研究内容の類似性に基づく科学技術マップの開発

Development of content-based map of science

国立研究開発法人科学技術振興機構　情報企画部主任調査員
川村　隆浩

株式会社東芝研究開発センター、カーネギーメロン大学客員研究員等を経て 2015 年より現職。電気通信大学客員准教授、大阪大学非常勤講師、法政大学非常勤講師を兼任。主にデータ分析、機械学習に従事。博士（工学）

国立研究開発法人科学技術振興機構　情報企画部システムエンジニア
江上　周作

2017 年より現職派遣。電気通信大学博士課程在籍。2018 年より日本学術振興会特別研究員（DC2）。ウェブ情報学の研究に従事。修士（工学）

国立研究開発法人科学技術振興機構　戦略研究推進部主査
渡邊　勝太郎

同機構情報企画部情報分析室を経て 2018 年より現職。修士（学際情報学）

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、既になくてはならないツールとなっている。さまざまな指標と共にマップを精査することによって、科学技術に関係する企業は今後予想される変化に気づくことができるかもしれない。また、既存の研究分野に基いて組織された研究所や大学は研究環境の変化を知ることができる。さらに、ファンディング機関や政策決定者はマップから定性的、定量的なメトリクスを得て、さまざまな分析を行い、知見を得ることができるだろう。しかしながら、引用・被引用分析に基づいて論文間の関係性を導く手法では、ファンディングプロジェクト情報や被引用情報が十分でない最新の論文や特許等をマップ上に表すことが難しかった。

そこで本論文では内容的な類似性に基づく科学技術マップの開発を目的とし、ニューラルネットワークを用いてテキスト情報を多次元ベクトルに変換してベクトル間の距離から内容的な類似性を算出する手法を提案し、既存手法よりも精度が良いことを示す。尚、本論文では先行発表[2, 3] に対して、クラスタリング手法の詳細やマップ分析機能の説明に加え、提案手法の有効性を検証するため既存の内容ベース手法との詳細な比較を行っている。

以下、2 章にて関連研究を示し、3 章にて独自の文書ベクトル化手法とその評価について説明する。さらに 4 章にて開発した科学技術マップを紹介する。最後に 5 章にてまとめと今後の課題を示す。

2 関連研究

Small らによる Maps of Science（http://mapofscience.com/）は、もっともよく知られた科学技術マップである。また、Borner らは Sci2 Tool[4] においてさまざまな分析ツールを提供している[5]。国内では、NISTEP によるサイエンスマップ（http://www.nistep.go.jp/research/science-and-technology-indicators-and-scientometrics/sciencemap）がしばしば活用されている。これらのサイト、ツールでは論文や特許間の類似性は主に引用・被引用情報を用いた共引用分析や直接引用分析、または書誌結合などに基いて計算されており、分野間の融合や協働をよく表すことができるが、被引用数が十分でない論文や、引用（参考文献）情報が付いていない（または十分ではない）プロジェクト情報や特許等には適用が難しい（但し、プロジェクト情報にはいずれ成果論文が登録され、それらが引用されることで対象となり得るだろう）。

一方、ファンディング機関や出版社はそれぞれ独自の分類体系を構築し、論文の内容に基いてタグ付けを行ってきた。一般に、1 論文に複数のタグが付けられるため、分野融合的な研究はそれらを見ることで判別できる。しかし、同じタグが付けられた論文同士であってそれらの類似性や関係性は測ることができない。ましてや複数のタグが付けられた分野融合的な論文が、既存のどの論文とどの程度似ているのかなどは判定できない。また、ファンディング機関や学術出版社、また特許における分類体系はそれぞれ異なるため、異なる機関・出版社の論文や特許間を比較することはできない。例えば、ACM（Association for Computing Machinery）タクソノミー（https://www.acm.org/publications/class-2012）が付与された論文を、Springer Nature 社の Springer Nature classification と比較するには別途対応付けが必要である。

そこで、論文の内容的類似性に基づくマップがこれまでもいくつか提案されている。例えば、pLSA（probabilistic Latent Semantic Analysis）[6] や LDA（Latent Dirichlet Allocation）[7] を用いた自動トピック抽出（分類）などが挙げられる。Griffiths らは、LDA を用いてあるトピックを表す 5 つの単語を抽出し、1 論文を複数のトピックの融合として表した[8]。これらの手法により、異なる機関・出版社間での論文比較は可能となったが、依然として同じタグが付けられた論文間の類似性や関係性は測ることができなかった。この問題に対して、NIH Visual Browser[9, 10] では、カルバック・ライブラー情報量[11] を用いて pLSA によって各トピックへの分類確率を融合し、プロジェクト間の類似性を計算している。しかしながら、この類似度は一定の規則に基づくトピック確率の融合であり、文のコンテキスト（文脈や語順）に基いて計算されたものとは言えない。他の研究も、いずれも TF-IDF（Term Frequency-Inverse Document Frequency）など論文内に含まれる語の集合（Bag-of-Words）間の類似性に基づく手法であり、文のコンテキストを考慮した手法ではなかった。Boyack らは内容的類似性に関して包括的な調査をしており[12]、大規模な医療文献セットをクラスタリングするために 9 つの類似性判定手法を比較している。その結果、タイトルと抄録文を対象とした BM25 手法[13] が他の TF-IDF や LSA、LDA に比べて優れていると結論づけている（但し、PubMed 固有の関連論文情報を使った手法を除く）。また、Waltman らもまた論文をクラスタリングする際の関連性を測る指標について比較を行っており[14]、やはり BM25 が他の内容ベースの関連性指標よりも正確であることを示している。そこで以下 3.2 節では提案手法と BM25 手法との比較を実施した。

