F-021

MEET 演算を用いた組合せ集合間の類似度の定義と応用

Similarity between Combinatorial Sets Using MEET Operation and Its Application

竹内 文登 † 鈴木 浩史 † 白石 恒介 ‡ 井上 祐馬 † 湊 真一 †
Fumito Takeuchi Hirofumi Suzuki Kousuke Shiraishi Yuma Inoue Shin-ichi Minato

1. イントロダクション

購買履歴や文書などの多くの現実のデータは組合せ集合とみなすことができる．このような組合せ集合に対して，それらの類似度や距離を定義することは，データマイニングや機械学習などの分野の重要な応用に繋がると考えられる．

集合間の類似度として Jaccard 係数が知られている．2 つの集合 F, G に対して Jaccard 係数は，$|F|$ を集合 F の要素数として，$\mathrm{Jaccard}(F, G) = |F \cap G|/|F \cup G|$ で定義される．しかしながら，Jaccard 係数は 2 つの集合の要素の有無のみを見ているため，組合せ集合に含まれる組合せの関係を考慮することができない．よって，組合せ集合間の類似度としては不十分であり，組合せの関係を考慮できる類似度の定義が望まれる．

一方で，正規化情報距離の考え方に基づいた組合せ集合間の距離尺度の定義が，細川らによって提案されている [3]．細川らの定義は，MEET 演算 [1] と呼ばれる組合せ集合間の演算を用いており，組合せ間の関係を考慮した距離尺度になっている．しかしながら，値が負値や不定値となる場合が存在し，距離尺度として望ましくない性質を持っている．

そこで我々は，MEET 演算を用いた組合せ集合間の類似度として望ましい性質を持った定義を与える．また，応用例として Twitter ユーザの分類に適用し，その妥当性を検証する．

2. 組合せ集合と MEET 演算

組合せ集合とは「n 個のアイテムから任意個を選ぶ組合せ」を要素とする集合である．例えば a, b, c, d という四つのアイテムに関して，$\{ab, d\}, \{abc, bd, cd\}$ は，いずれも組合せ集合の一例である．F の組合せのアイテム数の総和 $\|F\|$ を F の総アイテム数とよぶ．たとえば，$\|\{abc, bd, cd\}\| = 3 + 2 + 2 = 7$ である．

組合せ集合には，和集合や共通部分集合などを求める演算に加えて，本稿で扱う MEET 演算 [1] を定義することができる．MEET 演算は組合せ集合における二項演算として定義され，演算子として “$\sqcap$” を用いる．組合せ集合 F と G の MEET 演算の結果 $F \sqcap G$ は次式で定義される．

定義 1. *MEET* 演算

$$F \sqcap G = \{\alpha \cap \beta \mid \alpha \in F, \beta \in G\}$$

定義より MEET 演算は，F に含まれる組合せ α と G に含まれる組合せ β の任意のペアに対して共通部分集合を求めるという演算である．つまり，MEET 演算は F と G に共通して出現する組合せを列挙する演算であり，$F \sqcap G$ を計算することで F と G に共通して現れる組合せがどのくらいが存在するかを求めることができる．たとえば，$\{abc, ac\} \sqcap \{ab, cf\} = \{abc \cap ab, abc \cap cf, ac \cap ab, ac \cap cf\} = \{a, ab, c\}$ となる．

また，MEET 演算は和集合演算に対して分配的であることが知られている．つまり，

$$F \sqcap (G \cup H) = (F \sqcap G) \cup (F \sqcap H)$$

が成立する．

†北海道大学大学院 情報科学研究科
‡北海道大学工学部情報エレクトロニクス学科 (現在，株式会社エム・オー・シー)

3. MEET 演算を用いた組合せ集合間の類似度

二つの組合せ集合 F, G が共通して多くの組合せを持つならば，F と G は類似していると考えるのが自然である．共通して現れる組合せは MEET 演算を用いて計算することができるので，$F \sqcap G$ の総アイテム数 $\|F \sqcap G\|$ が多いならば，F と G は類似していると考えられる．たとえば，$F = \{abc, def\}, G = \{abc, de\}$ に対して $\|F \sqcap G\| = \|\{abc, de\}\| = 5$ であり，MEET 演算により多くのアイテムが出現する．このとき，F と G の違いは def という組合せと de という組合せ間におけるアイテム 1 つ分だけであり，これらは類似していると言えるであろう．そこで，$\|F \sqcap G\|$ を組合せ集合間の類似度に利用することを考える．

しかしながら，$\|F \sqcap G\|$ は F または G の総アイテム数が多いほど大きな値を取りうる．一方で，類似度は F や G の総アイテム数に影響されないことが望ましい．たとえば，$F = \{abc, def\}, G = \{a, b, c, d, e, f\}$ に対して $\|F \sqcap G\| = \|\{a, b, c, d, e, f\}\| = 6$ であり，多くのアイテムが出現する．しかし各アイテムについて見てみると，F では abc や def という組合せで出現しているのに対して，G では個別に出現しているため，F と G はあまり類似していないと判断したい．このため，$\|F \sqcap G\|$ を類似度に利用するためには，F や G の総アイテム数に影響されないように正規化を行う必要がある．

