STI Horizon 2016 Vol.2 No.4
http://doi.org/10.15108/stih.00056

ほらいずん

新たな予測活動の展開に向けてⅡ ー IARPA FUSE プログラムにみるホライズン・スキャニングの新手法ー

科学技術予測センター 客員研究官 七丈 直弘
（東京工科大学コンピュータサイエンス学部 教授）

【概　要】

近年、科学技術の急速な変化と社会課題の複雑化により、これまで以上に科学技術予測が困難となってきた。そこで、近年急速に発展を遂げている計算統計学、機械学習等の IT 技術を科学技術予測活動へ活用する試みが行われ始めてきた。本稿では、政府による科学技術予測活動への IT 技術活用の先進的事例として、米国 IARPA-FUSE プログラムの取組を紹介する。同プログラムでは、大学・シンクタンクからなるコンソーシアムに対して委託研究を行い、論文・特許・ニュース等の書誌情報・全文データを基に、ネットワーク解析、自然言語解析を行った結果を基に機械学習を実施し、科学研究における新興領域の特定が短期間で可能とした。日本でも同種の取組を通じ、より即時的な予測活動の実施が待望される。

1. はじめに

グローバル化や多極化が進む現在世界では、急速に変化・発展を遂げる社会環境や科学技術の変化の兆しを捉えて、迅速に政策に反映することが求められている。このような状況下で、科学技術政策の立案上、ホライズン・スキャニングに関心が持たれている。

ホライズン・スキャニングは特定の対象についての「その時点での考え方や計画に対する、潜在的脅威、可能性、あるいは将来の発展方向性の体系的評価」であり、科学技術政策への貢献を考えるならば、社会情勢や科学技術の情勢やその変化を広範に把握し、体系的評価を行う必要があり、膨大な作業が必要となる。しかし、近年の IT 技術の進展により、従来専門家の評価に任せるしかなかった上記の作業の一部を、IT の力を活用して遂行できる可能性が見えてきた。

自然言語処理に関する研究開発が進み、科学技術予測への適用が期待されるようになってきた 2011 年、米国情報高等研究開発活動（IARPA：Intelligence Advanced Research Projects Activity, 以降 IARPA と呼ぶ）では 5 か年の研究開発プログラムとして Foresight and Understanding from Scientific Exposition[1)]（以後、FUSE と呼ぶ）が開始された。FUSE は 2015 年に終了し、その成果の一部はジャーナル論文や学位論文として出版されており、その総数は 111 件に及ぶ[注 1]。また、FUSE の成果はベンチャー企業によって商用のホライズン・スキャニング事業として展開されてもいる。ただし、プログラム全体の成果は公表されていない。

本稿では、FUSE の概要と、その成果として出版された学術論文から分かる FUSE の成果について概観し、日本での科学技術予測活動での IT 技術活用の方向性について考察する。

2. FUSE プログラムとは

米国国家情報長官室は米国の情報機関を統括する部署である。同時多発テロ事件の後、米国の情報機関の間の連携不足を解消し、テロ予防を強化するために 2005 年に設置された。様々なインテリジェンス活動

注 1　2016 年 10 月 3 日時点での Google Scholar による検索結果を基にしている。謝辞に FUSE プログラムを研究資金とすることを示す ID が含まれているものを検索対象とした。

支援の一環として、情報技術を活用したインテリジェンス活動の支援も行っている。インテリジェンス活動支援を目的とした、公募型の研究開発プログラムにIARPAがある。米国国防高等研究開発局(DARPA)の情報機関版という位置付けであり、DARPA同様、プログラムマネージャーが3〜5年の任期の間に情報活動のニーズに基づく研究開発プログラムを立ち上げ、ハイリスク・ハイリターンな(不確実性が高いが、得られる利益が大きい)プログラムを行っている。2016年10月の時点でアクティブなプログラムは32個あり、その一つがFUSEである。2011年に開始されたFUSEは、科学・技術文献や特許などにみられるような公知の情報を基に、新技術の急速な出現を体系的かつ連続的な判断の自動化手法の開発を目標に掲げた[注2]。公募の結果、BAE Systems、コロンビア大学、Raytheon BNN Technologies Corporation、SRIインターナショナルによって率いられるチームと契約が行われ、2015年までの約5年間の間に研究が行われた。

