Story Signature:ストーリー展開特徴抽出による類似小説検索可視化方式の実現

仲程 凜太朗† 岡田龍太郎†† 中西 崇文†

†武蔵野大学データサイエンス学部データサイエンス学科 〒135-8181 東京都江東区有明 3-3-3

††武蔵野大学アジア AI 研究所 〒135-8181 東京都江東区有明 3-3-3

E-mail: †s1922069@stu.musashino-u.ac.jp, tnakani@musashino-u.ac.jp

††ryotaro.okada@ds.musashino-u.ac.jp

あらまし 本稿では，小説のストーリー展開に着目した小説コンテンツ類似探索方式について示す．近年，電子書籍やウェブ小説をはじめとする小説コンテンツの電子化が定着し，膨大な小説コンテンツが Web 上に散在している．これらの小説コンテンツを対象としてユーザの嗜好に合致するコンテンツの検索・推薦を効率的に実現することが重要となってきている．小説は全体としての内容だけでなく，その内容がどのように移り変わるのかというストーリー展開が重要であると考える．ストーリー展開は時系列な文脈変化と捉えることができる．本方式では，時系列な文脈変化を単位文章数あたりの日本語評価極性辞書による各極性値として抽出し，これを Story Signature と定義する．本方式は，各小説コンテンツの Story Signature について動的時間伸縮法(DTW)に基づく類似度計量を実現することにより，ストーリー展開の似た小説コンテンツを探索することを可能とする．

キーワード Story Signature，類似小説検索，DTW，時系列クラスタリング

1. はじめに

近年，書籍の電子化が進んでおり，膨大な量の小説コンテンツがインターネット上に散在している．例えば，青空文庫[1]のように著作権切れの小説をアーカイブ化し公開するサイトや，小説投稿サイト”小説家になろう”[2]で投稿されたオンライン小説が人気を博している．様々な種類の膨大な量の小説コンテンツにより，我々はそれらの膨大な小説コンテンツにアクセスし楽しむ機会が増大した一方で，これらの膨大な小説コンテンツの中から自分の趣味嗜好に合致した小説コンテンツを検索・推薦する機能の実現が重要となってきている．

現在，”小説家になろう”[2]において，キーワード(単語)パターンマッチングによる検索機能が提供されている．また，投稿者およびユーザが付与した一定のジャンルを表すキーワードタグによる整理もされており，それらを指定することにより，同じジャンルの小説コンテンツにアクセスすることが可能である．これらの機能により，ユーザが発する単語に合致する小説コンテンツを見つけることが可能となっている．

一方，小説コンテンツの中身を評価するためには，ストーリー展開に着目することが重要であると考える．そのため，小説コンテンツのテキストを分析するにあたっては，時系列情報を利用することが必要になる．すなわち，単語の出現頻度から内容を把握するといったことだけでなく，その内容がどのように移り変わるのかに着目すべきであると考える．

本研究では，ストーリーの大まかな構造として，どこに盛り上がりがあるかを抽出し可視化することを目指す．盛り上がりを示すための指標としては，文章に現れる単語がポジティブな語であるかネガティブな語であるかという情報を利用する．これによって小説全体のストーリー展開を時系列情報として可視化する．さらにここで，抽出された時系列情報は波として捉えることができる．そのため，波の類似度を比較する手法を用いることにより，抽出された構造同士の類似度を算出することが可能となる．これにより，ストーリー展開に着目した小説コンテンツの類似度計量方式を実現する．

また，小説のストーリー展開を把握することは，小説の読者のみならず小説の作家にとっても直感的に自分の書いた小説の構造はどのようなものであるかを，自身の文章の特徴を客観的に判断するためにも意義深いものである．そのため，そのストーリー展開，時系列な文脈変化の可視化についても重要であると考える．

