一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
HCS2016-77 (2017-01)

物語の類型に着目した絵本の類似探索手法に関する一検討

安尾　萌†,††　服部　正嗣††　藤田　早苗††　松下　光範†

† 関西大学大学院 総合情報学研究科　〒569–1095 大阪府高槻市霊仙寺町 2–1–1
†† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所　〒619–0237 京都府相楽郡精華町光台 2–4
E-mail: †{k290993,t080164}@kansai-u.ac.jp, ††{hattori.takashi,fujita.sanae}@lab.ntt.co.jp

あらまし　本研究の目的は，絵本を対象に，そのストーリー展開の傾向に基づいた検索を可能にすることである．既存の書籍検索サービスの多くは，書誌情報やレビューなど絵本に付随する情報に基づいて検索を行うため，絵本の内容自体に基づいた検索が難しく，「この本と同じような読後感を得られる絵本を読みたい」といったユーザの要求には応えられない．この問題を解消するため，本稿では絵本の中に出現する単語の評価極性に着目し，その推移を検索に用いることで類似した絵本を探索可能にする手法を提案する．提案手法では，本文に出現する単語から positive/negative な単語の出現量を取得し，それらを基にバタチャリア係数を算出することで絵本同士の類似度を測る．

キーワード　バタチャリア係数，単語の評価極性，絵本検索，類似検索

Similarity Search for Picture Books by Focusing on Their Story-types

Megumi YASUO†,††, Takashi HATTORI††, Sanae FUJITA††, and Mitsunori MATSUSHITA†

† Graduate School of Informatics, Kansai University　2–1–1 Rozenjicho, Takatsuki, Osaka 569–1095 Japan
†† NTT CS labs., NTT Corp.　2–4 Hikaridai, Seikacho, Sorakugun, Kyoto 619–0237 Japan
E-mail: †{k290993,t080164}@kansai-u.ac.jp, ††{hattori.takashi,fujita.sanae}@lab.ntt.co.jp

Abstract　The purpose of this research is to realize a system for searching picture books based on the unfolding trend of stories. Current book search systems seek books by referring “the outline” sources incidental to the books such as bibliographic information or reviews, thus content-based search is not performed. As the consequence, the existing book search systems do not answer the user’s request like “Please show me a picture book that gives the similar post-reading feeling as this picture book.” To handle such the user’s request, we focus on the emotional polarity of words: our proposed method enables similarity search of picture books by using the transition of emotional polarity of words that appear in the text of a picture book. In this method, positive and negative words appeared in the text were extracted and the similarity between picture books were calcurated by using Bhattacharyya coefficients.

Key words　Bhattacharyya Coefficient，sentiment polarity of words，picture book search, similarity search

1. はじめに

近年の出版点数の増加は著しく，2013 年には年間 80,000 点を超えるまでになっている [1]．そのため，自らの興味や嗜好に沿った書籍を見つけることが以前に比べて難しくなってきている．こうした現状に対処するため， 様々な書籍検索サービスが WEB 上で提供されている (e.g., 書籍横断検索(注1)，絵本ナビ(注2))．多くの書籍検索サービスは，出版社や著者などの書誌情報や，出版社や書店によって付与される書籍のキーワードやカテゴリに関する情報を用いることで検索を可能にしている．しかし，これらのサービスでは書籍のテキスト自体は利用しておらず，キーワードやカテゴリは必ずしもテキストすべてを表現できるわけではない．そのため，「ある本と似た内容の本が読みたい」「ハッピーエンドの本が読みたい」といった，書籍の内容やストーリー展開に直接関わる情報をクエリとして検索することが難しい．このような検索要求に応えるために，書籍自体ではなく書籍のレビューに着目して検索を可能にする試み [2] が行われている．しかし，新しく出版された書籍の場合は，そのレビューが投稿されるまでに時間を要するうえ，書籍の人気に応じてレビューの数に偏りが生じる傾向がある．そのため，上述の検索手法を用いると，新しく出版される書籍や，注目度の低い書籍は検索結果に現れず読まれにくくなるため，さらにレビュー数が偏ってしまう．このように，レビューを利用した

