拡散過程を考慮した学術文献におけるミーム指標の提案

鎌田麻衣子[1]　浅谷公威[2]　坂田一郎[2]
[1]東京大学工学部システム創成学科
[2]東京大学大学院工学系研究科

Abstract: With the increase of scientific publications and diversification of research areas, knowledge discovery from scientific publications has become difficult. A lot of methods which retrieve important concepts from scientific publications using natural language processing or machine learning have been proposed. However, it is still an unresolved problem to distinguish a widely spread concepts from that discussed in a narrow area. Hence, we propose a new indicator that evaluates how remarkable the new concept is. We investigated the difference between new concepts that spread to the scientific field as Meme and those that disappear with data from close to 67 million publication records from the 50 years of Scopus. To find out this, we introduced a network structure perspective to existing methods which analyzed memes in the scientific literature from the state of propagation. There, we identified scientific terms that have a strong impact on society and other research areas, such as scientific terms that represent concepts related to the Nobel Prize. We featured memes, such as jargon and Nobel Prize-winning groundbreaking concepts, were compared and found to have different network structures. This result suggests the conditions for spreading creative discoveries. Based on the results, we propose new expressions to evaluate new concepts.

1. 背景および目的

画期的な発見のために,研究者は広い知識を持つことが求められている.一方で近年、出版される学術文献数は年々指数関数的に増加するとともに,膨大な数の概念が生み出され続けている[1].研究者はこれらの中から有用な知識を広く見つけ出す必要があるが,分野が成熟するにつれて,研究者は他分野と融合させるような革新的な研究よりも自身の専門分野を掘り下げるような保守的な選択をしやすい[2]. 科学の細分化と専門化が進んだことにより,ある分野で発見されたことが他の分野では知られていないと言った知識の分断が起きることで研究者は自身の専門分野以外の知識を体系的に理解し,研究に取り入れることが難しくなり,その結果,自身の研究の周辺領域等に存在する潜在的に重要な情報を見逃してしまう.

その問題を解決する方法として注目されているのが Literature Based Discovery(LBD)の分野である[3].LBD とは電子出版やデジタルアーカイブの成長により収集が可能となった,文献に関する膨大な量の情報を用いて自動もしくは半自動的に新しい知識を発見するというものである.近年,引用ネットワーク上の被引用先の文献における科学用語の使用頻度に着目したミームスコア[6]が提案され,単語の遺伝に注目することで情報価値の高い科学用語を一定精度で抽出することに成功している.一方で,この手法はどの文献やコミュニティにおいて使用されたかに関しては考慮していないため,重要ではあるが狭い知識コミュニティで使われている科学用語を高く評価する傾向があり,知識コミュニティにおける影響力が大きな科学用語を評価することができていない.

本論文は,単語がどの程度広範なコミュニティで使用されているかを定量化することで,より多様な分野に伝播する科学用語を抽出する新たなミーム指標を開発することを目的とする.

2. 既存手法に対する提案手法の位置付け

文献から,科学的に有用な科学単語を抽出する際にはこれまで様々な方法が用いられてきたが[4][5],その中で,Kuhn ら[6]が定義したミームスコア M_m は”ミーム”の考え方を取り入れたことで,単語の登場回数の閾値の設定や専門知識などを必要としない,従来の手法の持つ課題を解決する手法であった.

ミームとは Dawkins が著書 The Selfish Gene[7]で提唱した概念であり,情報や文化が発生し、模倣によって伝達され淘汰されていく様子を遺伝子になぞらえた概念のことで,強いミームは何世代にも渡って広く伝播するのに対し,弱いミームは伝播せずに廃れていく.ある概念に対する発展の様子を引用関係から追うことができるという学術文献特有の特徴は”ミーム”の考え方と相性がよかったのである.ここで強いミームとは被引用先でも研究される情報価値の高い科学用語と考えることができる.

