平 成 3 1 年 度　卒 業 研 究 論 文

題目　電子書籍検索のための漫画特徴タグの Web 抽出に関する研究

指導教員　　服部 峻

提出者　　室蘭工業大学 情報電子工学系学科

氏　名　　村尾　和也

学籍番号　　16024151

提出年月日　　令和 2 年 2 月 13 日

目次

図目次

表目次

第 1 章

まえがき

インターネットの普及に伴って，紙でのみ作られ販売されていた書籍も，紙だけでなくインターネット上で読むことができる電子書籍として販売されることが多くなっている．現在ではスマートフォンの普及もあり，電子書籍は現実の場所を取らずどこでも読めるため多くのユーザに利用されている．

そして漫画の電子書籍を販売しているサイトの中には検索補助や漫画の内容を表すために，タグやキーワードを使ったサービスを提供しているサイトが存在する．しかしながら，これらの従来のタグは「ジャンル」や「ターゲット年齢層」,「メディア化の有無」などを表していることが多く，漫画の具体的なことについては把握できないため，内容の把握や購入するための参考にはあまり活用できないといった問題が現状存在する．

また，多くの電子書籍サイトでは漫画のあらすじの表示や漫画の推薦，試し読みという漫画の冒頭を読むことができるサービスをユーザに提供しているが，あらすじや試し読みでは内容をある程度しか把握することができないため購入の決め手としては弱く，推薦についても新たな漫画の発見には繋がるが推薦された漫画がどのような内容であるかは把握できない．加えて増税や社会への不安から消費も冷え込み，購入という行為について慎重な傾向にあると考えられる．そのため従来以上にユーザ個人の要求に近いものを探すことを可能にすることで購入意欲の促進に繋がる可能性がある．

そこで本研究では，従来の電子書籍販売サイトや出版社目線のタグとは異なり，読者の感想や評価が書かれているレビューを使用することで読者目線から見たその漫画を表す単語を Web 抽出し，各漫画のジャンルだけでなくその割合や，漫画の理解を手助けしてくれる内容を表す特徴語を抽出してその重要度も示すことで，作品についてよく知らないユーザであってもどのような作品か把握できる漫画特徴タグとして生成する手法を提案する．

第 2 章

関連研究

漫画に関する研究はいくつか既に行われている．

山下ら [1] は漫画のレビューから特徴語を抽出し，共起している単語で漫画の関連を示す情報アクセスのデザインを提案している．この研究では関連している漫画は視覚的にわかるようになっているが，どのような特徴語が関連しているか，どの特徴語の関連が強いかはユーザには把握できず，作品についてよく知らないユーザにとってはわかりづらいものとなっている．

また，村瀬ら [2] は利用者の好みのストーリーの作品を推薦するシステムとして，漫画に関する Wikipedia の記事から単語重要度や *cos* 類似度を求め，類似度の高い作品を推薦するシステムを提案している．この研究では同じ作者の作品や 1，2 個程度の特徴の合った作品を推薦することはできているが，多くの特徴の合った作品を推薦することはできてはおらず，また漫画に関する Wikipedia の記事は読者の感想が書かれているような文章ではないため，利用者の好みのストーリーを推薦するという点では不足している．

そこで本研究では，従来研究とは違い，Wikipedia ではなく読者の感想や評価が書かれているレビューを使用することで読者目線から見たその漫画を表す単語を Web 抽出し，また，ただ関連している漫画を表示するだけでなく各漫画のジャンルの割合や，漫画の理解を手助けしてくれる内容を表す特徴語を抽出してその重要度も示すことで，従来研究よりも作品についてよく知らないユーザにおいてもどのような作品か把握できる漫画特徴タグとして生成する手法を提案する．

第 3 章

提案手法

本章では，まず，漫画特徴タグを抽出するシステムの構成を図 3.1 に，また，漫画特徴タグを活用した電子書籍検索システムのイメージを図 3.2 に示す．図 3.1 から，最初に，収集した漫画のレビューをテキスト解析する．次にジャンル判定と類義語と類似語による補強を行いジャンルタグを，特徴語抽出を行い特徴タグを生成して，ジャンルタグと特徴タグから成る漫画特徴タグを生成し，最後に，電子書籍のページに表示するタグとしてユーザに示す．各処理の概要については後述する．

図 3.2 に関しては，まず，作者や出版社などの基本情報が表示されるタグとして基本タグがある．ジャンルタグは従来サイトにおいては 1 から 3 個ほどジャンルの単語が表示されているだけのものを本研究では割合として表すことで，あまり重要ではないが，しかし確かにその作品において作品の一部を構成するジャンルも網羅することができ，よりユーザ個人の好みに合ったジャンルの漫画を選んでもらうことができるタグとなっている．特徴タグは従来では登場人物や作品の特色を表すような単語がタグとして 0 から 3 個ほどであったが，本研究では特徴タグとして 10 個ほど単語を付与するものとし，提案手法の単語重要度のアルゴリズムを用いることで，求めたスコアを基にタグの大きさが変化するタグクラウドとなっている．この特徴タグクラウドにより，より直感的に，重要な特徴タグだけでなく，ニッチな特徴タグも効率的に把握して電子書籍検索に活かすことができる．

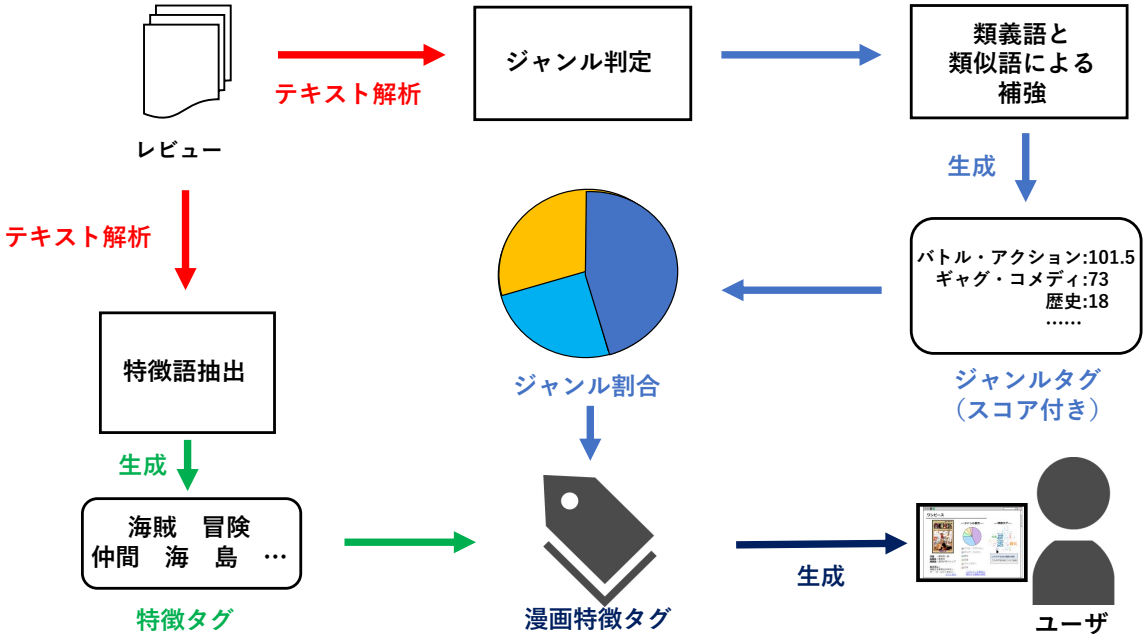


図 3.1　システム構成

図 3.2　システムイメージ

3.1　テキスト解析の概要

テキスト解析に用いるレビューに関しては，ユーザにより漫画レビューが多く書き込まれている「作品データベース」[*1]というサイトにおいて，評価数が 20 以上となる約 700 作品のレビューを収集した．これらの全レビュー集合を R と表し，全作品集合を M と表す．また，$|R| = 53178$，$|M| = 739$ である．

