レビューを利用した電子書籍立ち読み支援

村井　聡一†　　牛尼　剛聡††

† 九州大学芸術工学部　〒 815-8540 福岡県福岡市南区塩原 4-9-1
†† 九州大学大学院芸術工学研究院　〒 815-8540 福岡県福岡市南区塩原 4-9-1
E-mail: †1DS07206@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし　近年，書籍の電子化が進んでおり，膨大な電子書籍の中から，読者にとって価値のある書籍を，効率的に発見するシステムの必要性が高まっている．これまで，この目的のために，検索や推薦による書籍の絞り込みに関する研究が行われてきた．しかし，読者は絞り込まれた複数の書籍に対して，何らかの手法で個々の書籍の価値を判断し,「読む」か「読まない」か，を判断する必要がある．我々は，その判断の際に，現実の書店で行われている立ち読みが有効だと考えた．本研究では，電子書籍を対象とした，小説の立ち読み支援手法を提案する．本システムでは，web 上の書籍のレビューを利用することで，ユーザの興味を喚起する文を推定し，読み始めるべき文として提示する．
キーワード　電子書籍，立ち読み支援，レビュー，キーワード抽出

Support for Electronic Book Browsing Using Its Review

Soichi MURAI† and Taketoshi USHIAMA††

† Graduate School of Design, Kyushu University
4-9-1 Minami-ku, Fukuoka-shi, Fukuoka, 815–8540 Japan
†† Faculty of Design, Kyushu University
4-9-1 Minami-ku, Fukuoka-shi, Fukuoka, 815–8540 Japan
E-mail: †1DS07206@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

Abstract　Recently, e-book market is booming. Therefore, system for efficiently finding something worth reading from among vast numbers of e-books is demanded. It had been proposed that retrieval and recommendation to narrow down books. However user need select some books worth reading the full story from candidate at its discretion in some way. We would suggest stand reading on web. In this paper, we provide the supporting system to 'browse' e-books for novels. Our system shows an interesting sentence preferentially as a sentence that begins to be read. First, the interesting-keywords are defined by using review, and then, the degree interest of a sentence is calculated.
Key words　digital book, Support browsing, review, keyword extraction

1. はじめに

近年，iPad や Kindle などの電子書籍を閲覧できるデバイスの普及にともない，電子書籍市場は急成長している [1]．それにともない，ユーザにとって価値のある書籍を効率的に見つける機能が重要になっている．これまでに，大量の書籍の中から閲覧する候補を絞り込む事を目標として，検索・推薦に関する研究が行われてきた [2], [3]．しかし，これらのシステムを使用する際に，ユーザは検索や推薦された書籍をすべて読む訳ではない．ユーザは，検索された書籍や推薦された書籍の中から，自らの判断で読む書籍を選択することが一般的である．実世界では，人は，候補となった書籍を読むか読まないかを決定するために，書籍の本文を閲覧する「立ち読み」を行うことが多い．web においては，ユーザが読む書籍を選択する際に，出版社や「BOOK」データベース [4] の内容紹介，レビュー等が主に利用されているが,「立ち読み」を行うことで，ストーリー，レトリックによる雰囲気，読みやすさ等を把握することも有効である．

近年は，ユーザによる電子書籍の選択を支援するために，Web 上で「立ち読み」を実行可能とするサービスも提供されている [5]．しかし，Web 上での「立ち読み」は現実世界の「立ち読み」と比較するといくつかの制限が存在する．現実世界における立ち読みでは，好きな時間，好きなページを閲覧できるのに対して，商品の電子書籍における「立ち読み」では，閲覧可能な時間，もしくは閲覧可能なページが制限されている．このような制限により，書籍中にユーザの興味を引く文章があった

としても，それが目に触れないことで，ユーザが適切な判断を下せない場合がある．

立ち読みが利用できるサービスの一つに Google ブックス [6] がある．Google ブックスでは，ユーザがキーワードを指定することにより，提示されるページが変更される．利用者が指定したキーワードに基づいて，利用者に必要な情報があると思われるページを提示することで，効率的な立ち読みを支援している．このようなキーワードによるページの提示は，教科書や実用書など，調べたい事柄が分かっている場合では有効的である．一方，読んだことのない小説・エッセイ等においては，ユーザは興味のある部分のキーワードを思いつくことが難しいため，上記の手法は有効的でない．

インターネット電子図書館である，青空文庫 [7] では，約 9700 作ある作品 [8] を，無料で読むことができる．好きな時間だけ，好きなページを読むことが出来るが，利用者が自分にとって読む価値のある作品かを判別する際には，判別にかける時間，読む文章量は少ない方が効率が良い．

以上の背景より，本研究では，小説を対象として，電子書籍の「立ち読み」を支援するシステムを開発する．新書やビジネス書を「立ち読み」により，効率的に選別する際は,「目次を読み内容を把握する」,「節，章のまとめを読む」，等の手法があるが，小説の場合にはそのような手法はない．ユーザが読む小説を選ぶ基準として，自身が興味をもてるかどうかは重要な要素である．松田 [13] は，図書館または書店における利用者の行動の調査の結果,「書籍を適当に開いて飛ばしながら見るという行動は，観察中に最もよく見られた行動の一つであり，適当に開くということは初めからきちんと見るほど興味度あるいは特定度が高いわけではなく，漠然と中身に興味がある状態である」と考察している．web の立ち読みにおいて，閲覧できるページが指定されている場合，そのページは 1 ページ目から数ページであることが多い．そのため，現実世界の立ち読みのように，適当にページをめくり，その書籍が興味をひく内容であるか，を判断することができない．