ミームスコア M_m は文献中における単語 m の登場頻度 f_m と引用ネットワークにおける単語 m の遺伝度 P_m から構成されている(式 2.1). ミームスコア M_m は被引用先でも同じ概念について言及しており,かつ被引用先以外には現れない場合,つまり単語が引用によってのみ伝播しており,かつ伝播頻度が高い単語について高く評価される.

$$M_m = f_m P_m \tag{2.1}$$

しかし, 学術文献における科学用語およびそれによって表現される概念の情報価値を考えたとき,この手法は被引用先でも途絶えず議論されるような遺伝性の高さは評価できている一方で多くの分野に伝播する拡散力の強さを考慮できておらず,狭いコミュニティ内で深く議論されている単語を高くスコア付けしてしまう傾向があった.そこで本項では,単語の拡散力を定量化する拡散ミームスコア E を新たに提案する.

3. 提案手法

提案手法 E (拡散ミームスコア)は以下の式で表さ

れる.ここで, e は単語伝播のエッジの集合,$d(e_s, e_t)$は単語の伝播距離を表す.1 を加えているのは,単語の伝搬距離の総和が 1 であった場合のためである. Eを求めるための細かい手順については以下の節で述べる.

$$E = \ln\left(\sum_{e} d(e_s, e_t) + 1\right) \tag{3.1}$$

3.1. 文献引用ネットワークの作成

学術文献データより,各文献をノードとし,文献間の引用を引用元の論文から引用先へと向かうエッジとした重みなし有向ネットワークを作成する.このネットワークを文献引用ネットワークと呼ぶことにする.

3.2. 文献引用ネットワークのクラスタリング

3.1 節で作成した文献引用ネットワークに対して Leiden 法[8]を用いたクラスタリングを行い,文献を複数の部分集合に分割する.引用関係を持つ文献間にエッジを張っているため,文献引用ネットワークに対してクラスタリングを実行すると,関係の近い文献の集合がそれぞれのクラスタとして取得できる.このクラスター一つ一つを「研究分野」と定義する. 文献引用ネットワークのクラスタリングによって研究分野を設定する方法は将来的に主要な研究分野の特定のためにも用いられている[9][10].それぞれの文献は必ず属する研究分野を一つだけ持つ.この研究分野の分類を用いて 3.3 節以降の分野間の分散表現の獲得や単語の研究分野を超えた伝搬距離の計算を行う.

3.3. クラスタのエンベッディング

得られたクラスタ群について,それらが比較的似ている分野なのか全く違う分野なのかを表す指標として用いるために,クラスタ間引用ネットワークのエンベッディングによってグラフ構造をベクトル空間に射影した.

具体的な作業としては,まずクラスタ間の引用回数を重みとして持つ「クラスタ間引用ネットワーク」を作成し, DeepWalk[11]という手法を用いてグラフ構造を 128 次元の空間に射影した.これによりクラスタ毎に局所的な構造を保ちつつ、広域的な構造の情報を保持することができる.

3.4. 拡散ミームスコア E の算出

クラスタを超える伝播があった単語について, エンベッディングによって得られたクラスタのベクトル表現を用いて引用元文献の所属クラスタと引用先文献の所属クラスタ間の距離をコサイン類似度で算出し, その総和の自然対数をとったものを拡散ミームスコア E と定義した.

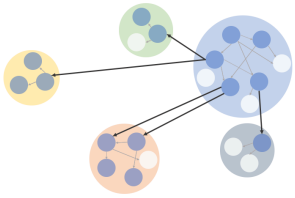

Fig.1 クラスタを超えた単語の伝播のイメージ

4. 実験

4.1. 使用データ

本論文では Scopus から抽出した約 6700 万件の医学カテゴリに属する文献データを使用した. このデータの中から,タイトル,アブストラクト,及び引用関係のデータを用いる.これら 49 年間のデータのうち,1970 年から 1974 年に出版された文献には現れず 1975 年以降に出版された文献内に新しく登場した単語のうち, 通算 20 回以上登場した年を起点とした 10 年間に 50 回以上登場した 29,832 単語を分析対象単語とした.

4.2 ベースライン

提案手法 E の比較手法として,登場回数,既存手法 M_m,ランダム抽出の 3 つを採用した.

4.3 ベースライン

手法を用いることで情報価値の高い単語を抽出できていたか検証する方法について述べる.

自身の専門分野外の分野において抽出したい科学用語とは,その分野特有の専門的な単語であり,かつ広く影響を与えうる注目度の高い単語であると言える.