本稿では，小説コンテンツのストーリー展開を表す新たなメタデータである Story Signature を定義する．

小説コンテンツから小説コンテンツが描く感情のポジティブ・ネガティブを表す特徴量として，日本語評価極性辞書[3][4]による極性値を単位文章ごとに抽出する．これらの値は 1 つの小説内のストーリー展開に応じて変化する時系列のメタデータとみなすことができ，これを本稿では Story Signature と定義する．Story Signature はそれ自体をプロットすることで小説のストーリー展開の構造を可視化することができる．さらに，各小説コンテンツから抽出された Story Signature について，動的時間伸縮法(DTW)に基づく類似度計量を実現することにより，小説コンテンツ同士のストーリー展開に基づく小説コンテンツを検索することを可能にする．

本稿の構成は以下の通りである．2 節では関連研究を紹介し研究の位置づけを 明確にする．また 3 節で動的時間伸縮法(DTW)について示し，4 節で本研究の提案する Story Signature 抽出およびそれを用いた小説間類似度検索の実現方式とその可視化方式について述べる． 5 節で評価した実験を行い，6 節でまとめを示す．

2. 関連研究

本節では，本方式に関連する研究について挙げる．小説コンテンツを対象とした検索・推薦システムは，テキストマイニングの文脈から多数の研究がされてきた．ここでは，特に小説コンテンツに着目した関連研究について挙げる．

2.1 表紙の構成要素の推薦に関する研究

川口ら[5]は，小説データからその内容の印象に沿った表紙の構成要素を推薦するシステムを提案している．川口らが参考にした既存研究では，小説全体に現れる単語を同じ重みで扱っていたため，バッドエンドの話であっても幸福な場面が多ければ明るい表紙になる可能性があった．そのため，この研究では小説全体の本文を解析し，読者の印象に残ると考えられる場面を抽出した後，その場面のみを用いて 色・フォント・象徴物の推薦を行う手法を提案している．その推薦された色・フォント・象徴物を用いることで，容易に表紙を作成することが可能となるとしている．象徴物については，tf-idf を用いて抽出場面に出現する特徴的な名詞を獲得している．

この研究では,読者が印象に残ると考えられる場面を抽出することを目的としており，その場面のみの表紙を生成することを可能としている．本稿で述べる提案方式では，一つ一つの場面ではなく，時系列の文脈変化に着目する点で異なる．

2.2 小説の類似度算出に関する研究

小説の類似度算出に関して，高田ら[6]は，語彙などの文体が重要であるとし，小説投稿サイトに投稿された小説から文体に該当するいくつかの指標を抽出し，マハラノビス距離を用いて小説間の類似度を算出するオンライン小説推薦手法を提案・実装し，利用者実験により提案手法の有効性を検証している．

この研究では，小説コンテンツから文体に関する特徴量(例えば，読点数，読みの文字数，文字数，品詞数，文節数，助詞数，漢字の割合，ひらがなの割合，カタカナの割合，句読点間の読みの文字数，品詞の割合，読点の前の品詞の割合，句点の前の品詞の割合，オノマトペ数，直喩数，Type-Token Ratio(TTR))を抽出し，これらの特徴量の類似度を求めている．これらの特徴量は，小説全体を集約した特徴量であり，小説コンテンツの特徴に合致した精度の良い検索・推薦を実現している．ただし，本研究で扱うような小説内のストーリー展開を表す時系列の特徴量は考慮されていない．

2.3 文体の類似度に関する研究

文章の特徴を数値化する研究は，計量文体学あるいは計量文献学と呼ばれる．その活用方法としては，著者推定や,特定の種類の文書の特徴分析が主流である．

金川ら[7]は，作家ごとに構文構造を調査することで、文体の類似性を数値化し，作家同士の構文構造の類似性を計量する手法を提案している．

この研究では，テキストコンテンツに含まれる全体の文体を特徴量として抽出することで，作家間の特徴表現の違いを見出したり，文体の違いによる表現の異なる文章の類似性を求めるものである．本稿で述べる提案方式であるストーリー展開を表す時系列の特徴量を導入することにより，この方法の精度向上に貢献できる可能性がある，