(注1)：http://book.tsuhankensaku.com/hon/
(注2)：http://www.ehonnavi.net/

検索手法はレビュー数の偏りを助長し，一部の書籍が検索されづらくなるという課題がある．

書籍の内容に関する検索をレビューを用いずに可能にするため，本研究では書籍の内容自体を対象とした検索手法について検討する．その端緒として，本稿では内容に基づく検索のニーズが高い「絵本」を対象とし，本文中に現れる評価極性を有する単語の出現頻度の推移を用いてストーリー展開の類似した本を検索する手法を提案する．

2. 先行研究

子どもの興味と発達段階に適した絵本を検索する支援を企図して，絵本検索システム「ぴたりえ」が提案されている [3]．「ぴたりえ」では書誌情報や絵本の本文（以下，絵本テキストと記す）から抽出した単語の出現頻度を特徴量とした類似探索を行っており， 2,400 冊を超える絵本データベースの中から，ユーザが入力したテキストと類似する絵本を探すことができる．例えば，多様な動物が登場する絵本テキストを入力すると，その絵本と同様に多様な動物が登場する別の絵本が結果として出力される．しかし，「ぴたりえ」は絵本テキストに含まれる単語の出現頻度やカテゴリに関する情報を検索に用いているため，物語のストーリー展開に基づく検索（e.g.,「ハッピーエンドの絵本を探す」）といった抽象度の高い検索を行うことは難しい．

佐々木は，絵本の主題を 6 つの大主題と 280 の主題からなる 2 階層のツリー構造に整理し，2100 冊以上の絵本について主題情報を付与した絵本データベース [4] を作成し，それに基づいた検索を提供している．大主題は絵本 1 冊に対して 1 種類のみ付与されるが，主題は複数付与されている．「ももたろう」（文：松居直，画：赤羽末吉，福音館書店，1965 年）を例に挙げると，大主題には「性格」が付与され，主題には，「ものごとをやり遂げる」「旅行する」「冒険遊び」「動物と遊ぶ」など，14 の項目が付与されている．佐々木の手法では，主題を利用することで，ストーリー展開に基づく検索を行うことができる．例えば，「ものごとをやり遂げる」が主題として付与されている絵本を探すことにより，ハッピーエンドで終わる絵本を検索できる．しかし，主題には順序に関する情報が付与されていないため，「ものごとをやり遂げる」「失敗する」など相反する主題が付与されていた場合，その物語がハッピーエンドで終わるのか，悲しい結末なのかが判別できない．また，新規に絵本が出版されるたびに，280 種類の主題の中から適切な主題を人手で付与することは大きな労力を要する．

松村らは，子どもの発する質問を利用したソーシャル絵本推薦システム「ぴくぶく」を提案している [5]．親が，子どもの発した質問とそのカテゴリを「ぴくぶく」に登録すると，他の「ぴくぶく」ユーザがそれを元に絵本を推薦する．登録された質問とそのカテゴリ，及びそれらに対して推薦された絵本はシステムに蓄積される．「ぴくぶく」は，同様の質問とカテゴリの組が入力されると，過去に推薦された絵本を元に自動推薦する機能を有している．この方式の場合，「楽しい（悲しい）ストーリー展開の絵本を知りたい」という質問を入力とすることで，意図したストーリー展開の絵本を見つけることが期待できる．また，事前に大規模な絵本データベースを準備しなくてもユーザの知識に基づいた絵本推薦が行えるというメリットがある．しかし，このシステムには，絵本知識のあるユーザが一定数参加しなければ機能しないという課題がある．

以上のように，既存の絵本検索・推薦システムには，ストーリー展開に関連した絵本の検索を実現するうえで種々の課題がある．そこで本研究では，絵本テキストから「ストーリー展開を表現する特徴量」を抽出・計数し，人手によらずにデータベースを構築することで，この課題の解消を図る．

3. 提案手法

3.1 基本方針

本研究では，「テキスト中のポジティブな単語（e.g., 楽しい，パーティ）やネガティブな単語（e.g., 悲しい，ケンカ）の出現傾向が類似している絵本同士はストーリー展開が類似している」という仮説を立て，絵本テキスト中の単語の「評価極性」に注目し絵本間の類似度判別を行った．評価極性とはテキストの内容分析などで用いられるアプローチのひとつであり，「単語が持つ，テキスト内に存在する事態が望ましいどうかの属性値」のことを指す [6]．このアプローチを応用し，絵本テキスト中の評価極性の付与された単語をページごとに計数して，その推移を「絵本のストーリーは初期状態から望ましい状態へ推移するものか否か」を判定する特徴として用いる．そして，推移パタンの類似した絵本同士はストーリー展開が類似していると考える．なお，本稿では評価極性を「ポジティブ (positive)」，「ネガティブ (negative)」，「どちらでもない (even)」 の 3 つに分類した．

