DEIM2021 F31-4

因果関係に基づく類似出来事の検索

澄川　靖信†

† 東京都立大学　〒 192–0364 東京都八王子市南大沢 1-1
E-mail: †ysumikawa@acm.org

あらまし　過去の出来事の因果関係を現代に類推させながら活用することが重要視されている。これまでの出来事どうしの類似度を測定する手法は、一つの文章で出来事が記述されている事を前提としており、因果関係を明示的に用いた出来事どうしの類似度を測定するものは少ない。本研究では、因果関係の類似度を考慮した出来事どうしの類似度を測定するアルゴリズムを提案する。本手法は、比較する出来事の因果関係間の類似度を重みとする二部グラフ上の最大重みマッチング問題を解く。本手法の有効性を評価するために先行研究と比較したところ、本手法の方が良い精度で類似度を測定できることを確認した。

キーワード　出来事検索、因果関係、二部グラフ、最大重みマッチング

1 はじめに

歴史をよく知り、その知識を現代社会で活用する能力の重要性が近年注目されている [1] [2]。実際、歴史を学習する授業は多くの国で小学校から開講されていることや、単に歴史を暗記するだけでなく現代社会で知識を活用できる能力を育成することを重視するカリキュラムや学習支援研究が行われていることから、歴史に重要性は広く認識されている。

「歴史は繰り返すのではなく、リズムが繰り返される」という言葉があるように、過去や現代で生じている出来事は、互いに類似する点だけでなく異なる点も含まれていることが多い。すなわち、過去の知見を現代で活用するためには、単に出来事の記述文で用いられている単語の類似度だけでなく、因果関係のような関係性の類似度を考慮して現代に類推させることができると望ましいと考えられている [3]。

出来事を検索する研究はこれまでに行われてきているものの、それらの多くはその記述文で使われている単語 [11] やそのカテゴリ [8] がどの程度似ているのかを分析しているので、因果関係は考慮していない。歴史教育の研究において因果関係の類似度を求めさせる学習方法も研究されているが、その学習テーマは研究者が選んだ特定のもののみが対象になるので、あらゆる出来事を対象にすることは難しい。

本研究では、長期的な因果関係に基づいた出来事の類似度を測定するためのアルゴリズムを提案する。本研究では出来事の因果関係を複数のサブ出来事の組みによって表現し、そのサブ出来事どうしの類似度の総和によって出来事どうしの類似度を求める。

本手法を適用するために類似度を求める二つの出来事に対して二部グラフを構築し、サブ出来事の類似度の総和が最大となる組み合わせを求めるために最大重みマッチングを本手法は解く。本研究で構築する二部グラフは出来事を表し、因果関係の類似度を求めることを目的としているので、最大重みマッチングの解となる辺には「互いに交点が存在しない」という制約を与える。この制約によって、サブ出来事どうしを因果関係の原因と結果に対応付けながらその関係性の類似を評価できる。

2 準　　備

2.1 定　　義

2.1.1 出来事、サブ出来事、因果関係

本稿では**出来事**を、複数の**サブ出来事**を統括するものと定義する。例えば、第二次世界大戦を用いてこの定義を考慮すると、出来事は第二次世界大戦、サブ出来事として真珠湾攻撃やポツダム宣言がある。本稿では**因果関係**を、2 つ以上のサブ出来事の集まりとする。なお、この集まりの中ではそれらが生じた時系列で全順序が定義できるものとする。

2.1.2 二部グラフ

二部グラフ G は二つの節集合 A と B の組み (A, B) である。二部グラフの辺集合 E は、A と B の節同士の類似度を重み辺 $e = (a_i, b_i)$ の集合である。

本手法を適用するとき、すべてのサブ出来事に対して節を定義する。

2.1.3 順序付き二部グラフ

順序付き二部グラフ G は、二つの節集合 A と B のおいて全順序を定義したものと定義する。すなわち、$a_i, a_j \in A$ の添え字 i, j に全順序を定義したものである。

2.2 最大重みマッチング

与えられたグラフ G に対して、マッチングとは、端点を共有しない辺の集合 $M \subset E$ である。最大重みマッチングは M として選択された辺の重みの総和が、最も高いもの $M' \in M$ を求める問題である。

