小説コンテンツを対象とした感情遷移に着目した類似小説検索方式

永田 智菜† 仲程 凜太朗† 岡田 龍太郎† 峰松 彩子† 中西 崇文†

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

E-mail: †{s2122044, s1922069}@stu.musashino-u.ac.jp ,
{ryotaro.okada,ayako.minematsu,takafumi.nakanishi}@ds.musashino-u.ac.jp

あらまし 本稿では，小説コンテンツを対象とした感情遷移に着目した類似小説検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して，多様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小説コンテンツの検索・推薦を効率的に実現することが重要となってきている．一般的に，小説は描かれるシーンごとに感情が移り変わるため，小説内の感情遷移を抽出することができれば，小説のストーリー展開に基づいた検索・推薦機能が実現できると考えられる．本方式では，文章ごとにテキスト感情分析に行うことにより，小説コンテンツの時系列感情遷移特徴を抽出し，その特徴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．これにより感情遷移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 類似小説，感情遷移，動的時間伸縮法，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ネット上に散在している．例えば，青空文庫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”[1]で投稿されたオンライン小説が人気を博している．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”[1]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，ユーザ(投稿者および閲覧者)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は可能である．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られる．

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．小説は単なる意味的な内容だけでなく小説コンテンツが描く感情のポジティブ・ネガティブなど，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．ここで，ストーリー展開とは，小説の文頭から文末までの文脈，感情の変化を順に取り出したものとする．

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．我々[2]はこれまで小説コンテンツ中のストーリー展開を時系列的な極性の変化(ポジティブ・ネガティブ)として抽出することで，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している．

本稿では，小説コンテンツを対象とした感情遷移に着目した類似小説検索方式について示す．本稿では，これまでの我々の研究[2]を拡張し，ストーリーの展開を表現する 10 個の感情パラメータごとに感情の盛り上がりを文頭から順に数値として抽出し可視化する．それぞれの感情値がどの程度の盛り上がりを示すかについては，各文章に現れる語彙をもとに ML-ASK[3]を用いた情報を利用する．これによって小説全体の感情遷移を時系列情報として可視化できる．また，抽出された感情遷移の時系列情報は波としてとらえることができる．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になる．これにより，ストーリー展開に着目した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の指標として用いることができる．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に着目

した，時系列な感情変化に基づく小説コンテンツ類似探索方式について示す．本方式は，小説コンテンツから小説コンテンツが描く感情とその強さを表す特徴量として，ML-ASK[3]によって示された感情とそのスコアを各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な感情変化と捉え，時系列メタデータとみなすことができる．各小説コンテンツから抽出された感情遷移グラフについて，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探索することを可能にする．

本研究の目的は，小説の構造の可視化と，その作品と他の作品の類似度を新たに求める指標を実現することである．

以下，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 ML-ASK を用いた感情遷移グラフ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なされてきた．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 小説のポジティブ・ネガティブを特徴量に用いた類似度検索方式に関する研究

我々の研究[2]は，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い，単位文章ごとに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したものを Story Signature として定義し，この Story Signature から小説間の類似度 DTW を用いて計量することにより類似小説を検索するシステムを提案している．

本研究は，小説の時系列のなかで 10 個の感情パラメーターごとにポジティブ・ネガティブを抽出し，類似性を求めるものである．本稿で述べる提案方式では，感情遷移を用いる点で異なる．

2.2 小説の類似度算出に関する研究

小説の類似度判定に関して，丸山ら[4]は，個人が考える小説の特徴が重要であるとし，仮想の個人を想定し．青空文庫に投稿された小説から単語ベクトルを抽出し，Linear Discriminant Analysis を用いて小説間の分離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単語を抽出し，これらの特徴量の分離度を求めている．