そしてレビューに形態素解析を行う．形態素解析エンジンにはオープンソースの汎用日本語形態素解析エンジン MeCab を，システム辞書には新語・固有表現に強い mecab-ipadic-NEologd を用いる．ただし，「名詞，一般」である単語以外は除去している．理由は電子書籍のタグとして用いられる単語は名詞が多いため，また作品についてよく知らない人にも解り易いタグを目指す上で，漫画特有の造語や登場人物の名前などの固有名詞は提案システムにおいてはノイズとなるためである．

3.2　ジャンル判定の概要

テキスト解析によって得られた結果を基に，ジャンルの判定を行う．ジャンルの判定には電子書籍販売サイトなどにおいてジャンルの単語として用いられる，「バトル」や「ミステリー」などの単語 17 語を用いる．ジャンル単語 17 語については表 3.1 に示す．ここでジャンル単語 17 語に含まれる「バトル」と「アクション」，「ギャグ」と「コメディ」，「ミステリー」と「サスペンス」に関しては区別をつけることが難しいジャンルであり，電子書籍販売サイトにおいても同じジャンルとして扱われることが多いため，それぞれのジャンルを「バトル・アクション」，「ギャグ・コメディ」，「ミステリー・サスペンス」としている，これらの 14 語のジャンル集合を G と表す．

3.3　類義語と類似語による補強の概要

ジャンル判定を補強するためにジャンル単語の類義語，ジャンル単語に対して類似度の高い単語（以下，類似語）を用いる．ここでは，ジャンル単語を Weblio 辞書 [3] から取得した類義語と，*cos* 類似度の高い単語を算出できる Word2Vec より得られた類似語を使用する．ただし，Weblio 辞書から取得した類義語で著者がジャンルの類義語として相応しくないと判断した単語や Word2Vec より得られた単語の中で「ストーリー」や「シーン」など漫画の単語として当たり前に使われる単語，ジャンル単語は除外している．また，Word2Vec で使用するモデルは収集したレビューを学習させ作成したモデルを用いる．

ここで漫画のジャンル単語の類似語の例を表 3.2 に示す．表 3.2 は，ジャンル単語の類似語として Word2Vec で取得した単語の例を示しているが，レビューにおいては辞書に登録されていない単語や，書き間違えられた単語なども存在するため，単語として意味の通らない「ポー

[*1] https://sakuhindb.com/（2019 年 10 月 14 日存在確認）

表 3.1　ジャンル単語 17 語

バトル，アクション，ギャグ，コメディ，ミステリー，サスペンス， 恋愛，ホラー，グルメ，スポーツ，SF，ヤンキー， ラブコメ，歴史，ファンタジー，ビジネス，日常

表 3.2　漫画のジャンル単語の類似語の例

ジャンル単語	恋愛		スポーツ		グルメ
類似語	ラブストーリー		スポーツ漫画		食材
	青春		競技		料理
	三角関係		スポ根		トリコ
	恋愛漫画		野球		細胞
	ラブ		サッカー		美味しく
	恋愛模様		ポーツ		猛獣
	恋愛関係		アメフト		食べる
	コメディー		剣道		美食
	ハーレム		メフト		おいし
	恋		卓球		料理漫画

ツ」や「メフト」といったノイズが混ざってしまっている．ノイズに関しては，辞書に載っている単語と照らし合わせることなどによって除去できる可能性がある．

3.4　特徴語抽出の概要

特徴語抽出では，「高校生」や「王様」などの登場人物の特徴や「海」や「魔法」などの漫画の舞台や特色を表すような単語を抽出し，特徴タグを生成する．また漫画をあまり知らないユーザにも解り易いように登場人物や漫画特有の単語は特徴タグに相応しくないものとして，できる限り除去するものとする．

第 4 章

ジャンル判定のアルゴリズム

本章では各漫画に対して，最も相応しいジャンルを判定したり，含まれるジャンルの割合を求めたりするアルゴリズムについて詳述する．ジャンル $g \in G$ の類義語，類似語も含む単語集合を $W(g)$，作品 m のレビュー集合を $r_i \in R(m)$ とする．また，あるレビュー r_i に対するある単語 $w \in W(g)$ の単語出現頻度である $\mathrm{TF}_{r_i}(w)$，ある作品 m のレビュー集合 $R(m)$ に対する単語 $w \in W(g)$ の文書（レビュー）頻度である $\mathrm{DF}_{R(m)}(w)$，あるレビュー r_i において一番スコアの高かったジャンルを集計して，あるレビュー集合 $R(m)$ において各ジャンルの相応しさを表すスコアを求めるアルゴリズムを RF として，3 種類のアルゴリズムの集合を $X = \{\mathrm{TF}, \mathrm{DF}, \mathrm{RF}\}$ とする．ある作品 m に対する，あるジャンル g の相応しさを表すスコアリングアルゴリズム 3 種類を以下のように定義する．

$$\mathrm{score}_m^{\mathrm{TF}}(g) = \sum_{r_i \in R(m)} \sum_{w \in W(g)} \mathrm{TF}_{r_i}(w)$$

$$\mathrm{score}_m^{\mathrm{DF}}(g) = \sum_{w \in W(g)} \mathrm{DF}_{R(m)}(w)$$

$$\mathrm{score}_m^{\mathrm{RF}}(g) = \sum_{r_i \in R(m)} \mathrm{judge}_{r_i}(g)$$

ここで，レビュー毎に最も相応しいジャンルは以下で決定する．

$$\mathrm{judge}_{r_i}(g) = \begin{cases} 1 & \text{if } \mathrm{score}_{r_i}(g) = \max\{\ \mathrm{score}_{r_i}(g') \mid \forall g' \in G\ \} \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathrm{score}_{r_i}(g') = \sum_{w \in W(g')} \mathrm{TF}_{r_i}(w)$$

また，ジャンル単語の「バトル・アクション」，「ギャグ・コメディ」，「ミステリー・サスペンス」に関しては例外的に構成する 2 つのジャンル単語，例えば「バトル・アクション」のスコアは「バトル」のスコアと「アクション」のスコアを足して 2 で割った数をスコアとし，これを「ギャグ・コメディ」，「ミステリー・サスペンス」に対しても同様の処理を行う．そして，3 種類のアルゴリズム TF，DF，RF において各作品 m に対して最も相応しいジャンル g は