これに対して、近年、自然言語処理研究においては単語および文書の分散表現として単語ベクトル、文書ベクトルが注目を集めている。Firth による単語の意味はコンテキストによって決定されるとする仮定[15]によれば、同様のコンテキストに現れる単語は同様の意味を持つと考えられる。そこで一般に、単語ベクトルはコーパス内のある単語 w の前後 c 語内に出現する語との共起頻度の行列として表される。その 1 つが Mikolov らによって提案された word2vec[16, 17] である。Word2vec とは、Skip-gram や Negative sampling によって改良された 2 層のニューラルネットワークであり、具体的には式(1)における目的関数 L（Likelihood）を最尤推定することによって求められる。T は単語の総数を表す。Word2vec では、ベクトル空間内で類似の意味を持つ単語は近い位置に配置される。

$$L = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \quad \text{式(1)}$$

一方で、Mikolov らは同様のニューラルネットワークを用いて任意の長さの文書をベクトル化した文書ベクトルも提案している[18]。単語ベクトルがコーパス全体を通して共通であるのに対して、文書ベクトルは文書毎に固有の仮想的な単語の単語ベクトルである。文書ベクトルのコンテキストは、当該文書内で一定の前後巾 c の語をスライディングさせることで求められる。

$$L = \sum_{t=1}^{T}\log p(w_t|w_{t-c},...,w_{t+c},d_i) \quad \text{式(2)}$$

式(2)において、d_i が単語 w_t を含む文書 i の文書ベクトルである。このため、文書ベクトルは文書毎に固有のトピックを表すと考えられ、コンテキストを考慮していることから LDA や pLSA などの Bag-of-Words モデルよりも文の内容をよく表すことができる。また、ベクトル化された文書は機械学習や統計解析パッケージに入力するのにも利便性が高く、異なる機関や出版社を跨いで論文やプロジェクト間の類似性を定量的に判定するのに役立つ。そこで、我々はプロジェクト情報や論文抄録などの自然文を文書ベクトルに変換することを試みた。

3 情報エントロピーを用いた文書ベクトル化

本章では、情報エントロピーを用いて改良した文書ベクトル化手法を紹介した後、文書間の内容的な類似性が正しくベクトル間の類似性として表されているかどうかについて評価する。

3.1 文書ベクトル化手法の提案

提案手法の導入に先立って、既存の文書ベクトル化手法[18]をファンディングプロジェクト情報に適用した。本研究では深層学習ライブラリ DL4J（https://deeplearning4j.org）を用いて文書ベクトルを作成した。データセットは、次章にて後述する米国 NSF（National Science Foundation）によるプロジェクト情報（およそ 10〜20 行からなるプロジェクト提案文）、約 3 万件である。各ハイパーパラメータは実験的に以下のように設定した。対象は最低 5 回以上出現した単語とし、次元数は 500、前後の単語巾 c は 10 とし、学習率と最低学習率はそれぞれ 0.025 と 0.0001 とした。また、学習には Adaptive Gradient algorithm を使用し、モデルは階層的ソフトマックスを用いた Distributed Memory を用いた。

しかしながら、既存の文書ベクトル化手法ではベクトル空間内でプロジェクト情報が散らばってしまい、何らかのテーマや分野によって意味のあるまとまりを構成しにくいことが分かった。多くのプロジェクトは少数のプロジェクトと僅かに類似性を見出すのみであり、例えば既存の分野と比較することで傾向を掴むことなどは難しかった。ベクトル空間内を詳細に見ていくと、このいわば unclustered 問題の原因としては、わずかな言葉遣いの違いによって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。事実、Mikolov ら[18]によれば多分類問題における分類精度は 50% 以下であったことが報告されている。

そこで、我々は文書ベクトルを構築する前に単語ベクトルを情報エントロピー[19]に基いてクラスタリングする手法を考案した。ベクトル空間内で類義語の単語ベクトルが集まり易いという性質は、単語の意味が一定の距離を持って空間内に広がっているということを意味している。この観測は、他の関連研究でも指摘されている[20]。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、科学技術用語シソーラス内の各概念の意味的な多様性に沿って単語ベクトルのクラスタを構成した。まず、JST 科学技術用語シソーラスおよび大規模辞書の Linked Data 版[21]（以下、JST シソーラス）から 1 つ以上の下位語を持つ上位語 19,685 語を抽出した。JST シソーラスは、JST が 1975 年より収集・管理してきた約 3600 万文献の索引語が階層的に整備されたものであり、計算機科学からバイオや土木工学まで 14 のカテゴリーに渡って 276,179 語の日英単語を含んでいる。また、Web 技術に関する国際標準化団体 W3C が規定する Simple Knowledge Organization System（skos）スキーマに従って、概念間の意味関係が skos:broader、