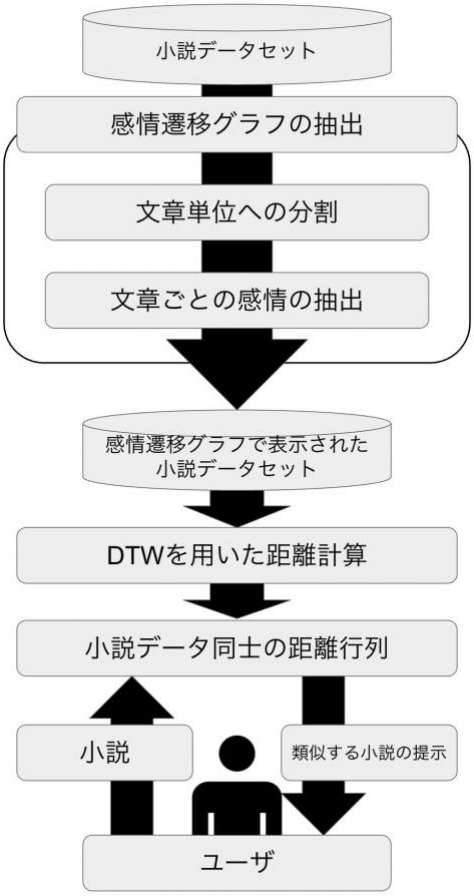

図 1　提案システムの全体像

本研究では，小説内のストーリー展開を表す時系列の特徴量を考慮している点が異なる．

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．

高田ら[5]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，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各感情を数値に変換する．そして，抽出した値から小説ごとの各感情繊維グラフ

を作製する．そして，小説間の類似度 DTW を用いて計量することにより，ユーザ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，感情遷移グラフを抽出し，その類似度を計算することで，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．これは，小説コンテンツ内の複数の感情遷移から各小節の類似度計量を行うことを意味する．つまり，小説コンテンツの細かい感情遷移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．

3. 動的時間伸縮法（DTW）

本稿は，小説コンテンツを感情遷移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，小説の感情遷移を感情分析ツール（ML-ASK）を用いて各感情を数値化し，グラフにする．そのグラフ同士の類似度を計量することで小説の感情遷移の類似検索を実現する．そのグラフ同士の類似検索に動的時間伸縮法（DTW）[6,7,8,9,10,11]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できる.DTW は 2 つの時系列データの各店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，例えば入力する小説の長さに大きな違いがあったとしても感情のグラフを波形として見た時にそれが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．

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

$f(0,0) = 0$,

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

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. ML-ASK 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，ML-ASK を用いた感情遷移グラフ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 節の構成について述べる．4.1 節では，提案手法の概要について述べる．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感情遷移グラフを定義し，小説データから感情遷移グラフを抽出する方法について述べる．4.3 節では，感情遷移グラフ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．

4.1 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，小説の感情遷移の類似性に基づく検索方式の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各感情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，各文に対して感情分析ツールである ML-ASK 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感情遷移グラフを表現する．さらに，時系列感情遷移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，感情遷移グラフを用いて，小説コンテンツ同士の各感情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度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できる．ユーザ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．

4.2 感情遷移グラフの抽出方式

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感情遷移グラフを小説データから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，入力された文章を単位文章ごとに分割するステップである．単位文章の単位として TopicTiling[12]などのテキストセグメンテ

ーション手法を用いて単位文章を設定することも考えられるが，本稿では，単純な単位文章を 1 文として設定する．単位文章をテキストセグメンテーション手法を用いて分割する手法の導入については今後の課題とする．

二つ目は，そうして得られた文章単位に対して，ML-ASK を用いて各感情を抽出し，感情遷移をグラフに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感情が変化する．これは一種の時系列データと捉えることができる．小説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，ML-ASK による感情分類と重みづけを採用した．これは，小説内のデータのある範囲がどの感情であるかと，そしてその感情の強さを表す値である．この ML-ASK を時系列に沿って並べたデータが感情遷移グラフである．

本研究では，時系列に沿って抽出する特性として，感情表現辞典[13]を用いた ML-ASK によ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どのような感情であるか，またどの程度の感情の強さであるかを抽出してくれるものである．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感情遷移グラフは，小説の文章を一文ごとに区切った上で，その文の感情とその強さの評価極性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

ML-ASK 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．ML-ASK 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,「喜・怒・昂・哀・好・怖・安・厭・驚・恥」の 10 種類の感情とその強さを抽出するものである．

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．感情遷移グラフは，文書単位の数を n とすると，各要素に 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 ML-ASK [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，感情遷移グラフ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．

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，提案システムを用いて感情遷移グラフの抽出の検証を行う．5.3 節では，提案システムを用いて，小説の類似度を計量する．また，5.4 節では，旧字旧かなと新字新かなの違いによる影響の有無を調査し，考察を行う．

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”[14]で取得した夏目漱石（1867 年-1926 年）の作家の掲載されているデータセット，110 編である．