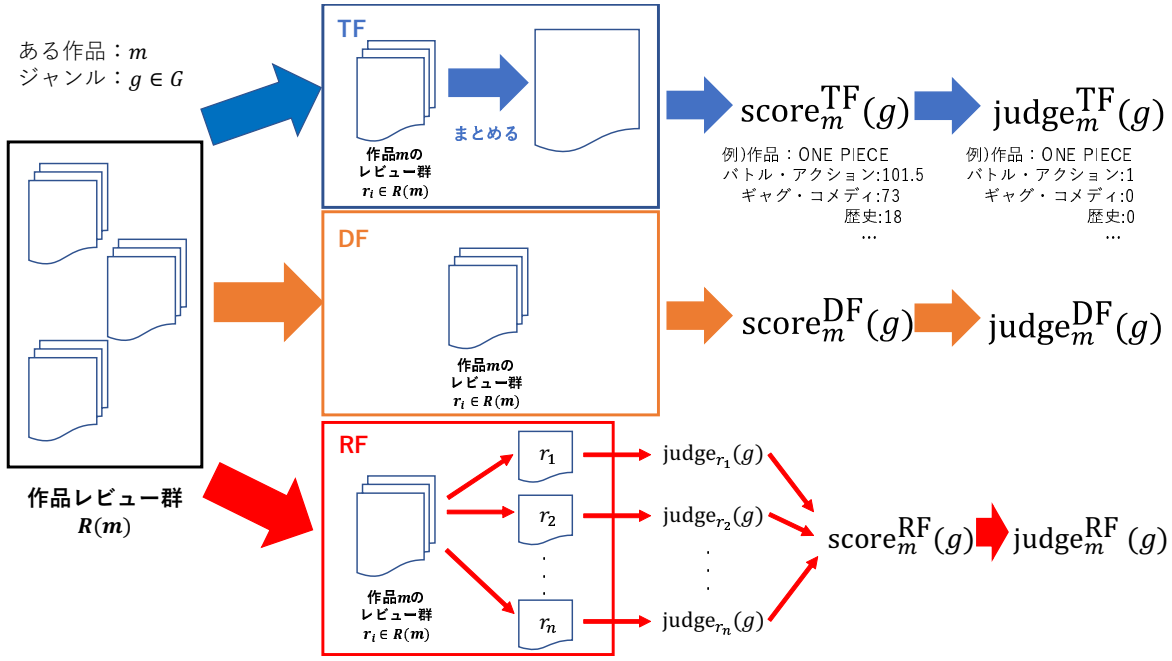


図 4.1　3 種類のジャンル判定アルゴリズム

以下の式によって決定される．また 3 種類のジャンル判定アルゴリズムのイメージを図 4.1 に示す．

$$\mathrm{judge}_m^X(g) = \begin{cases} 1 & \text{if } \mathrm{score}_m^X(g) = \max\{\ \mathrm{score}_m^X(g') \mid \forall g' \in G\ \} \\ 0 & \text{otherwise.} \end{cases}$$

そして，3 種類のアルゴリズム TF，DF，RF のそれぞれのスコア関数において，各ジャンル単語，類義語，類似語を含む単語集合 $W(g)$ を求める手法を以下の 6 種類用意した．

- 手法 $W1$：ジャンル単語のみ
- 手法 $W2$：ジャンル単語と類義語
- 手法 $W3$：ジャンル単語と Wikipedia のモデルを使用した Word2Vec より得られた類似語
- 手法 $W4$：ジャンル単語と漫画のレビューのモデルを使用した Word2Vec より得られた類似語
- 手法 $W5$：ジャンル単語と類義語，Wikipedia のモデルを使用した Word2Vec より得られた類似語
- 手法 $W6$：ジャンル単語と類義語，漫画のレビューのモデルを使用した Word2Vec より得られた類似語

ここで，Word2Vec で用いるモデルとして，手法 $W3$，$W5$ の Wikipedia のモデルは公開されている Wikipedia のモデル [4] を使用し，手法 $W4$，$W6$ の漫画のレビューのモデルは 3.1 節で収集したレビューから作成したモデルである．

第 5 章

特徴語抽出のアルゴリズム

本章では各漫画に対して，特徴語を抽出するためのアルゴリズムについて詳述する．作品 m のレビュー集合を $r_i \in R(m)$，また，あるレビュー $r_i \in R(m)$ に対するある単語 w の単語出現頻度である $\mathrm{TF}_{r_i}(w)$，レビュー集合 R，$R(m)$ に対するある単語 w の文書（レビュー）頻度である $\mathrm{DF}_R(w)$，$\mathrm{DF}_{R(m)}(w)$，各作品のレビューにある単語 w を含む作品頻度である $\mathrm{MF}(w)$ とすると，ある単語 w のある作品 m に対する特徴語らしさを測る尺度を以下の 4 種類定義する．また特徴語抽出アルゴリズムのイメージを図 5.1 に示す．

$$
\begin{aligned}
\text{TF-IDF}_m(w) &= \mathrm{TF}_m(w) \cdot \mathrm{IDF}(w) \\
\text{DF-IDF}_m(w) &= \mathrm{DF}_m(w) \cdot \mathrm{IDF}(w) \\
\text{TF-IMF}_m(w) &= \mathrm{TF}_m(w) \cdot \mathrm{IMF}(w) \\
\text{DF-IMF}_m(w) &= \mathrm{DF}_m(w) \cdot \mathrm{IMF}(w)
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{TF}_m(w) &= \frac{\sum_{r_i \in R(m)} \mathrm{TF}_{r_i}(w)}{\sum_{\forall w'} \sum_{r_i \in R(m)} \mathrm{TF}_{r_i}(w')} \in [0,1] \\
\mathrm{DF}_m(w) &= \frac{\mathrm{DF}_{R(m)}(w)}{\sum_{\forall w'} \mathrm{DF}_{R(m)}(w')} \in [0,1] \\
\mathrm{IDF}(w) &= \log_2 \frac{|R|}{\mathrm{DF}_R(w)+1} \\
\mathrm{IMF}(w) &= \log_2 \frac{|M|}{\mathrm{MF}(w)+1}
\end{aligned}
$$

これらは単語重要度を求める通常の TF-IDF 法をベースに改良したものであり，通常はレビュー 1 つ 1 つに対して単語出現頻度 $\mathrm{TF}_{r_i}(w)$ を算出するだけであるが，本研究では作品のレビュー群をまとめて 1 つの文書として見ており，この文書に対して $\mathrm{TF}_m(w)$ を求めている．また全レビュー集合における文書頻度 $\mathrm{DF}_R(w)$ ではなく，全作品集合における作品頻度 $\mathrm{MF}(w)$ を用いる狙いとして，どの作品にも出現するような単語の重要度を下げることや，作品毎に頻度を算出し，作品数でフィルターを設けることでほとんどの作品では出現しないような単語の除去を容易にしている．理由として，著者が考える特徴タグでは登場人物や漫画固有の単語はノイズとしているためである．

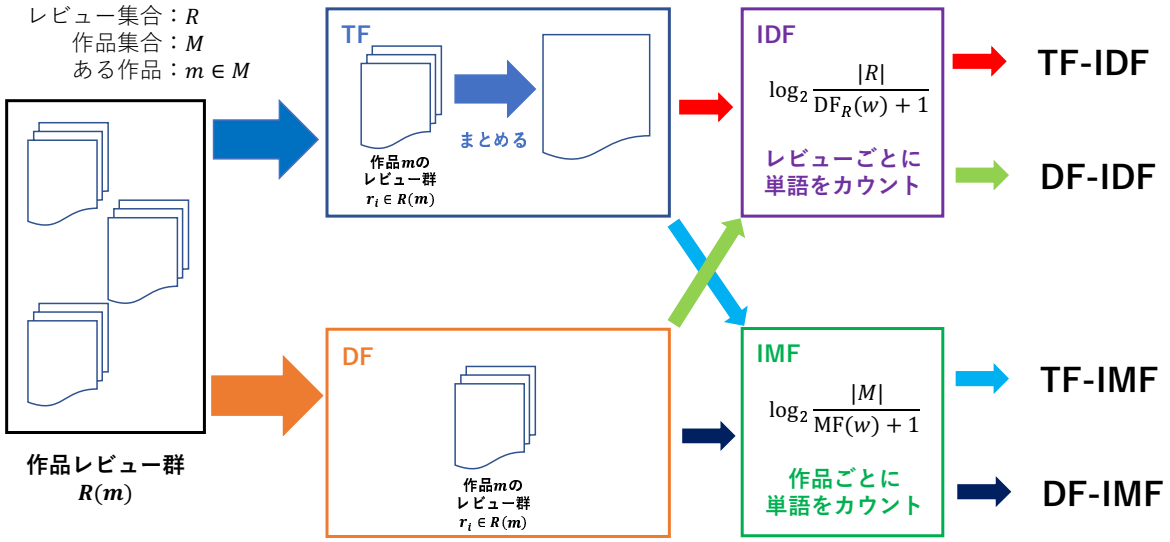


図 5.1　4 種類の特徴語抽出アルゴリズム

第 6 章

ジャンル判定の評価実験

本章では，提案した漫画作品のジャンル判定の精度評価として，各ジャンル g に関する単語集合 $W(g)$ を 6 種類に変化させ，ジャンル単語のみでジャンル判定を行った場合やジャンル単語の類義語，類似語も使った場合などの比較実験，及び，ジャンル単語のみでジャンルタグを生成した場合とジャンル単語に加え，類義語，類似語も加えてジャンルタグを生成した場合などの比較実験の 2 種類の評価実験を行った結果を示す．