ここで、その分野特有の専門的な単語を取り出せていたかの指標として人間の疾患の網羅的なデータベースである MalaCards を用いた[12].各指標によって抽出された単語のなかにこのデータベースにある疾患がいくつ含まれているかを専門的な単語の抽出精度として扱う.

広く影響を与えうる注目度の高い単語を取り出せていたかの指標としては世界最大級の百科事典である Wikipedia のタイトルデータを用いた

5. 結果および考察

5.1. 提案手法 E によって高く評価された単語

Table 1 は E で高く評価された単語上位 12 単語を表したものである.ここで+のついた単語はその単語に関連する薬等の発見者がノーベル賞を受賞したことを意味し, *のついた単語は発見者がクラリベイト・アナリティクス引用栄誉賞を受賞した単語である.この賞は「近い将来ノーベル賞を受賞する可能性の高い研究者」を発表することを目的とした賞である. この結果,上位 12 単語のうち 6 単語(重複のぞく)がノーベル賞受賞かそれに準じる発見だったことがわかった.今回得られた結果は,分野を超えた単語の伝播に注目することで社会にインパクトを与えた単語を抽出することができるという E の着想が正しかったことを裏付けるものである. ある概念に対する発展の様子を引用関係からたどることができるのは学術文献特有の性質であり,例えば SNS における単語の拡散などを分析する際に単語が誰から誰に伝わったのかを特定することは難しい.拡散ミームスコア E によって影響力の強い単語を抽出できたのは学術の引用を介した特殊な発展プロセスによるものだ. 今後、更に多くの発見や概念が分野の融合によって生まれ、分野を超えて広がるにつれて元の文献の被引用数だけでなく他

分野への影響に注目する傾向は進むと考えられる[13]. 科学用語とそれによって表される概念の他分野への影響力を定量化し、他分野への影響力が社会的な評価と強い関わりがあることを示した提案手法 E は学術の発展プロセスを明らかにする点においても有意義と言える.

Table 1 手法 *E* によって抽出された上位 12 単語

Helicobacter+	toll-like+	bevacizumab	IL-12
VEGF*	micrornas*	mirnas*	hiv-1+
chemokine	cox-2	leptin*	Infliximab

5.2. 提案手法 *E* の評価

Fig.2 は各指標の上位 n%に含まれる単語のうち,MalaCards に含まれている単語数を表したものである.縦軸は抽出できた MalaCards 記載単語数であり,これを見ると既存ミームスコア M_m が最も多くの疾患名を抽出できており,専門的な単語の抽出精度は既存手法 M_m が最も高かったことがわかる.

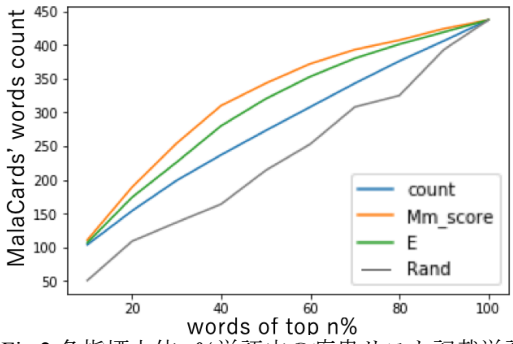

Fig.2 各指標上位 n%単語中の疾患リスト記載単語数

Fig3 は各指標上位 n%に含まれる単語のうち,Wikipedia に掲載されている単語の割合を表したものである.y 軸の値が大きいほど精度が高いことを意味している.これを見ると注目度の高い単語の抽出精度は拡散ミームスコア *E* が最も高い精度を示した.このことから,既存ミームスコア M_m は狭いコミュニティの中で密に議論されているような専門的な単語を評価することには適しているが,多くの分野に広く影響を与えた単語を評価することができないという仮説が正しかったことを示し,更に既存ミームスコア M_m では評価できていない注目度の高い単語を提案手法である拡散ミームスコア *E* では評価できることを示した.