また，本研究における絵本とは，絵と文字が相補的に用いられるコンテンツであり，見開きをひとつの単位としてデザインされるものを想定している．そのため，文字情報から評価極性を持つ単語を計数する際には，こうした背景も踏まえて計数の粒度を決める必要がある．今回は，文章単位ではなくページ単位で評価極性を計数し，その推移によって物語の類似度を測ることとした．

3.2 類似度の算出方法

テキスト同士の類似度の算出については様々な手法が考えられるが [7]，本稿ではページあたりの評価極性を持つ単語数の推移パタンの類似性に着目しているため，バタチャリア係数 (Bhattacharyya Coefficient) [8] を用いることとした．この手法では，総頻度が正規化され，同じ数のビンに分割されたふたつのヒストグラムの類似度を，対応するビン中の頻度の積を求めることでそれらの類似度を算出する．個数 n のビンに分割されたヒストグラム P および Q の類似度 $s(P,Q)$ は，

$$s(P,Q) = \sum_{i=1}^{n} \sqrt{P_i Q_i} \tag{1}$$

となる．ここで，P_i，Q_i は各々，ヒストグラム P，Q の i 番目のビンの頻度である．本研究では，ページ単位で positive および negative 極性を持つ単語の出現頻度を計り，それらにバタチャリア係数を適用して類似度を算出する．ページ数の異な

る絵本同士を比較する場合，物語の起承転結を考慮して類似性を比較するには，ページ数を正規化して比較する必要がある．そこで，比較対象の 2 冊の絵本各々について，指定の分割数でページを按分し，それに基いて計数した単語の評価極性の頻度を用いてバタチャリア係数を求めることとした．本稿で実装したプロトタイプは，分割数を 10 として式 (1) で得られた値を類似度とし，クエリとして入力された絵本に対して，この類似度の高い絵本から降順に出力することとした．

以上の指針に基づいて，本研究では，本文の内容から評価極性を持つ単語をページ単位で抽出・計数し，その推移パタンを指定の分割数単位のヒストグラムで表現し，評価極性の推移が類似している絵本を検索する．

3.3 実　　装

本稿では，評価極性判別のため，東北大学乾・岡崎研究室が公開している「日本語評価極性辞書」(注3)の用言編および名詞編を用いた．この辞書では，単語が持つ評価極性を positive と negative の 2 種類に分類している．今回は，この辞書に記載されていない単語の評価極性は，even と仮定して処理を行った．

本研究では，2418 冊の絵本について，登場する positive, negative, even の 3 種の極性を持つ単語をページごとに計数した．日本語評価極性辞書は日本語テキスト全般で使われている単語が登録されており，中には「軋轢」や「辟易」といった，絵本には登場する可能性が低い単語も多数存在している．そこで，絵本を対象とした処理の効率化を図るため，まず 2418 冊の絵本に出現する単語を列挙し，絵本に出現する単語のみで構成される日本語評価極性辞書のサブセットを作成した．ひらがな・カタカナなどの表記揺れを含めると，全絵本中には 14 万語近い単語が出現した．それら絵本に出現する単語のうちの 2958 単語について，日本語評価極性辞書に positive あるいは negative の評価極性が付与されていた．そこで実際の計数時には，これら 2958 単語からなる絵本用の評価極性データベースをサブセットとして用いた．つぎに，絵本テキストの形態素解析を行い，絵本ごとに記載されている単語を抽出した．絵本に出現する単語の抽出には，形態素解析器 Mecab [9]Ver2.1.2 を使用し，単語の原形および品詞情報を取り出した．抽出された単語それぞれについて，絵本用の評価極性データベースを参照し，positive あるいは negative の評価極性を持つ単語であった場合はその評価極性を持つ単語が出現したとして計数した．データベースに登録のない単語については even の評価極性を持つ単語として計数した．本稿では，ここで付与した評価極性を持つ語の語数を 3 種の「スコア」と呼び，positive 単語のスコアを p 値，negative 単語のスコアを n 値，even 単語のスコアを e 値と記す．日本語評価極性辞書に登録されている情報は自立語に限定されているため，極性評価を付与する単語も自立語を対象とした．最後に，抽出したスコアに関するデータを書籍ごとにページ単位で格納した．格納したデータは「単語の原形」「品詞」「読み」「評価極性」で構成されている．この際，「物語集」のような，一つの書籍に複数の物語が載っている書籍については，一つの話ごとにファイルを分けて格納した．