6.1　電子書籍サイトによる正解セットを用いた精度評価

6.1.1　実験概要

4 章で示した 3 種類のジャンル判定アルゴリズム TF，DF，RF に対して，各ジャンル単語，類義語，類似語から成る単語集合 $W(g)$ を求める手法 6 種類を変化させたジャンル判定結果と，正解セットを比較した精度評価を示す．また事前実験として手法 $W3$ から $W6$ において類似語の取得数について類似度の高い上位 1 から 50 単語までを取得し，各手法，各アルゴリズムにおいて最も正解率の高い中で類似語取得数が 3 種類のスコアリングアルゴリズムにおいて同数であればその類似語取得数を実験の比較対象とし，最も正解率の高い中で類似語取得数が同数の類似語取得数が無ければ，最も正解率の高い中で最小の類似語取得数を実験の比較対象とした．手法 $W3$ から $W6$ における 100 作品での類似語取得数に依る正解率の変化を図 6.1 から図 6.4 に示す．また各手法，各アルゴリズムにおける最も正解率の高い類似語取得数を表 6.1 に示す．ここで，同数の類似語取得数を選択するのは比較を解り易くするためである．また，最小のものを選択するのは類似語取得数は増えるほどジャンル単語と cos 類似度の低い単語が取得され，ノイズになる危険性も考慮したためである．

6.1.2　電子書籍サイトによる正解セットの作成

正解セットとして，現在電子書籍を販売しているサイト 6 箇所 [5–10] から漫画 100 作品のタグ，キーワードを取得して，ジャンル単語と一致する単語をカウントし，最も多かった 1 つまたは複数個のジャンルを正解とした．ここで「バトル」と「アクション」,「ギャグ」と「コ

メディ」,「ミステリー」と「サスペンス」については個別にカウントした上で各 2 つのジャンルの電子書籍販売サイトのタグとして付与されている数の多い方をそれぞれ「バトル・アクション」,「ギャグ・コメディ」,「ミステリー・サスペンス」のカウント数とした.

6.1.3　実験結果

図 6.1 から図 6.4 と表 6.1 より，Wikipedia のモデルを使用した Word2Vec より得られた類似語を用いている手法 $W3$ と手法 $W5$ では，類似語取得数 1 または 2 が最も精度が高くそれ以降低くなっていき，図 6.5 より，最も正解率の高い結果は類似語を取得していない手法 $W1$，手法 $W2$ の結果とほとんど変わっていないことから，Wikipedia のモデルを使用した Word2Vec より得られた類似語はジャンル判定においては相応しい類似語を取得できないということがわかる．また手法 $W4$，手法 $W6$ において最も正解率の高い時の類似語取得数が 13 から 21 の範囲にあることから十数語程度の類似語を用いることがジャンル判定の精度上昇に繋がるということがわかる．

図 6.5 より，手法 $W6$ のジャンル単語と類義語，漫画のレビューのモデルを使用した Word2Vec より得られた類似語でカウントする手法が最も正解率が高い結果となった．漫画のレビューをモデルとして使用し，Word2Vec から類似語を取得することで各ジャンルの判定に優位に機能したと考えられる．またジャンル単語と類義語とを用いた手法 $W2$ とジャンル単語とレビューから作成した Word2Vec により得られた類似語とを用いた手法 $W4$ の結果から，ジャンル判定を行う上で漫画のレビューから作成したモデルを使用した Word2Vec より得られた類似語の方が類義語よりもレビューからジャンル判定を行う上で相応しいということがわかるが，手法 $W3$ では類似語の取得数が増加していくに従って正解率が減少していくことから類似語の取得には使用するモデルが重要であるということがわかる．そしてジャンル単語とレビューから作成したモデルを使用した Word2Vec により得られた類似語とを用いた手法 $W4$ とそれに類義語が加わった手法 $W6$ の結果から，類義語を用いることも正解率の上昇に繋がることがわかる．しかし，スコアリングアルゴリズム TF において，手法 $W4$ と手法 $W6$ の正解率を比較した時に，手法 $W4$ の方が僅かに上回っている．このようになった理由としては，類義語を用いないことで最も正解率の高い類似語取得数に変化が生じ，その類似語取得数で正解となる作品の組み合わせに変化が生じたためであると考えられる．

そして，TF，DF，RF のスコアリングアルゴリズムで比較するとほとんど差が生まれない結果となり，また正解率も 70% に届かない結果となった．正解率が上がらない理由の 1 つとして考えられるものとして，3 種類のスコアリングアルゴリズムが漫画のレビューからジャンル判定を行うに当たり，適していない可能性が考えられる．そのため 3 種類のスコアリングアルゴルズムがジャンル判定に適しているかどうかの見直しや，類似語の *cos* 類似度の値であったり，TF-IDF 法の単語重要度を求める手法のように重みを付けられる手法によって，単語に対する重みを付けてジャンル判定の指標の 1 つとしてスコアリングに用いる手法の検討などが必要であると考えられる．また本研究では 17 単語をジャンル単語としてジャンル判定を行っているが，それ以外にもジャンル単語として相応しい単語が存在し，その単語を用いることで

ジャンル判定の精度が向上する可能性も考えられる．そして，正解セットとした電子書籍のタグが読者（被験者）の感じるものと合致していないのではないかという懸念も挙げられる．そこで，本研究では読者目線のタグを生成することを 1 つの目標として掲げているため，次節で正解セットを新たに用意し，再度実験を行った．

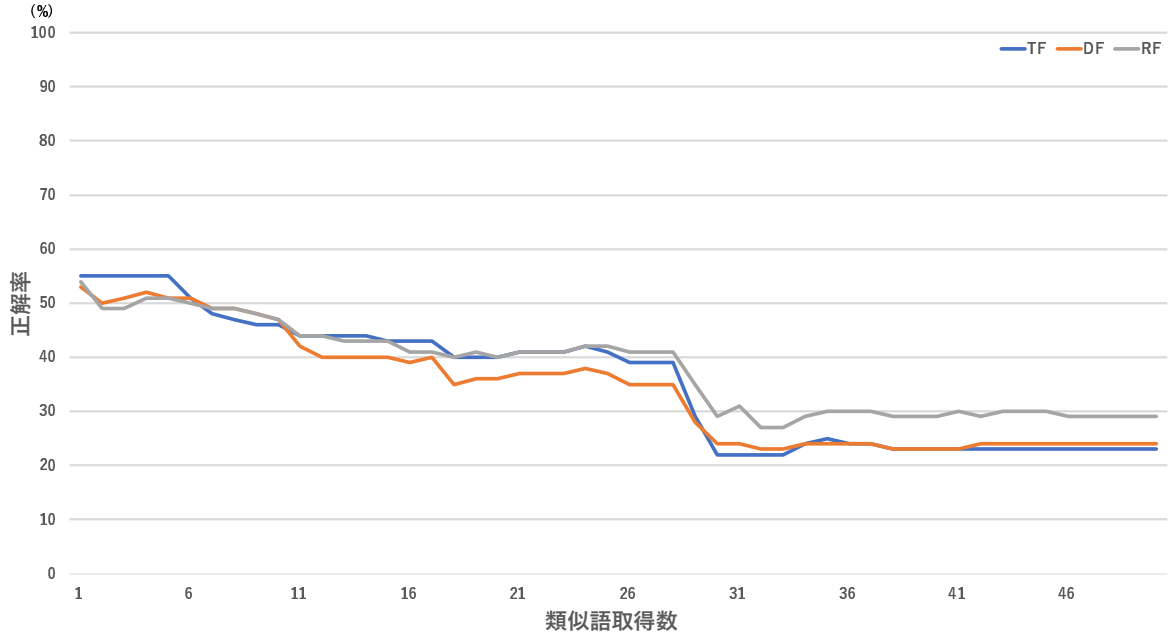


図 6.1　手法 $W3$ での類似語取得数に依る正解率変化 ($N = 100$)