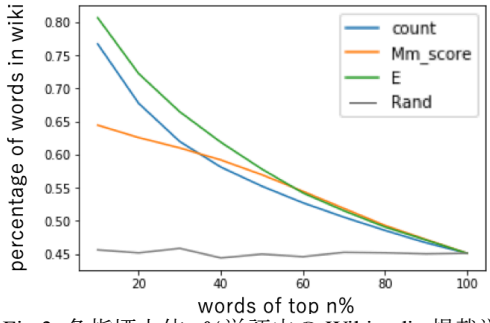

Fig.3 各指標上位 n%単語中の Wikipedia 掲載単語率

5.2.1 提案手法 *E* と登場回数の比較

登場回数は社会に伝播しているかどうかを確認する最も単純な指標であり、*E* と目的は同じである.しかし、登場回数は多いが *E* は低い単語が存在した.それらを確認すると,出版社名・雑誌名が多く含まれていた(Fig.4).

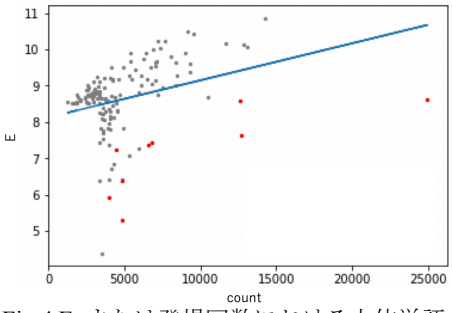

Fig.4 E または登場回数における上位単語

Fig.5 は出版社・雑誌関連単語の拡散ネットワーク図である.これを見ると,一つ一つのノード(クラスタ)が独立して存在しており,クラスタ間のエッジはほとんど見られなかった.つまり,これらの単語はクラスタ間の伝播がほとんどなく,それぞれのクラスタ内で個々に登場しているということが言える.

新しい物質や病気といった専門性の高い単語は全て,その物質や病気を初めて報告した文献が存在し,その文献を元に他の研究者も研究を進めていくため同じ研究対象を持つ文献は引用・被引用の関係となる.一方で出版社・雑誌名の様な専門性の低い単語は登場回数が多くても拡散によっては広がらない.登場回数を指標として用いた場合にはこの違いを判別できないため, 疾患名リストを用いた専門性のある単語の抽出精度が低くなったと考えられる. また,名詞の単数形と複数形などの違いについても,*E* では単数形の方が高く評価されやすいのに対し登場回数では順位があまり変わらず, 抽出できるユニークな単語の総数が減ってしまうため,Wikipedia 掲載単語の抽出精度が *E* よりも低くなったと考えられる.

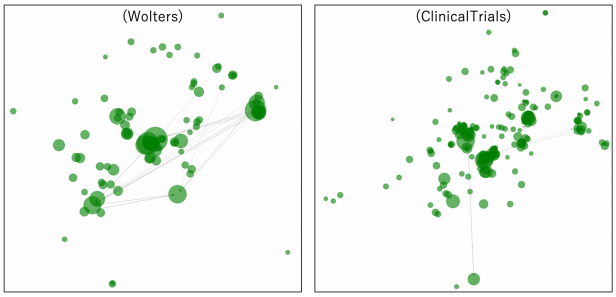

Fig.5 出版社・雑誌関連語の拡散ネットワーク図

5.2.1 提案手法 *E* と既存手法 M_m の比較

Fig6 は既存ミームスコア M_m または拡散ミームスコア *E* で高く評価された単語の散布図である.*E* において 100 位以内であったものを黄色,M_m において 100 位以内であったものを青色,両方で 100 位以内であったものを赤色でプロットしている.これを見ると *E* と M_m では高く評価される単語が全く異なることがわかる.

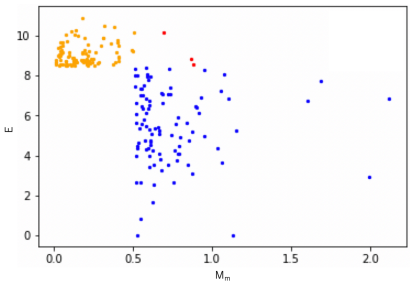

Fig.6 E 及び Mm で評価される単語の散布図

それぞれの指標で高く評価された単語の拡散ネットワーク図を示したものが Fig.7 である.これを見ると,既存ミームスコア M_m で高く評価された単語は大きなノードが一つ存在し,他のクラスタにはほとんど伝播がないことがわかる.一方で拡散ミームスコア E で高く評価された単語については登場クラスタ数が多く,それらのクラスタの中でも他のクラスタと複数のエッジを持ち拡散の中心となっているハブ的存在のクラスタとその周辺にハブクラスタよりは小さいが一定の大きさを持つクラスタの集合が確認できる.