そこで，本システムでは，web 上の書籍のレビューを利用することで，ユーザの興味を喚起する文を推定し，読み始めるべき文として提示する．提示された文の中から，読み始める文を選択し，決められたページ数だけ本文を読むことを可能とする立ち読みを提案する．本論文では，興味をひく文を抽出する手法について述べる．

本論文の構成は以下の通りである．第 2 章で関連する研究を紹介する．第 3 章では，提案する手法の概要を説明する．第 4 章では書籍本文，レビューに出現する単語の重要度を計算し，キーワードを抽出するアルゴリズムについて述べる．第 5 章では，Web 上にあるレビューについて考察し，実験で用いるレビューサイトを決定する．第 6 章 では，ユーザに提示する興味を喚起する文を求める手法について述べる．第 7 章では興味を喚起する文が抽出できたかについて被験者実験を行い，考察を述べる．第 8 章では今後の課題について述べる．第 9 章ではまとめを述べる．付録では，登場人物名を関連するキーワードに置きかえる手法について述べる．

2. 関 連 研 究

これまでに，文章の効率的な理解支援を目的として，キーワード抽出に関する研究が行われてきた．大澤ら [9] は「文章は著者独自の考えを主張するために書かれる」という仮説の下，筆者が新しく主張しようとする内容に関する単語を，単語の役割を「土台」,「屋根」,「柱」に分け，基本概念を形成している「土台」と，文章の出張点となる「屋根」を結ぶ強い「柱」（内容の主な展開）が多く集まった語をキーワードとする手法を提案している．一方，砂山ら [10] は，様々な観点から主題と密接に関連する単語を抽出するために，観点となる単語を特徴づける単語を抽出する手法を提案している．ここでは，文章のキーワードを役割に応じて周辺キーワード，中心キーワード，特徴キーワードの 3 種類にわけ，これらのキーワードを同時に含む文を重要文とする．

これらの研究は，論文等の情報伝達を目的に書かれた文書を対象としているのに対し，本研究では，小説という娯楽を目的とした文書を対象としている．小説は情報伝達を目的とした文書とは異なり，著者の主張を伝えるために簡潔に，理解しやすく書かれているわけではなく，物語は場面によって展開し，様々なレトリックを用いて書かれている．本研究では，小説の書き方の特徴を考慮に入れ，効率的な文章を理解するためではなく，興味を喚起する単語を抽出する．

Web 上には大量の口コミ，レビューが存在しており，レビューを利用したいくつかの研究がなされている．小倉ら [11] は，レビュー閲覧者のレビューに対する評価を考慮に入れた，ランキング手法を提案している．赤木ら [12] は，商品のレビューや口コミにおける評価属性語に注目し，既に閲覧したページには含まれない情報を含むページを検索することで，商品に関する幅広い情報を効率的に取得可能とする手法を提案している．ここでは，評価・レビュー中において評価属性値となる形容詞の近くに存在する名詞句を評価属性語として利用している．

これらの研究は，レビュー自体を効率的に利用することを目的としている．本研究では，書籍本文から興味を喚起する文章を推定するためにレビューを用いる．提示する結果は，レビュー対象の作品ではなく，書籍本文中の文章である点が異なる．

3. 立ち読み支援

3.1 立ち読みとは

松田 [13] は，図書館または書店における利用者の行動観察の結果,「ブラウジング」を「出会わなければ必要か判断できない情報を含む，曖昧さを持つ情報欲求を満たすため，何らかの期待を抱きながら，利用できる感覚全てを用いて広範で多量な情報源から何らかの基準で必要なものを選び取る情報獲得の一手法である」と定義している．本研究では,「立ち読み」を,『書籍の価値を判断するために，書籍の本文を「ブラウジング」 する行為』と定義する．ここでいう「ブラウジング」は，松田の定義した意味とする．また,『効率的な「立ち読み」』とは「利用者にとって，読む価値のある書籍かどうかを，少ない時間，少ない文章量で判断すること」とする．

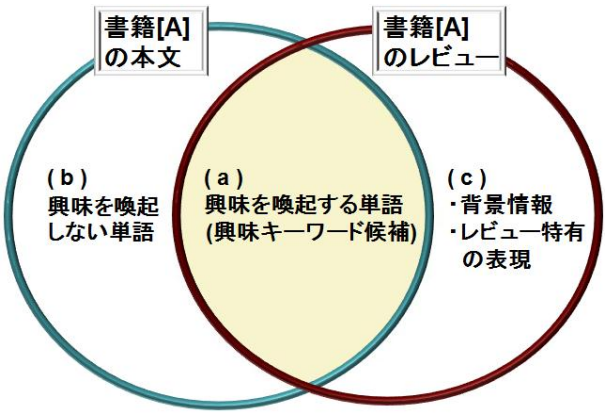

図 1　単語の分類

3.2　興味を引く文

本研究で提案する立ち読み支援は，実世界における立ち読みのように，パラパラとページをめくりながら，興味を引く文が目につくのを期待するのではなく，読み始めるべき興味を引く文を自動的に提示することで，立ち読みの効率を上げる．興味を引く文を決定するために，本手法では,「利用者の興味を喚起する単語を含む文は，利用者の興味を喚起する文である」と定義し，読者が興味を引くと思われる単語を自動的に抽出し，それに基づいて文の興味喚起度を求める．なお，本論文では，読者が興味を引くと思われる単語を「興味キーワード」と呼ぶ．

