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
HCS2017-86（2018-01）

トピック分類を用いた絵本の類似検索に関する検討

朴　　炳宣[†,††]　松下　光範[†]　　服部　正嗣[††]

† 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: †{k281401,t080164}@kansai-u.ac.jp, ††hattori.takashi@lab.ntt.co.jp

あらまし　本研究の目的は，絵本のストーリー展開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では，書誌情報やレビューなど絵本に付随する情報に基づいた検索機能が提供されているが，これらのサービスでは絵本の内容自体に基づいた詳細な検索が困難であり，「この本と同じようなストーリー展開を持つ絵本を読みたい」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，本稿では絵本同士が持つ共通の話題を推定し検索に用いることで，類似した展開を持つ絵本を検索可能にする手法を提案する．提案手法では，絵本本文から潜在的トピックの推定を行い，トピックに関連する語集合に対して各絵本のページ毎の出現ヒストグラムを作成することでストーリー展開を表現し，絵本同士の類似度はヒストグラム間のバタチャリア係数として算出する．
キーワード　トピック分類，LDA，バタチャリア係数，絵本検索，類似検索

Similarity Search for Picture Books Using Topic Classification

Byeongseon PARK[†,††], Mitsunori MATSUSHITA[†], and Takashi HATTORI[††]

† 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: †{k281401,t080164}@kansai-u.ac.jp, ††hattori.takashi@lab.ntt.co.jp

Abstract　The purpose of this research is to develop a system for searching picture books with storyline-related query. To carry out storyline-related queries, most existing search systems require costly manual preliminary processing; Outlines of books has to be written in reviews or annotated as bibliographic information by previous readers, and books without the information cannot be searched at all. Recently a search method that does not require the manual processing was proposed. By focusing on the number of positive / negative words that appear in each pages of books, the method support searches for books with basic story lines such as happy ending stories and heartbreaking stories. However, the variation of the storyline that can be searched by the method is limited due to the simplification. In this paper, we propose a expansion of this method to increase its expressiveness. We preform a latent topic classification to the picture book texts, and obtain topics of picture books and also word sets that are related to each topics. We then express storyline of a book by a page by page word frequency histograms of each word sets. By defining the similarity between picture books as the Bhattacharyya coefficients between histograms, picture books with the similar story lines can be searched.
Key words　Topic classification, Latent Dirichlet Allocation, Bhattacharyya Coefficient, 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

(注1)：http://book.tsuhankensaku.com/hon/, 2018 年 1 月 4 日確認
(注2)：http://www.ehonnavi.net/, 2018 年 1 月 4 日確認

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内容情報 (e.g., 絵, テキスト, ストーリー) 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絵本の内容を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」「旅をするお話が書かれた本を探したい」といった，書籍の内容に直接関わる情報をクエリとした詳細な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，書籍のストーリー展開に着目して検索を可能にする試み [5] が行われている．しかし，書籍のストーリー展開は登場人物やその行動，舞台となる背景など，様々な要素によって表現されている．これらの要素は著者によってテキストや絵として感性的に表現されており，同じ内容を含めた書籍であってもその表現が大きく異なることがある．そのため，ストーリー展開といった書籍の内容情報を用いた検索を実現するためには，表層的な表現が異なる書籍同士の関連を把握することが望ましい．

書籍の内容に関する検索を実現するため，本研究では書籍のテキストを用い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，絵本同士が持つ共通の話題を推定し検索に用いることで，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

2.1 絵本検索に関する研究

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.,「動物が幸せに遊ぶ物語」）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，2,100 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，それぞれの結果がどのような過程を通して得られた結果なのか判別できない．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．