3 関 連 研 究

3.1 出来事検索

出来事同士の類似性を評価する手法は、新聞記事で報道されたものを対象とする研究 [4], [6], [9]、Twitter 上にポストされた

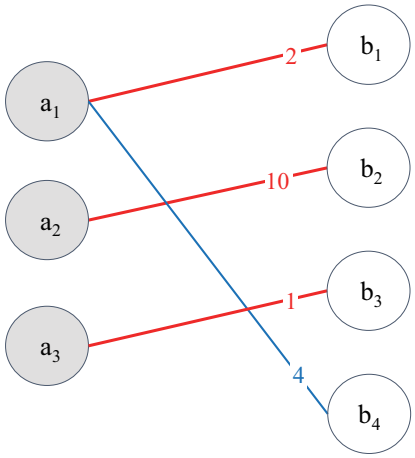

図 1　二部グラフの例.

ものを対象とした研究 [5]、などがこれまでに数多く行われてきた。これらの先行研究は、本研究で構築する二部グラフ G を用いて表現すると、G の 2 つの節集合 A と B の大きさをそれぞれ 1 とし、1 つしか無い辺の重みを求めるための手法である。一方、本研究の提案手法は辺の重みを計算するためのではなく、その値を用いて因果関係を考慮した出来事全体の類似度を測定する。すなわち、上記の先行研究と本研究は目的が異なり、本手法で用いる二部グラフの辺の重みを計算するために先行研究の手法を用いることができる。

3.2　因果関係抽出

出来事の因果関係を抽出する研究は、将来起こりうる出来事を予測するために過去の出来事の記述の中から因果関係を抽出する手法 [7]、現代社会における技術発展の影響を分析するための因果–効果関係を抽出する手法 [10] が提案されている。

これらの研究は本研究と同様に出来事の因果関係に注目しているが、本手法は出来事間の因果関係は既に与えられていることを仮定し、その因果関係を用いて出来事同士の類似度を測定する。

4　提 案 手 法

本節で本研究で解決する最大重みマッチング問題について説明するために提案手法の理論を最初に述べる。次に、提案手法の動的計画法による実現方法を述べる。

4.1　理　　論

本研究では従来の最大重みマッチング問題に対して、解となる辺には「互いに交点が存在しない」という制約を与える。図 1 にこの制約によって得られる解の違いを示す。

例：図 1 を用いて一般的な最大重みマッチング問題と本研究の解の違いを示す。なお、本研究ではグラフは平面空間上で定義することを仮定する。黒線はどちらの問題でも共通して解となる辺、青線は一般的な最大重みマッチング問題のみで解となる辺、赤線は本研究のみで解となる辺を示す。節 a_1 には 2 つの辺が接続している。節 b_4 にも接続している辺の重みの方が、もう

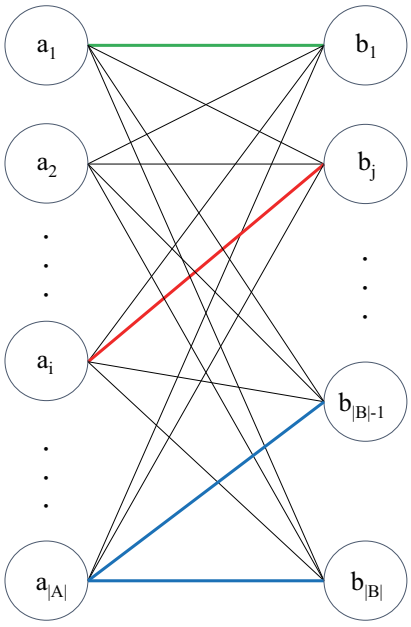

図 2　制約付き最大重みマッチングの例.

