
資源	資源、エネルギー、回収、鉱物資源、非在来型、在来型化石資源、CCS、廃棄、リサイクル、分離	◆脱化石燃料依存に適用可能な、CCSを組み入れたガス化による経済性ある発電および水素製造、合成燃料製造技術の実用 ◆多くの希少金属の必要安定量の50%以上が供給される、一般・産業廃棄物と焼却灰・飛灰から希少金属を合理的に回収・利用する技術
食料・生産支援	生産、産業、エネルギー、資源、生物、バイオ、管理、情報、作物、微生物	◆我が国における、食作物の50%以上を生産する効率的な大規模企業化農業(海外生産、室内生産を含み、50%は全国合計値)が普及 ◆中緯度温帯地域でもシステム成立可能な高収量かつ種作を可能とする草本系などのバイオマス商業生産技術
医療	医療、情報、バイオ、管理、健康、治療、感染、診断、デバイス、予防	◆ゲノム情報による疾患リスク診断技術 ◆カルテは動画を含み電子化され、患者個人の管理になり、検査その他の情報は全医療機関で共用され、それを元に健康管理エージェント業が成立
ナノバイオ	健康、生産、作物、細胞、品種、予防、遺伝、適応、窒産、チップ	◆細胞内・細胞内外間の物質循環互作用同定・機能解析技術 ◆半導体チップ上に数千～数万の反応容器を集積化、多くの生体反応の検出を一瞬で可能とするナノチャンバーアレイ
都市・インフラ	下水、資源、地下水、水資源、汚染、上下水道、インフラ、建築、大気、観測	◆地理情報と宮城圏、生態系および災害リスク分布の情報が統合された情報インフラが整備され、国土の計画や管理のための技術や制度が実用化 ◆貨物輸送効率化のため、鉄道、道路、港湾、空港の結節点における時間・コスト・環境負荷を平滑するシステムが開発
文化・ライフスタイル	生産、農業、災害、気候変動、グローバル、マネジメント、教育、コミュニケーション、心理	◆大学院教育から職業訓練においてリカレント教育が一般的になり、社会・経済の変動に対応して人材が流動するようになる ◆熟練者の判断過程や技能、ノウハウを明示化して、他者による再利用や学習を可能とするサポートシステム
ICT	データベース、情報、ネットワーク、情報管理	◆いつでもどこでも自身の情報環境に安全に自由にアクセスできる社会インフラとしてのユビキタス環境 ◆実時間データに基づき全地球的な気象・海洋・環境・生態系・伝染病・経済・人の動きなどをシミュレーションして予測するシステム

※表中の出現キーワードは、括りの中で類似性の高い主なもの

3－2. 科学技術の成果の貢献度が大きいと期待できる領域の抽出

専門家グループにより書かれた各シナリオの特徴キーワードをデルファイ調査のトピックのテキストと比較し、関係性の強いトピックから並べ、関係性の強さの強弱を赤から緑への色の変化で表した。



赤や黄色の部分が多いシナリオのテーマは、多くのトピックとマッチングがとれることから、これらは、現在の専門家集団の意識のなかにあり、または実際に進められている研究開発の成果が、将来的に社会還元される期待が大きく、「科学技術の成果の貢献度が大きいと期待できる領域」と考えることができる。また、種々の分野が学際的に取り組むことによって、将来の社会の課題が解決される可能性が高まるともいうことができる。以下の図は、どのような専門家が参画し、どのようなキーワードに注目して進めるとよいかを示す。

シナリオテーマ	デルファイ分科会 No.1（電子・通信）	No.2（情報）	No.3（バイオ）	No.4（医療）	No.5（宇宙・地球）	No.6（エネルギー）	No.7（資源）	No.8（環境）	No.9（材料）	No.10（製造）	No.11（マネジメント）	No.12（インフラ）	デルファイトピックに出現する主なキーワード
低炭素社会を実現するスマートグリッド	9		3			24			3	1		2	エネルギー、電池、電力、ネットワーク、太陽、効率、発電、供給、材料、再生可能
化石資源・鉱物資源の安全保障					4	6	20	6	1	3	1	1	資源、利用、エネルギー、回収、鉱物資源、在来型化石資源、化石資源、CCS、廃棄
健康長寿社会を支える世界最高水準の医療環境	3	12	28	68			1		31		3	2	医療、情報、バイオ、機能、治療、診断、デバイス、応用、基盤、通信
格差フリーのための健康情報インフラ	2	2	1	11				2	1	8	21	4	健康、管理、データベース、情報、教育、利用、サービス、医療、価値、活用
食料安定供給			12		1	1	9	3		12	9	5	生産、資源、エネルギー、利用、産業、生物、管理、バイオ、情報、作物
環境変化への適応策				1			3	8					下水、資源、地下水、水資源、汚染、利用技術、物質、化学、上下水道、進出

※図中の数字は、関連性の強いトピックの数。
※今回のデルファイ調査は12のNo分科会から成っている。分科会名はNoのみであり、それぞれが学際性に富む分科会であるが、各分科会は独自に視点を定めており、図中ではその視点を簡略化して（ ）内に1語で示している。

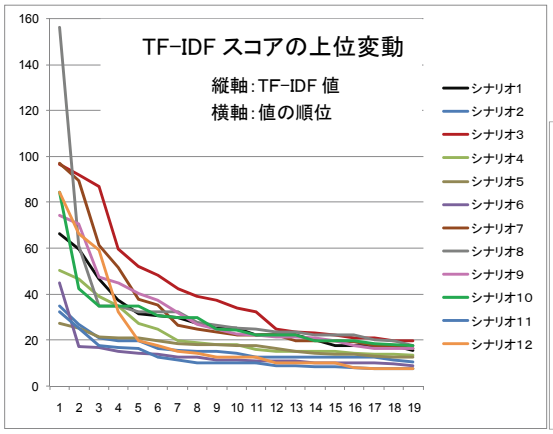
一方、緑部分の多いシナリオは、デルファイ調査のトピックとマッチングがとれないテーマであるが、その要因には、いくつかのケースが考えられる。