安尾らは，絵本のストーリー展開に基づく検索を実現するために，ストーリー展開を絵本テキストの positive / negative な状況を表す単語（e.g., うれしい，ケガ）のページごとの出現頻度のヒストグラムで表現する方法を提案している [5]．絵本間のストーリー展開の類似度は，positive / negative 単語ヒストグラム同士のバタチャリア係数の和で算出される．安尾らのシステムが実現すれば，「物語中盤に悲しい出来事が起こるが最後はハッピーエンド」というストーリー展開の絵本を入力として同様のストーリー展開の絵本を探すことが可能になる．しかし，安尾らの手法の評価実験では，「複雑な表現が多く含まれた絵本」や「場面転換の多い絵本」に関して，システムが類似度が高いと判断した絵本にユーザが違和感を覚える傾向が確認された．これは文脈に由来する単語の多義性があったことに加え，positive / negative という二極のみでは絵本のストーリー展開を十分に表現できなかったことが原因であると考えられる．

2.2 トピック分類を用いた検索

トピック分類を用いたコンテンツ検索として，山下らの「コミック探索システム」が提案されている [2]．山下らは，コミックの内容に関する情報を獲得するためにコミックのレビューに着目し，レビュー本文からトピック分類を行うことで，コミック同士の関連を生成する手法を提案した．渡邉らは，観光地の向上点発見を企図した可視化システムの実現のために，Web 上に存在する各観光地に関するテキストからトピック分類を行う手法を提案した [6]．これらの手法ではトピック分類を用いることで，記事や書籍よりも自由に書かれた Web 上のテキストから文書同士の関連性や類似性を生成している．

高木らは，e テスティングにおいて類似項目を自動検索することを目的とし，項目間の類似度の算出手法を提案しており，トピックモデルを用いることによって単語の出現頻度や共起に基づいた既存の手法よりも高い精度の類似度を実現した [7]．高木らの手法では，e テスティングに用いられる項目同士の類似度の向上のために，トピックモデルを推定する際に適した品詞や単語を絞り込むといった前処理を行っている．

2.3 本研究の位置付け

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．特に絵本のテキストを用いたストーリー展開の自動的分析においては，同じ意味をもつ単語同士であっても表層的な違いにより区別される困難が存在する．そこで本研究では，絵本テキストから潜在的トピック分類を用いて「絵本同士が持つ共通の話題 (トピック)」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，「絵本は話の流れが存在し，あるテーマ (e.g., 動

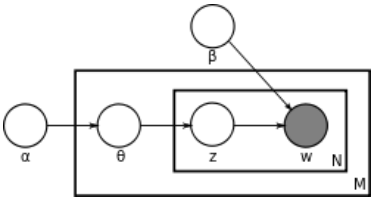

図 1　LDA のトピック生成過程

物の物語，家族の物語) を持っている」という仮説を元に，絵本に存在するトピックの推定を行った．トピックとは，スポーツ・政治・音楽といったテキストの内容を指す [8]．また，あるテキストに含まれたトピックが単一であるとは限らず，多重で存在する場合が存在する [8]．このようなトピックの特徴を踏まえ，絵本テキスト中のトピックをページごとに推定して，その推移を「絵本のストーリーの変化」を判定するアプローチとして用いる．そして，トピックの推移パターンが類似する絵本同士はストーリー展開が類似していると考える．

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，見開きを一つの単位としてデザインされるものを想定している．こうした背景も踏まえて，絵本のトピックの推移を推定する際には，絵本の分割の粒度を決める必要がある．本稿では，絵本を各ページに分割してページごとのトピックを推定し，その推移によってストーリー展開を表現する．さらにページ数の異なる二冊の絵本のトピックの推移を比較可能な関数に基づいて，絵本同士のストーリー展開の類似度を算出することとした．