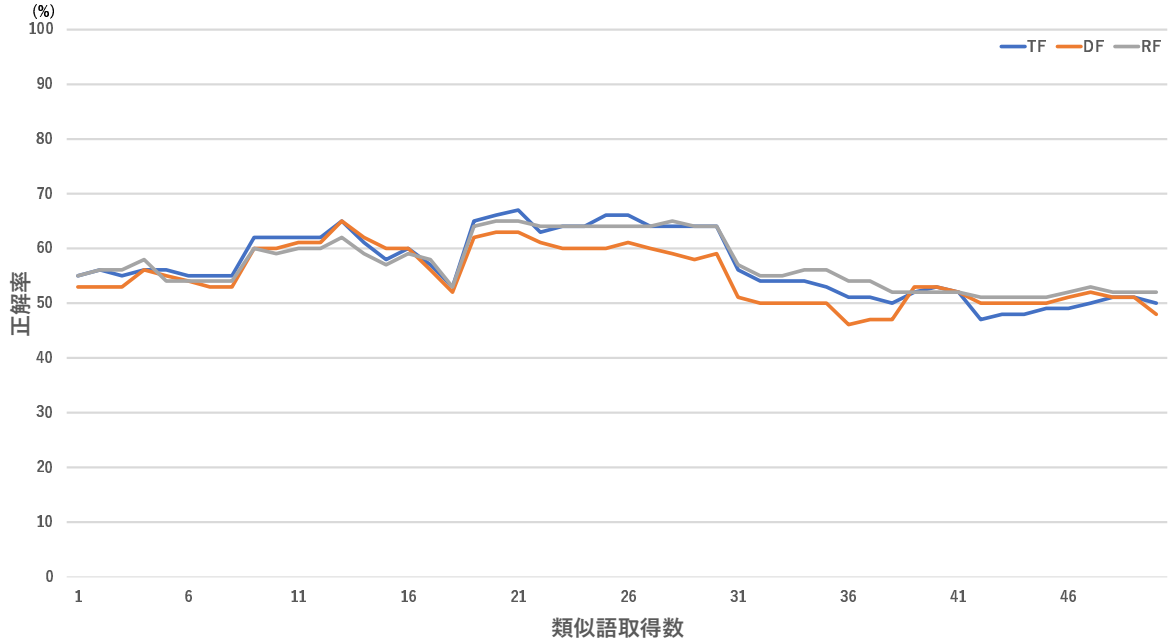


図 6.2　手法 $W4$ での類似語取得数に依る正解率変化 ($N = 100$)

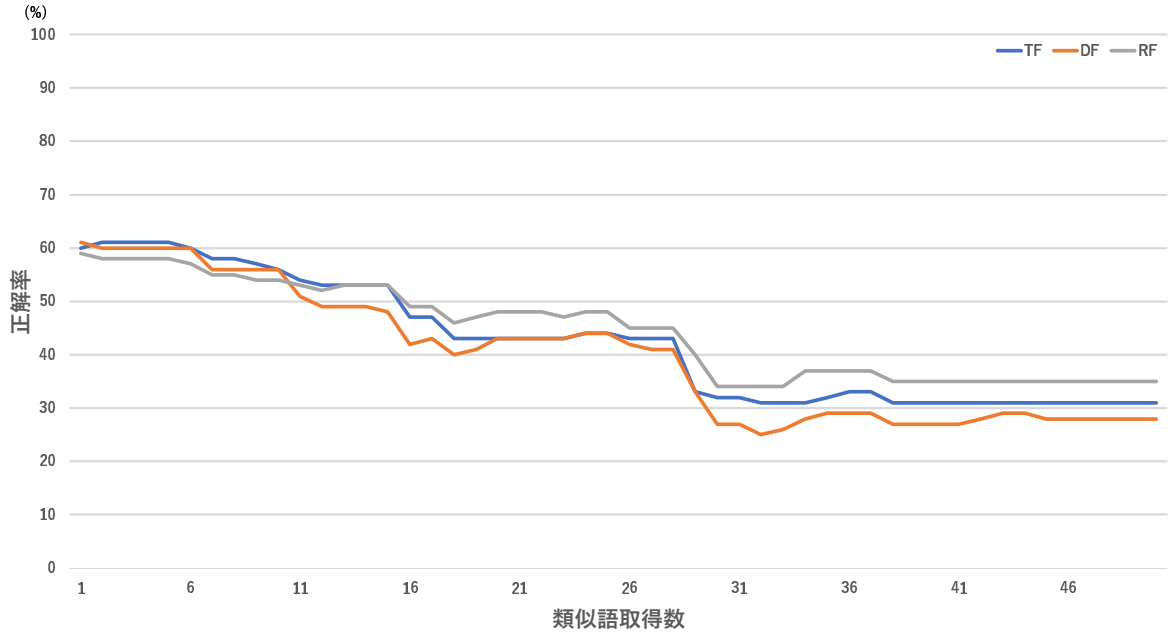


図 6.3　手法 $W5$ での類似語取得数に依る正解率変化 ($N = 100$)

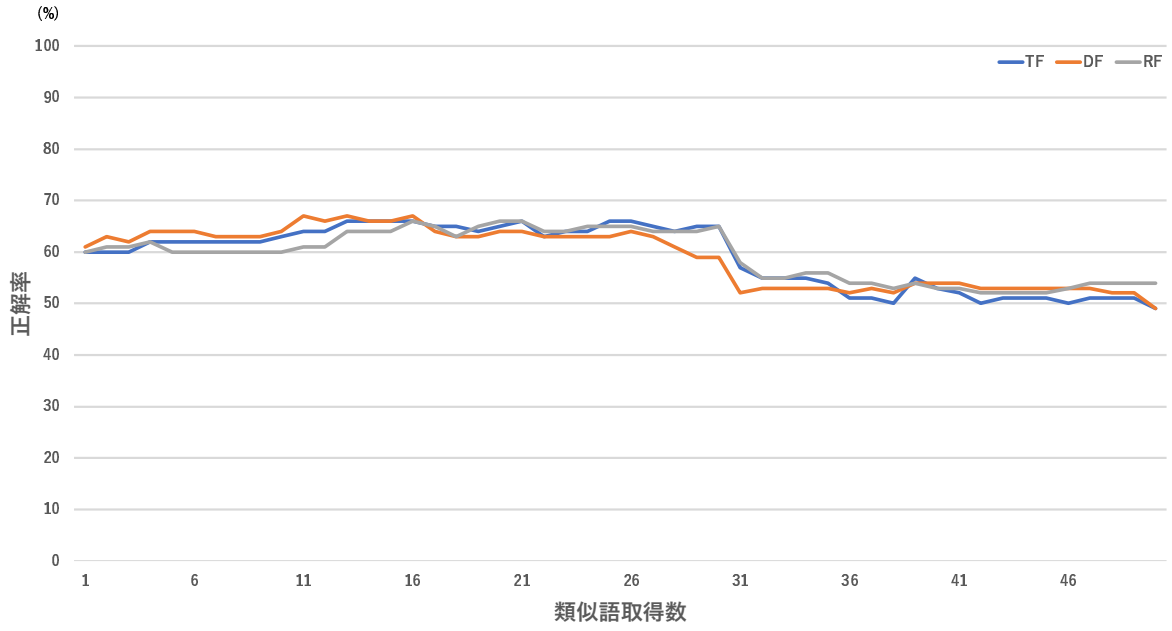


図 6.4　手法 $W6$ での類似語取得数に依る正解率変化 ($N = 100$)