単語の評価極性を取得するにあたり，「悪くない」のような，評価極性に否定表現が付与された単語は，評価極性を変更する必要がある．そのため，文章中に出現する否定文節については，評価極性を持つ単語の 3 単語先の単語を確認し，打ち消しの助動詞「ない」および形容詞「悪い」が出現した際に感情の評価を反転させる処理を行った．

次に，抽出したデータを用いて物語ごとの類似度の算出を行った．クエリとなる絵本テキストを与えると，そのテキストに含まれる評価極性のスコア推移を任意の分割数で分割し，データベースに格納されている他の物語の極性のスコア推移と比較して算出する．ページ数が奇数になる絵本テキストについては，分割したビンのうち，末尾のページを含むビンに畳み込む処理を行っている．

4. 評価極性の推移に基づく類似度算出法の検証

実装に使用した「日本語評価極性辞書」には，「狼」「キツネ」といった，一部の生物を表す単語にネガティブの極性が付与されている．これは，日本語文書においてこれらの単語が，貪欲あるいは狡猾な様子を「狼のよう」あるいは「キツネのようだ」というように比喩的に示す際に用いられているためであると考えられる．しかし，これらの単語は絵本において，主要登場キャラクタの名称として出現する場合がしばしばみられ，必ずしもネガティブな印象を伴わなかった．評価極性を持つ単語の計数の際，キャラクタが登場するたびにその単語の出現があったとして計数すると，全ページにわたってネガティブの極性を持つ語が多数計数されることになる．これでは読者の持つ印象とかい離してしまい正確なストーリー展開の抽出ができなくなってしまう．この課題に対処するため，絵本全編にわたって登場する単語の影響を低減する重みを考慮した単語出現頻度 (Term Frequency-Inversed Page Frequency; TF-IPF) の計数方法を導入する．絵本における単語 t の TF-IPF 値の算出方法を式 (2) に示す．

$$\text{TF-IPF}(t) = tf(t,p) \times \log \frac{N}{pf(t)} \tag{2}$$

この TF-IPF 値は，文書の特徴付けを行う際に用いられる TF-IDF (Term Frequency-Inversed Document Frequency) 値 [10] を応用したものである．TF-IPF 値を，ある書籍の全ページから，そのページを特徴づける単語の極性の重み付けに利用することで，要所に現れストーリー展開に重要な役割を果たしている単語と，全編に現れストーリー展開への影響が小さい単語の比重を変えることを企図している．

次に，even の評価極性を持つ (positive, negative いずれの評価極性も持たない) 単語の出現頻度 e 値の計数について述べる．評価極性を持つ単語の計数の際，positive と negative の評価極性だけを考慮すると，positive と negative いずれの評価極性も持たない単語の多寡は無視される．その結果，全単語に占める positive と negative の評価極性を持つ単語の割合はストーリー展開の類似度に反映されなくなり，読者の感覚と類似度の定義にずれが生じる懸念がある．そこで，positive, negative いず