3.2　Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) とは，文書中の単語は独立に出現しているのではなく，潜在的なトピックに基づいて出現するという仮定に基づいた文書生成モデルである [9]．なお，LDA では，各文書は複数のトピックで構成されており，各トピックの単語分布を合算した形で単語が生成されていると仮定しているアルゴリズムであるため，文書に含まれた複数のトピックを推定・把握することが可能である．図 1 は LDA の生成過程を示す．LDA を用いたトピックモデルの推定においては，語 w の列によって表現された文書の集合と，トピック数 K を入力として，各トピック $z_n(n = 1, ..., K)$ における語 w の確率分布 $P(w|z_n)(w \in V)$，及び，各文書 d におけるトピック z_n の確率分布 $P(z_n|d)(n = 1, ..., K)$ を推定する．LDA を用いることによって，表層的な表現の違いにより区別されていた語を関係付けることが可能となる．

絵本は小説や漫画など他の創作物と同様，様々な要素から構成されるので，絵本のストーリーは必ずしも一つのトピックで表現できるとは限らない．本稿では，LDA を用いて絵本のストーリーに含まれた複数のトピックを推定することで，より詳細にストーリーの展開を把握することを目指す．

3.3　類似度の算出

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 [10]，本稿ではページあたりのトピック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) [11] 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．ここて，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，ページ単位でトピックを推定し，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，それに基いて各トピックの関連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，本文の内容からページごとにトピックを推定し，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することで，トピックの推移が類似している絵本を検索する．

4.　実　　装

本研究では，3,104 冊の絵本を対象に，ページごとに現れる単語から絵本ごとのトピックを推定した．まず，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 [12] 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．全ての絵本から抽出した単語を元に，LDA によるトピックモデルを作成した．LDA を用いたトピック推定においては，トピック数 K を人手で与える必要があるが，今回は，事前に行った複数のトピック数を用いた場合の定性的な比較に基づいて，$K = 15$ を採用した．なお，LDA の適用に当たっては単語の品詞に注目し，助詞・代名詞などの出現頻度が多く意味を持たないとされる品詞の単語は除外し，名詞・動詞・形容詞・形状詞の単語のみを採用した．さらに，特定の一冊の絵本にしか現れない，あるいはほとんどの絵本に現れるなどの極端な文書出現頻度の単語の影響を排除するため，文書出現頻度が 1 の単語および 900 以上の単語 (全単語の 32%) も除外した．作成したトピックモデルの単語分布を表 1 に示す．作成したトピックモデルを用いて，絵本の各ページごとに登場する単語からトピック分布を求めた．図 2 は「かぐや姫の物語」(文：坂口理子, 原作：「竹取物語」, 角川書店, 2014 年) に現れたのページごとのトピック分布の変化の例を示す．図 2 を見ると，「かぐや姫の物語」では，自然に関連する単語が含まれた Topic 11 や，童話によく出現する単語が含まれた Topic 5 がトピック分布の多くを占めていることが確認できる．また，分布は一定ではなく，ページによって動物に関連する単語が含まれた Topic 2 も現れ

表 1　トピック分類結果 (上位 15 単語)

トピック	単語
Topic 0	王，大臣，卵，城，熊，象，部屋，国，兵隊，博士，コック，嘘，箱，時計，気
Topic 1	海，動物，魚，ライオン，水，象，長い，体，鳥，鼠，足，泳ぐ，虫，卵，仲間
Topic 2	狐，狼，兎，子，森，熊，父，山，空，虎，兄，歩く，友達，走る，風
Topic 3	ママ，パパ，赤ちゃん，花林，姉，童，豚，眠る，子，生まれる，凄い，御腹，遊ぶ，寝る，好き
Topic 4	婆，爺，セント，鼠，ラーメン，狸，太郎，山，豆，子，亀，村，兎，鶴，隣り
Topic 5	花，姫，王子，チム，国，娘，美しい，神，種，咲く，城，若者，実，子供，魔法
Topic 6	カー，車，走る，伯父，リトル，バス，乗る，自動，町，トラック，駅，主人，電車，運ぶ，小さい
Topic 7	雪，伯母，奥，蛙，父，降る，遊ぶ，山，寒い，団子，入れる，達磨，和尚，外，火
Topic 8	猫，パン，犬，伯父，困る，化け，入れる，子猫，美味しい，小さい，店，妹，驢馬，買う，鼠
Topic 9	先生，子，子供，父，靴，書く，学校，笑う，泣く，足，小さい，頭，歩く，嬉しい，歌う
Topic 10	メリー，ケーキ，父，ランプ，葉，クリスマス，クリーム，雪，入れる，魔法，男，海，城，部屋，チョコレート
Topic 11	使う，星，水，世界，マンデー，宇宙，地球，ゲーム，人間，空気，知る，体，場所，妹，考える
Topic 12	子豚，骨，原子，黒，元素，語，クレヨン，炭素，使う，酸素，バナナ，ドラゴン，駅，弁当，色
Topic 13	先生，本，気，訳，書く，おなら，戻る，部屋，獅子，読む，妖怪，金，トイレ，口，慌てる
Topic 14	鬼，猿，恐竜，鰐，帽子，船，川，海賊，島，女の子，坊や，雀，蛇，人間，金