これらの伝播構造の違いと,疾患名リストを用いた専門性の評価では既存ミームスコア M_m が精度が高く,wikipedia を用いた注目度の高さの評価では拡散ミームスコア E が高い精度を表していたことを合わせると,既存ミームスコア M_m は狭いコミュニティの中で密に議論されており,一つの研究分野として確立されているような単語を評価することには適しているが,一方で既存ミームスコア M_m が評価する単語の多くは狭い知識コミュニティ内に閉じており,その中で重要であってもコミュニティの外側の社会には広くは知られていない.反対に,拡散ミームスコア E ではできるだけ多く,もしくは遠くのコミュニティに伝播するほど高く評価されやすいため,wikipedia を用いた評価では拡散ミームスコア E の方が高い精度を示したと考えられる.

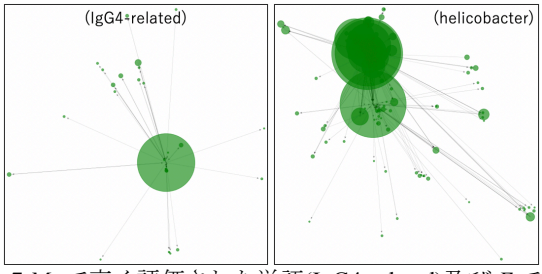

Fig.7 M_m で高く評価された単語(IgG4-related)及び E で高く評価された単語(helicobacter)の拡散ネットワーク図

5.2.2 E と M_m の複合についての検討

ここまでの分析を踏まえて次に,専門性のある単語の評価に有用な M_m と広範性のある単語の評価に有用な E を複合することで,情報価値の高い単語の抽出精度が上がるのか以下の 3 つの複合指標について検証を行った.

- $M_m + E$
- $M_m \times E$
- M_m における順位 + E における順位

M_m+E は最も単純な複合の方法である.$M_m \times E$ でのスコア付けは対数をとると $\ln(M_m \times E) = \ln(M_m) + \ln(E)$ であり,既存ミームスコア M_m と拡散ミームスコア E の対数の足し算と同じである.これにより,分布の偏りの影響を小さくした複合指標となると考えられる.順位の足し合わせは,分布の影響を全く受けない方法であるが,定量性が失われることに注意が必要である.

Fig.8 は疾患名リストおよび wikipedia を用いた評価により上記の 3 つの複合指標を比較したものである.これを見ると,既存ミームスコア M_m と拡散ミームスコア E を足し合わせたものは E だけの場合とほとんど変わらなかった.

一方で,既存ミームスコア M_m と拡散ミームスコア E を掛け合わせたものおよび順位を足し合わせたものについては,疾患名リストを用いた評価では既存ミームスコア M_m とほぼ変わらないか M_m よりも良い精度となり,wikipedia を用いた評価では拡散ミームスコア E には及ばないものの既存ミームスコア M_m よりも良い精度となった.

既存ミームスコア M_m 拡散ミームスコア E を足し合わせたものが拡散ミームスコア E だけの場合とほとんど変わらなかった理由としては,既存ミームスコア M_m と拡散ミームスコア E で分布が大きく異なり,既存ミームスコア M_m は 1 以下ものもがほとんどであったのに対して拡散ミームスコア E の数値が大きかったため既存ミームスコア M_m の影響がほとんど現れなかったのだと考えられる.

一方で,既存ミームスコア M_m を拡散ミームスコアと E を掛け合わせたものおよび順位を足し合わせたものについては,両方の指標で高く評価されたものが上位となるため,既存ミームスコア M_m と拡散ミームスコア E の両方の要素を反映することができる.特に順位を足し合わせたものは疾患名リストを用いた評価で既存ミームスコア M_m よりも良い精度で,かつ上位 1 割の単語のうち 75%以上が wikipedia 掲載単語であることから,既存ミームスコア M_m と拡散ミームスコア E の複合により専門性と広範性の両方の観点で情報価値の高い単語の抽出ができる指標となったことが示唆される.