同プログラムのゴールは「新技術の出現によって引き起こされる社会へのインパクトを軽減するために、複数の技術分野の融合によって生ずる新しい科学技術のコンセプトの出現を、主に科学文献の自動分析を通じ、信頼性のある方法で、早期に検出すること」とされ、その実現の方法として以下の項目が掲げられた:

- 英語及び中国語の科学文献と特許を自動要約し、数年内に出現すると予想されるコンセプトのパターンを発見する。
- 数千に及ぶ科学技術領域を可視化し、どのように分野が相互に影響を及ぼしているのかを時系列で把握する。

以上をゴールとして、以下の技術の開発・実装・試験を行う:

- 新技術出現の理論
- 技術用語抽出
- 新たな特徴技術(コミュニティ検出、研究トピックの持続性分析、センチメント分析を加味した引用分析、技術用語分類)
- 技術出現の指標
- 複数指標の複合利用と予測モデル
- 予測結果の説明可能性

これらの技術を総合し、2〜5年後の予測を行った結果、1〜20%の専門用語の盛り上がりを予測可能となったという。統合された技術の詳細は開示されていないが、公刊された論文を基にその概要を描像したい。

3. 主要な論文の概要

論文1. Boyack, K. W., et al. (2014). "Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science." Journal of Engineering and Technology Management 32: 147-159.[2)]

論文の被引用情報を基に、ブームとなっている研究領域の特定と、その時間変化について分析を行っている。論文データベースScopusに含まれる文献情報を年ごとに分け、共引用分析によりクラスタリングを行うことで、マイクロ・コミュニティ(MC)と呼ばれる研究分野の同定を行った[注3]。次に、年ごとに生成されたMCごとに、そこに含まれる論文が引用する文献がどれだけ重複するかを全ての組合せに対して算出することで、MCの年次変化を見た。MCの状態変化としては、生成、維持、分離、融合、消滅が存在する。このような状態変化をナノテクノロジー分野に対して分析した。その結果、MCはその生成後、翌年維持されるのは32%しかなく、翌年まで維持されたMCの58%は3年後まで継続し、9年間継続したMCの90%は更に翌年まで継続する、などの状況が判明した。

論文2. Breitzman, A. and P. Thomas (2015). "The Emerging Clusters Model: A tool for identifying emerging technologies across multiple patent systems." Research Policy 44(1): 195-205.[3)]

米国、欧州、中国、ドイツの特許データを基にした分析から、新技術の出現を予測するための手法(Emerging Clusters Model)が提案された。より多くの特許から引用を受ける特許をホット・パテントと呼び、年ごとにトップ5%の被引用数を持つ特許として定義した。これらの中には内容的に類似したものが多く含まれるため、被引用関係によってクラスタリングが行われた。その結果得られたホット・パテント・クラスタを引用する特許を次世代クラスタと定義した。

注2 原文では“seeks to develop automated methods that aid in the systematic, continuous, and comprehensive assessment of technical emergence using publicly available information found in published scientific, technical and patent literature.”となっている。

注3 このマイクロ・コミュニティは当研究所が作成しているサイエンスマップと類似した概念である。サイエンスマップは被引用数におけるトップ1%の論文間の共引用関係であるのに対し、マイクロ・コミュニティでは全論文に対する共引用関係を用いている点が主要な違いである。

次世代クラスタに対して、公的セクタの割合（クラスタに属する特許の発明者に、政府、大学、非営利団体が含まれる割合）、サイエンス指数（特許における非特許文献引用の数を、それが属するパテントクラスにおける非特許文献引用の平均値で除したもの）、オリジナリティ指数（当該特許が引用する文献がより多くのパテントクラスに分散しているか）、引用指数（当該クラスタに属する全特許の特許引用数平均の、当該年の全出願特許における特許引用数平均との比）を用いたスコアリングが行われ、過去の事例の対比から、このスコアがより高い次世代クラスタが将来ホット・パテント・クラスタとなる確率が高いことが示された。なお、上記のスコアリングにおけるパラメータは、NIST Advanced Technology Program（ATP）の成果である特許（これらは過去の分析により、他の特許よりも被引用数が高いことが知られる）に対して高いスコアが付与されるようにパラメータがチューニングされた。