2.4 本研究の位置付け

本研究では，小説コンテンツのストーリー展開に着目し，小説コンテンツのストーリーの流れを時系列データとして扱う．文章を単位文章に分割し，単位文章をそこに現れる単語を用いてポジティブ/ネガティブの極性値に変換する．これを Story Signature として定義する．さらに, Story Signature から小説間の類似度を DTW を用いて計量することにより,ユーザーが入力したその小説コンテンツとストーリー展開が類似した別の小説コンテンツを検索することが可能となる．

これまでの小説コンテンツの特徴量抽出に関する研究では，小説 1 作品全体を表す特徴量として抽出されることが主であった．他の作品と比較，類似度計量を行う際には，小説全体の特徴を捉えて行われることは可能になりつつある．一方，本研究では，小説 1 作品の中でもその場面ごとに特徴量が移り変わっていくということに着目し，そのストーリー展開の様相を表現する．Story Signature と呼ばれる小説コンテンツの特徴を抽出し，その類似度を計算することで，スト一

リー展開の類似性に基づく小説コンテンツの検索を実現する．これは，小説コンテンツ内の感情のポジティブ・ネガティブの変容から起因する盛り上がり方，落ち着き方の様相の類似度計量を行うことを意味する．つまり，小説コンテンツの盛り上がり方，落ち着き方の様相に着目した類似の小説コンテンツを検索することが可能となる．

また，一方で，抽出された Story Signature 自体を可視化することは，読者だけでなく，作家にとっても，自分の書いた小説の構造を，客観的に把握できるツールして応用できると考えられる．

3. 動的時間伸縮法（DTW）

本稿は，小説コンテンツをストーリー展開の類似性に基づいて類似検索することを目的としている．本研究では，小説のストーリー展開の構造を表現する特徴量である Story Signature を抽出したのちに，その Story Signature 同士の類似度を計量することで小説の類似検索を実現する．その Story Signature 同士の類似検索に動的時間伸縮法（DTW）[7,8,9,10,11,12,13]を用いる．

DTW(Dynamic Time Warping) は，音声識別などに使用されるパターンマッチングの手法で，周波数の異なる波形同士などの，長さの異なる 2 つの時系列データの距離をロバストに求めることが出来る．DTW は 2 つの時系列データの各点の距離を総当たりで求めた上で 2 つの時系列データが最短となるパスを見つける．このため，例えば入力とする小説の長さに大きな違いがあったとしても，Story Signature を波形として見た時に，それを拡大縮小して二つのデータが重なる時には，構造が似ているとして，高い類似度を示す．こうした性質が本研究に適していると考えた．

長さ$n_p \neq n_q$ の 2 つの時系列データ

$P = \left(p_1, p_2, \ldots, p_{n_p}\right)$, $Q = \left(q_1, q_2, \ldots, q_{n_q}\right)$の DTW 距離 $d(P,Q)$は以下の通りに定義される．