4.　興味キーワード

4.1　書籍本文，またはレビューに出現する単語の分類

本手法では，利用者の興味を引く単語を見つけるために Web 上のレビューを利用する．レビューには，読者が書籍を読んだ感想が含まれており，そこには読書中に興味を喚起された場面や人物に関する単語も含まれている．我々は,「あるレビューワが興味を感じた対象は，他の読者の興味を引く可能性が高い」と考え，書籍とレビューに出現する単語を図 1 の 3 種類のカテゴリに分類した．

それぞれのカテゴリの特徴を以下に示す．

- 書籍にもレビューにも出現する単語 (カテゴリ (a))　：興味を引く単語，すなわち興味キーワードに成り得ると考えられる．
- 書籍には出現するが，レビューには出現しない単語 (カテゴリ (b))　：興味を引かない単語であると考えられる．
- 書籍には出現しないが，レビューには出現する単語 (カテゴリ (c))　：作品の背景情報，印象，評価を表していると考えられる．

カテゴリ (a) における単語の中には，興味を喚起しない単語も含まれる．また，文の興味喚起度を計算する際に，全ての単語を用いることは，計算量のコストの観点から適切でない．そこで，単語の重要度を算出し，対象とする書籍のレビューおいて特徴的な単語のみを興味キーワードとして用いる．興味キーワード候補に関して，対象とする書籍のレビューにおける出現数が多く，他の書籍のレビューの出現数が少ない単語は，その書籍の特徴的な単語であると考え，興味キーワードとする．

興味キーワードの候補となるのは，書籍の本文とレビューの両方に出現する名詞（代名詞，接尾，数を除く），動詞（自立動詞以外を除く），形容詞である.「する」,「ある」,「よる」,「いる」,「なる」,「いう」,「みる」,「できる」の 8 形態素は出現頻度が高いが，他の語句の補助的な機能を持つ語句であることから，興味の強弱の判断材料として不向きであるため，ストップワードと設定した [14]．形態素に分ける際に，平仮名一文字，カタカナ一文字もストップワードとした．形態素解析には，オープンソースの形態素解析エンジン MeCab [15] を用いた．

4.2　興味キーワード候補の重要度

ある書籍においては出現頻度が高く，他の書籍においては出現頻度が低い単語は，その書籍の特徴を表していると考えられる．出現する単語の重要度を tf(Term Frequency) と idf(Inverse Document Frequency) の積である，tf-idf 法によって求める．tf は注目するドキュメントにおいて，出現数の多い単語を重要とする概念である．一方，idf は複数ドキュメントに出現する単語は重要でないとする概念である．式は以下のようになる．

$$tf\text{-}idf(t_i, d_j) = tf(t_i, d_j) \cdot idf(t_i) \tag{1}$$

この式は，ある書籍 d_j における，単語 t_i の特徴度を示している．ここで，tf は以下の式で求める．

$$tf(t_i, d_j) = \frac{n_i}{\sum_{i=k} n_k} \tag{2}$$

また，idf は以下の式で求める．

$$idf(t_i) = \log \frac{D}{|\{d : d \ni t_i\}|} \tag{3}$$

$|\{d : d \ni t_i\}|$ は単語 t_i を含むドキュメント数を表す．本手法では，ドキュメント D の単位を書籍とする．

本研究では，レビューにおいて特徴的な単語は，ユーザの興味を喚起する単語であると考え，その単語の重要度とする．カテゴリ (a) において，レビューにおける tf-idf 値が高い単語を，興味キーワードとして用いる．

4.3　単語の分類と，tf-idf 値が高い単語の例

例として，太宰治著『斜陽』の書籍の本文，レビューそれぞれにおいて tf-idf 値が高い単語の上位 15 単語を表 1 に示す．書籍の本文は青空文庫 [7] にあるものを，レビューは Booklog [16] にあるものを用いた．

この例において，カテゴリ (b) に含まれる「お座敷」,「叔父」,「ギロチン」,「お方」は，レビューに出現しないことから，レビューワの興味を引かず，印象に残らなかった単語と考えられる．カテゴリ (c) に含まれる単語には,「太宰」,「斜陽」といった書籍名や著者名,「没落」,「滅びる」といった印象，書籍本文において「お母さま」と書かれているところを，レビューに「お母様」と書いた誤植が含まれる．