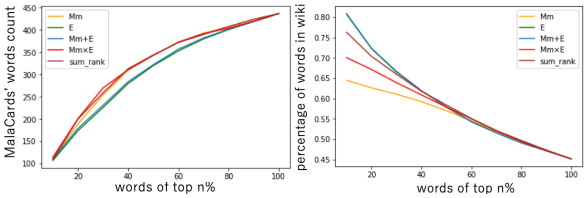

Fig.8 M_m, E および複合指標における上位 n%単語中の疾患リスト記載単語数(左)と Wikipedia 掲載単語率(右)

5.2.3 拡散ミームスコア E の妥当性

他の論文との引用関係を持つ全文献約 5700 万件を対象にクラスタリングをした後,1000 以上の文献が所属するクラスタについてはさらにクラスタリングを行うという処理を 2 度行った結果,24,908 のクラスタに分割された.

例えば,クラスタ番号が「0-1-2」であった場合,1 回目の全文献を対象にしたクラスタリングでクラスタ番号 0 番に分類され,クラスタ番号 0 番に分類された文献だけを対象にした 2 回目のクラスタリングでクラスタ番号 1 番に分類され,さらにクラスタ番号が 1 番に分類された文献だけを対象にした 3 回目のクラスタリングでクラスタ番号 2 番に分類されたということを表す.ここで,1 回目のクラスタリングの結果を大分類,2 回目のクラスタリングの結果を中分類,3 回目のクラスタリングの結果を小分類と呼ぶこととする.

24,908 のクラスタのうち,分析対象単語を含むクラスタは 9,172 件であった.Fig.9 は分析対象単語を含む文献数上位 10 クラスタについて,クラスタ内から文献を 100 件ランダムに抽出し,クラスタ番号とその題目に含まれる頻出単語を python の wordcloud というライブラリを用いて可視化したものである.多くの文献に登場している単語ほど,大きい字で表示される.例えば,4-7-0 番クラスタは主にマイクロ RNA を研究しているクラスタであり,7-4-0 番クラスタは HIV について研究しているクラスタと言える.

これを見るとそれぞれのクラスタを代表する特徴的な単語がクラスタ間で異なっていることが確認でき,小分類までクラスタリングを行ったことで研究分野ごとにまでコミュニティを細分化できたことがわかる.