論文 3. McKeown, K., et al. (2016). "Predicting the impact of scientific concepts using full-text features." Journal of the Association for Information Science and Technology: (to appear).[4)]

科学における特定の概念のインパクトの予測を行うシステムを提案している。ここでの概念とは論文全文データに表れる技術用語であったり、科学分野やカテゴリなどといった論文書誌情報中に出現する情報であったりする。与えられた概念の将来のインパクトを予測するために、書誌情報を基に分析した情報、論文の全文データから抽出した情報が用いられる。書誌情報から生成される情報としては、概念に属する論文を執筆する著者や、その所属、あるいは論文共著ネットワークにおける指標（中心性、クラスタリング、次数相関）などが含まれる。論文全文データから抽出される情報としては、固有表現（アルゴリズム、データセット、遺伝子、ウイルスなどといった各種科学概念）、研究資金提供機関、新規性の言及（新手法の提案なのか、既存手法の改良なのか、あるいは新しいデータに基づくものか、既存のデータに基づくものか）などが含まれる。また、概念がどのようなコンテキストで言及されているかも検出している。論文構造の解析から、各固有表現が「研究の目的」「自らの過去の研究」「研究の背景」「対比」「基礎」「その他」の一つに分類される。さらに引用文献のセンチメントの同定も行った。本文中に出現する引用文献への言及の仕方を「肯定的」「否定的」「中立的」のいずれかに分類している。以上のような多様な情報は時系列で表現され、最終的に時系列分析を行い、各種のバイアス除去を行った結果、予測結果が提示された。

論文 4. Wolcott, H. N., et al. (2016). "Modeling time-dependent and -independent indicators to facilitate identification of breakthrough research papers." Scientometrics 107(2): 807-817.[5)]

論文被引用数の分布は大きくゆがんでおり、ごく一部の論文が多くの被引用数を集める傾向にある。極端に多くの被引用数を集める論文をブレークスルー論文として定義し、その論文の各種属性情報（引用文献、被引用文献、ページ数、著者数、著者の所属国数、著者の所属組織数、助成機関、著者の研究履歴、論文共著ネットワークにおける中心性など）を基に、その論文が将来ブレークスルー論文となるかを出版から半年内に判定するアルゴリズムの開発が目標とされた。具体的には、がん研究を対象とし、米国がん研究学会（AACR）及び米国臨床腫瘍学会 (ASCO) が高評価した高被引用論文と、Nature Medicine が 2011 年に出版したがん研究特集号で言及された高被引用論文の計 283 本をブレークスルー論文として抽出し、これと対照集団（腫瘍学あるいは医学の分野で出版された論文 2,500 件）の 2 群間の属性の相違を機械学習によって分析した。

論文 5. Babko-Malaya, O., et al. (2013). "Modeling Debate within a Scientific Community." in International Conference on Social Intelligence and Technology (SOCIETY).[6)]

新しい科学分野の発生を同定しその推移を観察するため、科学コミュニティの中でのディベート構造が分析された。発展しつつある新しい科学分野に対しては、意見の不一致や不確実性が多い。本研究ではディベートを以下の 3 種に分類している。(i) 非整合型（提示する事実に既往論文との不一致がある）、(ii) 能動的な不一致型（能動的に他の研究と自らの研究を対照させる）、(iii) 内容不確実性型（その他）の 3 種に分類した。対象とした文献は、Web of Science, PubMed Central, Lexis-Nexis Patent data に及ぶ。ディベートの認識は、文献やコンセプト間の対比を抽出することによって行われる。さらにディベートは、その文献やコンセプトが属する科学分野、表出した媒体の特性（レビュー論文か、通常の論文か）、出現した位置（イントロダクション、バックグラウンド、手法など）によって分類された。専門家によって作成された、1981～2000 年までの間にディベートが存在した科学分野のリスト（常温核融合、メタマテリアル、組織工学など）とその期間を作成し、同じ分野を対象として上述のアルゴリズムによってディベートの有無を判定した結果、再現率 0.97、適合率 0.8 となった。