本実験では，夏目漱石の 110 編をそれぞれ比較し，類似度を求め，得られた結果について考察する．5.3 節で提案したシステムを実装し，夏目漱石 110 編について互いの類似度を計量した．

また，夏目漱石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）のバージョンと，新字新かな（現代訳文）のバージョンが両方存在している作品が合計 4 作品存在している．4 作品を検証するだけでは不十分と考え，同時代を生きた芥川龍之介の 5 作品を加えて検証を行う．提案手法で用いている ML-ASK 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感情遷移グラフが変化してしまう．そのため，旧字旧かな版と新字新かな版のどちらを使うかで

図 2 吾輩は猫であるの感情の感情遷移グラフ

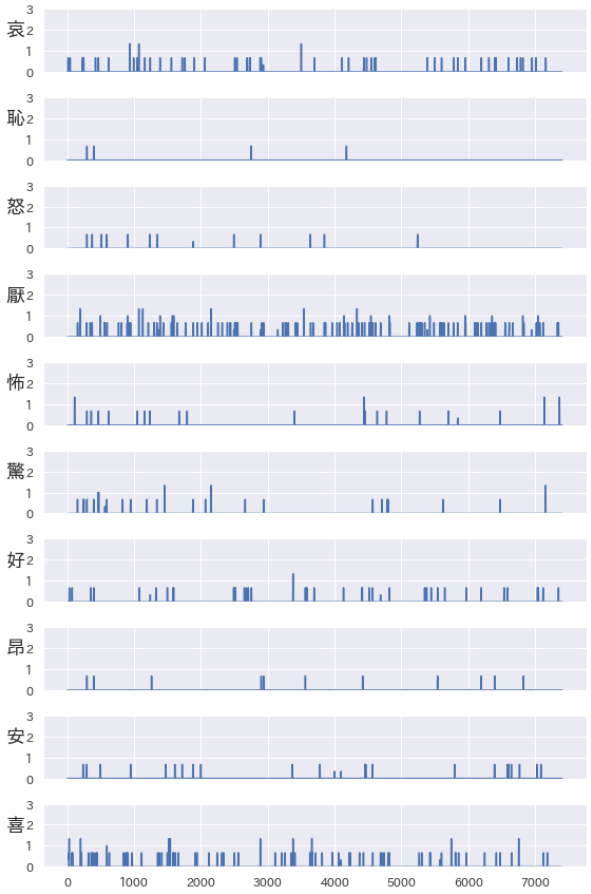

図 3　三四郎の感情遷移グラフ

類似度には差が出てしまう．この，バージョンが違うことによる影響について調査する．5.4 節にて，同一作品の旧字旧かな版と新字新かな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．この実験では，芥川龍之介（1892 年-1927 年）の 5 編と夏目漱石 4 編について，類似度を計量し比較を行った．

5.2 実験 1　感情遷移グラフの抽出の検証

5.2.1　実験目的

本節では，4 節で提案したシステムを実装し，小説のストーリー展開を直感的に把握するために感情遷移グラフを可視化する．本手法では評価極性スコアの抽出に，ML-ASK というライブラリを利用している．ML-ASK は，感情表現辞典を用いて単語および文章の感情判定を行うライブラリである．これにより抽出した 10 種の感情を可視化し，正しく抽出できているかを検証する．

5.2.2　実験結果

図 2 に「吾輩は猫である」から抽出された感情遷移グラフを示す．図 3 に「三四郎」から抽出された感情遷移グラフを示す．図に示されるように，10 種の各感情について，物語文章の出現順に沿ってその感情がどれだけ表れているのかを抽出し可視化することが出来た。

5.2.3　考察

図 2 の「吾輩は猫である」の感情遷移を可視化したグラフを見ると，「厭」「喜」の感情が多く出ていることが見て取れる．「恥」はあまり出ていないことも分かる．10190 行目付近では驚きの感情が強く出ている．この部分に当たる文章を表 1 に示す．10188 文目の「はっと我に帰った」の部分を示していると考えられる．また，同タイミングで現れる「安」の感情に関しては，10189 文目の「安心した」や，「胸を撫でおろす」などが反映されていると考えられる．これらから文章に表れる感情を抽出できていることが分かる．