表 6.1　3 種類のアルゴリズムの手法 $W3$ から $W6$ における最も正解率の高い類似語取得数

手法	$W3$	$W4$	$W5$	$W6$
TF	1*	21	2	16
DF	1*	13	1*	16
RF	1**	20	1*	16

*：類似語取得数 0 と正解率が同じ

**：類似語取得数 0 より正解率が低い

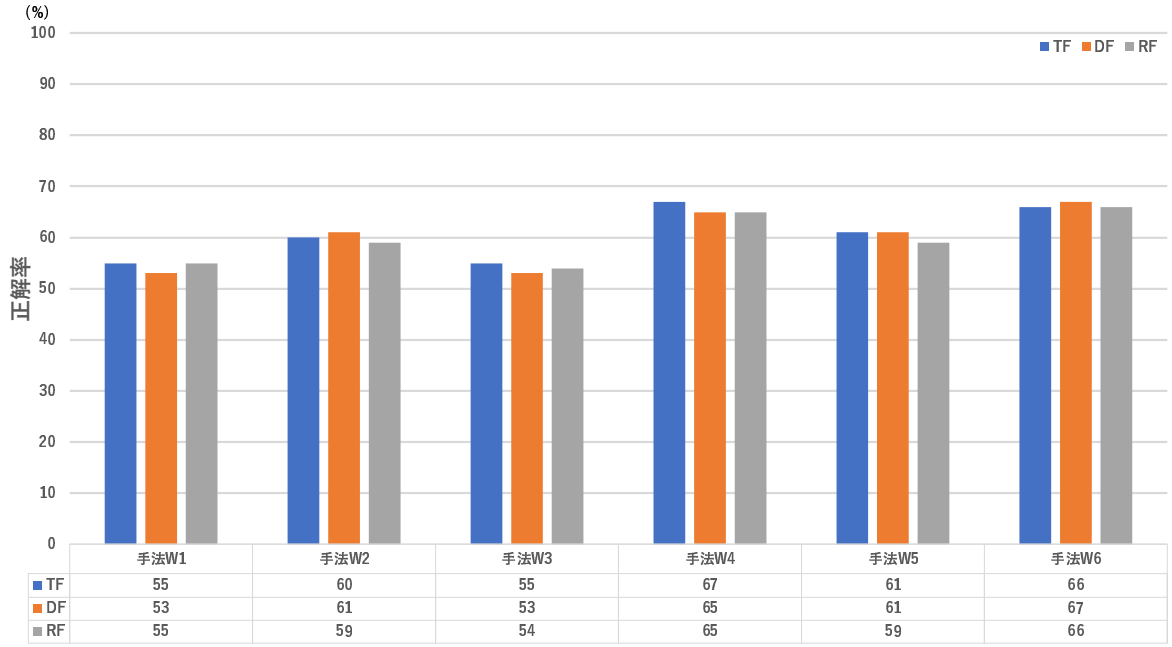


図 6.5　ジャンル判定の正解率の比較結果 ($N = 100$)

6.2　人による正解セットを用いた精度評価

6.2.1　実験概要

漫画に相応しいジャンルに関して被験者に質問を行い，その中で，漫画に対して何らかジャンルが回答された回答率が 50% 以上であった 74 作品と，回答率が 80% 以上であった 54 作品について 6.1 節と同様のアルゴリズム，手法で各々の正解率を比較した．

また，6.1 節と同様に，事前実験として手法 $W3$ から $W6$ において類似語の取得数について類似度の高い上位 1 から 50 単語までを取得し，各アルゴリズム，各手法において最も正解率の高い中で類似語取得数が 3 種類のスコアリングアルゴリズムにおいて同数であればその類似語取得数を実験の比較対象とし，最も正解率の高い中で類似語取得数が同数の類似語取得数が無ければ，最も正解率の高い中で最小の類似語取得数を実験の比較対象とした．手法 $W3$ から $W6$ における類似語取得数に依る正解率変化を回答率 50% 以上の作品，74 作品で求めた結果を図 6.6 から図 6.9 に，回答率 80% 以上の作品，54 作品で求めた結果を図 6.10 から図 6.13 に示す．また各アルゴリズム，各手法における最も正解率の高い類似語取得数をそれぞれの回答率に対して集計したものを，表 6.2，表 6.3 に示す．

6.2.2　人による正解セットの作成

著者を含む男性 11 名で 100 作品に対して，各漫画タイトルで最も相応しいと思うジャンルをジャンル単語 17 個の中から 1 つ，または，その漫画タイトルを知らない場合は「わからない」を選んでもらい，最も割合の多かった 1 つまたは複数のジャンルを各漫画の正解とした．また 6.1.2 項と同様に，「バトル」と「アクション」，「ギャグ」と「コメディ」，「ミステリー」と「サスペンス」については個別にカウントした上で各 2 つのジャンルの選択された数の多い方をそれぞれ「バトル・アクション」，「ギャグ・コメディ」，「ミステリー・サスペンス」のカウント数とした．

6.2.3　実験結果

図 6.6 から図 6.13 と表 6.2，表 6.3 より，どちらの回答率の結果においても，Wikipedia のモデルを使用した Word2Vec より得られた類似語を用いている手法 $W3$ と手法 $W5$ では，類似語取得数 1 で最も精度が高くそれ以降低くなっていき，図 6.14，図 6.15 より，最も正解率の高い結果は類似語を取得していない手法 $W1$，手法 $W2$ の結果とほとんど変わっていないことから，Wikipedia のモデルを使用した Word2Vec より得られた類似語はジャンル判定においては相応しい類似語を取得できないということがわかる．また手法 $W4$，手法 $W6$ において最も正解率の高い時の類似語取得数が 11 から 28 の範囲にあることから十数語程度の類似語を用いることがジャンル判定の精度上昇に繋がるということがわかる．

図 6.14，図 6.15 より，最も正解率が高い手法は 6.1 節での評価実験と同様に手法 $W6$ の

ジャンル単語と類義語，漫画のレビューのモデルを使用した Word2Vec より得られた類似語でカウントする手法となった．また 6.1.3 項と同様に，ジャンル単語と類義語とを用いた手法 $W2$ とジャンル単語とレビューから作成した Word2Vec により得られた類似語とを用いた手法 $W4$ の結果から，ジャンル判定を行う上で漫画のレビューから作成したモデルを使用した Word2Vec より得られた類似語の方が類義語よりもレビューからジャンル判定を行う上で相応しいということがわかる．そしてジャンル単語とレビューから作成したモデルを使用した Word2Vec により得られた類似語とを用いた手法 $W4$ とそれに類義語が加わった手法 $W6$ の結果から，類義語を用いることも正解率の上昇に繋がることがわかる．この結果から，ジャンルを判定することに関して類義語や類似語を使用することは一定の効果があることがわかる．

また，3 種類のスコアリングアルゴリズム TF，DF，RF に関しては大きな変化を得られる結果には至らなかったが，6.1 節と比べて正解率が全体的に上昇していることから，レビュー分析による提案手法は，漫画に対してユーザの感じるジャンルを判定することには成功しているものと考えられる．