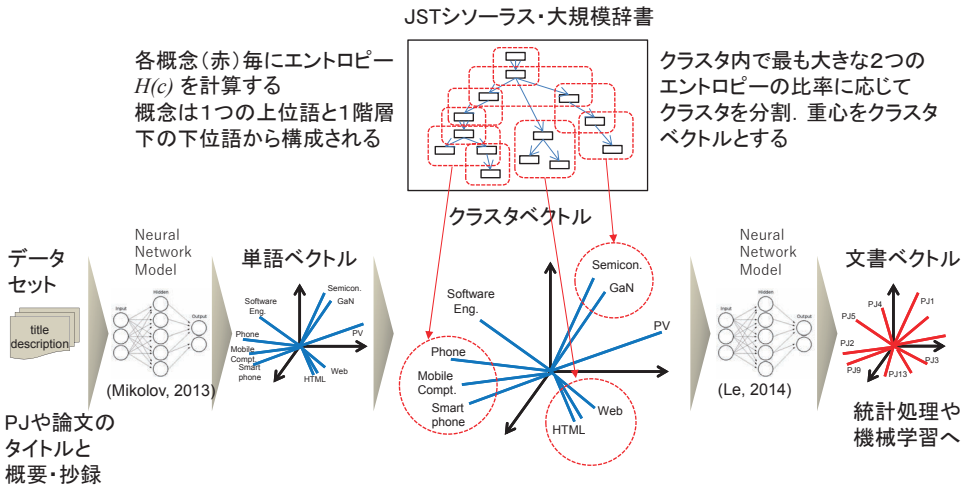

図１　クラスタベクトルからの文書ベクトルの構築

skos:narrower、skos:related などで定義されている。上位下位関係は主に包摂関係 is-a を表しているが、地理、体組織など一部においては部分全体関係 part-of を表している。JST シソーラスは、J-GLOBAL（http://jglobal.jst.go.jp/）サイトにて Web API で公開している。また、可視化ツールとしてシソーラスマップも公開されている。次に、データセット内における各概念の情報エントロピーを求めた。Shannon の情報エントロピーはある事象の情報量を表している。我々は先行研究[22] らを参考に概念の意味的な多様性を表す指標としてエントロピーを用いた。そして、エントロピーの大きさに沿って単語ベクトルのクラスタを構成し、同一クラスタ内の全ての単語ベクトルを 1 つのクラスタベクトルに集約し、これらに基いて文書ベクトルを構築した。一連の処理の流れを図 1 に示す。

以下、単語とはデータセット内の単語を、用語とはシソーラス内の語を指し、用語は上位語、下位語、および同義語の 3 つに分類される。また、概念とは上述の通り、1 つ以上の下位語を持つ上位語であり、図 2 中、赤で囲まれた範囲を指す。用語 T_i から構成されるシソーラスが与えられた場合、概念 C のエントロピーは上位語 T_0 とその下位語 $T_1 \cdots T_n$ の出現頻度を事象発生確率として計算される。用語 T_i の同義語 $S_{i0} \ldots S_{im}$ の出現頻度は、対応する概念に集約される。尚、同義語 S_{ij} は、用語 T_i の統制語を含むものとする。

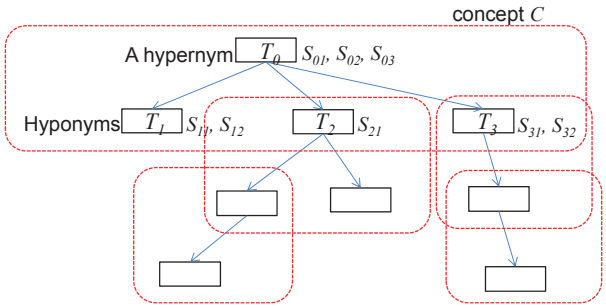

図２　シソーラス内における概念の定義

$$H(C) = -\sum_{i=0}^{n}\left(\sum_{j=0}^{m} p(S_{ij}|C)\cdot \log_2 \sum_{j=0}^{m} p(S_{ij}|C)\right) \quad \text{式(3)}$$

式(3)において、$p(S_{ij} \mid C)$ は、概念 C、用語 T_i における同義語 S_{ij} の確率を表す。シソーラス内の各概念について、データセットからエントロピー $H(C)$ を計算した。エントロピーは事象確率が等しくなればなるほど上昇し、反対に特定の事象のみが発生する場合に情報量は低くなりエントロピーは減少する。したがって、概念のエントロピーは、その概念を構成する上位語や下位語などがデータセット内でいずれも一定頻度をもって出現していれば上昇する。このことから、エントロピーの大きさは、概念の意味的な多様性を表している。さらに、エントロピーの大きさとその概念の単語ベクトル空間内での空間的大きさが一定程度比例していることを仮定し（予備実験では、概念のエントロピーは少なくともエントロピーが高い範囲においては概念内の用語間の最大ユークリッド距離と高い相関 $R = 0.602$ を示した）、我々は単語ベクトル空間を複数のクラスタに分割した。具体的には、式(4)に示した条件が満たされるように空間の分割を繰り返した。

$$Cl(w_k) = \begin{cases} Cl(w_i) & \left(\frac{H(C(w_i))}{H(C(w_j))} > \frac{\|w_k - w_i\|}{\|w_k - w_j\|}\right) \\ Cl(w_j) & (\text{otherwise}) \end{cases} \quad \text{式(4)}$$

この条件は、単語ベクトル $w_0 \ldots w_T$ は 1 クラスタ内の

もっとも高い２つの概念エントロピー $H(C(w_i))$、$H(C(w_j))$ の比に応じて２つのクラスタに分割されることを意味している。初期クラスタは空間全体とする。$C(w_i)$ と $C(w_j)$ は、それぞれ単語 w_i と w_j が属する概念を表す。単語 w_i、w_j は、その基底形がシソーラス内の用語と合致したデータセット内の単語である。シソーラス用語に該当しない単語のエントロピーは考慮されないが、全ての単語はいずれかの概念クラスタの要素となるため、タイトルや抄録内の単語がシソーラス用語に合致しない場合でも文書ベクトルは生成される。また、$Cl(w)$ は単語 w のベクトルが分類されるべきクラスタを意味する。

ベクトル空間はエントロピーの上位 1.5% に相当する 0.25 を下回るか、クラスタ内の単語数が 10 を下回るまで分割を行った。これらのパラメータは実験を通して設定した。結果として、66,830 の単語ベクトルから 1,260 のクラスタが生成され、クラスタ毎に全ベクトルの重心をクラスタベクトルに設定した。式(5)に示す通り、文書ベクトルは、単語ベクトルからではなく、これらクラスタベクトルから構築した。

$$L = \sum_{t=1}^{T} \log p(Cl(w_t)|Cl(w_{t-c}), ..., Cl(w_{t+c}), d_i) \quad \text{式(5)}$$

