なぜ単語分散表現に代数的構造が現れるのか
― 二部グラフを用いた単語共起関係の類型化の提案 ―

Why does the algerbaic structure emerge in word distributed representations?

前田 晃弘 † 鳥居 拓馬 ‡ 日髙 昇平 †
Akihiro Maeda, Takuma Torii, Shohei Hidaka
† 北陸先端科学技術大学院大学, ‡ 東京電機大学
Japan Advanced Institute of Science and Technology, Tokyo Denki University
akihiro.maeda@jaist.ac.jp

概要

単語分散表現は，そのベクトル演算が単語の類推関係に対応するなどの数理的な性質を示す．この性質は単語共起分布に由来するが，共起分布の数理的構造は未解明である．本研究は，二部グラフを用いる分析手法により，バイクリークと呼ばれる完全二部部分グラフが，単語ベクトル間の関係性を数理的に特徴付けるとともに，言語上の意味関係に対応することを示す．さらに共起関係が二項関係であることに起因して，単語共起分布に代数構造が現れる機序を明らかにする．

キーワード：単語ベクトル，単語共起行列，二部グラフ，バイクリーク，代数構造，形式概念分析，意味関係

1. はじめに

自然言語処理で用いられる単語分散表現（単語ベクトル）が数理的な性質を持つことが知られている．例えば，類推関係にある四単語に対応するベクトルが平行四辺形をなす word2vec[1] の例のほか，類義語など関連する単語ベクトル間のコサイン類似度が高い [2]，上位語の単語ベクトルは下位語の射影で表される [3]，単語ベクトルが加法構成性を示す [4] などである．ニューラルモデルにより生成される単語ベクトルは，単語共起頻度の PMI（自己相互情報量）を成分とする単語共起行列の行列分解を近似する可能性が論じられており [5]，また，機能的にも共起頻度を成分とする単語共起行列と同一視できると理解されている [6]. 従って，単語ベクトルの数理的性質は単語共起分布の構造を反映したものと考えられる．単語ベクトルを理論的に根拠づける分布仮説を唱えた Harris も，言語の分布に代数的構造が内在することを示唆している [7]．しかしながら，これまで共起分布自体の数理的構造や性質は十分に解明されたとは言えない [6, 8].

本研究では，単語共起分布の数理的な構造解明のために，二部グラフを用いる分析手法を提案する．まず，単語共起行列をグラフの隣接行列としてみなし，共起関係を二部グラフの頂点間の辺として定式化する．バイクリークと呼ばれる完全二部部分グラフが単語ベクトル間の関係性を数理的に特徴付けるとともに，言語上の意味関係に対応することを示す．次に，バイクリークに形式概念分析を適用する．形式概念分析は，束論と呼ばれる数学の応用で，集合論並びに代数学と密接な関係を持つ [9]．バイクリークは，その間に定義される順序関係のもとで，束と呼ばれる代数系（構造を備えた集合）を構成する [10]．バイクリークを用いる提案分析手法により，単語の共起関係が二項関係であることに起因して，単語共起分布に代数構造が現れる数理的機序の解明が期待できる．

2. 単語共起行列中のバイクリーク

二部グラフは，グラフ理論において定義される**グラフ**のうち，その全ての頂点が，交わりを持たない二つの集合 V_1, V_2 のいずれかに属し，かつ，それらの同じ集合に属する頂点同士は隣接しないグラフのことである．辺の集合を E とすると，二部グラフは $\mathcal{G} = (V_1, V_2, E)$ と表され，V_1, V_2 を**独立集合**という．

独立集合の部分集合 $A \subseteq V_1, B \subset V_2$ の間で，任意の頂点が隣接するグラフ $(A \times B \subseteq E)$ を完全二部部分グラフ, または**バイクリーク**といい，(A, B) と表す．バイクリークのうち他のバイクリークの部分グラフでないものを**最大バイクリーク**という．