一方の辺よりも高い値なので一般的な最大重みマッチングでは解として選ばれる。しかし、辺 (a_2, b_2)、(a_3, b_3) と交点を持つので本研究では解とせず、もう一方の辺 (a_1, b_1) を解とする。

4.1.1　二部グラフの構築

本研究で定義する二部グラフ $G = (A, B)$ は因果関係を表す出来事なので、節集合 A, B それぞれの内部の節は、その添え字番号に対して順序を定義している。また、因果関係は複数の出来事の集合 $SubG$ で表現しているので、因果関係の類似度は $SubG$ のサブ出来事どうしの順序と内容の類似度に帰着させることができる。

因果関係を表すサブ出来事を節とし、節となる出来事どうしの類似度を辺の重みとする順序付き二部グラフを定義する。

4.1.2　制約付き最大重みマッチング

本手法の順序付き二部グラフはサブ出来事を節とし、因果関係はサブ出来事を列挙することによって表現するので、因果関係に基づいた類似度はサブ出来事どうしの類似度を、一対一で求めながら、その順序を守ることと言い換えることができる。

本稿では、因果関係に基づいた出来事どうしの類似度を、第 4.1.1 節で定義した二部グラフ上で、交点を持たない辺の組み合わせの中で、それらの重みが最大となるものと定義する。

$e = (a_i, b_j)$ と $e' = (a_k, b_l)$ を解をなす辺 $(e \neq e')$ とする。このとき、解をなす辺集合が互いに交点を持たないを以下のように定義する。

- もし $k < i$ ならば $l < j$ である。
- もし $l < j$ ならば $k < i$ である。

4.2　実　　装

本手法は動的計画法によって実現する。本節では、まず、図 2 と図 3 を用いて本研究で提案した二部グラフ上での制約付き最大重みマッチング問題の動的計画法による解法の例を示す。その後、本手法を一般化するために定義やアルゴリズムを示す。

	1	2	3	⋯	j	⋯	|B|
1	70	66	99	57	56	76	94
2	2	18	73	10	82	69	3
3	27	26	13	96	79	89	22
⋯	58	85	54	38	46	67	30
i	8	55	14	78			
⋯							
|A|							

図 3　動的計画法による解法.

	b_1	b_2	b_3	b_4
a_1	2	0	0	4
a_2	0	10	0	0
a_3	0	0	1	0

(a) W

0	0	0	0	0
0	2	2	2	6
0	2	12	12	12
0	2	12	13	13

(b) DP

図 4　動的計画法による図 2 の解.

例：図 2 と図 3 は解の対象となるか解析を終えた辺を緑、現在解析している辺を赤、赤を解として選択した場合にこれから解析対象とする辺を青で示す。

本手法は二部グラフ $G = (A, B)$ の節集合 A, B に対して添字の昇順で解析する。このとき、もし現在解析している辺を解として選択したなら、それまでに解として選択した辺集合に加えて重みの総和を表にメモする。すなわち、図 3 の矢印が表すように、緑の領域から選択済みの辺の重みを赤色のセルへ流していき、最終的にその値が最も高い辺を解とする。このとき、「辺が交点を持たない」という制約を満たすために、選択した辺の添字よりも大きい値の添字は解として選択できないように対象から除外する。

本手法は辺の重みを記録する表 W と、選択済みの辺の重みの総和を記録する表 DP を用いる。

例：図 1 に対して本手法で作成する 2 つの表を図 4 に示す。図 4(a) に示すように、辺が未定義の場合には表に 0 を記録する。表 DP は選択済みの辺集合の重みの総和を記録するので、各セルには、表 DP の左、左上、上の 3 つの値の中から最大のものと、現在のセルに対応する W の値を足した値を記録する。これを全てのセルに値を記録するまで反復する。この結果、表 DP の右下に本手法で選択した辺集合の重みの総和が記録できるので、その構成要素となる辺を逆方向に走査することで実際の辺集合を求める。

Algorithm 1 動的計画法による実装

Input: A weighted matrix W, tables DP and $range\_max$
Output: A set of edge $SubE$