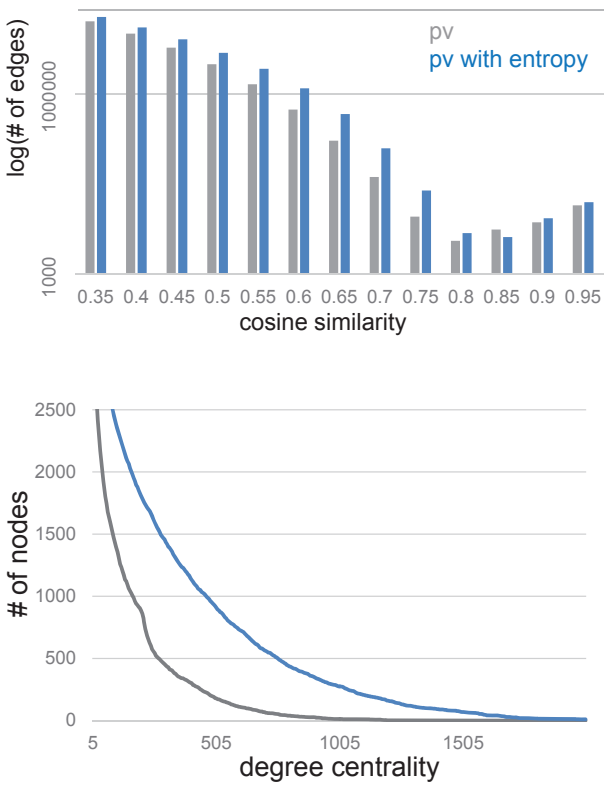

図３　文書ベクトルにおけるクラスタ形成度合いの比較

3.2　文書ベクトルの類似性に関する評価

次に、以下のデータセットから実際に構築したマップを対象に文書ベクトルの評価を行う。今回データセットとして、2012 年から 2016 年に発行された Scopus 収録の IEEE 国際会議論文 266,772 編のタイトルと抄録 2,290,743 文、および同期間の NSF プロジェクト 34,192 件のタイトルと概要 730,563 文を用いた。但し、IEEE において journal, transaction, symposium, workshops は含めていない。また、NSF の分野は IEEE に合わせて Computer & Information Science & Engineering、Mathematical & Physical Sciences、および Engineering とした。

まず、上述の unclustered 問題について、提案手法によるベクトル群は既存の文書ベクトル[18] に比べて複数のまとまりを形成していることを確認した。図３に、ノード（論文またはプロジェクト）間の cosine 類似度と該当するエッジ数の関係、および cosine 類似度 0.35 以上に限定した場合の各ノードの次数中心性（degree centrality）と該当するノード数の関係の比較を示す。結果として、提案手法はより高い cosine 類似度を持つエッジが多く、より高い次数中心性を持つノードも多いことが確認できた。これは科学技術的な観点から意義のある高エントロピー概念を文書ベクトルの主な構成要素としつつ、同時にシソーラスに定義されていない新規または未知の概念や同義語はもっとも近しいクラスタベクトルに集約させることで、文書ベクトルがよりまとまり、共通性が増したことが理由と思われる。いわば科学技術的な類似性を強調した文書ベクトルとなっている。また、クラスタ内の重心をクラスタベクトルとすることは、ベクトル空間内で互いの違いを明確にするため各クラスタベクトルをできる限り分離する効果がある。

また、文書ベクトルの cosine 類似度が元文書の内容的類似性を正しく表しているかどうかを２つの方法で確認した。著者らの知る限り、科学技術文書間の類似性を評価するための Gold Standard データは見つかっていない。そこで、まずサンプリング手法を用いて人手で評価した。cosine 類似度 0.5 以上の全ペア（2 つの論文またはプロジェクトのタイトルと抄録からなる）の中から 10、全体の分布におよそ等しくなるようにランダムに 100 ペアを抽出し、類似度を weak

(0.5≦*cos*.<0.67)、middle (0.67≦*cos*.<0.84)、strong (0.84≦*cos*.) の 3 段階に分けた。その上で、JST のメンバー 3 名によって各ペアの類似度を評価した。メンバーには、事前にマップの用途や評価例についての説明を与えている。また、各メンバーの専門知識はバイオサイエンスから認知科学、情報工学とさまざまであるが、同じ評価セットを与え、正解は 3 名の多数決で選んだ。

さらに、関連研究で述べた BM25 手法と提案手法との比較を行った。BM25 は TF-IDF に似た Bag-of-Words モデルの一種であり、検索エンジンにおいて検索語に近い文書をランク付けするために広く用いられている。単語 w_i と文書 d_j の類似度は以下の式(6)で計算される。

$$BM25\ Score(w_i, d_j) = \frac{f_{ij} \cdot (k_1 + 1)}{f_{ij} + k_1 \cdot \left(1 - b + b \cdot \frac{F_j}{F_{ave}}\right)} \log \frac{|D| - |D_i| + 0.5}{|D_i| + 0.5}$$

式(6)

f_{ij} は文書 d_j 内の単語 w_i の頻度であり、F_j は文書 d_j の単語数、F_{avg} は全文書セットにおける平均文書長、$|D|$ は全文書セットにおける文書数、$|D_i|$ は単語 w_i を含む文書の数、k_1 と b は定数であり、経験的にそれぞれ 2.0、0.75 と設定されている。式の後半は単語 w_i の IDF に相当している。提案手法と比較するため、JST シソーラスに含まれる用語の BM25 値を並べたベクトルを構成した。全文書セット内で頻度 0 の単語の次元は削減し、次元数は 3722 であった。

評価結果を表 1 に示す。提案手法における F1 値の平均は 78% であった。誤判定された例としては、異なる意味を持つ同じ略語を持つ無関係のペアや、まだ数は少ないが近年注目されている単語を含むより関係性の強いペアなどが存在した。尚、BM25 手法による F1 値の平均は 20% に留まった。既存の文書ベクトル化手法による F1 値の平均は 21% であった。また、メンバー 3 名の一致度は Fleiss' Kappa で 0.29 (fair agreement) であった。更に、評価結果が順序尺度であることから {強、中、弱、関係なし} の 4 段階を {3, 2, 1, 0} に変換して Ebel の級内相関係数を求めたところ、一致度 0.59、95% 信頼区間でも下限 0.49 であり、moderate agreement であった。

