２Ｃ０５

Mapping Science
～文書ベクトルを用いた科学技術マップの作成と萌芽領域の抽出～

○川村 隆浩，渡邊 勝太郎，松本 尚也，江上 周作，治部 眞里（JST 情報分析室）

1. はじめに

1965 年，Price らが科学的な手法による科学技術の調査・研究を提唱して以来[1]，科学技術論文や特許等の関係性を把握するべく，さまざまな科学技術マップ（Map of Science）が作られてきた．これら科学技術マップは，科学技術政策やファンディングに関する検討において，種々の Scientometrics と共に現在なくてはならないツールとなっている．しかしながら，論文や特許間の関係性を導くための手法は引用・共引用分析に基づくことが多く，結果としてファンディングプロジェクト情報や引用が十分でない最新の論文等をマップ上に表すことが難しかった．

そこで我々は，昨今，進展が著しいニューラルネットワークを用いた統計的な自然言語処理技術を用いて，プロジェクト情報や論文抄録など自然言語で書かれたプレーンテキストを多次元ベクトルに変換し，ベクトル間の定量的な距離から内容的な類似性を算出することで，ファンディングプロジェクトと最新の論文等を同時に表すことのできる独自のマップを構築した. 更に, IEEE 論文および米国 National Science Foundation のプロジェクト情報約 30 万件から実際にマップを描き，それらの時間的な変化を辿ることで科学研究における萌芽領域が発展していく過程に関していくつかの事例を抽出した．

以下, 2 章にて文書ベクトル化手法について簡単に説明し, 3 章にて我々が開発した Mapping Science ツールを紹介する. 更に, 4 章にて Internet of Things 等に関する研究領域の変遷をツール上で確認し，最後に 5 章にてまとめと今後の課題を示す.

2. 情報エントロピーを用いた文書ベクトル化手法の提案

従来までも論文の内容的類似性に基づくマップはいくつか提案されている[2,3]．しかし，いずれも TF･IDF や Latent Dirichlet Allocation など論文内に含まれる語の集合（Bag of Word）間の類似性に基づく手法であり，文章のコンテキスト（文脈や語順）を考慮した手法ではなかった．そこで我々は，2013 年に Google が発表したニューラルネットワークを用いた単語ベクトル化技術 word2vec[4]，および word2vec に基づく文書ベクトル化技術 paragraph vector[5]を用いてプロジェクト情報や論文抄録などの自然文から多次元ベクトルに変換することを試みた．しかしながら，paragraph vector をそのまま使用するだけでは，十分な精度が得られなかった．原因としては，わずかな言葉遣いの違いによって同義語でも異なる単語ベクトルが構築され，結果的にほぼ同じ内容の文書であっても異なる文書ベクトルが生成されてしまうケースや，反対に技術的な観点では異なる内容の文書でも科学技術用語ではない一般語の共通性によって近しい文書ベクトルが生成されてしまうケースが散見された．

そこで技術的な用語以外を排除しつつ, かつ, 技術的に同義な概念を表す同義語を集約するため, JST が 1975 年より整備してきた科学技術用語シソーラス Linked Data 版[6]（以下，JST シソーラス）を参照しつつ，文書ベクトルを構築する手法を開発した[7,8]．文書ベクトル化の処理を図 1 に示す．具体的には，単語ベクトル空間内において類似の単語群はクラスタを構成することから，一定の意味を持つ単語ベクトルの空間内での広がり（超球面）が，意味的な多様性を表す指標であるシャノンの情報エントロピー[9]に比例すること仮定した（実際に高エントロピーの範囲において，エントロピーとベクトル間の距離に相関があることを確認している）．そこで，JST シソーラスに含まれる約 2 万の概念 $C(w)$（1 つの上位語と 1 つ以上の下位語から成る）の情報エントロピー$H(C)$を計算し，エントロピーの大きさに