$$d(P,Q) = f\left(n_P, n_Q\right).$$

ここで，$f(i,j)$は次の様に再帰的に定義される．

$$f(i,j) = \left\|p_i - q_j\right\| + min(f(i,j-1), f(i-1,j), f(i-1,j-1)),$$

$$f(0,0) = 0,$$

$$f(i,0) = f(0,j) = \infty.$$

ただし，再帰の中には同じ項が多数現れるため，実際に計算する際はボトムアップに計算を行うことで，計算量を削減することが出来る．

4. Story Signature を特徴量に用いたストーリー展開に基づく小説の類似検索方式

本節では，提案方式である，Story Signature を特徴量に用いたストーリー展開に基づく小説の類似検索方式について述べる．

4.1 提案手法の概要

本節では提案手法の概要を述べる．提案システムの全体像を図 1 に示す．本研究の目的は，小説コンテンツを対象に，ユーザーの嗜好に合致するコンテンツの検索・推薦を効率的に実現するための手法として，小説のストーリー展開の類似性に基づく検索方式を実現することである．本研究では，小説のストーリー展開を，状況がポジティブであるかネガティブであるかの時系列的な変化と仮定する．そこで本方式では，文章を一文ごとに区切り，各文に対して日本語評価極性辞書[3,4]による評価極性のスコアを算出し，それを時系列順に並べたデータを抽出することでストーリー展開を表現する．この時系列データを Story Signature と定義する．さらに，ストーリー展開の類似性に基づく小説コンテンツの検索を実現するために，Story Signature で表現された小説コンテンツ同士の類似度を計量することを考える．本方式では，時系列データ同士の類似性を計量する手法として，3 節で紹介した動的時間伸縮法(DTW)を採用した．DTW を用いることで，システムに小説のデータセットを与えると，各小説コンテンツ同士の距離を計量することが出来る．ユーザーが本システムを小説の類似検索システムとして用いる際は，入力として小説コンテンツを選ぶことで，システムはその小説に類似する小説を距離の近い順にソートして提示する．

4 節の構成について述べる．4.2 節では，ストーリー展開の構造を表現する特徴量である Story Signature を定義し，小説データから Story Signature を抽出する方法について述べる．4.3 節では，Story Signature として表現された小説同士の類似度を計量する方法として，3 節で述べた動的時間伸縮法（DTW）を用いた計量方法について述べる．4.4 節では，ユーザーが本システムを小説の類似検索方式として利用する際に必要な処理について述べる．

4.2 Story Signature の抽出方式

本節では，小説のストーリー展開の構造を表現する特徴量である Story Signature を定義し，小説データから Story Signature を抽出する方法について述べる．この抽出方式は，二つのステップによって実現される．一つ目は，入力された小説の文章を，一文ごとの文章単位に分割するステップであり，二つ目は，そうして得られた文章単位に対して，日本語評価極性辞書を用

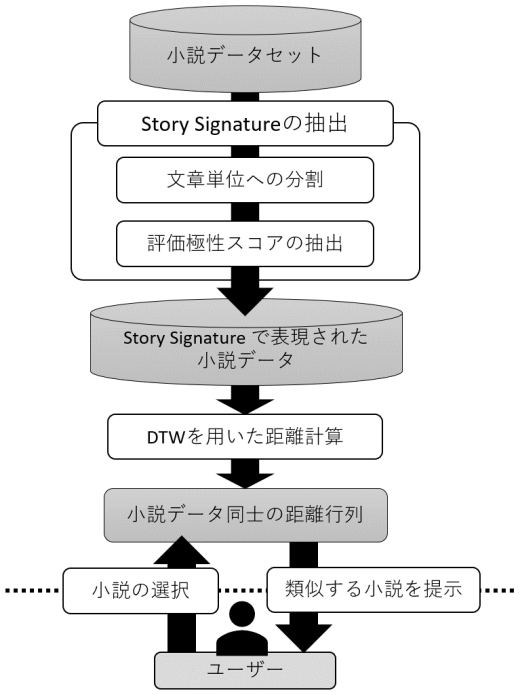

図 1　提案システムの全体像

いて評価極性スコアに変換するステップである．小説は基本的に前から順に一方向に読まれるメディアであり,それに沿ってストーリーが進行していく.そして,ストーリーの進行に従って場面や登場人物の心情などが変化する．これは一種の時系列データと捉えることが出来る．小説の特徴を抽出する際には，時系列を捨象して小説全体としての特徴を抽出するだけでなく，こうした時系列に沿った変化をストーリーの構造として捉えることが有用であると我々は考える．本研究では，時系列に沿って抽出する特性として，日本語評価極性辞書における評価極性スコアを採用した．
これは，小説内のデータのある範囲の状況がポジティブであるか，あるいはネガティブであるかを表す値である．この評価極性スコアを時系列に沿って並べたベクトルデータを Story Signature として定義する．Story Signature は，小説の文章を一文ごとに区切った上で，その文のポジティブ・ネガティブの評価極のスコアを時系列順に並べたベクトルデータとなる．