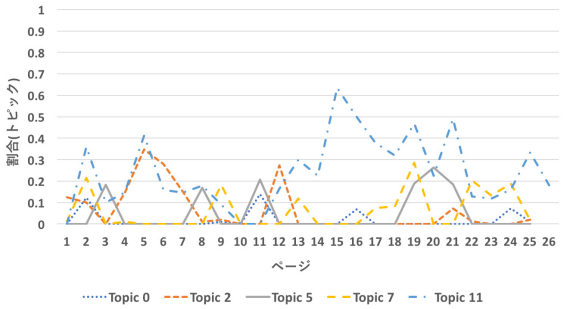

図 2　トピックの推移の例 (「かぐや姫の物語」より)

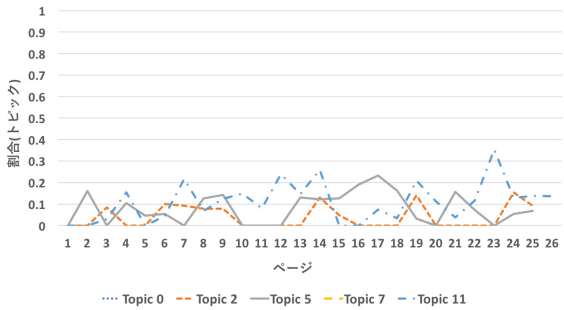

図 3　トピックの推移の例 (「はじめての古事記日本の神話」より)

ていることから，トピックの推移によってストーリーの展開が表現されていると考えられる．

次に，作成したデータを用いて絵本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，そのテキストからトピック分布の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トピックの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．

5.　トピックの推移に基づく類似度算出法の検証

トピックモデルの作成に用いた 3,104 冊の絵本に対して，トピックの推移に基づく類似度を計算を行った．

同一作品でありながら異なる書籍として出版されている絵本 5 作品 (e.g., まてまてタクシー, ちょっとだけまいご) では，提案手法による類似度は 100% であった．このことから，トピックの推移による類似検索から内容において詳細は異なるが同一テーマの作品を検索できることが確認された．

同一作品以外で類似度が高く現れた例として，「かぐや姫の物語」と「はじめての古事記日本の神話」が挙げられる．図 3 は「はじめての古事記日本の神話」のページごとのトピックの推移を示す．両作品の類似度は 95% という高い類似度を見せており，特に昔話によく出現する単語が含まれた Topic 5 において高い類似度を示し，動物に関連した単語が多く含まれる Topic 2 においても一定程度の類似度を示した．両 Topic 共に二つの作品を特徴付ける重要な要素である．この結果により，提案手法を用いることにより，似たテーマを持ちつつも異なる内容と表現を持つ二つの作品を結びつけることが可能であることが示唆された．