①シナリオの記述が科学技術に関しては具体策が提示されていない、あるいは議論の範囲が狭いなど、シナリオ側にミスマッチの要因がある場合

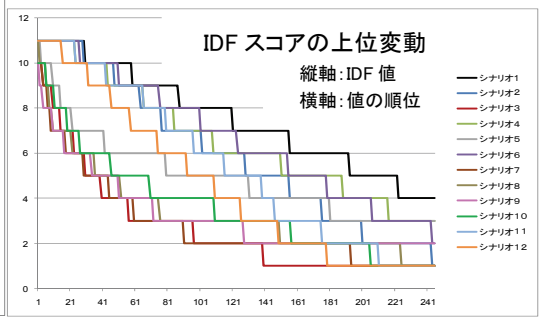
②デルファイ調査のトピックの選出に偏りがある、あるいは科学技術者の意識が希薄でトピックが挙げられていない、などデルファイ側にミスマッチの要因がある場合

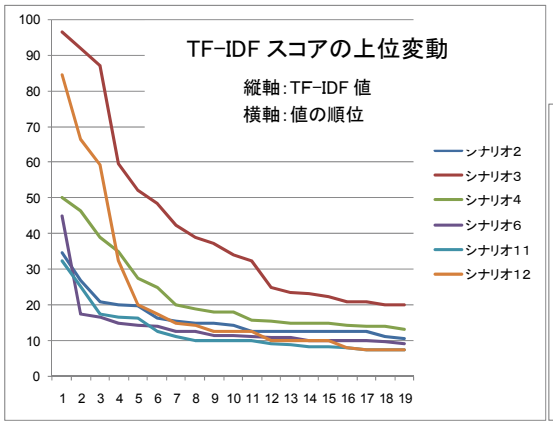
③「科学技術の成果の貢献度は低い領域」と考えられる場合、例えば、科学技術とは関係性の薄い社会的問題である、あるいは、現在の科学技術で考えられている範囲を大きく超えている、などの場合

これらを要因分析するため、各シナリオのキーワードの特徴を、TF－IDFとIDFのスコアの様子を図示する。なお、ここでは、シナリオのテーマから連想される既成概念を排除して比較するため、シナリオテーマ名を伏せて比較する。

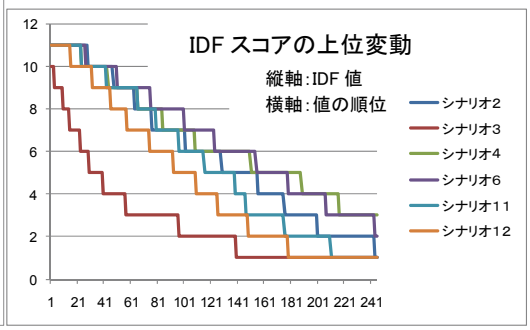


専門家グループによって書かれた
12 のシナリオの特徴





上記のうち、デルファイ調査のトピックとの
マッチングが少ないシナリオの特徴



○ シナリオ3は、非常に特徴的キーワード（他と非常に違った用語）を用い、他との共通用語が少ない（TF-IDF のスコアが高く IDF の高い言葉が少ない）テキストである。これは、内容がかなり狭い領域を扱っている可能性がある。
○ シナリオ12もかなり特徴的なキーワードが出ているが、その数が少なく、他との共通語は普通程度である。話題としては特徴がありそうだが、具体性に欠ける可能性がある。
○ シナリオ2、6、11は同じような特徴を持っており、その特徴はシナリオ3とは対照的である（TF-IDF のスコアが低く、IDF の高い言葉が比較的多い）。これらは、特異性の少ない文章であり、一般的な話で終わっている可能性がある。
○ シナリオ4はさほど特異なテキストではない。デルファイ調査側にトピックが少なかったことがマッチングのとれない理由ではないかと考えられる。もう少し将来イメージを具体化した形での研究開発が議論されるとよいということかもしれない。

上記の分析から、シナリオ側に具体策が提示されていないなどの要因がある場合が多かった。デルファイ調査側のトピックの不足が原因でマッチングがとれないのは農林水産業の総合産業化というテーマであり、このような領域の推進には、社会への成果還元をより重要視する形で研究開発を進めるような意識向上が効果的かもしれない。

4、結言

潜在意味分析を用いて、恣意性を排除した形で、科学技術予測のテキストから、科学技術政策の方向性を議論するためのいくつかの分析結果を導き出すことを試みた。今後の世界の科学技術政策やイノベーション政策において、「社会のための科学」が重要視されるようになると考えられるが、そのような議論の前提になるものとして、本研究のような方法の開拓の必要性が増していくのではないかと考えられる。

参考文献
1） 例えば、Bryan S. Cofffman, Weak Signal Research (1997)
2） 奥和田、白井、小関、“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”、2E09、研究技術・計画学会第22回年次学術大会(2007.10)
3） 例えば、James Surowiecki, The Wisdom of Crows (2005)（ 小高尚子訳、「みんなの意見」は案外正しい (2006)）
4） 例えば、欧州委員会第 7 次研究枠組み計画(FP7)プログラム、iKnow Project
5） 科学技術政策研究所、「将来社会を支える科学技術の予測調査」、NISTEP REPORT No.140,141,142 (2010.6)