5.　レビューサイトの種類と特徴

5.1　レビューサイトの分類

本研究では，興味を喚起するキーワードを，Web 上にあるレ

表 1 「斜陽」に出現する単語の分類

本文に出現		
順位	単語	本文 tf-idf 値
1	お母さま	0.0476
2	直治	0.017
3	上原	0.0103
4	おっしゃる	0.0061
5	ひと	0.0053
6	かず子	0.0053
7	言う	0.0044
8	事	0.004
9	お座敷	0.0038
10	叔父	0.0038
11	ギロチン	0.0035
12	ほう	0.0033
13	お方	0.0033
14	無い	0.0031
15	お家	0.0029

レビューに出現		
順位	単語	レビュー tf-idf 値
1	太宰	0.0231
2	貴族	0.0225
3	直治	0.0217
4	かず子	0.0189
5	斜陽	0.0136
6	没落	0.0118
7	弟	0.0095
7	革命	0.0095
9	お母様	0.0078
10	上原	0.0076
11	治	0.0065
12	お母さま	0.0051
13	失格	0.0049
14	母	0.0043
15	滅びる	0.0043

本文とレビューの両方に出現					
順位	単語	本文 tf-idf 値	順位	単語	レビュー tf-idf 値
1	お母さま	0.0476	1	貴族	0.0225
2	直治	0.017	2	直治	0.0217
3	上原	0.0103	3	かず子	0.0189
4	おっしゃる	0.0061	4	弟	0.0095
5	ひと	0.0053	4	革命	0.0095
6	かず子	0.0053	6	上原	0.0076
7	言う	0.0044	7	治	0.0065
8	事	0.004	8	お母さま	0.0051
9	ほう	0.0033	9	母	0.0043
10	無い	0.0031	10	姉さん	0.0041
11	お家	0.0029	11	遺書	0.0031
12	お金	0.0027	12	娘	0.0026
13	山荘	0.0026	13	読む	0.0024
13	洋画	0.0026	14	戦闘	0.0023
15	不良	0.0025	14	開始	0.0023
15	奥さま	0.0025	14	麻薬	0.0023

ビューを用いて決定する．レビューが書かれているサイトは，「コミュニティサイト」,「通販サイト」,「個人の HP や blog」に分類できる．用いるレビューによって結果が異なることが考えられるので，3 種類のサイトの違いについて比較をし，目的にあったものを用いる．コミュニティサイトの代表として，web 本棚サービスである Booklog [16] を，通販サイトの代表として Amazon [17] を利用した．それぞれのサイトに投稿されているレビューには，以下のような特徴がある．

- Booklog

– Booklog は，ブックレビューサイトの一つであり，web 上に仮想的な本棚を作成し，書籍の感想やレビューを書くことができ，他のユーザの本棚や，本の感想をチェックもできる．Booklog には気軽にレビューを投稿できるため，一言のレビューを投稿しているユーザも多い．レビューの内容は，未読者に勧めるというよりも，自分の感想を述べるユーザが多い．投稿数が多く，多数の意見を集めやすい．一方,「○○がよかった」と一言述べるのに留まる簡単な感想，読み終わった日をメモ代わりに残す用い方も見られる．2010 年 12 月 3 日に，レビュー内にネタバレを含む場合、そのことを明記するネタばれ機能が搭載された．それ以前に投稿されたレビューには，ネタばれを含んでいそうなものもあるが，今後はネタばれと明記されていくと思われる．他のユーザのレビューをお気に入りに登録することもできるが，それほど熱心に行われている訳ではない．

- Amazon

– レビューは未読者を対象に書かれており，レビューを書く前に簡単なあらすじや，小説が書かれた背景などを述べるレビューワーもいる．文章の長さも長すぎず短すぎないものが多い．多くの人が結末を知っているような名作以外では，結末をばらすというネタばれ行為は，自粛されている傾向がある．レビューを評価することも普及しており，良質のレビューを書こうとするモチベーションの一つになっている．その影響か，著者や小説が書かれた背景について調べて書かれている，内容を考察している，読み応えのあるレビューも存在する．

- 個人の HP や blog

– レビューにはネタばれを明示しているものも多く，既読者を対象に書かれていることが多い傾向がある．読書家と思われる管理人の考察を含むレビューから，とにかく面白おかしく伝えようとするレビューなど，内容は管理人よって様々であった．文章の長さは，長めである．1 つのページに複数人のレビューを掲載しているサイトもあり，1 ページ＝一人の感想，というわけではない．

レビューを抽出するサイトによって，興味キーワードが異なる可能性が考えられる．抽出されるキーワードに違いが出るかどうかについて，実験を行い確認した．

5.2 実　　験

青空文庫に登録されている作品の中から，10 作品をレビューの対象とする作品として手動で選び，Booklog，Amazon，個人の HP や blog それぞれにおいて，その作品に対するレビューの,「総語数」,「語彙数」,「書籍の本文にも出現する語彙数」,「レビューワー数」,「1 レビューワー（1 ページ）の平均の文数」の 5 項目について調べた．総語数，語彙数には，興味キーワードの候補となる単語のみを数えた．文数の平均を出すにあたって，キーワード候補を一つも含まないレビューはカウントしなかった．実験に用いる，個人の HP や blog は,「書籍名　感想」をキーワードとして Web 検索エンジンを実行し，検索結果の中から，書籍のレビューが含まれているものを，手動で 5 ページ回収した．検索エンジンには google を用いた．