(注3)：http://www.cl.ecei.tohoku.ac.jp/（2016/12/27 確認）

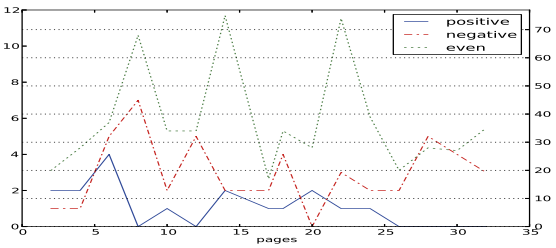

図 1 「かわいそうなぞう」の TF

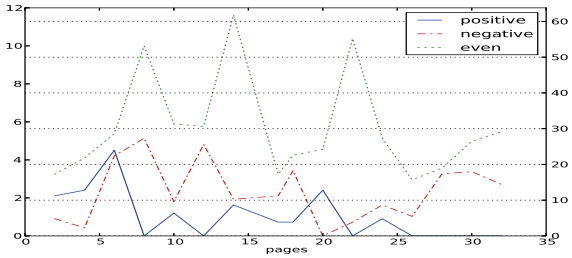

図 2 「かわいそうなぞう」の TF-IPF

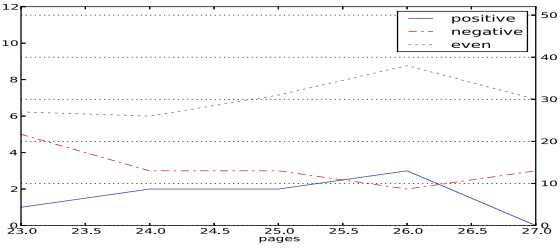

図 3 「おおかみと七ひきのこやぎ」の TF

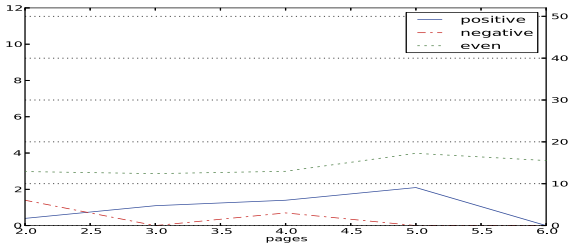

図 4 「おおかみと七ひきのこやぎ」の TF-IPF

表 1 「かわいそうなぞう」にストーリー展開が類似する絵本の検索結果

ランク	TF		TF-IPF	
	e 値なし	e 値あり	e 値なし	e 値あり
1	かたきを うった きつね	かたきを うった きつね	かたきを うった きつね	かたきを うった きつね
2	ふるやのもる	ふるやのもる	ごんぎつね	ごきげんななめのてんとうむし
3	あしたもともだち	きつねと つる	ごきげんななめのてんとうむし	ごんぎつね
4	しゃっくり百万べん	おおかみと七ひきのこやぎ	さると かに	さると かに
5	きつねと つる	こうもり	なぞなぞライオン	ざしきにいる一つ目こぞう
6	こうもり	あしたもともだち	魔法使いのチョモチョモ	ふるやのもる
7	おおかみと七ひきのこやぎ	ふるでらの しゃみせんひき	ふるやのもる	なぞなぞライオン
8	ごきげんななめのてんとうむし	しゃっくり百万べん	おれはレオ	かっぱの生きばり
9	おおかみと 7 ひきのこやぎ	ごきげんななめのてんとうむし	しゃっくり百万べん	七夕天人
10	ふるでらの しゃみせんひき	七夕天人	七夕天人	魔法使いのチョモチョモ

れの評価極性も持たない単語は even の評価極性を持つとして e 値を計測し，e 値の有無による検索結果の比較を行うこととした．また，合わせて出現頻度 (Term Frequency; TF) を用いて類似度定義を行った場合と前述の TF-IPF を用いた類似度定義を行った場合の比較を，実際の検索例を用いて検討する．

算出した類似度に基づくランキングの出力結果を表 1 に示す．この表は，「かわいそうなぞう」をクエリとして得られた類似度のランキング（上位 10 話分）を示している．類似度算出の際の分割数は 10 に統一した．表 1 における出現頻度 (Term Frequency; TF) の結果において，「きつね」「おおかみ」「こうもり」がキャラクタとして出現する物語は，e 値を考慮した場合は 4 話，e 値を考慮しない場合は 5 話だった．これらの単語はいずれも「日本語評価極性辞書」において negative 極性が付与された単語である．この結果に対して，「TF-IPF」の結果では，「きつね」「おおかみ」「こうもり」が出現する物語は，いずれも 2 話にとどまった．このことから，絵本の場合には TF-IPF による類似度算出が適していることが示唆された．この結果を受けて，TF-IPF による結果の変動についてより詳細に検証するため，評価極性をヒストグラムに出力し，TF-IPF による結果の変化を観察した．図 1 および図 3 は，「かわいそうなぞう」および「おおかみと七ひきのこやぎ」における TF の推移を表している．「おおかみと七ひきのこやぎ」では，「おおかみ」という単語が複数回使用されており，その影響で n 値が比較的高く算出される傾向が見られた．TF-IPF を用いることで，「おおかみと七ひきのこやぎ」の評価極性の推移は図 4 へと変化した．その結果，物語全体にわたって p 値と n 値の度数が逆転する様子が確認された．