コーパスから単語共起行列を作成するには，ターゲットとする単語 w（ターゲット単語という）に対して，その文脈（通常，ターゲット単語の前後 n 語の範囲）にある単語 c（コンテクスト単語という）との組 (w, c) の出現頻度をカウントする．ターゲット単語とコンテクスト単語の共起は二項関係なので，これを二部グラフの隣接関係（辺）として定式化できる．本稿では，単語ペアが一度でも共起すれば，それらの単語を表す頂点の間に辺があるものとする．この時，単語共起行列は，ターゲット単語の語彙を V_1，コンテクス

ト単語の語彙を V_2 とする $|V_1| \times |V_2|$ の二値行列であり，グラフの隣接行列と見做される．

単語の部分集合の組 $(A \subseteq V_1, B \subseteq V_2)$ がバイクリークであるとき，二つのターゲット単語 $w_i, w_j \in A$ は，いずれも全てのコンテクスト単語 $\forall c_k \in B$ と共起している．w_1, w_2 がバイクリークに属するコンテクスト単語のみと共起する極端なケースでは，共起行列の行ベクトルを単語ベクトルとすると，w_1, w_2 のそれは同一となる．その時，自然言語処理でよく用いられるコサイン類似度は 1 となる．単語ペアのコサイン類似度が高いほど，より多くのコンテクスト単語を共有するので，それぞれの共起分布内において単語ペアを含むバイクリークの比率が高くなる関係にある．

コサイン類似度の限界は，単語のペアに対する量であるため四項類推関係のような 3 以上の単語を携えるより大きな構造を捉えることには必ずしも適さないことと，また，あくまでも二単語の共起分布の間の重なりのみに着目するので，類義関係や反意関係の識別が難しい [2] ことである．

他方，本稿で取り上げるバイクリークは，比較対象の単語だけではなく，複数の単語を跨ぐ構造を捕捉することができる．特に，類義関係や反意関係などの意味関係を特徴づけるには，ペア以外の単語も含めた多様なコンテクスト単語との関係や構造を捉える必要があると思われる．そこで，次にバイクリークの構造を調べるための方略として，形式概念分析を導入する．

3. 形式概念によるバイクリークの構造化

二部グラフの（最大）バイクリークは，形式概念分析において定義される**形式概念や前概念**と同じものである [11]. 形式概念はガロア接続の構造を持ち，順序関係を導入することで束構造をなす [9]．また，前概念の演算を定義して代数系を構成できる [10].

単語共起分布を特徴づけるバイクリークに，形式概念による解釈を適用して，単語の意味関係を抽出できることを示す．形式概念分析を適用するため，共起行列を表す二部グラフ $\mathcal{G}(V_1, V_2, E)$ において単語の部分集合 $A \subseteq V_1, B \subseteq V_2$ に対する二つの導出演算子 $\uparrow, \downarrow$ を定義する．

$$A\uparrow := \{c \in V_2 \mid (w, c) \in E \quad (\forall w \in A)\} \tag{1}$$

$$B\downarrow := \{w \in V_1 \mid (w, c) \in E \quad (\forall c \in B)\} \tag{2}$$

演算子 $\uparrow$ は，ターゲット単語の集合に対してそれらが共有するコンテクスト単語の集合を返し，演算子 $\downarrow$ はその逆を行う．

集合の対 (A, B) が式 (3) を満たすとき形式概念 (formal concept) といい，また，式 (4) を満たすとき前概念 (preconcept) という．

$$A = B\downarrow \quad \text{and} \quad B = A\uparrow \tag{3}$$

$$A \subseteq B\downarrow \iff B \subseteq A\uparrow \iff A \times B \subseteq E \tag{4}$$

前概念には，次のように順序関係 $\subseteq^2$ が定義される．

$$(A_1, B_1) \subseteq^2 (A_2, B_2) \text{ iff } A_1 \subseteq A_2,\ B_1 \subseteq B_2 \tag{5}$$