5.3 結　　果

Booklog，Amazon，個人の HP や blog，について調べた結果について，太宰治著『斜陽』の特徴量を表 2 に,10 作品の平均値を表 3 に示す．レビューワーの数は，Booklog，Amazon，個人の HP や blog の順に多く，一人のレビューワーの平均の文数は，個人の HP や blog，Amazon，Booklog の順に多かった．

5.4 レビューサイトの特徴に関する考察

Booklog，Amazon，個人の HP や blog のうちの，どのサイトのレビューを用いるかについて考察する．それぞれの結果を比較してみると，良い結果が出たサイトは，書籍によって異なっており，判断が困難であった.. 例えば，太宰治著の『斜陽』の場合，Booklog においては「遺書」，個人の HP や blog においては「恋」といった，深く読まなくても印象に残る，重要な場面で出現する単語が高い tf-idf 値をとった．一方，Amazon においては，深く読む際に重要な単語の一つである「蛇」という

表 2 『斜陽』の特徴量

書籍名	Booklog	Amazon	個人の HP や Blog
レビューの総語数	10403	3122	1972
レビューの語彙数 (1)	2387	1210	913
書籍の本文にも出現する単語の語彙数 (2)	929	485	556
(2)/(1)*100	38.92	40.083	60.90
レビューワー (サイト) 数	572	45	5
レビューの平均文数	3.5	8.5	49.8

表 3　10 作品の平均値の特徴量

書籍名	Booklog	Amazon	個人の HP や Blog
総語数	8177.8	3233	2377.9
レビューの語彙数 (1)	1633.2	1183.3	932.9
書籍の本文にも出現する単語の語彙数 (2)	633.2	451.2	457.8
(2)/(1)*100	38.13	37.33	46.80
レビューワー (サイト) 数	314.4	45.9	5
レビューの平均文数	3.73	9.08	57.2

単語が高い tf-idf 値をとった.

次に，レビューを収集していた時に気付いたことについて述べる．システムで，個人の HP や blog のレビューを使用する問題点の一つに，自動的なレビューの収集がある．書籍に対する話題には，感想だけでなく，その書籍を用いた授業の指導内容，映像化された作品の感想等，様々な種類がある．Web 全体を検索する時には，書籍の感想以外の話題を拾う可能性が，特定の SNS や通販サイト内から抽出する時よりも高くなる．今回は，書籍の感想以外を話題とするページは，手動で除去したが，自動収集する際には関係のない話題に関する単語まで抽出する恐れがある．今回は，HP と blog のレビューを利用するのは難しいと考えた．それは，それらを自動的に抽出することが困難であるからである.

Booklog，Amazon においては，それぞれの作品を固有の ISBN，もしくは ASIN を指定することにより，明確に同定できる．ISBN は「International Standard Book Number（国際標準図書番号）」の略で，書籍を識別する 10 桁，13 桁の番号であるである。「ASIN」は,「Amazon Standard Identification Number」の略で，Amazon グループが取り扱う，書籍以外の商品を識別する 10 桁の番号であり，雑誌やビデオなど，書籍以外の商品の詳細ページに記載されている [18]．これらの固有の番号を指定することにより，ほぼ確実に書籍に対するレビューを抽出することができる．ここで，完全に抽出できないのは，作品が短編集に収録されている場合は，他の収録作品の感想も混じるからである.

Booklog と Amazon のレビューを混在させて使用する方法も考えられるが，Booklog に含まれるレビューが，純粋に読者の感想が述べられており，レビュー数が多く，ネタばれ防止に対応する可能性があるという点から，本研究では Booklog にあるレビューを対象とする.

6. 興味を喚起する文

文の興味喚起度は，文に出現する興味キーワードの，レビューにおける tf-idf 値を足し合わせることで決定する．また，興味キーワードを複数含む文は，利用者の興味をより喚起しやすい，と考え，出現するキーワード数も掛け合わせた．t_k を文番号 i に出現する興味キーワードとする時，文番号 i の興味喚起度 $Interest(i)$ を以下の式で求める.

$$Interest(i) = \text{文に出現するキーワードの種類} \times \sum_{t_k \ni i} review\text{-}tf\text{-}idf_{t_k} \quad (4)$$

$review\text{-}tf\text{-}idf_{t_k}$ はレビューにおける t_k の $tf\text{-}idf$ 値を表す.

ユーザには $Interest(i)$ が高い順に，文を提示する．しかし，特に高い tf-idf 値を持つ単語を含む文が，複数上位にあがり，提示する文に多様性がない．そこで，$Interest(i)$ の値が一番高い文番号に出現する興味キーワードを，興味キーワードから除外し，再び興味喚起度を求める．この作業を，全ての興味キーワードが除外されるまで繰り返した.

7. 実　　験

本節では，興味キーワードにより，興味を喚起する文を抽出できたかを，被験者実験によって確かめる.

7.1 実験環境

実験に用いる書籍は，書籍本文のデータが青空文庫に存在し，Booklog にレビューがある作品の中から，10 作品を筆者が手動で選択した．開発環境は perl，データベースには MySQL を利用した.

実験に用いる書籍本文として，以下の 10 作品を利用した.