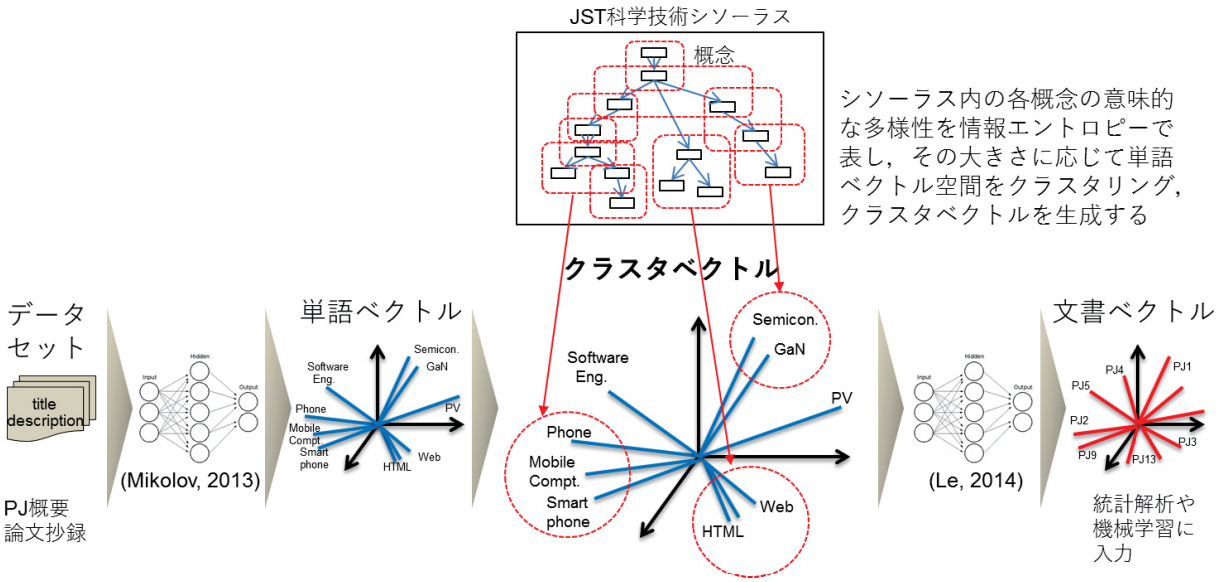

図1. 文書ベクトル化処理の流れ

基づいて単語ベクトル w のクラスタリング Cl を行った．

$$Cl(w_k)=\begin{cases} Cl(w_i) & \left(\frac{H(C(w_i))}{H(C(w_j))} > \frac{\|w_k - w_i\|}{\|w_k - w_j\|}\right) \\ Cl(w_j) & (\text{otherwise}) \end{cases}$$

これによって似た意味を持つ単語ベクトル群は，科学技術的に重要な概念を表すクラスタベクトル（クラスタの重心に設定）にまとめられる．そして，このクラスタベクトルから文書ベクトルを生成することで，科学技術的な類似性を強調した文書ベクトルを構築した．これはシソーラスにおける明示的なセマンティクスと文書ベクトルによって表された暗黙的なセマンティクスを融合させるアプローチであるとも言える．

図 2 に簡単な評価結果を示す．現状，本評価に適した Gold Standard データは見つかっていないため，ランダムに抽出した 1000 のプロジェクト情報の内，一定割合をランダムに別のプロジェクト情報に置き換えた人工データを作成し，元のプロジェクト情報の文書ベクトルとの cosine 類似度を測った．図中，横軸は置換した文数の割合を縦軸に cosine 類似度を表す．結果として，内容的な類似性と文書ベクトルの cosine 類似度の間には明らかな相関 R^2=0.88 があることを確認できた．より詳細な手法，評価については文献[7,8]を参照してほしい．