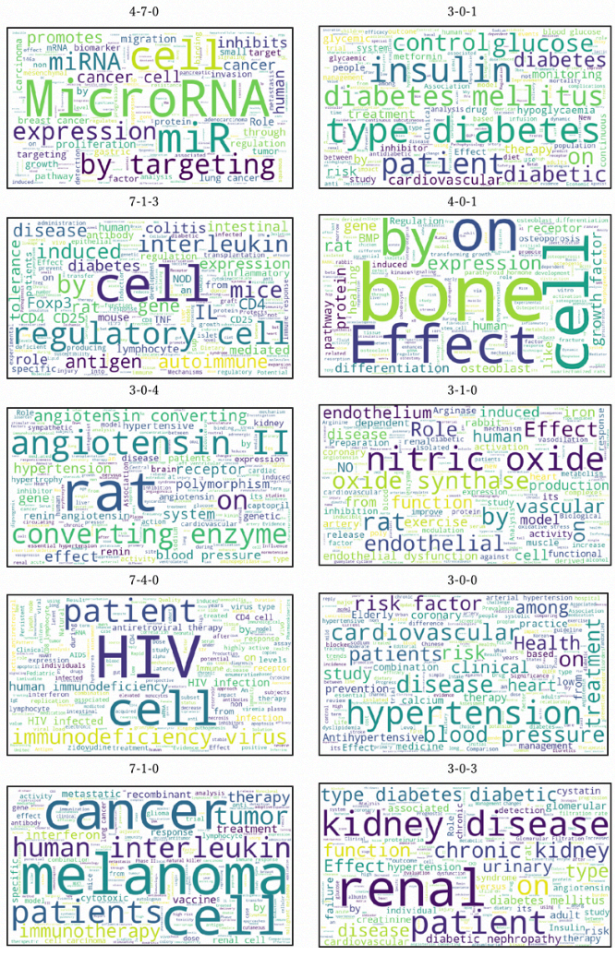

Fig.9 分析対象単語頻出クラスタにおける代表単語

Fig.10 は,所属クラスタ数が上位 20 位である大分類 0 番から 19 番に所属する 15,173 のクラスタの位置関係をプロットしたものである.ここで,クラスタの位置とはエンベッディングによって得られた 128 次元のベクトル表現を t-SNE によって 2 次元に圧縮したもののことを指す.

各クラスタに所属する文献のうち,最も多くの文献が所属していたジャーナルのカテゴリで色分けした Fig.11 と Fig.10 を比較すると,クラスタリングで得られた大分類とカテゴリは大まかに一致しており,特に多くの文献が所属する Medicine カテゴリ(MEDI)についてはクラスタリングの際に複数の大分類に分けられているものと考えられる.

また Fig.10 を見ると,同じ大分類に所属するクラスタ同士が近いところにプロットされていることが確認でき,研究分野の近さがエンベッディングの位置関係に反映されていると考えられる.

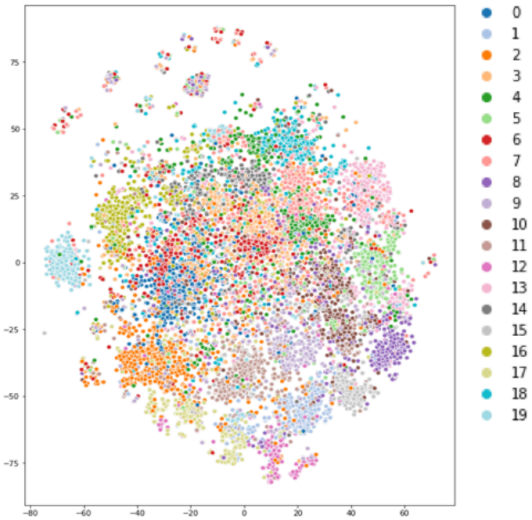

Fig.10 エンベッディングによるクラスタの配置

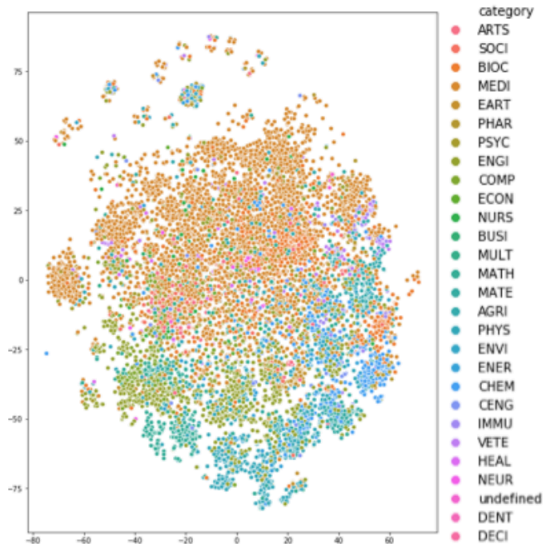

Fig.11 カテゴリの分布

6. 結論

高度に複雑化・多様化した学術界において,研究者が効率的に他分野の知識を取り入れるためには,情

報価値の高い知識を抽出する指標が必要である.

本研究では,単語の情報価値の評価にミームの考え方を取り入れ,「被引用先でも途絶えず議論される遺伝性の高さ」と「多くの分野に伝播する拡散力の強さ」という二つの軸で科学用語(単語)を評価することが重要であると考えた.既存手法のミームスコア M_m は前者を評価できるが,後者の評価はできていない. そこで,Scopus の文献の引用情報を元にクラスタリングを行いクラスタ毎にベクトル表現を得ることにより,文献を研究分野コミュニティに分割して研究分野間の距離を表現し,それを元に単語の伝播の広さを定量化する新たなミーム指標 E を提案した.