表 2 と表 3 は，ストーリー展開を考慮した書籍検索の先行研究である安尾らの手法と提案手法の類似度検索結果を比較したものである．比較する標本として，安尾らの手法によって提示された結果に対してユーザからの評価が最も高い結果が現れた「ももたろう」と最も低い結果が現れた「フランダースの犬」を用いた [5]．表 2 に現れた結果に対して本の内容を確認するなど定性的に評価した結果，提示された絵本の中で最も対象とする絵本と類似していたのは安尾らの手法によって提示された「お話してよ，もう一つコルウェルさんのお話集」であった．この際，対象の絵本と類似していると判断した基準は，「起承転結が存在する」点である．一方，提案手法によって提示された結果には「ももたろう」と類似している絵本は現れなかったものの，どれも類似度が低く現れていた．しかし，安尾らの手法による結果の中には，明らかに対象絵本と大きく内容や展開の異なる絵本であっても 80% 以上の高い類似度を示していた．さらに，提案手法によって提示された絵本における各トピックごとの類似度を確認した結果，　各絵本が動物や食べ物を連想させる Topic 8 に対して高い類似度を示していた．この結果により，提案手法では安尾らの手法よりも絵本の内容より細分化された類似検索結果を示していることが示唆された．表 3 の結果においては，提示された絵本の中で最も対象とする絵本と類似

表 2 「ももたろう」をクエリとした際の類似度検索結果 (上位 5 位)

	安尾らの手法		提案手法	
順位	作品名	類似度	作品名	類似度
1	お話してよ，もう一つコルウェルさんのお話集	**92%**	みんなうんち	64%
2	ぼく，だんごむし	90%	きょうのおべんと	63%
3	てぶくろをかいに	90%	どうぶつ	63%
4	おさるのジョージアイスクリームだいすき	89%	つのはなんにもな	62%
5	チキン・サンデー	89%	りんご	62%

表 3 「フランダースの犬」をクエリとした際の類似度検索結果 (上位 5 位)

	安尾らの手法		提案手法	
順位	作品名	類似度	作品名	類似度
1	たんけんたいと消防たい	97%	続・しごとば	80%
2	本だらけの家でくらしたら	97%	本だらけの家でくらしたら	78%
3	王さまなぜなぜ戦争	97%	一年生の漢字えほん	78%
4	おしゃべりなたまごやき	97%	はじめての古事記日本の神話	78%
5	しらゆきひめ	97%	忘れないよリトル・ジョッシュ	**78%**

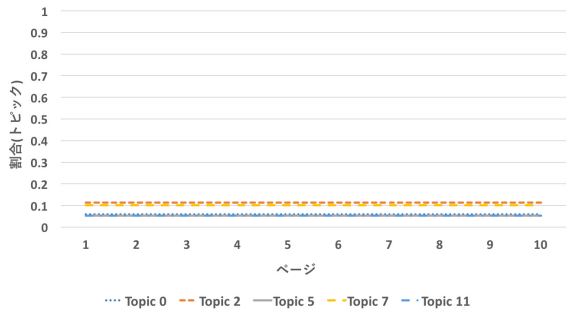

図 4　トピックの推移の例 (「おかあさんだ」より))

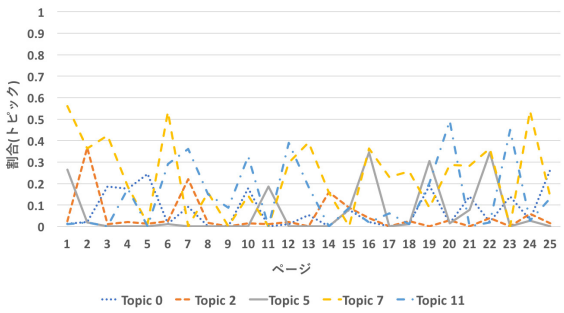

図 5　トピックの推移の例 (「一年生の漢字えほん」より))