表 1　サンプリングによる類似度評価 (%)

手法	提案手法			BM25 手法		
類似度	*weak*	*middle*	*strong*	*weak*	*middle*	*strong*
適合率	77.5	83.3	100.0	65.0	60.0	100.0
再現率	98.6	33.3	83.3	18.6	20.0	66.7
F1 値	86.8	47.6	90.9	28.9	30.0	80.0

但し、本手法は利用するシソーラスの正しさや更新頻度に依存する点に留意が必要である。今回利用した JST シソーラスは少なくとも四半期に一度は更新されており、維持更新等の継続的運営に対する信頼度や収録範囲からも科学技術用語シソーラスとして現状ではもっとも適したリソースと考えられる。更新にあたっては、一定期間毎に区切った論文集合から特徴語句を抽出し、それらの経時変化から追加すべき語句を自動的に抽出するといった作業者に依存しないプロセスでも運用されている。しかし、シソーラスとしての正しさや十分性の検証[23] は十分ではなく、今後の課題としたい。また、本手法は科学技術シソーラスを活用しているため、対象が科学技術文献に限定される点も留意が必要である。

最後に、一定割合を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。ランダムに抽出した 1,000 の論文またはプロジェクトを対象に、その内の 10, 20, ..., 100% をランダムに選択、やはりランダムに選んだ別の論文またはプロジェクトの中からランダムに選んだ同量の文で置換した。そして、元文書から生成したベクトルと置換した文書から生成したベクトルとの cosine 類似度を測った。図 4 に置換した文数の割合と cosine 類似度の関係を示す。結果として、文書の相違度と文書ベクトルの 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。一方で、BM25 手法は R^2=0.76 に留まった。これは、BM25 手法では類似度の範囲が狭くなり、提案手法よりも解像

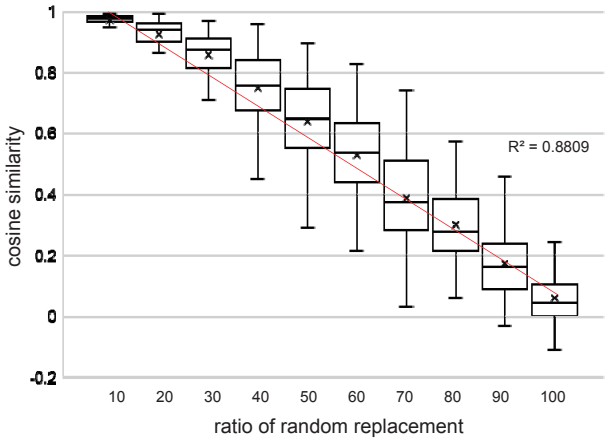

図 4　文書の相違度と cosine 類似度との関係

度が低くなることを意味している。また、ばらつき、特に置換割合が低い場合のばらつきが提案手法よりも大きく、置換割合 10-50% の標準偏差の平均は、提案手法が σ=0.08 であったのに対して、BM25 手法では σ=0.12 であった。既存の文書ベクトル化手法に関しても同様の傾向を持ち、R^2=0.87, σ=0.09 であった。

4 科学技術マップの開発

本章では、実際に開発した内容ベースの科学技術マップ[24, 25, 26] について説明する。初めにインターフェイスの概要について紹介した後、マップ上における論文・プロジェクトのレイアウト方法、およびマップが提供する分析機能について説明する。なお、本マップは Web 上で公開している（https://jipsti.jst.go.jp/foresight/）。

4.1 インターフェイスの概要

図 5 にマップの画面イメージを示す。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 つに分かれている。

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、Information、Mathematics & Physics、Communication、Electronics & Mechatronics、Power & Energy の 5 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・プロジェクトをクラスタリングした結果である。レイアウト方法の詳細は次節にて述べる。NISTEP サイエンスマップ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 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。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、1 ノードが 1 論文または 1 プロジェクトに相当する。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（citing→cited）をエッジとして表し、BM25 で抽出した 2 論文間の特徴語をエッジラベルとして表示している。尚、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画面下部に表示する。

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またはプロジェクト情報を全文検索し、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。

図 5　科学技術マップの画面イメージ

4.2　論文・プロジェクトのレイアウト方法

本節では、ポートフォリオビューから詳細ビュー作成までについて述べる。ポートフォリオ内の 1 分野に絞っても前節のデータセットにおいては、その中には最大 16 万ものノード（論文またはプロジェクト）が含まれ、この中から特定のトピックを見つけ出すのは容易ではない。そこで、事前にこれらをクラスタリングすることで意味のあるまとまりを構成し、クラスタ単位でユーザが各種分析を行うことを想定している。

クラスタリングからマップレイアウト作成までの課題としては、500 次元空間内のノードからの 2 次元ネットワークの作成（次元削減）が挙げられる。クラスタリングあるいは次元削減に関してはこれまでさまざまな手法が提案されているが、例えば多次元尺度法（multi-dimensional scaling, MDS）では予め任意の 2 ノード間の全ての距離を計算しておく必要があるなど、ノード数の 3 乗オーダー $O(n^3)$ で計算量が増加していくため計算コストが高いことが知られている。そこで本研究では現実的な計算時間に収めるために、各ノードの近傍（cosine 類似度が大きい）30 ノード、かつ、cosine 類似度 0.5 以上のエッジのみからマップを作成することとした。前述した Sci2 Tool[5] においても、各ノードの近傍 15 ノードのみを用いてマップを作成しており、意味のある特徴が失われていないことが確認されている。