（1） ごん狐　著者：新実南吉
（2） 銀河鉄道の夜　著者：宮沢賢治
（3） 走れメロス　著者：太宰治
（4） 吾輩は猫である　著者：夏目漱石
（5） 蟹工船　著者：小林多喜二
（6） それから　著者：夏目漱石
（7） 斜陽　著者：太宰治
（8） 無人島に生きる十六人　著者：須川邦彦
（9） こころ　著者：夏目漱石
（10） 少女地獄　著者：夢野久作

7.2 実験内容

被験者実験用の web ページを制作し，被験者には，10 作品の中から，評価したい本を自由に選んでもらった．被験者には，選択した書籍中に含まれる文を提示し，それぞれの文に対して「この文に興味をひかれるか」という質問に,「興味をひかれる」,「どちらかと言えばひかれる」,「わからない」,「どちらかと言えばひかれない」,「ひかれない」の五段階で評価してもらった．また，文に出現する興味キーワード集合を提示し,「単語集合に興味をひかれるか」についても 5 段階で評価してもらった.

提示する文は，(タイプ 1)：書籍と Booklog の両方に出現する単語のうち，レビューにおいて高い tf-idf 値をとる上位 15 単

表 4　実 験 結 果

		文に興味をひかれるか		単語集合に興味をひかれるか	
id	書籍名	提案手法 (タイプ 1)	本文のみ (タイプ 2)	提案手法 (タイプ 1)	本文のみ (タイプ 2)
1	ごん狐	0.7222	0.1667	0.9444	0
2	銀河鉄道の夜	0.5	0.1667	0.6667	-0.0556
3	走れメロス	0.5556	1.3889	0.2222	0.7222
4	吾輩は猫である	0.9524	0.5238	0	0.0952
5	蟹工船	1.0556	0.4444	0.6111	-0.3889
6	それから	0.7222	0.6667	-0.0556	-0.3333
7	斜陽	1.1333	0.6	0.1333	-0.6
8	無人島に生きる 16 人	1.3333	0.5333	1.2	0.2667
9	こころ	0.6	1	0.6	0.2667
10	少女地獄	1.2667	1.2	0.7333	0.0667
	平均	0.88413	0.66905	0.50554	0.00397

表 5　被 験 者 数

		書籍を読んだことがあるか			
id	書籍名	ある (内容もよく覚えている)	ある (内容はよく覚えていない)	ない	被験者数
1	ごん狐	4	2	0	6
2	銀河鉄道の夜	0	3	3	6
3	走れメロス	5	0	1	6
4	吾輩は猫である	2	1	4	7
5	蟹工船	0	0	6	6
6	それから	0	2	4	6
7	斜陽	1	0	4	5
8	無人島に生きる 16 人	0	0	5	5
9	こころ	5	0	0	5
10	少女地獄	0	1	4	5
	合計	17	9	31	57

表 6　t 検定における有意水準

	文に興味をひかれるか	単語集合に興味をひかれるか
f 値	0.389215405	0.779519908
p 値	0.199064523	0.011616371

語を興味キーワードとし求めた文，と (タイプ 2)：レビューを利用せず，書籍本文において高い tf-idf 値をとる上位 15 単語をキーワードとし求めた文，の二種類である．(タイプ 2) は提案手法との比較のために用意した．それぞれ，興味喚起度が高い文のうち，上位三文を被験者に提示した．(タイプ 1) と (タイプ 2) において同じ文が提示された場合は，その文を提示する文から除外し，次に高い興味喚起度をとる文を繰り上げて提示した．

文の提示順番によって，結果が異なることも考えられるので，提示する文の順番は，被験者ごとに，プログラムを用いてランダムに並び替えた．

7.3　実 験 結 果

「興味をひかれる」,「どちらかと言えばひかれる」,「わからない」,「どちらかと言えばひかれない」,「ひかれない」の評価を，順に 2 点，1 点，0 点，-1 点，-2 点，と点数に換算して，書籍ごとに平均点を出した．書籍ごとの平均点を表 4 に示す．また，被験者数を表 5 に示す．(タイプ 1) と (タイプ 2) の間に，興味において有意な差があるかを t 検定で調べた結果を表 6 に示す．文の興味喚起度においては，有意な差は見られなかった．単語集合の興味喚起度においては，5 %水準で有意な差がみられた．

7.4　考　　察

被験者実験の結果における「文の値」と「その文に出現する単語集合の値」の差を，5 段階に分類したものを表 7 に示す．文に出現する単語集合の興味喚起度が高ければ，その文の興味喚起度も高いというわけではなかった．

表 7　「文の値」と「単語集合の値」の差

「文の値」−「単語集合の値」	<=-1	-1〜0	0	0〜1	1<=
組み合わせの数	2	9	5	26	18

「文の値」が「単語集合の値」よりも高くなる文の一つに，『斜陽（著者：太宰治)』の『それでね、直治が帰って来て、お母さまと、直治と、かず子と三人あそんで暮していては、叔父さまもその生活費を都合なさるのにたいへんな苦労をしなければならぬから、いまのうちに、かず子のお嫁入りさきを捜すか、または、御奉公のお家を捜すか、どちらかになさい、という、まあ、お言いつけなの」』がある．この文に出現するキーワードは (お母さま, かず子, 直治)，被験者実験による値は，文:0.60, キーワード集合：-0.40 である．この文は，キーワード以外の単語により，興味をひいたと考えられる．