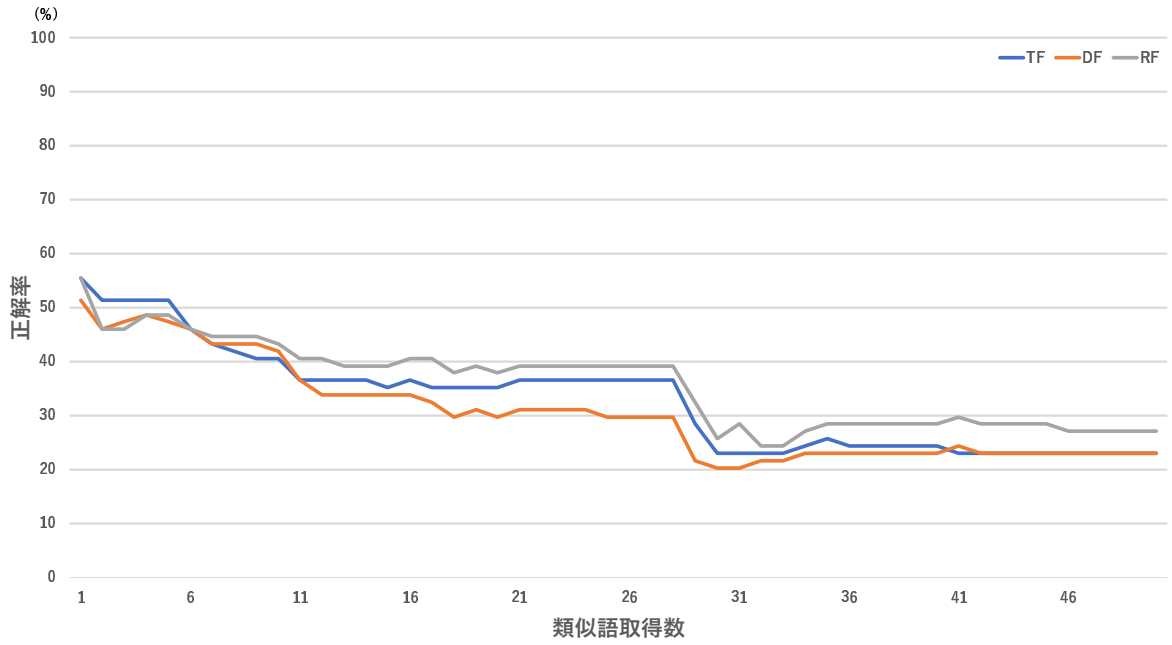


図 6.6　回答率 50% 以上における手法 $W3$ での類似語取得数に依る正解率変化 ($N = 74$)

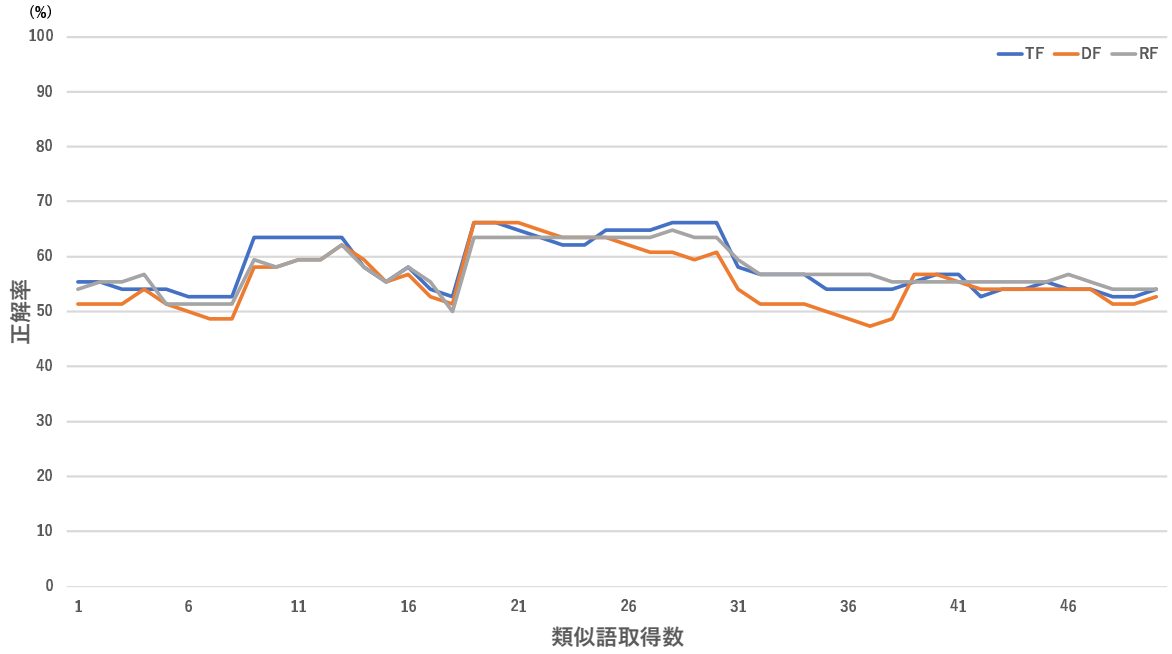


図 6.7　回答率 50% 以上における手法 $W4$ での類似語取得数に依る正解率変化 ($N = 74$)

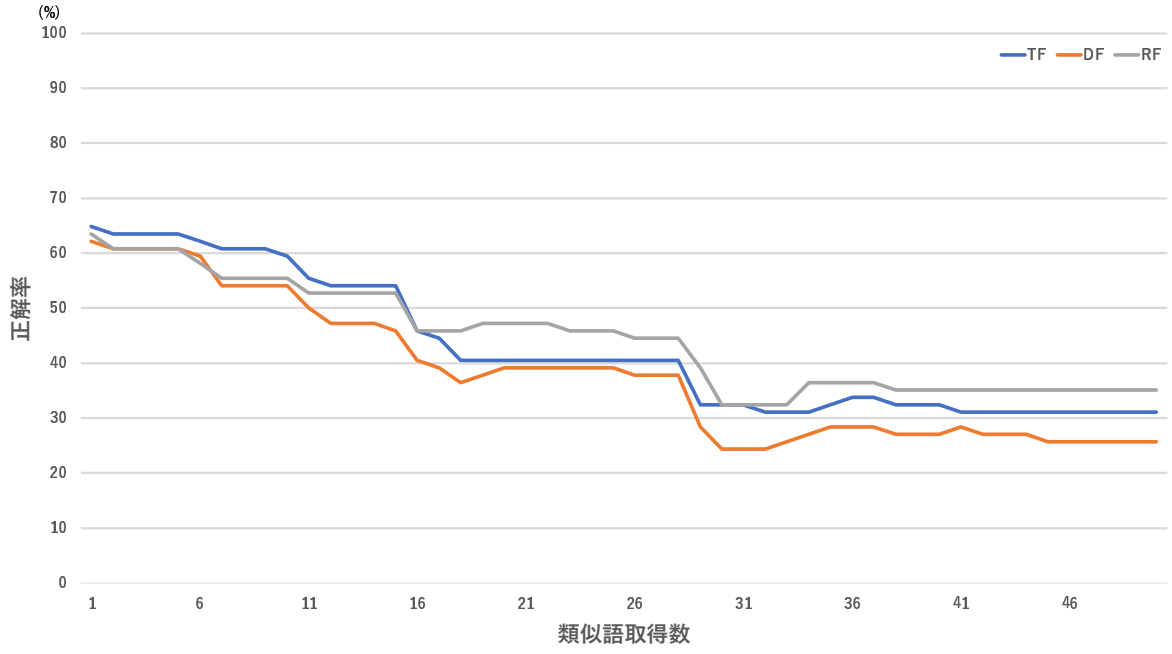


図 6.8　回答率 50% 以上における手法 $W5$ での類似語取得数に依る正解率変化 ($N = 74$)