図 3 の「三四郎」の感情遷移を可視化したグラフを見ると，「厭」「喜」の感情が多く出ていることが見て取れる．「恥」「怒」「昂」はあまり出ていないことも分かる．5000 行目付近ではすべての感情において 0 となっていた．この部分に当たる文章を表 2 に示す．文章

には感情を表す言葉は出現していないため，正しく抽出できていると言える。しかし，表面的な感情表現はないものの行動から読み取ることのできる感情は表現されているとも考えられ，そうした表現には対応していないことがわかる．

図 2 と図 3 を見比べると感情遷移グラフの感情の値が全体的に吾輩は猫であるが大きく，感情遷移グラフも多くの感情において密であるということが読み取れる．

以上から物語文章から，提案方式によって感情の遷移を抽出できていると考えられる。

5.3 実験 2 小説の類似度検索の検証

5.3.1 実験目的

提案システムを用いて感情に基づく作品の類似検索を行う際に，システムが正しく類似する作品を提示できることを検証するために，内容が類似していると仮定した作品を対象として，それらの作品同士の類似度が，その他の作品との類似度より高くなるかどうかを調査する．

夏目漱石の作品には「三四郎」「それから」「門」で構成される前期三部作と,「彼岸過迄」「行人」「こころ」で構成される後期三部作が存在する．この三部作に着目し，三部作同士の類似度が別の三部作の作品より高くなるのではないかと仮定し，前期と後期の三部作で類似度の比較を行う．

5.3.2 実験結果

三部作それぞれの作品から，他の 5 作品との距離を計量した．前期三部作における他の作品とその距離をランキングしまとめたものを表 3 に、後期三部作についてまとめたものを表 4 に示す．表の中で，その作品と同一の三部作に含まれる作品には網掛けを付けて示している．

5.3.3 考察

表 3 から，前期三部作と他の作品との距離を計量した結果を見ると,「こころ」が上位に来てはいるものの，その次に上位に来ているのは前期三部作に含まれる作品であり，仮定した通り前期三部作の作品同士は似ているという結果が示されている．表 4 から，後期三部作についての結果を見ると，前期三部作とは違って，後期三部作に含まれる作品が上位に表れたとは言えない結果となった．

前期三部作について想定通りの結果が出ていることから，本システムは作品の類似度を計量するという目的においてある程度の有効性を持っていると考えられるが，後期三部作については想定した結果とはならなかった．この原因としては，提案システムに改善の余地があるとも考えられるが，夏目漱石の後期三部作がそもそも前期三部作の作品同士より似ていないとも考えられる．前期三部作は一連のストーリーになっているが，後期三部作は共通のテーマはあるものの独立した作品となっており，三部作とは言っても性質の違いがあると考えられる．そのため，検証実験における仮定の置き方についてさらなる検討が必要であると考えている．

表 1　吾輩は猫であるの 10190 行付近の文章

行番号	本文
10187	「いよいよ出たね
10188	「その声が遠く反響を起して満山の秋の梢を、野分と共に渡ったと思ったら、はっと我に帰った……
10189	「やっと安心した
10190	と迷亭君が胸を撫でおろす真似をする
10191	「大死一番乾坤新なり
10192	と独仙君は目くばせをする
10193	寒月君にはちっとも通じない
10194	「それから、我に帰ってあたりを見廻わすと、庚申山一面はしんとして、雨垂れほどの音もしない

表 2　三四郎の 5000 行付近の文章

行番号	本文
4995	人に目立たぬくらいに、自分の口を三四郎の耳へ近寄せた
4996	そうして何かささやいた
4997	三四郎には何を言ったのか、少しもわからない
4998	聞き直そうとするうちに、美禰子は二人の方へ引き返していった
4999	もう挨拶をしている
5000	野々宮は三四郎に向かって、「妙な連と来ましたね
5001	と言った
5002	三四郎が何か答えようとするうちに、美禰子が、「似合うでしょう
5003	と言った
5004	野々宮さんはなんとも言わなかった
5005	くるりとうしろを向いた

5.4 旧字かなと新字新かなの違いの影響の検証

旧字旧かなのバージョンと，新字新かなのバージョンが両方存在している 9 作品に対して，同一作品の旧字旧かな版と新字新かな版の類似度を提案システムを

表 3　前期三部作類似順

三四郎		門		それから	
こころ	753.0	こころ	808.7	門	850.0
門	844.3	三四郎	844.3	こころ	868.3
それから	874.0	それから	850.0	三四郎	874.0
行人	1207.7	彼岸過迄	948.0	彼岸過迄	1126.3
彼岸過迄	1256.0	行人	1257.0	行人	1219.3