4.2.1 文章単位への分割

小説コンテンツの文章を時系列データとして捉えるために，文章全体を短い文章単位に分割する．本研究では，一文を一つの文章単位とすることとした．本研究では入力する小説の文字列として日本語の文章を対象としているため，句点をセパレータとして文章を分割する．

4.2.2 評価極性スコアの抽出

日本語評価極性辞書を用いて，各文章単位に対して評価極性スコアを算出する.日本語評価極性辞書では，まず入力文章を形態素解析し，単語に分割する．その上で，単語ごとに，それがポジティブな語であるか，ネガティブな語であるか，あるいはそのどちらでもないかを判定する．その際，未登録の語はどちらでもないと見なされる．例を示すと,「遅刻したけど楽しかったし嬉しかった。」という文章を入力とした場合,「遅刻」という単語がネガティブな語として,「楽しい」と「嬉しい」という単語がポジティブな語として抽出される．その他の語は評価極性には関係ないとして無視される．そして，ポジティブな語は+1.0，ネガティブな語は-1.0 の値に変換する．したがって，上記の例に日本語極性辞書を用いて評価極性を抽出すると，出力は [-1.0, 1,0, 1.0] という，要素数 3 のベクトルとなる．さらに，文全体の評価極性スコアは，上記のベクトルの平均を取った値となり，すなわち，0.33…となる．
ここで我々が入力としているのは，4.2.1 節で抽出した，時系列順に並んだ文書単位であるので，この文書単位ごとに評価極性スコアを算出することで，出力されるのは，文書単位ごとの評価極性スコアが時系列順に並んだベクトルとなる．このベクトルを，Story Signature として定義する．Story Signature は，文書単位の数を n とすると，各要素に-1.0～1.0 の値を取る n 次元のベクトルとなる．
実際の計算には，評価極性スコアを計量する Python 用ライブラリとして公開されている oseti [2,3]を用いた．

4.3 動的時間伸縮法(DTW)を用いた距離計算

3 節で述べた動的時間伸縮法（DTW）を用いて，Story Signature として表現された小説同士の距離を計量する．距離の短い小説同士を類似度が大きいと見なす．この距離は用意した小説データセットに存在するすべての小説コンテンツに対してあらかじめ計算しておくことが出来る．その際には，各行と各列にそれぞれ小説コンテンツが対応し，各要素にはその小説コンテンツ同士の距離を持つ，距離行列が出力される．

4.4 類似する小説の検索

4.3 で用意した距離行列を用いて，ユーザーが入力として選択した小説コンテンツに類似する小説コンテンツを提示する．距離行列から，選択された小説コンテンツに対応する行あるいは列を抜き出し，要素である距離の短い順にランキング化して，要素に対応する小説コンテンツと，その距離をユーザーに提示する．以上によって，ストーリー展開の類似性に基づく小説の類似検索を実現する．

表 1　新字体と旧字体の影響の有無

著者名	作品名	類似度
夏目漱石	京に着ける夕	14.33
夏目漱石	子規の絵	9.02
森鴎外	文づかい	38.04
森鴎外	花子	18.93
森鴎外	じいさんばあさん	16.70
森鴎外	舞姫	34.16
森鴎外	最後の一句	38.51
森鴎外	高瀬舟	29.27
森鴎外	高瀬舟縁起	8.00

5. 評価実験

本節では，本手法の評価実験について述べる．5 節の構成について述べる．5.1 節では，本手法の評価方法について述べる．5.2 節では，実験環境について述べる．5.3 節では，旧字体と新字体の影響の有無を調査し考察を行う．また，5.4 節では，提案システムを用いて森鴎外「青年」と夏目漱石「三四郎」の類似度を計量し，提案システムの有効性について論ずる． 5.5 節では小説コンテンツを Story Signature に変換した上で可視化する例を提示する．また，Story Signature の抽出が適切に行われているか検証するために，その Story Signature と算出箇所との照応を行う．