領域ビューで表示されるクラスタ（詳細ビューの表示単位）は、分割評価関数としてモジュラリティ[27] を用いる手法の 1 つ、infomap 法[28] を用いて構成されている。モジュラリティが大きくなるようにネットワークを分割することで、クラスタ内では各ノードが密に結合し、クラスタ間では疎に結合するように分割される。そのため、分割されたクラスタはノード間の cosine 類似度が高い、特定の意味を持つ集団として形成される。しかし、infomap 法の単純な適応では 1 分野（information や communication）毎に最大約 2800 ものクラスタに分割されてしまい、領域ビューとしては数が多すぎると判断した。そこで、ノード数が 50 未満のクラスタについては類似度が最も高いペアを含むクラスタと結合する（single linkage clustering）ことでクラスタ数を減少させた。これにより分割精度は落ちる（モジュラリティは低下する）が、詳細ビュー表示時には他クラスタから取り込まれたノードの多くは独立した集合を形成するため、実際上の問題はないものと考えた。また、領域ビューではこの single linkage clustering の距離をクラスタ間の類似度（距離）として描画している。

詳細ビューのレイアウト計算には、ノード間の cosine 類似度を距離として、他の科学技術マップでもしばしば用いられる力学的（force-directed）ネットワークレイアウトアルゴリズムである OpenOrd[29] を用いた。なお、インターフェイス上はパラメータおよびレイアウトアルゴリズムを変更する機能を設け、利用者が状況に合わせてマップを変更できるようにした。

4.3　マップが提供する分析機能

本マップでは 4.1 章で述べた基本的な機能に加えて、(1)論文抄録／プロジェクト概要の翻訳機能、(2)統計情報表示機能、(3)特徴語サマリー機能、(4)SPARQL 検索機能とエクスポート機能、(5)レイアウト変更機能、(6)カスタムマップ生成機能などを備えている。以下に各機能について説明する。

1．論文抄録／プロジェクト概要の翻訳機能

詳細ビューでは、ノード（論文／プロジェクト）の詳細情報（タイトル、抄録／概要、著者／提案者、発行年／提案年）を画面下部に表示することが可能である。表示される情報の内、論文抄録およびプロジェクト概要については文尾に付いている Translate ボタンで日本語に翻訳することが可能である。日本語で流し読みする際に有用であり、元の英文も併記しているため誤訳が疑われる場合に確認することも容易である。

2．統計情報表示機能

詳細ビューで表示している IEEE 論文の統計情報を視覚的に確認することが可能である（図 6）。表示可能な統計情報は、引用評価指標（Impact Factor, SJR, CiteScore）、出版年別の論文数の推移、出版年別の被引用数の推移、論文数トップ 5 ヶ国であり、それぞれウィジェットとして表示する。また、ノード集合を矩形選択して該当部分のみの統計情報を表示することも可能である。国別の統計情報は上段が整数カウント法、下段が分数カウント法[30] を用いて集計したものである。整数カウント法は国単位での論文への関与の有無の集計であり、例えばある共著論文の著者所属機関が A 国、B 国、

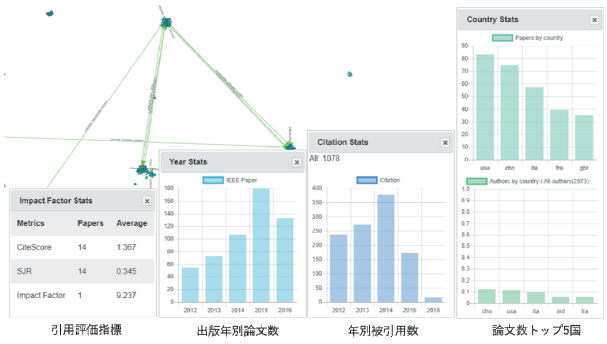

図 6　統計情報の表示（左から、引用評価指標、出版年別論文数、年別引用数、論文数トップ 5）

B 国の場合に A 国 1 件、B 国 1 件として数える。分数カウント法は重み付けを用いた集計であり、先の例では A 国 1/3 件、B 国 2/3 件として数える。

3. 特徴語サマリー機能

クラスタ単位での特徴語表示に加えて、矩形選択されたノード集合内に頻出する特徴語をワードクラウドで表示することも可能である（図 7）。ワードクラウド表示のために、予め BM25 を用いてクラスタ毎に全ノードの特徴語を最大 10 語まで算出している。矩形選択されたノード群の特徴語集合の内、出現回数の多い語ほどワードクラウド上で中心に大きく表示される仕組みである。これらの機能を用いて、詳細ビュー内に出現する島（ノードが密集した場所）の特徴理解に役立てることができる。

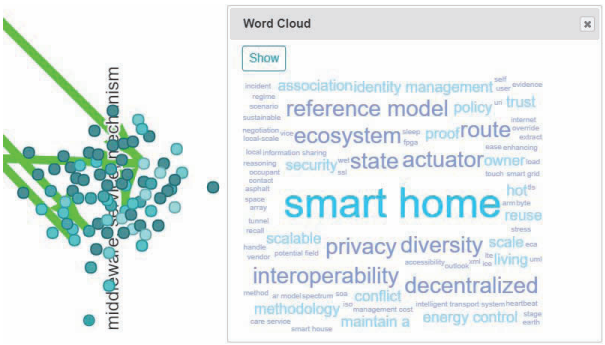

図 7　選択ノード集合の特徴語表示

4. SPARQL 検索機能とエクスポート機能

マップ上に表示されているデータは、W3C が規定する RDF（Resource Description Framework）形式（https://www.w3.org/RDF/）の Linked Data としてグラフ DB に格納されている。詳細ビューでは、このグラフ DB に対して直接、クエリー言語 SPARQL を用いて高度な検索を実行し、検索結果をビュー上で確認できる機能も有している（図 8）。例えば、「Impact Factor10 以上の論文から 100 回以上引用されている論文に cosine 値 0.9 以上で類似した論文（広く知られていないが有用な論文である可能性が高い）」といった検索を実行することが可能である。検索結果として表示されたテーブル内のノード ID をクリックすると、自動的にビューを視点移動し、ノードをハイライトする。さらに、ユーザが作成した SPARQL クエリーをマクロとして登録することも可能である。登録されたマクロはワンクリックで実行できる。これにより、SPARQL に精通したユーザが作成した分析用クエリーを他のユーザも利用することができる。