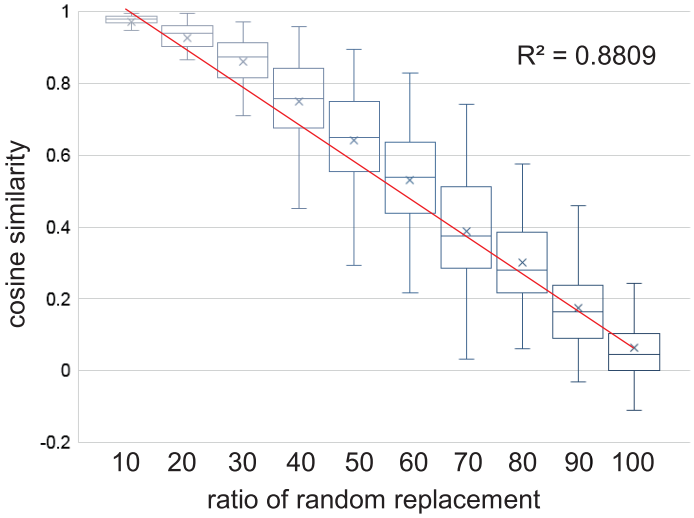

図 2. 内容的類似性と cosine 類似度の相関

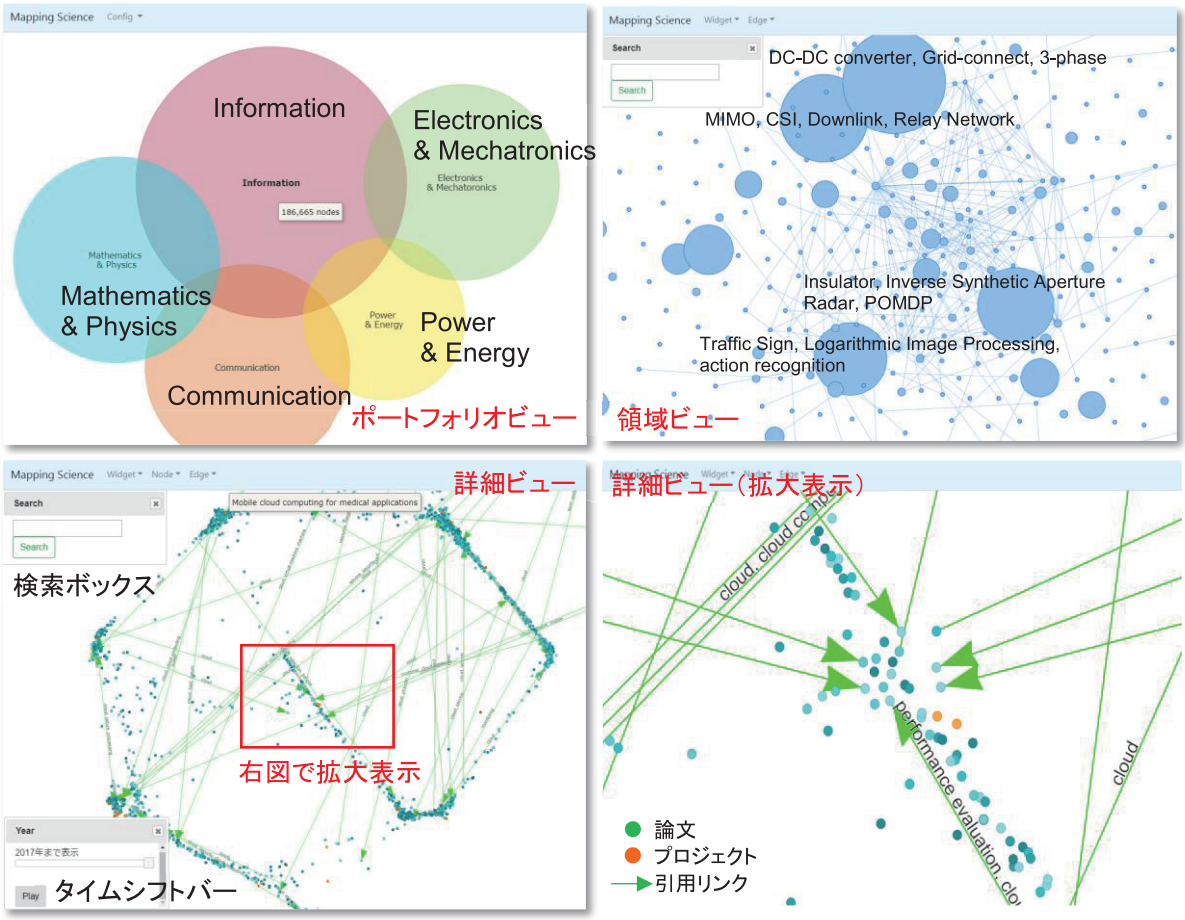

図 3. Mapping Science の画面イメージ

3. Mapping Science ツール

本章では，今回我々が開発した文書ベクトルに基づく Mapping Science ツール（図 3）について説明する．インターフェイスは大きくポートフォリオビュー，領域ビュー，詳細ビューの 3 つに分けている．