していたのは提案手法によって提示された「忘れないよリトル・ジョッシュ」であった．この際，対象の絵本と類似していると判断した基準は，「人間である主人公と動物が中心になる内容である」，「起承転結が存在する」点である．「フランダースの犬」と「忘れないよリトル・ジョッシュ」の各トピックごとの類似度を確認した結果，自然を連想させる Topic 11 や動物や食べ物を連想させる Topic 8 に対して最も高い類似度を示し，乗り物や町を連想させる Topic 6 と王国を連想させる Topic 0 に対しても他の絵本と比べて高い類似度を示していた．しかし，提案手法による結果において，「忘れないよリトル・ジョッシュ」は類似度の順位で 5 位に位置しており，「忘れないよリトル・ジョッシュ」より上位の絵本は「フランダースの犬」とはかけ離れた内容の絵本のみであった．

6. 議　　論

6.1 トピックの推移に基づく類似度に関する議論

5. 章の結果により，同じ内容を持つ絵本同士が高い類似度を示したことや表現は異なるものの扱うテーマが類似している絵本同士が特定のトピックにより関連づけることが可能だったことから，トピックの推移よるストーリー展開に基づいた類似検索が可能であることが示唆された．さらに，安尾らの手法との比較による定性的評価では，各トピックごとの類似度を把握することによって安尾らの手法よりも絵本の内容に基づき細分化された類似度を提示することが可能であることが期待できる．

しかし，一部の結果においてはクエリとした作品との関連を見出すことができない絵本が高い類似度を示していることが確認された．複数の絵本に対して異なる内容でありながらも高い類似度を示していた絵本の例として，「おかあさんだ」と「一年生の漢字えほん」があげられる．図 4 と図 5 は両絵本のトピック推移の例を示す．絵本の内容を確認した結果，「おかあさんだ」は低年齢向けの絵本であり，テキストによる描写が少ない傾向を持つ．今回の実装のために設けた単語の選定基準により，各ページを表現する単語が「母」という一つのみしか残っておらず，その結果各ページごとのトピック推定を行う際，どのトピックにも分布されなかった．ある絵本が特定の内容のトピックに多く分布している場合，ほかの 14 種類と高い類似度を持つため，全体的類似度が高くなることから，どの絵本とも高い類似度を示していると思われる．これにより，1 ページに現れる単語の量が極端に少ないことに起因して関連するトピックの変化が乏しい絵本は，ストーリー展開による類似検索に適さないと考える．表 4 は，ページごとの単語数の最小値が 1 となる絵本を取り除き提案手法を適応した際の類似検索結果を示す．クエリとして「2 歳からはじめるよみきかせ絵本日本の名作」に掲載された「おおかみと 7 ひきのこやぎ」を用いた結果，安尾らの手法や本提案手法による結果では異なる内容を持つ絵本のみである一方，改良案による結果では異なる書籍に掲載されている同一内容の作品が最も高い類似度を示していた．

また，「一年生の漢字えほん」は小学生を対象とした学習教材

表 4 「おおかみと 7 ひきのこやぎ (2 歳からはじめるよみきかせ絵本日本の名作)」をクエリとした際の類似度検索結果

	安尾らの手法		提案手法		改良案	
順位	作品名	類似度	作品名	類似度	作品名	類似度
1	親指トム （金のがちょうのほん 四つのむかしばなし）	93%	かいけつゾロリの きょうふの宝さがし	65%	**おおかみと 7 ひきのこやぎ （こどもに人気のめいさくたからばこ）**	**96%**
2	だるまちゃんとかみなりちゃん	89%	かいけつゾロリの クイズ王	65%	はなさかじいさん （2 歳からはじめる よみきかせ絵本日本の名作）	86%
3	リサとガスパールのしんがっき	88%	おばけのコッチピピピ	65%	おしゃべりなたまごやき	85%
4	エアポートきゅうこうはっしゃ!	86%	ぐぎがさん，ふへほさん， おつきみですよ	64%	くまの子ウーフ	85%
5	給食番長	85%	わかったさんのショートケーキ	64%	ミリー・モリー・マンデーとともだち	83%

