ペナルティを付与する因果関係の類似度測定

町澤　大輔　　澤畑　尚希　澄川 靖信
拓殖大学

１はじめに

過去の起きた出来事を分析することの重要性は様々な分野で知られている．例えば，類似する過去と現在の因果関係の違いの分析は，現代の課題に対する解決策を考える土台となることが知られている．このような分析を有効にするためには単一の出来事のみを対象とするのではなく，一連の出来事をタイムラインとしてまとめたものを対象とすることが重要である．

過去の出来事は，Web 上に膨大な量が蓄積されているため，入力したタイムラインに類似するタイムラインを出力する検索技術は重要であり，そのアルゴリズムとして ECM [1]が提案されている．しかし，ECM は出来事同士の類似度を，順序を考慮しながら計算するために最大重みマッチングを求めるため，似ていない出来事が存在しても類似度の結果に反映されないことがある．

本研究では，2 つのタイムラインの類似度を評価するとき，出来事同士の類似度とその順序だけを求めるのではなく，類似しない出来事の存在や出来事数の不一致に対して類似度を低減するために Needleman-Wunsch (NW) [2]アルゴリズムを拡張した手法である PASS を提案する．

本手法の有効性を Wikipedia の出来事データを使用して評価したところ，先行研究よりも入出力間のタイムラインには似ていない出来事が少ないことを確認した．

２提案手法

図 1 に本手法によるタイムライン間の類似度を測定するときに似ていないデータに対してそのスコアを低減する例を示す．この例ではクエリとして入力したタイムラインには含まれない「電気系統からの出火」という出来事がタイムライン B に含まれている．その前後の出来事はクエリと一致していているため，クエリのタイムラインの 3 番目に空出来事を意味するギャップを挿入する．

このようなギャップを挿入することで，同じ出来事が同じ順序で発生していることと不一致が起きていることの両方を解析できる．

Timeline Similarity with Penalty Adjustment
Disuke Machizawa, Takushoku University
Naoki Sawahata, Takushoku University
Yasunobu SUmikawa, Takushoku University

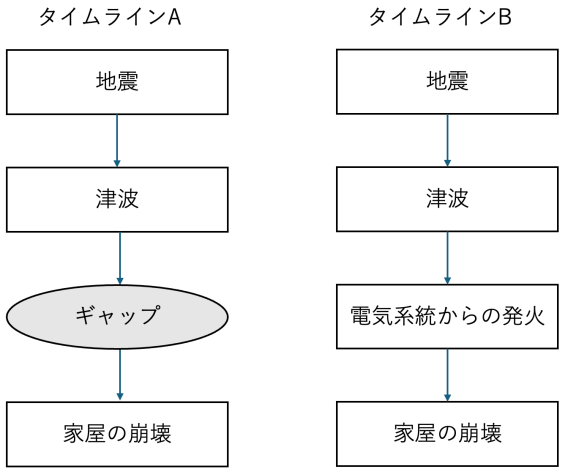

図 1 本手法によるタイムライン類似度計算

ギャップの挿入位置は，ペアワイズアライメントを実現する NW アルゴリズムと同様に求める．NW アルゴリズムは DNA 配列を構成するアルファベット同士が完全一致するかを解析するので，本研究ではタイムラインの要素が出来事の文章である配列とみなし，NW アルゴリズムを文章同士の類似度を求めるように拡張する．本手法は NW アルゴリズムと同様に動的計画法を用いて計算するので，まず，表 DP を定義する．類似度を比較する 2 つのタイムラインをそれぞれ $X=\{x_1,x_2,x_3 \ldots x_n\}$,$Y=\{y_1,y_2,y_3 \ldots y_m\}$とすると，DP の 1 行目に X，1 列目に Y を配置して式(1)によって各セルの値を初期化する．

$$\begin{cases} DP_{0,0} = 0 \\ DP_{i,0} = gap * i \\ DP_{0,j} = gap * j \end{cases} \quad (1)$$

ここで，*gap* は 2 とする．その他の各セルの値は式(2)によって計算される．

$$DP_{i,j} = \max \begin{cases} DP_{i-1,j-1} + s(x_i, y_j) \\ DP_{i-1,j} - gap \\ DP_{i,j-1} - gap \end{cases} \quad (2)$$