正規化として，類似度を 0 以上 1 以下の値にすることを考える．つまり，0 に近いほど類似しておらず，1 に近いほど類似していると判断する．ここで，MEET 演算が和集合演算に対して分配的であることから $F \sqcap G \subseteq (F \cup G) \sqcap (F \cup G)$ であることがわかる．したがって，$\|(F \cup G) \sqcap (F \cup G)\|$ は $\|F \sqcap G\|$ の上界となっている．そこで我々は，$\|F \sqcap G\|$ を $\|(F \cup G) \sqcap (F \cup G)\|$ で割ることで正規化を行い，次式で与えられる類似度を定義する．

定義 2. 組合せ集合間の類似度

$$S_{\mathrm{MEET}}(F, G) = \frac{\|F \sqcap G\|}{\|(F \cup G) \sqcap (F \cup G)\|}$$

先述の例を用いると，$F = \{abc, def\}, G = \{abc, de\}$ に対して $S_{\mathrm{MEET}}(F, G) = \|\{abc, de\}\|/\|\{abc, de, def\}\| = 5/8$，のように，組合せ集合間の類似度が算出される．

この類似度には次の性質がある．

1. $0 \leqslant S_{\mathrm{MEET}}(F, G) \leqslant 1$
2. $F = G \Longrightarrow S_{\mathrm{MEET}}(F, G) = 1$
3. F と G に共通して現れるアイテムが存在しない $\Longleftrightarrow S_{\mathrm{MEET}}(F, G) = 0$

また，特別な場合として F, G に含まれる全ての組合せが 1 つのアイテムからなるとき，我々の定義は Jaccard 係数に一致する．たとえば，$F = \{a, b, c\}, G = \{b, c, d\}$ などの場合である．このとき，$F \sqcap G = F \cap G, (F \cup G) \sqcap (F \cup G) = F \cup G$ であり，$\|F \cap G\| = |F \cap G|, \|F \cup G\| = |F \cup G|$ であるから，$S_{\mathrm{MEET}}(F, G) = |F \cap G|/|F \cup G| = \mathrm{Jaccard}(F, G)$ となる．

我々の定義では，MEET 演算により組合せの関係を考慮することができるので，二つの組合せ集合が同じアイテムから構成される組合せ集合であっても，それらの類似度が 1 とは限らない．また，MEET 演算の結果には同じアイテムが

複数の組合せに現れることがある．これは，そのアイテムが元の二つの組合せ集合に複数回現れるときであり，この性質からアイテムの頻度を考慮できると考えられる．

4. 実験：Twitter ユーザー分類への応用

4.1 目的と準備

提案した類似度が，組合せを考慮した妥当な類似度であるかを検証する．そのために，応用例の一つとして考えられる，Twitter* ユーザーの分類へ適用し，妥当性を検討する．

実験には 100 人のそれぞれ 200 ツイートを用いた．100 人の内訳は，声優，TV タレント，サッカー選手，野球選手，経営者それぞれ 20 人である．これらのツイートを MeCab** を用いて形態素解析を行い名詞だけの単語の組合せとし，各ユーザのツイート集合を組合せ集合とした．このとき，指示語やひらがな一文字のみ，アルファベットのみからなる単語は削除した．

本実験では，すべてのツイート集合間の類似度を計算し，スペクトラルクラスタリングと呼ばれる手法を用いて，k-means によりクラスタリングを行った [2]．この手法には，距離行列（または非類似度行列）が必要となるため，非類似度 (二つの組合せ集合が一致するとき値が 0 となる尺度) として $1-\mathrm{S}_{\mathrm{MEET}}(\mathsf{F},\mathsf{G})$ を算出してこれを距離行列として用いた．実験では，5 つのクラスタにクラスタリングを行い正解率を計算する．ここで，正解率とは，声優，TV タレント，サッカー選手，野球選手，経営者のそれぞれ 20 人ずつを正解のクラスタとし，100 人のうち何人が正解のクラスタに分類されるかの値である．

また，我々の定義が組合せを考慮していることを確かめるために，組合せを考慮せずにツイート集合を単なる単語の集合として Jaccard 距離を算出した場合との比較を行った．Jaccard 距離は $1-\mathrm{Jaccard}(\mathsf{F},\mathsf{G})$ で定義される．Jaccard 距離は組合せ集合を扱うことができないため，ツイート集合に含まれるすべてのツイートを繋げて一つのツイートとみなし，そこに現れる単語の集合を扱った．

実験には，CPU : 1.7GHz Intel Core i5，Memory : 4GB，OS : OSX Yosemite 10.10.3 のマシンを用いた．

4.2 実験結果

得られた非類似度行列を表現したヒートマップを図 1. に示す．ヒートマップは，各行または各列は一人のユーザの組合せ集合に対応し，上または左から 20 人ずつがそれぞれ声優，TV タレント，サッカー選手，野球選手，経営者に対応している．交差する点には二人のユーザの組合せ集合間の非類似度を白黒の濃淡で表しており，非類似度の値が小さければ黒，大きければ白に近づくようになっている．