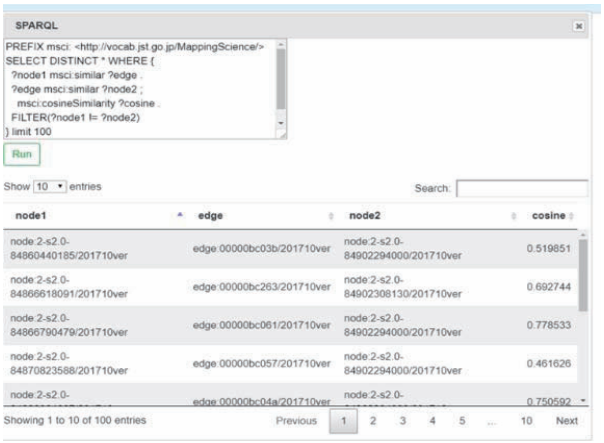

図 8　SPARQL 検索ウィジェット

尚、マップで表示されている情報を手元に保存して Excel 等のツールで確認したいというニーズに応えるため、クラスタ内の全ノードや矩形選択したノードの詳細情報を CSV 形式でエクスポートする機能も備えている。SPARQL 検索の結果を CSV 形式でエクスポートすることも可能である。

5. レイアウト変更機能

前章で述べた通り、詳細ビューのレイアウトは OpenOrd、エッジカット値 0.91 を用いて事前に計算されているが、ユーザが他のレイアウトを指定して再描画することも可能である。現状、OpenOrd アルゴリズムでエッジカット値を 0.94、0.88 にした場合、および Large Graph Layout[31]、Fruchterman-Raingold[32]、Kamada-Kawai[33] の計 5 種類のレイアウトで再描画が可能である。ユーザがレイアウトを指定すると、既に計算済みの場合は即座に表示され、未計算の場合はリアルタイムにレイアウト計算が実行され、結果が保存されて表示可能となる。尚、レイアウト計算時間はノード数とアルゴリズムにより異なるが、数秒から数分程度である。

6. カスタムマップ生成機能

最後に前述の通り詳細ビューは infomap 法で構成されたクラスタ毎に用意されているが、ユーザが任意のキーワードを指定してクラスタを新たに構成することもできる。ポートフォリオビューで専用ウィジェットからキーワードを指定すると 5 分野に含まれる全ノードから全文検索を行い、ヒットしたノードの cosine 類似度情報から OpenOrd でレイアウト計算を行い、カスタム詳細ビューを構成する。例えば、複数分野に跨っているであろう Neural Network や Artificial Intelligence［AND］［NOT］Neural Network といった条件式でカスタム詳細ビューを作成し、目的に応じて分野横断的な分析を行うことができる。レイアウト変更と同様に計算時間はノード数により異なり、完了までは数秒から数分程度である。一度作成されたカスタム詳細ビューのレイアウト情報は保存されるため、2 回目以降はレイアウト計算を行わずに表示できる。カスタム詳細ビューでは通常の詳細ビューと同様に、ノード検索、統計表示、時間遷移アニメーション表示やレイアウト変更が可能である。

5 まとめと今後の課題

本研究では、引用・被引用分析の適用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、特許等を対象とした内容の類似性に基づく科学技術マップを開発した。本論文では、まず JST シソーラス内の概念毎に情報エントロピーを求め、単語ベクトルを概念に集約することで科学技術観点での類似性を強化した文書ベクトル化技術を提案した。その上で複数の関連研究において最もよいとされる BM25 を用いた内容ベース手法との比較を行い、提案手法が文書間類似度の精度においてより優れていることをサンプリング実験および人工データを用いた実験を通じて示した。加えて、一連のマップ開発において欠かせない論文・プロジェクト群のクラスタリング手法、2 次元レイアウト手法、およびマップが提供する分析機能についても技術的透明性、および分析作業の再現性を保証するために詳述した。

今後は、まず引用分析に基づくマップとの比較や、データセットへの特許情報の追加などを行っていきたい。また、日英 JST シソーラスと文書ベクトルを介して日本のファンディング情報と海外のファンディング情報を重ね合わせ、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。さらには、研究領域の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たい。既に、米国情報高等開発活動（Intelligence Advance Research Projects Activity, IAPRA）では、2011 年から 2015 年に渡って実施された FUSE（Foresight and Understand from Scientific Exposition）プログラム（D11PC20152）において、複数の科学技術マップから得られた指標に基づいて新技術の急速な出現を予測する研究を進めている。4.3 節 SPARQL 検索機能とエクスポート機能を用いることで既にマップからはさまざまな特徴量を抽出することができる。今後、JST も統計処理や機械学習技術を適用して、新しい科学技術コンセプトの出現を早期に検出することを目指していきたい。

謝辞

本稿の執筆にあたり有益なご助言を戴いた文部科学省 科学技術・学術政策研究所 治部 眞里様、（株）ジー・サーチ 松本 尚也様に感謝いたします.