一方，e 値の推移を類似度判定に用いることについて，ランキングの出力における大きな影響は見られなかったが，ヒストグラムの比較による定性的な評価では，e 値の影響が p 値，n 値と比較して，より強く影響する様子が見られた．この影響は，特に e 値が多く出現する物語において確認された．この結果は，評価極性の推移で類似度を判定するのではなく，単なる文字の出現量によって類似度を判定してしまう懸念がある．そのため，人手によるランク付けの実験を行う際には，TF-IPF による p 値及び n 値を類似度算出に用いることにした．

5. 人手によるランク付けの実験

前章で採用した類似度定義に基づき，ある絵本に対して類似であるとされた絵本が，実際に人の感性に照らしても適切かを実験的に検証する．実験を開始する前に，まず任意の絵本（以下，クエリの絵本と記す）を選択し，提案法を用いて選択した

絵本それぞれに類似する絵本を出力する．クエリの絵本は，(1) 実験をスムーズに進行させるため，およそ 30 ページ前後の本に統一する，(2) 実験対象の物語を参加者が誤解するのを避けるため物語集を省く，という二つの指標を元に，人手で選択した．出力した絵本の中から，関連度のランキングが 50 位以内に位置する絵本，101 位以下に位置する関連度の低い絵本，その中間に位置する絵本（以下，比較対象の絵本と記す）を 2 冊ずつ，合計 6 冊選択する．

実験の手順を以下に示す．実験参加者は，提示されたクエリの絵本，および比較対象の絵本を読み，内容を把握する．いずれの絵本に関しても，絵本を読み返す回数に制約は設けず，参加者が任意のタイミングで何度も内容を確認することを可能にした．最後に，参加者は読んだ絵本の物語の内容が似ている順に 6 冊の比較対象の絵本をランク付けする．最後に 2 分程度の半構造化インタビューを行い，終了とした．

実験の結果は，Spearman の順位相関を用いて分析した．表 2 は，クエリの物語と比較対象の絵本についての実験参加者のランク付けの結果を示している．表 2 中の要素は，提案法によって算出された正解データのランクになっている．実験の結果，0.4 から 0.5 程度の相関が見られた．一方，「かわいそうなぞう」および「フランダースの犬」については，他の書籍に比べて著しく相関係数が低くなった．とりわけ，正解データにおいて，提案法によって 1 位にランク付けされた絵本が，実験参加者には 6 位にランク付けされる，もしくは提案法によってもっとも類似していない 6 位であるとランク付けされていた絵本が，実験参加者によって 1 位にランク付けされるといった，提案法による評価と実験参加者による評価に大きな隔たりがある絵本が確認された．そこで，この 2 冊の絵本を対象に，絵本の内容に関する追加アンケートを行った．

アンケートは，絵本の内容について，算出された類似度と人による評価の間の差異を確認するために行った．アンケートの対象にした絵本は，前述の 2 冊のクエリの絵本と，それぞれについて最も正解データから離れた位置にランク付けされた絵本である．それぞれをクエリの絵本によって 2 群に分けた．具体的には，クエリの絵本「かわいそうなぞう」とその比較対象の絵本である「どろぼうがっこう」(提案法での類似順位=1 位) を群 1，クエリの絵本「フランダースの犬」とその比較対象の絵本である「かたあしだちょうのエルフ」(提案法での類似順位=6 位) を群 2 とした．

アンケートに先立ち，まず回答者に群ごとの書籍 2 冊を読んでもらい，内容を把握してもらった．次に，それぞれの書籍について「この物語は明るい話でしたか，暗い話でしたか」という設問に，5 段階で答えてもらった．回答欄は 1 に近づくほど「暗い話」と評価され，5 に近づくほど「明るい話」と評価される形式になっている．参加者は 10 名で行い，群ごとに 5 つの回答を収集した．