5.1 本手法の評価方法

実験に使用した小説データセットは，”青空文庫”[16]で取得した夏目漱石（1867 年-1926 年）と森鴎外（1962-1922）の作家の掲載されているデータセット，夏目漱石 105 編，森鴎外 79 編である．この作家二人は明治時代から生き，現代においても共に日本近代文学の代表的存在でもある．

また，森鴎外は夏目漱石論を著すなど，二人の関係性は深く，夏目漱石の「三四郎」に影響を強く受け森鴎外が「青年」を書くなど，共に影響を与え合った作家としても有名である．

本実験では，森鴎外が夏目漱石の「三四郎」をオマージュし「青年」書き上げたというエピソードに着目し，森鴎外「青年」と夏目漱石「三四郎」の類似度が本手法を用いて計量した際に類似度が高くなると想定する．その前提の上で，「青年」，「三四郎」と両作家全作品との本手法を用いた類似度を求め，得られた結果について考察することで DTW を用いた類似度計量が適切に抽出できるかを検証する．

また，夏目漱石と森鴎外が明治時代の作家であることから，旧字旧かな（明治時代の文語文）のバージョンと，新字旧かな（現代訳文）のバージョンが両方存在している作品が合計 9 作品存在している．提案手法で用いている日本語評価極性辞書は，登録されている単語に対してしか評価極性を判断することはできないため，現れる単語が異なると抽出される Story Signature が変化してしまうことで，文語文と現代訳文のどちらを使うかで類似度には差が出てしまう可能性がある．このバージョンが違うことによる影響の大きさについて検証することで，類似度計量が適切に行えるのかを調査する．本稿では，5.4 節にて，同一作品の文語文版と現代訳文版の類似度を本提案システムを用いて計量することにより，その距離を測り，影響を評価する．

最後に，Story Signature の抽出が適切に行われているか検証するために，太宰治の「走れメロス」を用いて Story Signature と算出箇所との照応を行う．

5.2 実験環境

4 節で提案したシステムを実装し，夏目漱石 105 編，森鴎外 79 編，計 184 編のすべての小説データについて，互いの類似度を計量した．

本手法では評価極性スコアの抽出に oseti[2,3]というライブラリを利用している．oseti は，日本語評価極性辞書を用いて単語および文のネガポジ判定を行うライブラリである．oseti は，文章を入力として，句点をセパレータとして文に分割するが，それ以外にも，空白と改行が連続して出て来たときにそこを文の区切りと認識してしまう問題が発生する．本実験では，空白文字しか含まれない文字列について分析処理を行わないという例外処理を追加した．

5.3 旧字体と新字体の影響の有無

旧字旧かな（明治時代の文語文）のバージョンと，新字旧かな（現代訳文）のバージョンが両方存在している 9 作品に対して，同一作品の文語文版と現代訳文版の類似度を本提案システムを用いて計量することにより，その距離を測り，バージョンが違うことによる影響を評価する．結果を表 1 に示す．結果を見ると，ここでの距離の平均は 23.00 であった．これは作品間の距離の平均である 526.90 に比べれば小さな値であるため，誤差の範疇であると考えて良さそうである．つまり，旧字体と新字体でバージョンが違うことの影響は小さいと考えて良さそうである．