Function $OGMM(W, DP, range\_max)$
  // Table calculation
  for i = 1 to $ColumnSize(W)$
    for j = 1 to $RowSize(W)$
      $prev\_max \leftarrow Val(range\_max_{i-1,j-1})$
      $DP_{i,j} \leftarrow prev\_max + W_{i,j}$
      $range\_max_{i,j} \leftarrow tuple\_val\_max(range\_max_{i-1,j}, range\_max_{i,j-1})$
      if $DP_{i,j} > val(range\_max_{i,j})$
        $range\_max_{i,j} \leftarrow (DP_{i,j}, (i, j))$
      end if
    end for
  end for
  // Answer determination
  $i, j \leftarrow ColumnSize(W), RowSize(W)$
  $SubE \leftarrow \emptyset$
  while $i \neq -1$ and $j \neq -1$
    $SubE$.append($e_{i,j}$)
    $i, j \leftarrow Index(range\_max_{i-1,j-1})$
  end while
  return $SubE$

W は二部グラフ $G = (A, B)$ の各辺の重みを要素とする行列とする。すなわち、引数で与えられた辺の重みを返す関数 $weight$ を用いて、W の各要素を、任意の $e_{a_i} \in A, e_{b_j} \in B$ に対して、$W_{i,j} = weight(e(a_i, b_j))$ と定義する。

本手法を動的計画法として定義するために表 DP を定義する。この表に上記の例で示したように値を帰納的に記録することで制約付き最大重みマッチング問題を解く。

第 4.1.2 節の定義より「もし $e(a_i, b_j)$ を解とするとき、直前までに解として選ばれていた辺に接続される節の添え字は i, j のそれぞれよりも小さい」となるので、もし $e(a_i, b_j)$ が解をなすならば、a_i や b_j に接続される他の辺を選択するよりも $e(a_i, b_j)$ を選択し、その重みを (0,0) から (i-1,j-1) の範囲までに求めた解集合の辺の重みの総和に足した結果は大きくなる。

この特徴を利用して、辺 $e(a_i, b_j)$ を解とした場合の計算結果を表 DP に記録するためには、$DP_{i-1,j-1}$ に $W_{i,j}$ を足した結果が、$DP_{i,j-1}$ や $DP_{i-1,j}$ よりも値が大きくなるはずである。さもなければ、$e(a_i, b_j)$ を解としない場合の方が値が大きくなっている。この考えを用いると、$DP_{i,j}$ の計算式は以下のようになる。

$$DP_{i,j} = max(DP_{i-1,j-1} + W_{i,j}, DP_{i-1,j}, DP_{i,j-1}) + W_{i,j} \quad (1)$$

Algorithm1 に本手法のアルゴリズムを示す。本手法を適用する前に、表 DP を 1 行目と 1 列目のすべての成分を 0 で初期化しておくことを仮定する。また、表 DP を埋めながら解となる辺を記録するために表 $range\_max$ を用いる。$range\_max$ は、解となる辺の重み $value$ と接続されている節の添え字 $index$ の組を成分とする。2 行目から 12 行目が表 DP の各成分に式 1 の計算結果を記録している。3 行目の $ColumnSize$ は引数の表の行数を返す関数、4 行目の $RowSize$ は引数の表の列数を返す関数、

表 1　評価用データセットの統計情報.

地震	222
地震と津波	57
地震と地崩れ	3
ウィルスの発生とワクチン開発	5
ウィルスの発生と治療	6
計	293

表 2　**p@1**.

コサイン類似度	DTW	提案手法
0.567	0.753	0.858

5 行目の *Val* は引数として与えられた値と添え字の組の中から値を返す関数、7 行目の *tuple_val_max* は引数として与えられた値と添え字の組の中から値が最大のものを返戻する関数、18 行目の *Index* は引数として与えられた値と添え字の組の中から添え字を返す関数である。このアルゴリズムでは、*DP* を埋めるのと同時に、解として選択した辺の添え字を *range_max* に記録している。解となる辺の添え字は昇順に並ばれているので、13 行目から 19 行目で示すような *range_max* の添え字を逆順で辿ることによって本手法で求める辺集合を得られる。

計算量. *DP* 行列と *range_max* を同時に計算できるので各 $DP_{i,j}$、$range\_max_{i,j}$ は定数時間で求めることができる。すなわち、時間計算量は $O(|A||B|)$ となる。一方、空間計算量は $O(|A||B|)$ である。