形式概念は二部グラフにおける最大バイクリークに，前概念は同じくバイクリークに対応しており，従って，形式概念は他の前概念に含まれない前概念である．

以下では，バイクリークと形式概念分析の考えを用いることで，複数の単語間の関係や構造がどのように記述できるかを論じる．

上位・下位関係 (Hyponym) animal-dog のような上位下位関係を，形式（前）概念を用いて，仮説的に次の様に特徴づける．上位語 w_1 と下位語 w_2 は，バイクリークを形成して，共通のコンテクスト単語群 $\{w_1\}\uparrow \cap \{w_2\}\uparrow \neq \emptyset$ をもつ．その上で，下位語 w_2 は，このバイクリークに含まれない追加的な共起単語を持つ．すなわち，$\emptyset \subset \{w_1\}\uparrow \subset \{w_2\}\uparrow$．これは，上位語 w_1 が w_2 を含む下位語に共通する性質を持つ一方で，下位語 w_2 は固有の性質を持つことを反映する．形式概念の導出演算子を用いれば，$\{animal\}\uparrow \subseteq \{dog\}\uparrow$，すなわち単語のコンテクスト間の包含関係として定式化される．この時，ガロア接続により，$\{animal\} \supseteq \{dog\}$ であり，animal が dog を包含する概念である状況を特徴づける．つまり，上位下位関係は，共起単語群を集合とする包含関係として定義できる．

類義関係 (Synonym) と反意関係 (Antonym) 類義関係の例として large-huge, 反意関係の例として large-small の対を考える．二組の対のコサイン類似度はいずれも高く，意味関係の種類を識別することが困難である．これに対し，導出演算子を適用して抽出された単語ペアを含むバイクリークを例示すると，共有するコンテクストの違いが式 (6),(7) のように現れる．前者は，「大きいもの」の集合であり，後者は「(大小を）測定するもの」である．

$$\{large, huge\}\uparrow = \{amassed, crowd, debts, \ldots\} \tag{6}$$

$$\{large, small\}\uparrow = \{dose, scale, portion, \ldots\} \tag{7}$$

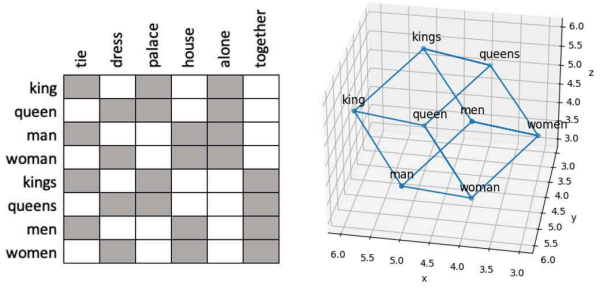

図 1　前田ら [12] のトイコーパスの単語共起行列（灰色=1, 白色=0）と，その各行に対応する単語ベクトルが成す平行六面体

さらに，それぞれのペア内の単語を差異化するコンテクスト単語を次のように抽出することができる．

$$\{large\}\uparrow \setminus \{large, huge\}\uparrow = \{intestine, \ldots\} \quad (8)$$

$$\{huge\}\uparrow \setminus \{large, huge\}\uparrow = \{ego, \ldots\} \quad (9)$$

$$\{large\}\uparrow \setminus \{large, small\}\uparrow = \{elephant, \ldots\} \quad (10)$$

$$\{small\}\uparrow \setminus \{large, small\}\uparrow = \{mouce, \ldots\} \quad (11)$$

式 (8),(9) は文脈の違い，すなわち，中立的・定量的な大きさと，抽象的・感覚的な大きさの違いであり，式 (10),(11) は同じ文脈内の対比，すなわち同じ集合内の部分集合間の補完関係として特徴づけることが考えられる．前概念を用いた代数系では，opposition と呼ばれる次の演算が定義されており [10]，

$$\lrcorner(A, B) := \Big((V_2 \setminus B)\downarrow, (V_2 \setminus B)\Big) \quad (12)$$

この演算を用いて，コンテクスト単語群が補集合を成すような二つの単語が反意関係にあると特徴づけることができる．すなわち反意関係を，同じ文脈内にある対比的な関係として定義できたことになる．

一方，類義語は，類似する意味的作用を持ち，一般に異なる文脈で用いられる（式 (8,9) の例）．こうした状況を特徴づけ，類義関係の定義を確定するためには，類義関係の背景にある，コンテクスト単語が属する複数のバイクリークの関係性や構造を明示的に取り扱う必要がある．

共有されるコンテクスト（コンテクスト単語の集合）と，差異化をもたらすコンテクストの関係性や構造を共起分布全体の観点から見通すために，次にトイモデルを用いた単語共起分布の分析を示す．