5.4 「青年」と「三四郎」に着目した提案システムの有効性の検証

「青年」が「三四郎」をオマージュしたという経緯に着目し，その類似性が高いと想定した上で，他の作品との比較でそれが現れているかを確認する．

「青年」と「三四郎」の距離は 1052.95 であり，全

表 2 「走れメロス」スコア対応箇所

抽出部 L=句読点数	原文	スコア
L1:5	メロスは激怒した。～笛を吹き、羊と遊んで暮して来た。	-0.6
L11:14	メロスには竹馬の友があった。～久しく逢わなかったのだから、訪ねて行くのが楽しみである。	0.2
L36:40	いいえ、乱心ではございませぬ。～御命令を拒めば十字架にかけられて、殺されます。	-0.3
L85:90	「私は約束を守ります。～あれを、人質としてここに置いて行こう	0.4
L96:100	それを聞いて王は、残虐な気持で、そっと北叟笑んだ。～人は、これだから信じられぬと、わしは悲しい顔して、その身代りの男を磔刑に処してやるのだ。	0
L116:119	初夏、満天の星である。～そうして、うるさく兄に質問を浴びせた。	0.33
L126:130	妹は頬をあからめた。～結婚式は、あすだと。」	-0.33
L146:150	祝宴は、夜に入っていよいよ乱れ華やかになり、人々は、外の豪雨を全く気にしなくなった。～ちょっと一眠りして、それからすぐに出発しよう、と考えた。	0
L166:170	「仕度の無いのはお互さまさ～花婿は揉み手して、てれていた。	-0.4
L191:196	メロスは額の汗をこぶしで払い、ここまで来れば大丈夫、もはや故郷への未練は無い。～そんなに急ぐ必要も無い。	0
L206:210	流れはいよいよ、ふくれ上り、海のようになっている。～あれが沈んでしまわぬうちに、王城に行き着くことが出来なかったら、あの佳い友達が、私のために死ぬのです。」	-0.2
L224:230	「何をするのだ。～その、たった一つの命も、これから王にくれてやるのだ。」	-0.6
L241:245	愛する友は、おまえを信じたばかりに、やがて殺されなければならぬ。～約束を破る心は、みじんも無かった。	-0.3
L280:285	セリヌンティウス～私だから、出来たのだよ。	0.4

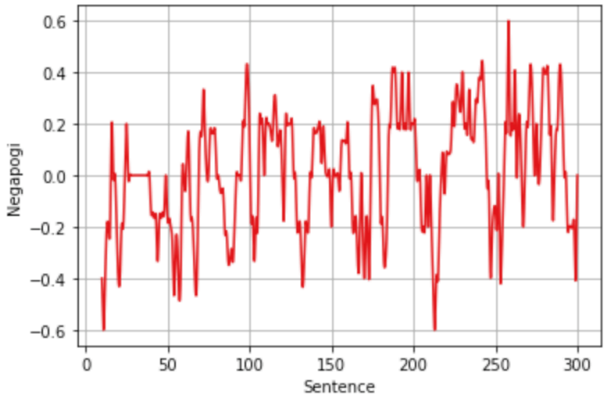

図 2 走れメロスの Story Signature

186 の作品同士の組み合わせがある中で距離が近い順で 36 位であった．また，森鴎外「青年」から最も近い作品は夏目漱石「草枕」で距離が 646.90 であったのに対し，「三四郎」は森鴎外「みちの記」が一番近い作品になり距離は 995.46 となった．

この結果は，「青年」と「三四郎」は想定したほどの類似性を示さなかったと言えるが，類似性自体はまずまず検出されていると考えることが出来る．

この結果の原因として夏目漱石と森鴎外は互いに参照しあっていた可能性があり，これはつまり三四郎と青年だけが近いという仮説が正しくなかったと考えられる．この結果を鑑みるとむしろより距離が近い小説がある可能性が存在することになる．そこで，新たに森鴎外「青年」と夏目漱石「三四郎」からそれぞれ一番距離が近いとされている作品を調査してみると，森鴎外「青年」からは夏目漱石「草枕」が最もストーリーの展開が近く，一方で，夏目漱石「三四郎」からストーリー展開が最も近い作品が，森鴎外「みちの記」であった．この結果により，一番類似度が高い作品が自身の作品ではなく、互いの作品群の小説であるため、互いが互いの作品を意識していたことが現れていると考えられる．

このように，提案システムで類似性が高いと示された作品同士を改めて分析することで，今まで知られていなかった影響関係について新たな知見を得ることが可能になると考えられる．

5.5 Story Signature 可視化

小説のストーリー展開を直感的に把握するために Story Signature を可視化し，その可視化した Story Signature と文章の照応箇所を表 2 に示す．本実験では，太宰治の「走れメロス」を可視化した例を示す．また，太宰治の「走れメロス」において，Story Signature の各文のスコア部分の対応箇所を「抽出部，原文，スコア」の三点において編纂している．