5 実　　験

5.1 データ収集

本稿では、Wikipedia に記載されいている地震とウィルスに関する出来事を用いて提案手法を評価する。特に、表 1 の 6 種類の Wikipedia 記事を収集し、各サブ出来事が記述された章・節・段落のいずれかを手動で抽出して二部グラフの節を定義した。

5.2 比 較 対 象

本稿では以下の手法を比較対象とする。

- TF-IDF+CosSim（コサイン類似度）: 特徴ベクトルを作成する際に TF-IDF を用いる。この特徴ベクトルに対してコサイン類似度を用いて文章間の類似度を評価する。
- TF-IDF+動的時間伸縮法（DTW）: DTW は時系列データ同士の類似度を、２つの時系列の各点の距離を総当たりで求め、全て求めた上で２つの時系列が最短となるパスを見つけることによって類似度を測る。

5.3 評 価 基 準

本稿では、上述した比較対象と本手法を、各手法を適用した結果の上位 1 件の結果が正解かどうかを表す p@1 によって評価する。

5.4 結　　果

表 2 に各手法の p@1 の結果を示す。この結果から提案手法が一番良い結果が得られたことがわかる。また、コサイン類似度が最も精度が低いので因果関係を考慮する際にはサブ出来事の順序を考慮するのが出来事同士の類似度を評価する際には重要であることがわかる。

6 おわりに

本稿では因果関係に着目した出来事同士の類似度を測定するための手法を提案した。本手法は因果関係を表すサブ出来事を節、各サブ出来事同士の類似度を辺の重みとする二部グラフ上で、解となる辺は交差しないという制約を与えた最大重みマッチング問題を解く。本手法を実現し、先行研究と比較したところ、本手法の方が良い精度が得られることを確認した。

今後の課題としては次の 2 つが考えられる。1 つ目は、因果関係を含む他の出来事に対して網羅的に評価を行う。2 つ目は、木構造で表現した因果関係を用いることによって問題の一般化を行う。

謝辞 本研究の一部は科研 (#19K20631) の助成を受けたものである.

文　　献

[1] Abelson, R.P., Levi, A.: Decision making and decision theory, handbook of social psychology. pp. 231–309 (1985)
[2] Gilovich, T.: Seeing the past in the present: The effect of associations to familiar events on judgments and decisions. Journal of Personality and Social Psychology **40**(5), 797 (1981)
[3] Ikejiri, R.: Designing and evaluating the card game which fosters the ability to apply the historical causal relation to the modern problems. Japan Society for Educational Technology **34**(4), 375–386 (2011). (in Japanese)
[4] Lei, Z., Wu, L.d., Zhang, Y., Liu, Y.c.: A system for detecting and tracking internet news event. PCM'05, pp. 754–764. Springer-Verlag, Berlin, Heidelberg (2005)
[5] Manaskasemsak, B., Chinthanet, B., Rungsawang, A.: Graph clustering-based emerging event detection from twitter data stream. ICNCC'16, pp. 37–41. ACM, New York, NY, USA (2016)
[6] Qi, Y., Zhou, L., Si, H., Wan, J., Jin, T.: An approach to news event detection and tracking based on stream of online news. pp. 193–196 (2017)
[7] Radinsky, K., Davidovich, S.: Learning to predict from textual data. J. Artif. Int. Res. **45**(1), 641–684 (2012)
[8] Sumikawa, Y., Jatowt, A.: System for category-driven retrieval of historical events. JCDL '18, pp. 413–414. ACM, New York, NY, USA (2018)
[9] Tan, Z., Zhang, P., Tan, J., Guo, L.: A multi-layer event detection algorithm for detecting global and local hot events in social networks. Procedia Computer Science **29**, 2080–2089 (2014). 2014 International Conference on Computational Science
[10] Zhang, Y., Jatowt, A., Tanaka, K.: Causal relationship detection in archival collections of product reviews for understanding technology evolution. ACM Trans. Inf. Syst. **35**(1) (2016)
[11] 池尻良平, 澄川靖信: 真正な社会参画を促す世界史の授業開発 – その日のニュースと関連した歴史を検索できるシステムを用いて–. 社会科研究 **84**, 37–48 (2016)