4.　単語分布に内在する代数構造

著者らの研究 [12, 13] は，4 単語の平行四辺形（四項類推）関係を拡張して，名詞 8 単語がそれらを

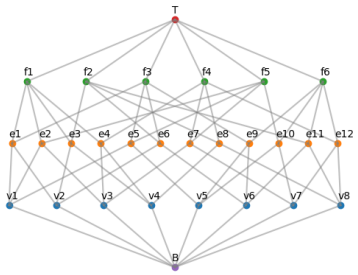

図 2　トイコーパスの共起分布に内在する束構造

頂点とする平行六面体を成す条件を探るため，図 1 に示した共起行列を分析した．各行は類推関係にある 8 つのターゲット単語 *king, queen, man, woman, kings, queens, men, women* に対応し，各列は各単語を特徴づける 6 つのコンテクスト単語 *tie, dress, palace, house, alone, together* に対応している．この共起行列を二部グラフとして見ると，行や列を入れ替えて灰色部分が長方形をなす箇所がバイクリークに対応する．例えば，1,3,5,7 行目と 1 列目を見ると，$(\{king, man, kings, men\}, \{tie\})$ が形式概念である．また，この形式概念は，男性の属性 (**masculine**) と解釈できる．この共起分布には，以下に代表される 3 タイプを含め 5 タイプの形式概念が存在する．

$$f_1 := (\{king, man, kings, men\}, \{tie\}) \quad (13)$$

$$e_1 : (\{king, man\}, \{tie, alone\}) \quad (14)$$

$$v_1 := (\{king\}, \{tie, palace, alone\}) \quad (15)$$

形式概念間に式 (16) の順序関係を定義することができ，この順序関係のもとで共起行列にある形式概念は束構造をなす（図 2）．

$$(A_1, B_1) \leq (A_2, B_2) \text{ iff } A_1 \subseteq A_2 \ (\text{or } B_1 \supseteq B_2) \quad (16)$$

図 1 に示す平行六面体に対応させると，i 番目の頂点に対応する形式概念 v_i と，それを含む辺に対応する形式概念 e_j は順序関係にあり，同様に，辺に対応する形式概念は，それを含む面に対応する形式概念 f_k との間に順序関係がある．すなわち形式概念が構成する束構造が，立体における単体間の関係（頂点，辺，面の包含関係）に対応している．

各頂点にはターゲット単語がそれぞれ付与され，その形式概念が識別される一方，面にはコンテクスト単語が対応し，次式に示すように，**masculine, feminine, royal, common, single, plural** の六つの意味上の属性を表す形式概念が対応すると解釈できる（式中の単語は先頭文字で略記）．

$$f_1 = (\{k, q, m, w\}, \{a\}) \qquad \textbf{single} \quad (17)$$

$$f_2 = (\{q, w, qs, ws\}, \{d\}) \quad \textbf{feminine} \quad (18)$$

$$f_3 = (\{k, q, ks, qs\}), \{p\} \quad \textbf{royal} \quad (19)$$

$$f_4 = (\{m, w, ms, ws\}, \{h\}) \quad \textbf{common} \quad (20)$$

$$f_5 = (\{k, m, ks, ms\}, \{t\}) \quad \textbf{masculine} \quad (21)$$

$$f_6 = (\{ks, qs, ms, ws\}, \{tg\}) \quad \textbf{plural} \quad (22)$$

この時，$V_1 = \{k, q, m, w, ks, qs, ms, ws\}$ として，形式概念の補完を次式により定義する．

$$\neg(A, B) := \Big((V_1 \setminus A), (V_1 \setminus A)\uparrow\Big) \quad (23)$$

すると，面に対応する形式概念の間には，$\neg f_1 = f_6$; $\neg f_2 = f_5$; $\neg f_3 = f_4$ の 3 組の補完関係が成り立つ．これは単数複数，男女性別，王平民の三軸が作る三桁の二進法であり，3-cube と呼ばれるブール代数を構成する．これにより，類推関係にある 8 単語のなす平行六面体の背景には，単語共起分布に内在する代数構造があることが明らかになった．