として構成された絵本であり，ストーリーを持つ絵本ではない．「一年生の漢字えほん」でテキストによって表現されるものは教材としての説明が主な内容である．説明文に含まれた単語により，ページごとのトピック推定の結果，統一性を持たずあらゆるテーマのトピックに分布していた．これにより，あるトピックと高い類似度を持つ特定の絵本に対して局所的に類似することにより，ストーリーを持つ絵本よりも高い類似度を示していると考えられる．このような絵本は，本提案手法のアプローチ対象としていた「ストーリー展開を持つ」絵本とは異なり，既存の手法のように単語の出現頻度や共起を用いた検索を用いることで分類が可能である．本手法の類似度検索の精度向上のためには，ストーリーの有無を事前に把握し，ストーリー展開による類似検索に適した対象の絞り込みが有効であると考える．

6.2 今後の展望

本稿によって提案された手法により，絵本の内容を考慮した類似検索が可能になることが期待できる．一方で，ユーザによる内容に関する情報要求がある際，「ストーリー展開」や「雰囲気」といった抽象度の高いクエリを生成することは，ユーザにとって負担となり得る．そこで，今後の課題として，本提案手法を用いたシステムを実装する上で，提案手法に適した検索方法が必要であると考える．現在検討を行っている検索方法として，「好きな絵本をクエリとする方法」や「キーワードをクエリとする方法」といった案が挙げられる．「好きな絵本をクエリとする方法」とは，ユーザが求める内容と最も近い既知の絵本をクエリーとして用いるという方法であり，ユーザのクエリーを生成する負担が少ないことが期待される．「キーワードをクエリとする方法」とは，ユーザが求める内容を表す単語の組み合わせからトピックを推定した結果と類似する絵本を提示する方法であり，ユーザが思いがけない様々な絵本を提示できることが期待される．また，本稿における提案手法では，LDA やバタチャリア係数の実装においてハイパーパラメータや分割数などを定性的に設定しているため，各パラメータの変化によるトピックの分布やユーザの反応を確認するといった定量的評価を行うことで適切なパラメータの選定を試みる．

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，絵本に含まれたトピック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トピック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手法について検討した．評価の結果，既存の手法よりも内容に関する詳細な類似検索が可能になることが期待できることが確認された．今後の展望としては，トピック推定を活用した検索結果について定量的な実験を行うほか，内容情報に基づく詳細な類似検索に適したインタフェースの実装を目指す．

文　献

[1] 総務省統計局, “第六十六回日本統計年鑑”, 2016.

[2] 山下 諒, 朴 炳宣, 松下 光範, “コミックの内容情報に基づいた探索的な情報アクセスの支援,” 人工知能学会論文誌, vol.32, no.1, pp.1–11, 2017.

[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] 安尾萌, 服部正嗣, 藤田早苗, 松下光範, “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,” 電子情報通信学会技術研究報告, vol.116, no.436, pp.103–108, 2017.

[6] 渡邉小百合, 吉野孝, “観光地間の類似性を基にした向上点発見のための観光情報可視化システム,” 分散協調とモバイルシンポジウム論文集, pp.1357–1362, 2016.

[7] 高木輝彦, 高木正則, 勅使河原可海, 田中健次, “e テスティングにおける LDA を用いた項目間類似度の算出,” 情報処理学会論文誌, vol.55, no.1, pp.91–104, 2014.

[8] 上田修功, 斉藤和巳 “多重トピックテキストの確率モデル–テキストモデル研究の最前線,” 情報処理, vol.42, no.2, pp.184–190, 2004.

[9] D. M. Blei, Andrew Y. Ng and M. I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol. 3, pp.993–1022, 2003.

[10] 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,” 情報処理学会論文誌, vol.49, no.3, pp.1426–1436, 2008.

[11] 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.

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