「単語集合の値」が「文の値」よりも高くなる文の一つに『こころ（著者：夏目漱石)』の『いずれにしても先生のいう罪悪という意味は朦朧としてよく解らなかった。』がある．出現する興味キーワードは (罪悪, 先生)，被験者実験による値は，文:0.00, キーワード集合：1.40 である.「先生」と「罪悪」という単語のみを与えられた場合は，様々な場面を想像できたが，文として単語と単語の関連が見えた時に面白みがなくなった，と考えられる．

8.　今後の課題

8.1　興味を引くキーワードの抽出

本手法では，興味を引く単語を出現頻度に基づいて決定し，それを足し合わせることで文の興味喚起度を決定した．しかし，単純に単語の興味をひく程度の値の合計値が，そのまま文の興味をひく程度を反映しているわけではなかった．文は単語と単語の結びつきで構成されており，それを考慮に入れる必要がある．例えば,「ある対象を説明する単語」，と「動作主と動作の受け手，複数の登場人物をつなぐ単語」という，単語による人物間の結びつきである．また，同じ単語でも，組み合わされる単語によって，面白さが変わることも考えられる．今回は，単語の出現頻度に注目した単純な手法で，ある程度の精度を出せたが，今後はより精度を上げるために，キーワードを抽出する際に，文として提示された時にどうなるか，を考慮に入れたキーワードの抽出を行っていきたい．

8.2　レビューにおける書籍本文の引用箇所

レビューワが興味を持った書籍本文の内容を，レビュー内に引用することは，一般的に行われている．今回提案した手法では，レビュー中の引用に対して，特別な処理は行わなかったが，今後，引用部分のデータは増えていくと予想される．引用は，書籍中の興味をひかれた文を，ピンポイントで示すが，引用を行った意図がわからないという欠点もある．提示方法も含め，有効的な活用法を模索していきたい．

8.3　登場人物名に代わるキーワード

語の主要な登場人物の名前は，書籍本文とレビューの両方で tf-idf 値が高くなる傾向がある．これには，大きく 2 種類の理

由が考えられる．一つは，登場人物名が，書籍の本文においては主語や動作の対象となることが多いためである．もう一つは，レビューにおいて，登場人物名は感想を述べるために必要な語として使われているためである．未読の作品の登場人物の名前は，tf-idf 値は高くなるが，それ自体は読者の興味を喚起しないことが多い．例えば，提示される登場人物の名前が「春子」から「花子」に変わっても，利用者が感じる興味の度合いには，それほど影響を与えない．登場人物名を代理のキーワードに置き換える手法を提案したが，今回は被験者実験までは間に合わなかった．付録として，提案手法のみを記す．

8.4 興味を引く文

本研究では，ユーザの興味をひく文を「利用者の興味を喚起する単語を含む文」と定義している．しかし，抽出された文を読み，文の面白さには，比喩等のレトリックも重要な要素だと感じた．また，前文を読まなければ，意味が分かりづらい文も抽出された．レトリックの重みづけ，文章の区切りの良さ等は今後の課題とする．

9. ま と め

本研究では，検索・推薦によって得られた候補の中から，効率的に小説を選択するために，興味を引くと思われる文を提示する立ち読み支援を提案した．「利用者の興味を引く単語を含む文もまた，利用者の興味を引く」と仮定し，興味を喚起する単語を抽出し，それを用いて文の興味喚起度を求める手法を提案した．被験者実験により，提案手法により，興味を喚起するキーワードを抽出できたことを確認した．今後は，単語間の関係やレトリックを考慮に入れて，興味を喚起する文を抽出する手法を考えていきたい．

文　献

[1] http://ebook.itmedia.co.jp/ebook/articles/1012/21/news076.html

[2] 見並史彬，小林幹門，伊藤孝行，“ 概念辞書を利用した目的指向書籍推薦システムの試作 ”，日本ソフトウェア科学会第 24 回大会（2007 年度）論文集

[3] 丸川雄三，阿辺川武，“ 横断的連想検索サービス「想－ IMAGINE」　データベース連携が拓く新たな可能性 ”，情報管理，pp.198-204，Vol.53(2010) , No. 4

[4] http://www.nichigai.co.jp/dcs/index3.html

[5] http://www.kadokawa.co.jp/book/tachiyomi.html

[6] http://books.google.co.jp/bkshp?hl=ja&tab=wp

[7] http://www.aozora.gr.jp/

[8] http://ja.wikipedia.org/wiki/%E9%9D%92%E7%A9%BA%E6%96%87%E5%BA%AB

[9] 大澤幸生，Benson, N.E.，谷内田正彦，“ KeyGraph: 単語共起グラフの分割統合によるキーワード抽出 ”， 電子通信学会誌論文誌 J82-D1 No.2, pp. 391-400， 1999.

[10] 砂山渡，谷内田正彦，“ 文章の特徴を表すキーワードを発見して重要文を抽出する展望台システム ”， 電子情報通信学会論文誌. D-I, 情報・システム，I-情報処理，pp.146-154, 2001

[11] 小倉達矢，宍戸開，“ レビューサイトにおける良質なレビューの特性とそれを考慮した. 評判情報の抽出に関する一考察 ”，DEWS 2008.