また，100 人のユーザをクラスタリングしたときの正解率を表 1. に示す

4.3 考察

実験結果より，5 つにクラスタリングしたときの精度は Jaccard 距離を用いた場合より高いことが分かる．これには次の二つの要因が考えられる．まず一つ目の要因として，サッカー選手間や TV タレント間の類似度が Jaccard 係数を用いたときに比べて，MEET 演算を用いた類似度の方が高くなっていることが挙げられる．これは，我々の定義がアイテムの頻度を考慮していることが理由であると考えられる．たとえば，$\mathsf{F}=\{\mathfrak{abc},\mathfrak{abd}\},\mathsf{G}=\{\mathfrak{ab},\mathfrak{abd},\mathfrak{abcde}\}$ に対して，$\mathsf{F}\sqcap\mathsf{G}=\{\mathfrak{ab},\mathfrak{abc},\mathfrak{abd}\}$ となり，複数回 $\mathfrak{ab}$ が数えられることがある．一方で，Jaccard 係数は，アイテムが複数回数えられることはない．このように，ある単語がどちらのツイート集合にも複数回現れる場合は，Jaccard 係数よりも類似度が高くなることがある．二つ目の要因として，声優と TV タレントの間の類似度が Jaccard 係数よりも，MEET 演算を用いた類似度の方が比較的低くなっていることが挙げられる．これは，二つの組合せ集合が同じアイテムをもつ組合せ集合でも，組合せが異なれば，類似度が低くなることが理由であると考えられる．たとえば，$\mathsf{F}=\{\mathfrak{ab},\mathfrak{cd}\},\mathsf{G}=\{\mathfrak{ac},\mathfrak{bd}\}$ の場合，どちらの組合せ集合も $\mathfrak{a},\mathfrak{b},\mathfrak{c},\mathfrak{d}$ のみからなるので，Jaccard 距離は 0 となってしまうが，$\mathsf{F}\sqcap\mathsf{G}=\{\mathfrak{a},\mathfrak{b},\mathfrak{c},\mathfrak{d}\}$ より，$1-\mathrm{S}_{\mathrm{MEET}}(\mathsf{F},\mathsf{G})=1-1/3=2/3$ となり，これらの類似度は低くなっている．

*https://twitter.com/

**http://mecab.sourceforge.net/

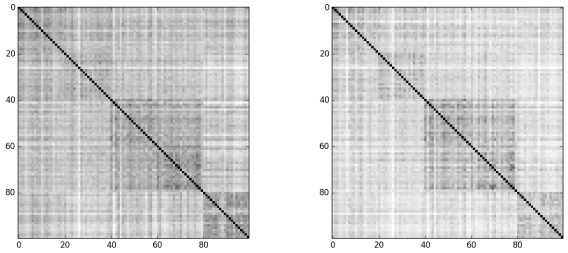

図 1: Jaccard 距離（左）と MEET 非類似度（右）のヒートマップ

表 1: クラスタリングの正解率の比較

クラスタ数	Jaccard 距離	MEET 非類似度
5	69%	82%
3	98%	88%

また，声優と TV タレントは芸能関係として，サッカー選手と野球選手をスポーツ選手としてまとめて考えることができるので，芸能関係 40 人，スポーツ選手 40 人，経営者 20 人の 3 つのクラスタを正解とし，同様の実験を行った．結果を表 1. に示す．結果より Jaccard 距離を用いた場合の 3 クラスタへの分類と 5 クラスタへの分類実験において正解率に大きな差が出ている．この原因の一つとして，声優と TV タレント，サッカー選手と野球選手の区別ができていなかったことが挙げられる．一方で，我々の定義が 3 クラスタと 5 クラスタの分類実験において，正解率にあまり変化がないことから，声優と TV タレント，サッカー選手と野球選手の区別ができており，より細かな分類ができることが期待できる．一方で，我々の定義がクラスタ数によらず一定の誤差があることに関しては,「年，月，日」などの職業とは無関係に頻出するアイテムの組合せに影響を受けているのではないかと考えている．

5. まとめ

本稿では MEET 演算を用いた組合せ集合間の類似度を定義した．Twitter ユーザの分類の実験から，我々の定義が，Jaccard 係数よりも組合せの関係を考慮した類似度の定義となっていることが確認された．

なお本研究は，JST 湊離散構造処理系プロジェクトからの助成による．

参考文献

[1] D. E. Knuth. *The Art of Computer Programming, Volume 4, Fascicle 1: Bitwise Tricks & Techniques; Binary Decision Diagrams.* Addison-Wesley Professional, 12th edition, 2009.

[2] J. Poland and T. Zeugmann. Clustering the google distance with eigenvectors and semidefinite programming. In *Knowledge Media Technologies, First International Core-to-Core Workshop*, pages No.21 61–69, 2006.

[3] 細川拓也. ZDD とコルモゴロフ複雑性を利用した Twitter のユーザ分類. 情報処理学会第７６回全国大会, 2014.