アンケートの結果を表 3 に示す．1 群は提案法で出力した正解データによると数値が近くなることが期待されたが，アンケートの結果では平均値に 2.8 の差が出た．また，2 群では出力した正解データによると数値が大きく離れることが期待されたが，アンケートの結果では 0.8 の差にとどまった．これは人手によるランク付けの実験の結果を裏付けるものである．

実験後の半構造化インタビューでは，「最も似ている，もしくは似ていないとランク付けした書籍について，その判断の決め手になった特徴は何か」「ランク付けの際に最も悩んだ書籍とその特徴は何か」の 2 点についてインタビューを行った．前者についての回答としては，ストーリー展開の他に「絵本の登場キャラクタの人数」や「場面転換の多さ」といった回答が得られた．後者については，「似ているかどうかは判断できたが，ランク付けが難しかった」「似ているかどうかの判断基準が自分で定まらなかった」という回答があった．また，インタビューの中で，「絵本のレイアウトを意識した」「図鑑のような．読む順序が決まっていない書籍について悩んだので，そういうものは大体ランクが下になるようにした」という回答が得られた．このことから，図鑑形式の書籍のような，読む順序が一意に定まらない書籍については，この手法は不適であるといえる．また，アンケートに用いた絵本のうち 2 冊を定性的に評価した結果，「かたあしだちょうのエルフ」は具体的な情景描写によってシーンが描かれており，アンケートの結果に対して negative な評価極性を持つ単語が比較的少なかった．また，「どろぼうがっこう」については「笑いとしての罵倒表現」のような，単語単独で意味することと文脈上意味することが逆である場合が見受けられた．いずれも，現行の手法のみでストーリー展開を把握することは難しいと言える．

6. 議　　論

6.1 使用したデータに関する議論

今回の実装に使用した「日本語評価極性辞書」は，人手で集められたデータであり，positive，negative，そのどちらでもないものを示す even の 3 種類の評価極性の評価を付与している．一方，自然言語の意味を考慮したシステムを開発する際，シソーラスや意味辞書など，語と語の関係性を示したデータベースがしばしば活用される．意味辞書に関する研究として，NICT で取り組まれている日本語 WordNet などが挙げられる [11]. これは日本語の意味辞書であり，ある語について複数の語を「関係 (e.g., 上位関係，全体部分関係)」で分類し，概念，語義，定義文といった情報が付与されている．こうしたデータベースを併用することで，物語に出現した単語の関係を知ることは，類似する単語の評価極性をたどることが可能になり，より多くの単語を，ストーリー展開を測るための手掛かりとして使用することが期待できる．

評価極性の分類についても検討の余地がある．今回使用した評価極性は，日本語テキスト全般用の評価極性であり，前述した「一部の生物名に対する評価極性」に関する問題など，絵本特有の表現について，現状の評価極性が適していない様子が確認された．以上のことから，絵本のストーリー展開の傾向を把握するためには，絵本用に特化した評価極性の分類が必要である．絵本のストーリー展開に特化した評価極性の分類を明らかにすることで，より内容が類似する絵本の検索が可能になると考えられる．

表 2　実験参加者の評定結果

クエリの絵本	実験参加者	1 位	2 位	3 位	4 位	5 位	6 位	相関係数	同じ組の平均
かわいそうなぞう	user1	3	5	6	2	4	1	−0.486	−0.057
	user2	2	3	5	4	1	6	0.371	
ないたあかおに	user3	1	6	2	5	4	3	0.200	0.429
	user4	2	1	4	6	3	5	0.657	
14 ひきのとんぼいけ	user5	1	5	4	3	2	6	0.429	0.486
	user6	3	4	1	2	5	6	0.543	
ももたろう	user7	2	3	6	4	1	5	0.200	0.5434
	user8	1	3	2	4	6	5	0.886	
フランダースの犬	user9	6	5	4	1	2	3	−0.771	−0.571
	user10	4	6	1	5	2	3	−0.371	

表 3　追加アンケートの結果

		回答者						
群	書籍のタイトル	1	2	3	4	5	平均値	中央値
1	かわいそうなぞう	1	2	1	2	1	1.4	1
	どろぼうがっこう	4	3	4	5	4	4.2	4
2	フランダースの犬	2	2	2	1	3	2.0	2
	かたあしだちょうのエルフ	1	1	2	1	1	1.2	1