ポートフォリオビューは事前に設定した検索式に沿って対象データ・セットを全文検索し，複数分野に分けたものである．円の大きさは含まれる論文・プロジェクトの数に対応している．

領域ビューはポートフォリオビューにおけるいずれかの分野をクリックすると開くビューであり，当該分野内に含まれる全論文・プロジェクトを infomap 法のコミュニティ検出（cosine 類似度 0.5 以上，1 ノード最大 30 エッジ）でクラスタリングしたものである．NISTEP サイエンスマップ（http://www.nistep.go.jp/archives/30357）に相当するビューであり，分野内の技術を概観するためのものである．尚，BM25 手法を用いて領域毎に特徴語 10 語を抽出し，ラベリングしている．領域間の距離は含まれる論文・プロジェクトの重心間の距離である．

詳細ビューは領域ビューにおけるいずれかの領域をクリックすると開くビューであり，ノードは 1 論文または 1 プロジェクトに相当する．Open Ord（www.sandia.gov/~smartin/presentations/OpenOrd.pdf）を用いて edge-weighted でレイアウトされており，ノード間の距離は可能な範囲で cosine 類似度に比例している．更に，論文間の直接引用関係（cite→cited）をエッジとして表し，2 論文間の特徴語をエッジラベルとして表示した．ノードをクリックすると該当する論文・プロジェクトの詳細情報を表示する．また，画面下部には含まれる論文・プロジェクトの統計情報（被引用数や IF など）を表示する．

それぞれのビューでは，左上の検索ボックスから現在のビューに含まれる論文を全文検索し，該当するノードをハイライト表示することができる．更に，詳細ビューにおいては発行年毎の論文，プロジェクトの累積的な増加をアニメーション表示で確認することもできる（次章にて例を示す）．