引用ネットワーク上での単語の伝播距離を定量化する指標を提案し,既存手法では抽出できない広範囲の分野に影響力を持つような概念の抽出に有用であることを分野における専門的な単語の抽出精度の検証に人間の 疾患データベースを,社会に定着している単語の抽出精度の検証として Wikipedia のデータを用いて示した.

単語の拡散過程をネットワークで可視化することで,科学的な単語と出版社名や雑誌名などの科学的でない単語のネットワーク構造に違いがあることを確認し,提案手法では引用によって伝播しない科学的でない単語を排除できることが確認できた.

また,既存ミームスコア M_m では専門性の高い単語の,今回提案した拡散ミームスコア E では広範性の高い単語の抽出精度が高いという違いがあるが,この差が生まれるのは各手法で評価される単語が異なる拡散ネットワーク構造を持っているためだと考えられる.これらの性格の異なる 2 つのミーム指標を組み合わせることで,専門性と広範性の両方の観点で情報価値の高い単語の抽出ができる指標となることが示唆された.

また,提案手法を用いた評価では,ノーベル賞を受賞する様な学術界において価値の高い単語を抽出できることが確認された.このことから,多くの分野に伝播し影響を与えていることが学術界における評価と密接な関わりがあり,提案手法は学術の発展プロセスを明らかにする上でも有用なアプローチであることが示唆された.

今後の課題としては,1 単語で表現されていない科学用語を含めた評価を行うことが挙げられる.

そのために分析対象を 1 単語ではなく熟語に広げて検証を行う必要がある.

今回,対象を新出単語に絞ったことで,既存の単語で構成されている新しく生まれた概念(grid cell など)を分析対象とすることができていない. n-gram などの方法を用いて熟語も分析対象とすることでより広い範囲で単語評価精度の検証が可能になるだろう.

また,指標の評価方法についても,今回用いたデータベース内の単語の有無の様な定量的な指標に加えて,専門家によるアノテーション作業を行うことで,妥当性検証がより強固なものになると考えている.

本研究は,ある概念に対する発展の様子を引用関係からたどることができるという学術文献特有の性質を,単語の影響力の定量評価の指標に用いたで新規性がある.一方で,対象単語の選定や対象期間の設定,既存手法との複合方法など,議論が十分でない点も多い.これらについては実用性を見据えて有識者の意見も含めた改善が必要だろう.

参考文献

[1] Milojevic, Stasa. (2015). Quantifying the Cognitive Extent of Science. Journal of Informetrics. 9. 10.1016/j.joi.2015.10.005.

[2] Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and Innovation in Scientists' Research Strategies. American Sociological Review, 80(5), 875–908.

[3] Henry, S., & McInnes, B.T. (2017). Literature Based Discovery: Models, methods, and trends. Journal of biomedical informatics, 74, 20-32 .

[4] Cohen, T., Widdows, D., Schvaneveldt, R. W., Davies, P., & Rindflesch, T. C. (2012). Discovering discovery patterns with predication-based semantic indexing. Journal of biomedical informatics, 45(6), 1049-1065.

[5] Bodenreider O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic acids research, 32(Database issue), D267–D270. doi:10.1093/nar/gkh061

[6] Kuhn, T., Perc, M., & Helbing, D. (2014). Inheritance patterns in citation networks reveal scientific memes. *Physical Review X*, *4*(4), 041036.

[7] Richard Dawkins. The Selfish Gene. Oxford University Press, Oxford, England

[8] Traag, V. A., Waltman, L., & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports, 9(1), 1-12.

[9] Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I., & Matsushima, K. (2009, August). Detecting emerging research fronts in regenerative medicine by citation network analysis of scientific publications. In PICMET'09-2009 Portland International Conference on Management of Engineering & Technology (pp. 2964-2976). IEEE.

[10] Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation, 28(11), 758-775.

[11] Perozzi, B., Al-Rfou, R., & Skiena, S. (2014, August). Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 701-710).

[12] Espe, S. (2018). MalaCards: the human disease database. Journal of the Medical Library Association: JMLA, 106(1), 140.

[13] Alexander J. Gates, Qing Ke, Onur Varol & Albert-László Barabási(2019). Nature's reach: narrow work has broad impact. Nature 575(7781), 32-34