[12] 赤木法生，大島裕明，小山聡，田島敬史，田中克己，“ レビューページ例からの属性抽出に基づくレビューページ検索 ”，DEWS2006

[13] 松田千春，“ 情報探索におけるブラウジング行動 : 図書館と書店における行動観察を基にして ”， Library and information science No.49, p.1- 31, 2003

[14] 沢井康孝，山本和英，“ 文書に対する大衆の興味の強さの推定 ”，自然言語処理 = Journal of natural language processing Vol.15, No.2, pp.101-136, 2008-04-10

[15] http://mecab.sourceforge.net/

[16] http://booklog.jp/

[17] http://www.amazon.co.jp/

[18] http://www.amazon.co.jp/gp/help/customer/display.html?ie=UTF8&nodeId=747416

付　　録

1. 登場人物名に代わるキーワード

本手法では，登場人物名は興味を喚起するキーワードになるとは考えられないが，登場人物と密接にかかわる単語は興味を喚起する可能性が高いと考え，それらを登場人物の代替キーワードとして利用する．

なお，今回は，登場人物名の同定は人手で行うことを前提とした．人物名でも，書籍本文において出現頻度が少ない単語は登場人物名としない．また，作中において，名前の代わりに「お母さま」や「先生」等のように代名詞を用いて呼ばれる頻度が高い場合は，それらを登場人物名として扱った．

登場人物名と関連する単語には，物語全体を通して関連する単語と，ある場面においてのみ関連する単語，の二種類が考えられる．今回は，興味キーワードの登場人物名の代わりに用いる，代理キーワードを抽出する手法について述べる．

1.1 物語全体を通した代理キーワード

1.1.1 代理キーワード抽出のアプローチ

書籍の本文において，同じ文に出現する頻度が低くい単語でも，登場人物との関連が高ければ，レビューワの頭の中で結び付き，複数のレビューにおいて同一の文に書かれることがある．レビューにおいて，登場人物名と同一の文に出現する頻度から，単語の重要度を計算する．重要度の高い単語を，物語全体を通しての登場人物像に関連する代理キーワードとして，興味キーワードにおける登場人物名の代わりに用いる．

1.1.2 代理キーワードの抽出手法

代理キーワードの重要度を求める手法について述べる．代理キーワードの候補となる単語は，レビューにおいて登場人物名と同じ文に出現する名詞（代名詞，接尾，数，人名，登場人物名とした単語，を除く），動詞（自立動詞以外を除く）である．我々は，登場人物名と同一の文に出現する単語ほど，関連が高いと考えた．また，キーワードは登場人物を特徴づけるために出現すると考え，登場人物名とキーワード候補の出現数が近い単語の重要度を上げた．キーワード候補 t_i の重要度 $weight(t_i)$ を，以下の式で求める．

$$weight(t_i) = review\text{-}idf_{t_i} \cdot \frac{sentence(t_p \cap t_i)}{1+ \mid t_p - t_i \mid} \qquad \text{(A·1)}$$

t_p は登場人物名，t_i はキーワード候補となる単語，$sentence(t_p \cap t_i)$ は t_p と t_i が同じ文に出現する回数を表す．$review\text{-}idf_{t_i}$ はレビューにおける t_i の idf 値を表す．

1.1.3 代理キーワードの抽出例

太宰治著『斜陽』を例にみてみる．表 A·1 に登場人物である「上原」に対して，提案手法において高い重要度を出した単語の上位 5 単語を示す．なお，正誤は筆者が Booklog のレビューを読んで，登場人物の関連が高いと判断したら「○」，やや関連があると判断したら「△」，関連がないと判断したら「×」とした．ある特定の登場人物だけに関わる単語と，複数の登場人物に関連する単語がある.「上原」の例において,「作家」は上原の職業を表す単語であり，上原だけに関連がある．一方「手紙」は,「かず子」から「上原」に宛てられた物であり，上原とは関係はあるが，動作の受け手なので,「かず子」と「手紙」との関係よりも弱いと考え「△」と評価した．

表 A·1 「上原」と関連の高い単語

順位	正誤	単語	重要度
1	×	母	0.551782
2	○	作家	0.477681
3	△	麻薬	0.342556
4	△	手紙	0.33791
5	×	貴婦人	0.326986

1.1.4 代理キーワード結果の考察

誤回答を抽出した原因として，注目する登場人物と，他の登場人物に対する感想が同一の文に書かれていることがあげられる．精度を上げるためには，文章中での単語の位置や，係り受けを考慮にいれる必要がある．また,「お母さま」（斜陽）において高い重要度をとる単語「幽か」,「さじ」,「叫び声」,「食堂」は，一つ一つを見れば,「お母さま」と関連がない．しかし,『朝、食堂でスウプを一さじ、すっと吸ってお母さまが、「あ」と幽かな叫び声をお挙げになった』という本文をレビュー中に引用している人が複数人いたため，重要度が高くなっている．単語の重要度を個別に求めるのではなく，他の単語との関連を考慮に入れることを，今後の課題としたい．