参考文献

[1] Price, D.: Networks of Scientific Papers. Science, 149, 510-515 (1965)

[2] 川村隆浩，渡邊勝太郎，松本尚也，江上周作，治部眞里 : Mapping Science―飛躍が期待される科学技術領域の抽出―. In: 第 14 回情報プロフェッショナルシンポジウム（INFOPRO2017）予稿集, 119-124 (2017)

[3] 川村隆浩，渡邊勝太郎，松本尚也，江上周作，治部眞里 : 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―. In: 研究・イノベーション学会 第 32 回年次学術大会予稿集（2017）

[4] Borner, K.: Sci2: A Tool of Science of Science Research and Practice. In: Tutorial of the 10th International Conference on Scientometrics and Informetrics（ISSI 2011）(2011)

[5] Boyack, K., Klavans, R., Borner, K.: Mapping

the Backbone of Science. Scientometrics, 64(3), 351-74 (2005)

[6] Steyvers, M., Griffiths, T.: Probabilistic Topic Models. Laurence Erlbaum (2007)

[7] Blei, D., Ng, A., Jordan, M.: Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022 (2003)

[8] Griffiths, T., Steyvers, M.: Finding Scientific Topics. In: Proceedings of the National Academy of Sciences, vol. 101 (suppl. 1), 5228-5235 (2004)

[9] Talley, E., Newman, D., Mimno, D., Herr II, B., Wallach, H., Burns, G., Leenders, A., McCallum, A.: Database of NIH Grants Using Machine-learned Categories and Graphical Clustering. Nature Methods, 8, 443-444 (2011)

[10] Herr II, B., Talley, E., Burns, G., Newman, D., LaRowe, G.: The NIH Visual Browser: An Interactive Visualization of Biomedical Research. In: Proceedings of 13th International Conference on Information Visualisation (ICIV 2009), 505-509 (2009)

[11] Kullback, S., Leibler, R.: On Information and Sufficiency. Annals of Mathematical Statistics, 22, 79-86 (1951)

[12] Boyack, K.W., Newman, D., Duhon, R., Klavans, R., Patek, M., Biberstine, J., Schijvenaars, B., A., S., Ma, N., Borner, K.: Clustering More Than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-based Similarity Approaches. PLoS ONE, 6(3), 1-11 (2011)

[13] Jones, K.S., Walker, S., Robertson, S.E.: A Probabilistic Model of Information Retrieval: Development and Comparative Experiments. Information Processing and Management, 36(6), 779-808 (2000)

[14] Waltman, L., Boyack, K.W., Colavizza, G., Van Eck, N.J.: A principled methodology for comparing relatedness measures for clustering publications. In: Proceedings of the 16th International Conference on Scientometrics & Informetrics (ISSI 2017), 691-702 (2017)

[15] Firth, J.R.: A Synopsis of Linguistic Theory 1930-1955. Studies in Linguistic Analysis, 1952-59, 1-32 (1957)

[16] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation ofWord Representations in Vector Space. In: Proceedings of Workshop at the International Conference on Learning Representations (ICLR 2013) (2013)

[17] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed Representations of Words and Phrases and Their Compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS 13), 3111-3119 (2013)

[18] Le, Q., Mikolov, T.: Distributed Representations of Sentences and Documents. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014), vol. 32(2), 1188-1196 (2014)

[19] Shannon, C.: A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379-423, 623-656 (1948)

[20] Vilnis, L., McCallum, A.: Word Representations via Gaussian Embedding. In: Proceedings of International Conference on Learning Representations (ICLR 2015), 1-12 (2015)

[21] Kimura, T., Kawamura, T., Watanabe, K., Matsumoto, N., Sato, T., Kushida, T., Matsumura, K.: J-GLOBAL knowledge: Japan's Largest Linked Data for Science and Technology. In: Proceedings of the 14th International Semantic Web Conference (ISWC 2015) (2015)

[22] Santus, E., Lenci, A., Lu, Q., Walde, S.: Chasing Hypernyms in Vector Spaces

with Entropy. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014), 38-42 (2014)

[23] Klavans, R., Boyack, K.W.: Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge?. Journal of the Association for Information Science and Technology, 68(4), 984-998 (2017)

[24] Kawamura, T., Watanabe, K., Matsumoto, N., Egami, S., Jibu, M.: Funding Map for Research Project Relationships using Paragraph Vectors. In: Proceedings of the 16th International Conference on Scientometrics & Informetrics (ISSI 2017), 1107-1117 (2017)

[25] Kawamura, T., Watanabe, K., Matsumoto, N., Egami, S., Jibu, M.: Science Graph for characterizing the recent scientific landscape using Paragraph Vectors. In: Proceedings of the 9th ACM International Conference on Knowledge Capture (K-Cap 2017), 9-16 (2017)

[26] T. Kawamura, K. Watanabe, N. Matsumoto, S. Egami, M. Jibu: Funding Map using Paragraph Embedding based on Semantic Diversity, Scientometrics, Springer, 116 (2), pp. 941-958, 2018. https://doi.org/10.1007/s11192-018-2783-x

[27] Newman, M.E.J.: Modularity and community structure in networks. In: Proceedings of the National Academy of Sciences of the United States of America (PNAS 2006), 103(23), 8577-8582 (2006)

[28] Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. In: Proceedings of the National Academy of Sciences of the United States of America (PNAS 2008), 105(4), 1118-01123 (2008)

[29] Martin, S., Brown, W.M., Klavans, R., Boyack, K.: OpenOrd: An Open-Source Toolbox for Large Graph Layout. In: Proceedings of SPIE, Visualization and Data Analysis (VDA), 786806 (2011)

[30] 村上昭義、伊神正貫：科学研究のベンチマーキング 2017―論文分析でみる世界の研究活動の変化と日本の状況―. 科学技術・学術政策研究所、http://hdl.handle.net/11035/3177 (2017)

[31] Adai, A.T., Date, S.V., Wieland, S., Marcotte, E.M.: LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. Journal of molecular biology, 340(1), 179-190 (2004)

[32] Fruchterman, T.M.J., Reingold, E.M.: Graph Drawing by Force-directed Placement. Software - Practice and Experience, 21(11), 1129-1164 (1991)

[33] Kamada, T., Kawai, S., An algorithm for drawing general undirected graphs. Information Processing Letters, 31(1), 7-15 (1989)