4. FUSE プログラムの成果

プログラム全体を通じて、基礎的内容から応用的内容まで幅広く成果が出ていることが分かった。概要は提示していないが、自然言語処理の基礎技術[7,8]に属する成果や、論文 1[2]のような計量書誌学に属する実験成果、論文 5[6]のような応用言語学に属する成果もあれば、論文 2〜4[3〜5]のように他の成果を応用してブレークスルーの判別を行うという応用的内容もある。このように、基礎から応用と成果の性質は異なるもの同士が、ミッション達成のために、相互に密接につながっている。これは、プログラムディレクターがプログラム実施に先立って、目標達成に用いる要素技術とその組合せをプランニングしていたことを推測させる。また、最終的な目標達成においては複数のチーム（論文 2 は 1790 Analytics、論文 3 はコロンビア大・ミシガン大を中心とした複合体、論文 4 は Thomson Reuter と ÜberResearch）に異なる手法で取り組ませるなどしており、コンペティションによってより良い手法を選ぼうとしたのではないかと想像できる。これらの成果のうち、SRI International が受託していた分について、その成果は論文として公開されていないものの、Meta というスタートアップ企業がライセンスを受け、商用ホライズン・スキャニングサービスとして既にサービス提供を開始している[9]。

5. 日本でのホライズン・スキャニングの実施に向けた課題

FUSE プログラムの成果全体は公開されていないものの、本稿が引用した論文を解析することで、同等のシステムを作ることは不可能ではない。引用した文献の多くには、各アルゴリズムのパフォーマンスも記載されていることから、どれかをスターティングポイントとして開発するのが一つの選択肢となる。ただし、世界の学術研究・特許のトレンドから注目すべき変化を抽出するのであれば、英語文献だけでも良いが、日本の科学のトレンドや、日本政府の科学技術政策との関連を調べるには、日本語と英語との間の固有表現や概念に関するコンコーダンス（対応表）が必要となり困難性が増す。逆に、日本語で得られる情報だけに特化するならば、コンコーダンスの問題を避けて、日本における科学分野の隆盛を予測することも可能である。さらに、予算の面では、FUSE プログラムの総予算は明らかにされていないものの、参加企業の一つである BBN 社は $1.7M の契約をしたと表明[1]しており、総額は 10 億円を超えている公算もある。日本での実施には、用途を絞ることでコストを抑える、あるいは国際連携を視野に入れることが求められるだろう。

参考文献

1) Rahmer, R. *Foresight and Understanding from Scientific Exposition (FUSE)*. 2016 [cited 2016 10/12]; Available from：https://www.iarpa.gov/index.php/research-programs/fuse
2) Boyack, K.W., et al., *Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science*. Journal of Engineering and Technology Management, 2014. 32: p. 147-159.
3) Breitzman, A. and P. Thomas, *The Emerging Clusters Model: A tool for identifying emerging technologies across multiple patent systems*. Research Policy, 2015. 44(1): p. 195-205.
4) McKeown, K., et al., *Predicting the impact of scientific concepts using full-text features*. Journal of the Association for Information Science and Technology, 2016: (to appear).
5) Wolcott, H.N., et al., *Modeling time-dependent and -independent indicators to facilitate identification of breakthrough research papers*. Scientometrics, 2016. 107(2): p. 807-817.
6) Babko-Malaya, O., et al., *Modeling Debate within a Scientific Community*. in International Conference on Social Intelligence and Technology (SOCIETY), 2013.
7) Wick, M., et al., *A joint model for discovering and linking entities*, in *Proceedings of the 2013 workshop on Automated knowledge base construction*. 2013, ACM: San Francisco, California, USA. p. 67-72.
8) Singh, S., et al., *Joint inference of entities, relations, and coreference*, in *Proceedings of the 2013 workshop on Automated knowledge base construction*. 2013, ACM: San Francisco, California, USA. p. 1-6.
9) SRI International. *Meta and SRI International Announce Agreement to Bring IARPA FUSE Predictive Intelligence to the Scientific World*. 2016 [cited 2016 10/12]; Available from：https://www.sri.com/newsroom/press-releases/meta-and-sri-international-announce-agreement-bring-iarpa-fuse-predictive