6.2　検索手法に関する議論

絵本の中には右側に絵，左側にテキスト，というような，一定の形式をもつものが存在する．こうした絵本をページ単位で扱うことは，評価極性をもつ単語の数の推移が振動的になる．このことは，バタチャリア係数による類似度算出の際，分割数による影響を大きく受ける懸念がある．この影響を低減するためには，ローパスフィルタを用いて単語数の推移を平滑化し，その上で類似度を算出する方法が考えられる．また，物語の中には，「冒頭が比較的長い」「ごく短く結末が述べられる」など，ストーリーの展開の比重が異なる状況が考えられる．現行の手法では，クエリの絵本と，比較対象の絵本のページ数の差を正規化した際，結果が線形的に処理され，ストーリー展開の起承転結の比重の差が考慮されない．このような起承転結が異なる絵本を扱うには，動的時間収縮法 (Dynamic Time Warping) [12] を用いるなど，長さの異なるデータ間の類似度を測る手法を適用する必要がある．

絵本間の類似度の算出方法についても，複数の手法が考えられる．本研究で使用した絵本のデータベースの中には，「物語集」のような，ごく短いページ内に物語が記載されている物語が見受けられた．今回はページ単位で取得したデータに，バタチャリア係数を使用することで極性ごとの類似度を算出したが，こうした 1 ページに物語が収まっている形式の童話など，極端にページ数の少ないもののストーリー展開を分析するには，類似度の定義が別途必要となる．同様に，前述した「読む順序が一意に定まらない書籍」に関しても類似度の再定義が必要である．アンケートで得られた「言外の意味を含む記述」については，記述されたテキスト情報のみを用いて類似度判定をする手法で対応することは難しい．そのため，ストーリーに関するクラスタリングと併用するなど，大まかにストーリーの分類を行う方法と併用する必要がある．

7. おわりに

本研究は，絵本のストーリー展開傾向に基づいた検索を可能にするための枠組みづくりを目指している．本稿では，単語の評価極性に着目して物語のストーリー展開を把握し，バタチャリア係数を用いて物語の評価極性の推移の類似度を算出することで，ストーリー展開の類似検索を行う方法について検討した．実験の結果，高度な文脈理解を必要とする物語や，ページ数の極端に少ない物語，読む順番が決まっていない物語など，幾つかの書籍においてより検討が必要であることがわかった．今後の展望としては，現行の評価極性の分類および類似度の算出方法について検討するほか，書籍のクラスタリングなどと併用することで，検索の精度向上を目指す．

文　献

[1] 総務省統計局，“第六十六回日本統計年鑑，” 2016.

[2] R. Yamashita, K. Okamoto, and M. Matsushita, “Exploratory search system based on comic content information using a hierarchical topic classification,” Proc. ACIS2016, pp.310–317, 2016.

[3] 服部正嗣，小林哲生，藤田早苗，奥村優子，青山一生，“ピタリエ：興味・発達段階にピッタリの絵本を見つけます，” NTT 技術ジャーナル，pp.54–59，2016.

[4] 佐々木宏子，新曜社，“絵本の心理学 子どもの心を理解するために，” 2000.

[5] 松村　敦，濱沖肯志郎，榎本祐季，三島悠希，“ソーシャル絵本推薦システムにおける自動推薦機能導入の試み，” 情報知識学会誌，vol.26，no.2，pp.211–216，2016.

[6] 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，pp.584–587，2008.

[7] 相澤彰子，“大規模テキストコーパスを用いた語の類似度計算に関する考察，” 情処論，vol.49，no.3，pp.1426–1436，2008.

[8] S.-H. Cha, “Comprehensive survey on distance/similarity measures between probability density functions,” Int. J. Math. Mod. Meth. Appl. Sci., vol.1, no.4, pp.300–307, 2007.

[9] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying conditional random fields to japanese morphological analysis,” Proc. EMNLP2004, pp.230–237, 2004.

[10] 三木光範，加藤恒昭，自然言語処理，情報工学テキストシリーズ，共立出版，2014.

[11] H. Isahara, F. Bond, K. Uchimoto, M. Utiyama, and K. Kanzaki, “Development of japanese wordnet,” Proc. LREC2008, pp.2420–2423, 2008.

[12] H. Sakoe and C. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” IEEE Trans. Acoust., vol.26, no.1, pp.43–49, 1978.