6. まとめ

本稿では，小説コンテンツのストーリー展開に着目した，小説コンテンツの類似検索方式について述べた．本方式は，小説コンテンツをそのストーリー展開に対応する時系列データとして捉え Story Signature という特徴量として抽出する．さらにそれに対して動的時間伸縮法(DTW)に基づく類似度計量をすることによって，小説コンテンツ同士のストーリー展開に基づく類似度検索を実現した．

また，本方式を検証するための実験システムを構築し，著名な小説を使って本方式の有効性を検証する実験を行った．

さらに，小説コンテンツから本方式に基づいて抽出される Story Signature の可視化方式についても示した．

小説のストーリー展開の類似性に着目した検索方式を実現したことによって，ユーザーの趣味趣向に合致した小説コンテンツの取得機会を増大させることができると考えらえる．

今後の課題として，小説コンテンツの他の要素を特徴とした類似度計量方式の実現と，それらと本方式との連携による新たな統合検索方式の実現，ストーリー展開に基づく小説コンテンツ検索におけるユーザインタフェースの実現，多種多様で膨大な小説コンテンツを対象とした評価実験が挙げられる．また，Story Signature を抽出する前にダウンサンプリングし，さらにサンプル選択バイアスを取り除くことで，DTW の想定していない高周波成分が多い波は類似性が旨く計量できないという懸念をダウンサンプリングすることが可能になると考えられる．

参　考　文　献

[1] 青空文庫, https://www.aozora.gr.jp/

[2] 小説家になろう, https://syosetu.com

[3] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一．意見抽出のための評価表現の収集．自然言語処理，Vol.12, No.3, pp.203-222, 2005. / Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi. Collecting Evaluative Expressions for Opinion Extraction, Journal of Natural Language Processing 12(3), 203-222, 2005.

[4] 東山昌彦，乾健太郎，松本裕治，述語の選択選好性に着目した名詞評価極性の獲得，言語処理学会第 14 回年次大会論文集，pp.584-587, 2008. / Masahiko Higashiyama, Kentaro Inui, Yuji Matsumoto. Learning Sentiment of Nouns from Selectional Preferences of Verbs and Adjectives, Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing, pp.584-587, 2008.

[5] 川口 晴会，鈴木伸崇，物語展開を考慮した小説データからの表紙の自動生成, DEIM 2019 ,2019

[6] 高田 叶子，佐藤 哲司,文体の類似度を考慮したオンライン小説推薦手法の提案,DEIM 2017,2017

[7] 金川絵利子，佐原諒亮，岡留剛．作家の文体の類似性：情報量木 カーネルの導入による構文間距離を用いた分析．人工知能学会全国大会論文集，Vol. 29, pp. 1–4, Sep. 2015.

[8] Sakoe, H. and Chiba, S.: Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Transaction on Acoustics, Speech, and Signal Processing, Vol. ASSP-26, No. 1, pp. 43–49 (1978)

[9] Berndt, D.J. and Clifford, J.: Finding Patterns in Time Series: A Dynamic Programming Approach, Advances in Knowledge Discovery and Data Mining, pp.229–248, AAAI/MIT(1996).

[10] Rabinar, L. and Juang, B.-H.: Fundamentals of Speech Recognition, Englewood Cliffs, N.J. (1993)

[11] Method for Content-based Music Retrieval via Acoustic Input, Proc. ACM Multimedia, pp.401–410 (Sept./Oct. 2001)

[12] Mount, D.W.: Bioinfomatics: Sequence and Genome Analysis, Cold Spring Harbor, New York (2000).

[13] Rabinar, L. and Juang, B.-H.: Fundamentals of Speech Recognition, Englewood Cliffs, N.J. (1993).

[14] 櫻井　保志，吉川正俊，ダイナミックタイムワーピングのための類似探索手法", 情報処理学会論文誌，2014