さらに，代数構造（厳密には閉包系）であることから，任意の頂点は三つの面すなわち属性の meet（共通部分）として表すことができる．

$$v_1 = f_1 \wedge f_3 \wedge f_5 \quad (24)$$

代数構造のもとで，各単語は属性の共通部分から構成され，個々の属性の間にある補完関係が単語間の意味関係を特徴づけている．前節で論じた類義関係も代数構造の中で特徴づけること可能であると期待されるが，今後の研究課題である．

5. 考察と結論

本研究は，単語共起関係を二項関係と捉え，単語共起行列において，同定されるバイクリークを形式（前）概念として解釈し，そこに代数構造が内在することを示した．また，類義語，反意語，類推などの単語間の意味関係がバイクリークの組み合わせパタンにより特徴づけられる可能性を示した．

その意義としては，第一に，バイクリークによる定式は，単語の意味を形式概念のなす代数構造の中で特徴づけるため，コサイン類似度よりも豊かな関係性を記述可能にする．

第二に，単語共起分布に代数構造が内在することから，単語の意味空間を代数系として構成する道筋が見えた．今後，単語の意味合成を代数演算として定義できれば，言語の意味理解や意味操作が計算と対応づけされ，大規模言語モデルの内部解明も期待できる．

第三に，大規模データからバイクリークへ分解する手法の研究は他分野において蓄積されており，実コーパスから形式概念を定量的に抽出可能である．

一方で, 今後の研究課題としては, 実データから解釈可能なバイクリークを選別することがある．ターゲット単語の前後の範囲にある単語を一律に共起単語とみなすと意味的に好ましくない共起ペアも含まれる．そのため確率的重み付きグラフを導入するほか，多義性に対応したサブコンテクストへ分解し，コーパス中の言語構造を取り出すモデルの開発が必要である．

謝辞

本研究は科研費基盤研究 B(一般)JP23H0369, JST さきがけ JPMJPR20C9 の助成を受けて行われた。

文献

[1] Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013) “Efficient Estimation of Word Representations in Vector Space”, International Conference on Learning Representations

[2] Jurafsky, D. & Martin, J. H. (2009) “Speech and language processing”, Pearson Prentice Hall

[3] Fu, R., Guo, J., Qin, B., Che, W., Wang, H., & Liu, T. (2014) “Learning Semantic Hierarchies via Word Embeddings”, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 1199–1209.

[4] Arora, S., Li, Y., Liang, Y., Ma, T. & Risteski, A. (2018) “Linear Algebraic Structure of Word Senses, with Applications to Polysemy”, Transactions of the Association for Computational Linguistics, Vol. 6, pp. 483–495.

[5] Levy, O. & Goldberg, Y. (2014) “Neural Word Embedding as Implicit Matrix Factorization”, Proceedings of the 27th International Conference on Neural Information Processing Systems, Vol. 2, pp. 2177–2185.

[6] Lenci, A. (2018) “Distributional Models of Word Meaning”, Annual Review of Linguistics, Vol. 4.

[7] Harris, Z. (1954) “Distributional structure”, Word, Vol. 10, No. 2-3, pp. 146–162.

[8] Gastaldi, J. L. (2021) “Why Can Computers Understand Natural Language?”, Philosophy and Technology, Vol. 34, No. 1, pp. 149–214.

[9] Ganter, B. & Wille, R. (2012) “Formal concept analysis: mathematical foundations”, Springer

[10] Wille, R. (2004) “Preconcept Algebras and Generalized Double Boolean Algebras”, Concept Lattices, pp. 1–13.

[11] Chiaselotti, G., Ciucci, D. & Gentile, T. (2015) “Simple undirected graphs as formal contexts”, Formal Concept Analysis: 13th International Conference, ICFCA 2015, Vol. 13, pp. 287–302.

[12] 前田晃弘, 鳥居拓馬, 日髙昇平, (2022) “単語共起行列の内部構造解明のための構成論的アプローチ” 2022 年度日本認知科学会第 39 回大会

[13] Torii, T., Maeda, A & Hidaka, S. (2022) “Embedding parallelepiped in co-occurrence matrix: simulation and empirical evidence”, Joint Conference on Language Evolution (JCoLE2022),