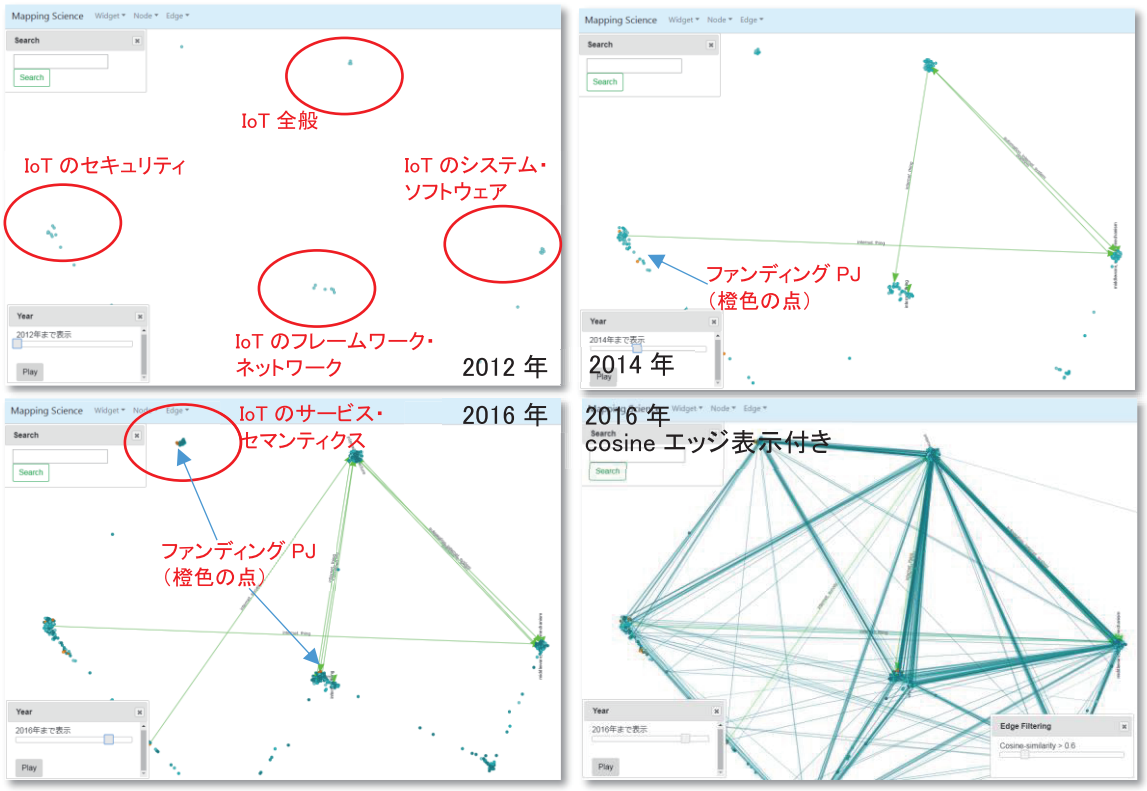

図 4.萌芽領域の形成例（IoT）

4. 萌芽領域抽出に関する事例研究

今回, 2012 年～2016 年に Scopus に収録された IEEE 論文誌論文, および国際会議論文 266,774 編, さらに同期間の NSF における 3 分野（Computer & Information Science & Engineering, Mathematical & Physical Sciences, Engineering）34,192 件のプロジェクト情報を文書ベクトル化し, Mapping Science を用いてマップを描画した. マップ上では, いくつかの萌芽的な研究領域が形成されていく様子（ネットワーク構造の時系列的な変化）を確認することができた. 本章では, Internet of Things (IoT) に関する領域, および Brain Computer Interface (BCI)に関する領域について述べる.

図 4 は, Information 分野内の IoT に関する領域の詳細ビューであり, 2016 年時点で 574 のノードを含んでいる. この領域の 2012 年からの 5 年分と最後に 2016 年のビュー上に一定の cosine 類似度を表すエッジを表示させたものである. 2012 年時点では, IoT に関する主にフレームワーク・ネットワークやシステム・ソフトウェア, セキュリティなどに関して 4 つの島（ノードが密集した場所）を辛うじて見つけることができる（主なテーマは目視で抽出した）. 2014 年になると, 2013 年にいち早くファンド（橙色の点）が投下されたセキュリティに続いてフレームワーク・ネットワークの島にもファンドが投下され, 各島の論文数が増加している. 同時に異なる島（いわば研究コミュニティ）の研究者が互いの研究活動を認識し, 島間に相互引用（緑色の線）が引かれ始めている. 2016 年にはその流れが加速し, 島の巨大化, 密集化が進むと同時に相互引用数も増えている. 更に, 左上の IoT サービス・セマンティクスに関する島など, 当初 4 つの島以外も徐々に大きくなり, 中にはファンドを付けられることで論文数を大きく増やす島も出てきている. 最後に, 右下の図は 2016 年のビュー上に一定の cosine 類似度（0.6～0.7）を表すエッジを表示させたものである. 0.6～0.7 は一般に弱い類似性に相当しており, IoT 領域全体としては複数の特化したテーマに関する島が内容的にも引用関係的にも弱く繋がりながら, 各テーマを発展させつつ, IoT という領域を形作ってきたという経緯を読み取ることができる.

一方, 図 5 に BCI に関する領域の 2016 年の詳細ビューを示す. ここでは, 図中, 上部の島が医療や神経科学, ロボットや AR, 脳波などそれぞれ特化したテーマの論文を引用しながら発展してきたことが見て取れる. これにより BCI という領域は, 複数の異なる従来テーマから同時多発的にアプローチされ, 統合的に研究発展してきたと考えられる. このように時系列的に本マップを見ていくことで萌芽領域の形成過程を捉えられることが確認できた.