表 4 後期三部作類似順

こころ		彼岸過迄		行人	
三四郎	753.0	門	948.0	こころ	1163.3
門	808.7	それから	1126.3	三四郎	1207.7
それから	868.3	こころ	1133.3	それから	1219.3
彼岸過迄	1133.3	三四郎	1256.0	門	1257.0
行人	1163.3	行人	1431.3	彼岸過迄	1431.3

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．

実験結果を表 5 に示す．結果を見ると，ここでの距離の平均は 140.4 であり，これは表 2 に示されるような作品間の距離に比べれば小さな値である．しかし，表を見てもわかるように，小さいものは誤差といえるほどの差しかないが，差の大きいものだと 500 程度の差がある．これは違う作品との距離と比較すると，距離の近い作品の距離と同程度かやや低い程度の値である．そのため，全体的には影響は軽微であるものの，作品によっては無視できないほど大きな影響があることが分かった．この一因としては，ML-ASK が旧字旧かなの表現方法にまだ対応できていないことが影響していることが考えられる．なお，距離が大きかった二つの作品について感情ごとに距離を見ると,「厭」の感情の距離が特に大きな数値を示していることが分かった．

5.5 実験全体の考察

本節では小説の感情遷移に着目した類似検索方式の検証のために実験を行った．5.2 節では感情遷移グラフの抽出を行い，図 2，図 3 のように各感情の粗密を視覚的に表すことができ，三四郎では 5000 文付近で感情表現がなくなっていることが分かった．また，「吾輩は猫である」と「三四郎」では感情表現の量の違いが見て取れる．5.3 節では夏目漱石の三部作についての類似度計量を行ったところ，前期三部作同士は類似していると示されたが，後期三部作は類似しているとはならなかった．5.4 節では旧字旧かなと新字旧かなのバージョンの違いについての影響の検証を行ったところ，無視できない影響がある作品が存在することが分かった．

表 5 旧字旧かな版と新字新かな版の類似度

著者名	作品名	距離
芥川龍之介	アグニの神	19.3
芥川龍之介	地獄変	76
芥川龍之介	河童	109
芥川龍之介	羅生門	4
芥川龍之介	トロッコ	14
夏目漱石	子規の絵	7
夏目漱石	それから	523.3
夏目漱石	門	503.3
夏目漱石	京に着ける夕	8

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を提案した．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感情遷移グラフという特徴量として抽出することを可能にする．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感情遷移グラフ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，小説コンテンツに含まれる感情以外の場所，時間，事象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，旧字旧かなの文章への対応，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．

参　考　文　献

[1] 小説家になろう, https://syosetu.com

[2] 仲程凛太朗，岡田龍太郎，中西　崇文，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，DEIM 2020, 2020.

[3] M. Ptaszynski, P. Dybala, W. Shi, R. Rzepka, K. Araki, “A system for affect analysis of utterances in Japanese supported with web mining”,　日本知能

情報ファジイ学会誌, 21 (2), pp194 - 213, 2009.

[4] 丸山 正人, 竹川 高志, 個人の特性を反映した文章の類似度判定による小説推薦, DEIM 2020, 2020.

[5] 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.

[6] H. Sakoe, S. Chiba, “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.

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

[8] L. Rabinar, B.-H. Juang, “Fundamentals of Speech Recognition, Englewood Cliffs”, Prentice-Hall, Inc., 1993.

[9] J.-S. R. Jang, L. Hong-Ru “Hierarchical filtering method for content-based music retrieval via acoustic input.” Proceedings of the ninth ACM international conference on Multimedia, pp.401–410, 2001.

[10] D.W. Mount, “Bioinfomatics: Sequence and Genome Analysis”, Cold Spring Harbor, New York ,2000.

[11] 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

[12] M. Riedl, C. Biemann, TopicTiling: a text segmentation algorithm based on LDA. In Proceedings of ACL 2012 Student Research Workshop (ACL '12). Association for Computational Linguistics, USA, pp.37–42, 2012.

[13] 中村明：感情表現辞典, 東京堂出版, 1993.

[14] 青空文庫, https://www.aozora.gr.jp/