$s(x_i,y_j)$は x_i と y_j の文章が類似しているかどうかを判定する関数である．もし一致していれば 1，さもなくば-1 を返戻する．この結果，セル $DP_{n,m}$ には，X と Y 全体の類似度が記録される．

本研究では X と Y の各要素は文章であること

を仮定している．そのため $s(x_i,y_j)$では文章同士の類似度の評価を行う．まず，ストップワードの除去といった前処理を行い，特徴ベクトルを生成する．その後，コサイン類似度などの文章間類似度を求める手法を適用して類似度を数値で求める．この値が事前に設定した閾値以上ならば x_i と y_j は一致しているとみなし，さもなくば一致していないとみなす．

３データセット構築

様々な種類のタイムラインを用いて本手法の有効性を評価するために，本研究では Wikipedia の current portal に記録されている出来事を使用してタイムラインを手動で作成した．まず，2016 から 2017 年のすべての出来事データを収集した．このデータには，Wikipedia 編集者が手動で割り付けた Wikipedia current portal で定義されている出来事カテゴリが各出来事に 1 つずつ付与されているので，このカテゴリ情報も収集した．

このタイムライン作成は 3 人の作業者によって行なった．まず，2 名が 1 年分のすべてのデータを確認しながら別々にタイムラインを作成した．その後，3 人目の作業者が同様にすべてのデータを確認しながらタイムラインを作成した．この結果，2 人の意見が異なるタイムラインが発生したとき，両者が議論してタイムラインを修正した．この結果，5,204 件の出来事を収集し，2,879 件のタイムラインを作成した．

４実験

本稿では各検索アルゴリズムの出力結果ランキングの上位 10 件，50 件，100 件に対して Mean Squared Error (MSE) を用いて評価した．MSE は，予測値と真の値の差の二乗の平均を表す指標であり，値が小さいほど予測精度が高いことを示す．本実験では，タイムライン間の構造的類似性を適切に反映するため，類似する出来事は同じ出来事カテゴリに属するものと仮定し，各出来事に付与されたカテゴリの一致性によって MSE の値を算出した．具体的には，入力とみなしたタイムラインの各出来事のカテゴリ分布と各検索アルゴリズムによるランキング結果との誤差で MSE を求めた．

本手法の関数 $s(x_i,y_j)$の実現方法として文章間類似度を求める手法として広く利用されている Jaccard 係数の値で評価した．本稿ではこの値を PASS_Jaccard と表記する．また，タイムライン同士の類似度を評価する手法として， ECM を比較対象とした．さらに，出来事間の順序を考慮せずにタイムライン全体を 1 つの文章とみなして類似度を求める Jaccard 係数 (jaccard)，コサイン類似度も比較対象として実装した．コサイン類似度は TF-IDF によって生成された特徴ベクトルを用いて求めた．ECM では出来事同士の類似度求める必要があるため TF-IDF によって特徴ベクトルを生成し，コサイン類似度で算出した．

表 1 各手法の k=10, 50, 100 に対する MSE の値

手法	k=10	k=50	k=100
コサイン類似度	18.88	20.58	19.86
ECM	107.52	96.93	80.02
Jaccard	9.56	16.57	20.35
PASS_Jaccard	**8.34**	**12.89**	**14.95**

表 1 に各手法の MSE の値を示す．この結果から，提案手法の PASS_Jaccard は，全ての k の値において，比較対象よりも良い値が得られることを確認できる． ECM の結果と比較すると，k の値が 10 のときは 73.6 ポイントも改善している．これは似ていない出来事が含まれているときには類似度を低減することが良い結果を出力することを示す．また，提案手法は，Jaccard 単独での評価に比べて，MSE の値が改善している．すなわち，タイムライン検索においては類似する出来事が同じ順序で出現していることを解析することが重要であることを示している．

５まとめ

本研究では NW アルゴリズムを拡張することで，似た出来事が同じ順序で出現することを解析しながら，類似しない出来事が含まれるときには類似度を低減してタイムラインの類似度を測定する手法を提案した．この拡張によって，先行手法よりも高い精度で類似するタイムラインを検索できることを確認した．今後の課題として，出来事間の関係をタイムラインではない異なるグラフ構造で表現されていても適切に類似度を評価できる手法の提案が考えられる．

参考文献

[1]: Y. Sumikawa. Event causal relationship retrieval. In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT ’21, 318–325, New York, NY, USA, 2022. Association for Computing Machinery.

[2]: S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology, 48(3), 443–453, 1970.