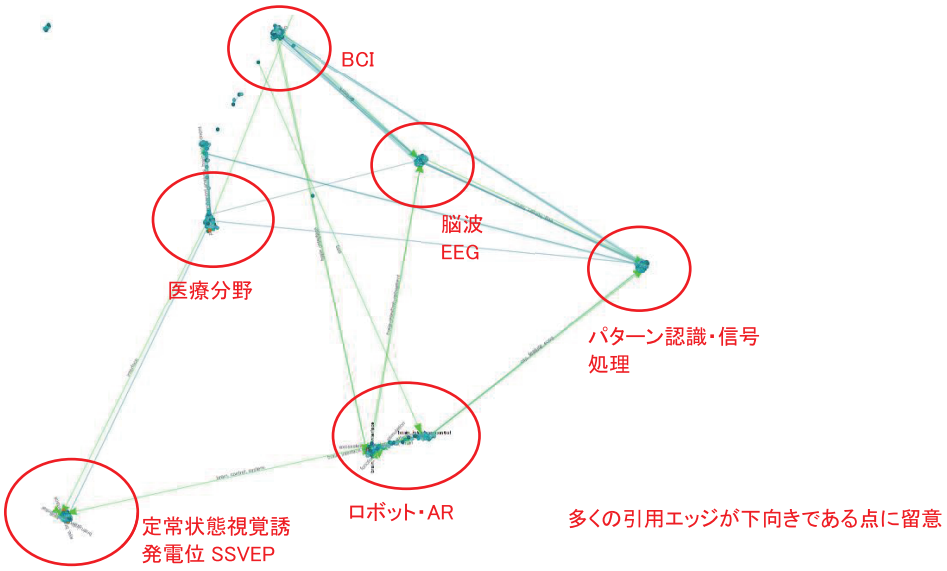

図 5.萌芽領域の形成例(BCI)

5. まとめと今後の課題

本研究では，引用分析が難しいファンディングプロジェクト情報や最新の論文を対象に，独自に開発した文書ベクトル化技術を適用し，研究内容の類似性に基づく科学技術マップを開発した．また，萌芽領域の発展の様子が構造変化として捉えられることを確認した．

今後は，そうしたネットワーク構造の時間的変化を数値的に捉えることを試みる．そして，それら特徴量に基いて今後の伸びてくるであろう萌芽領域を予測するため，統計処理や機械学習技術の適用を検討していく．また，データ・セットに特許情報を加えることや，引用分析に基づく従来マップとの比較も行っていきたい．更に，JST シソーラスと文書ベクトルを介して日本語で書かれたファンディング情報や論文と，英語による海外のファンディング情報，論文を重ね合わせることで，国内外のファンディング傾向の違いなどを明らかにしていきたい．

参考文献

[1] Price, D.J.D. Networks of scientific papers. Science. 1965, vol. 149, p. 510-515.
[2] Talley, E. M. et al. Database of NIH grants using machine-learned categories and graphical clustering. Nature Methods. 2011, vol. 9, p. 443–444.
[3] Wang, S.; Koopman, R. Clustering articles based on semantic similarity. Scientometrics. 2017, vol.111, no. 2, p. 1017–1031.
[4] Mikolov, T. et al. Distributed representations of words and phrases and their compositionality. 2013, In Proc. of NIPS 26, p. 3111–3119.
[5] Le, Q.; Mikolov, T. Distributed representations of sentences and documents. In Proc. of ICML 2014. 2014, vol. 32, p. 1188–1196.
[6] Kimura, T. et al. J-GLOBAL knowledge: Japan's largest linked data for science and technology. In Proc. of ISWC 2015. 2015.
[7] Kawamura, T. et al. Funding Map for Research Project Relationships using Paragraph Vectors, In Proc. of ISSI 2017. 2017.
[8] Kawamura, T. et al. Science Graph for characterizing the recent scientific landscape using Paragraph Vectors, In Proc. of K-Cap 2017. 2017.
[9] Shannon, C. A mathematical theory of communication. Bell System Technical Journal. 1948, vol. 27, p. 379–423.