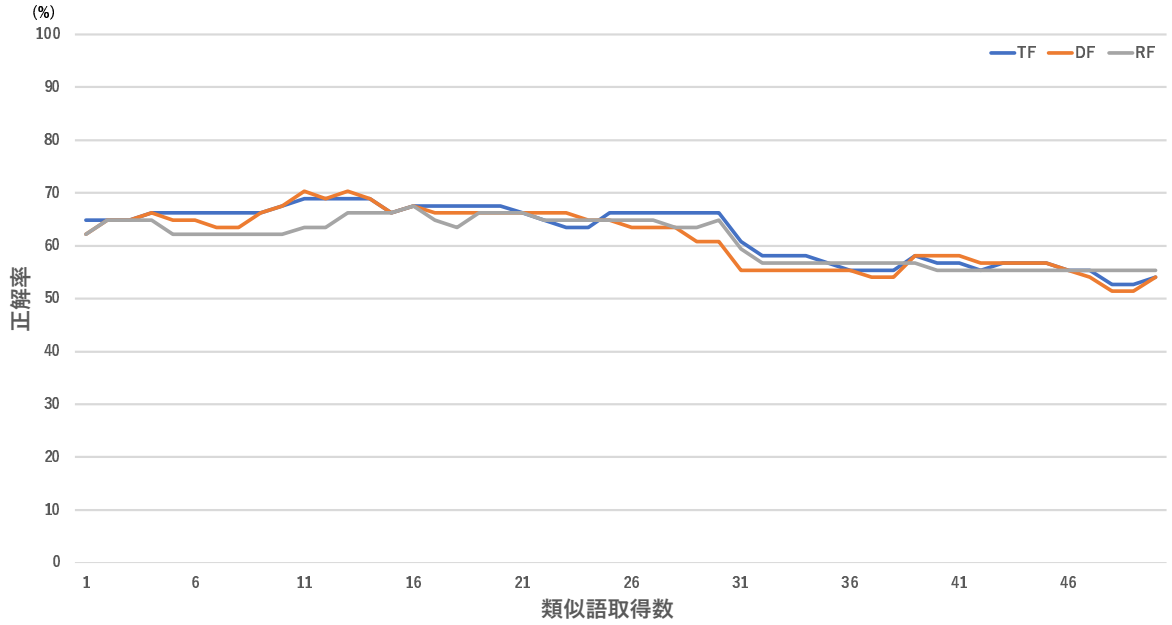


図 6.9　回答率 50% 以上における手法 $W6$ での類似語取得数に依る正解率変化 ($N = 74$)

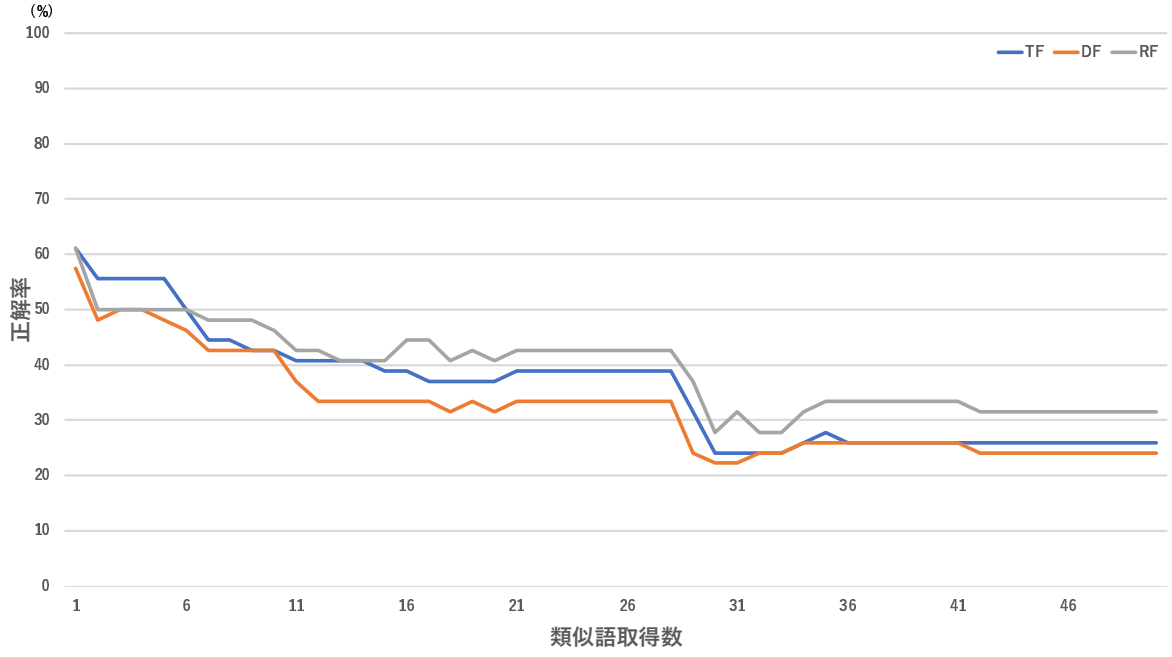


図 6.10　回答率 80% 以上における手法 $W3$ での類似語取得数に依る正解率変化 ($N = 54$)

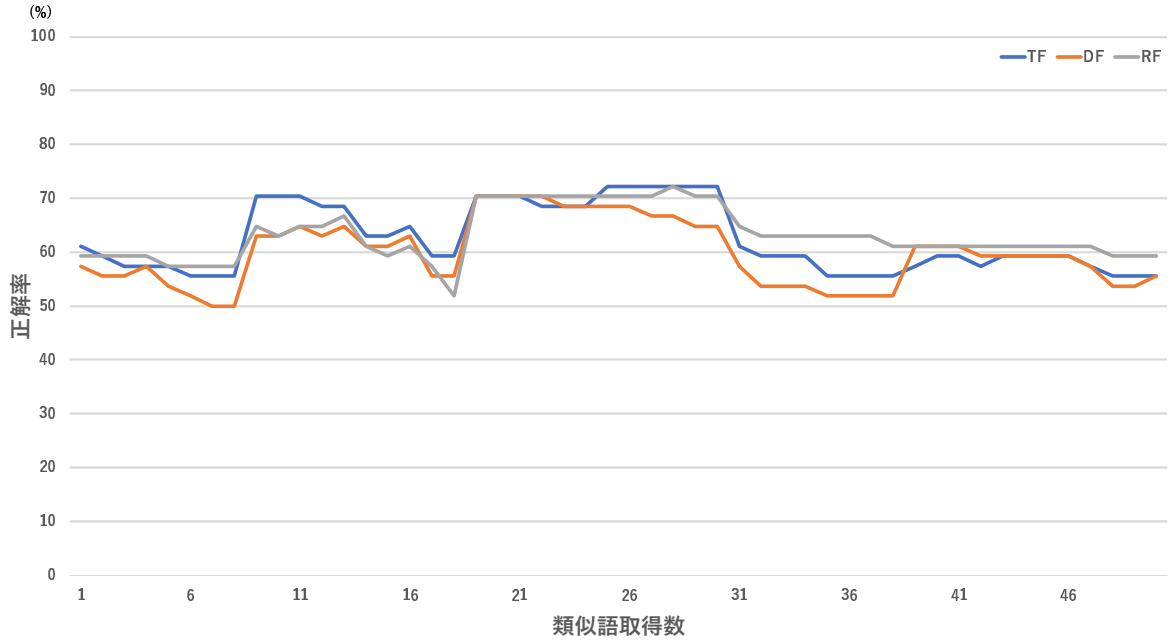


図 6.11　回答率 80% 以上における手法 $W4$ での類似語取得数に依る正解率変化 ($N = 54$)

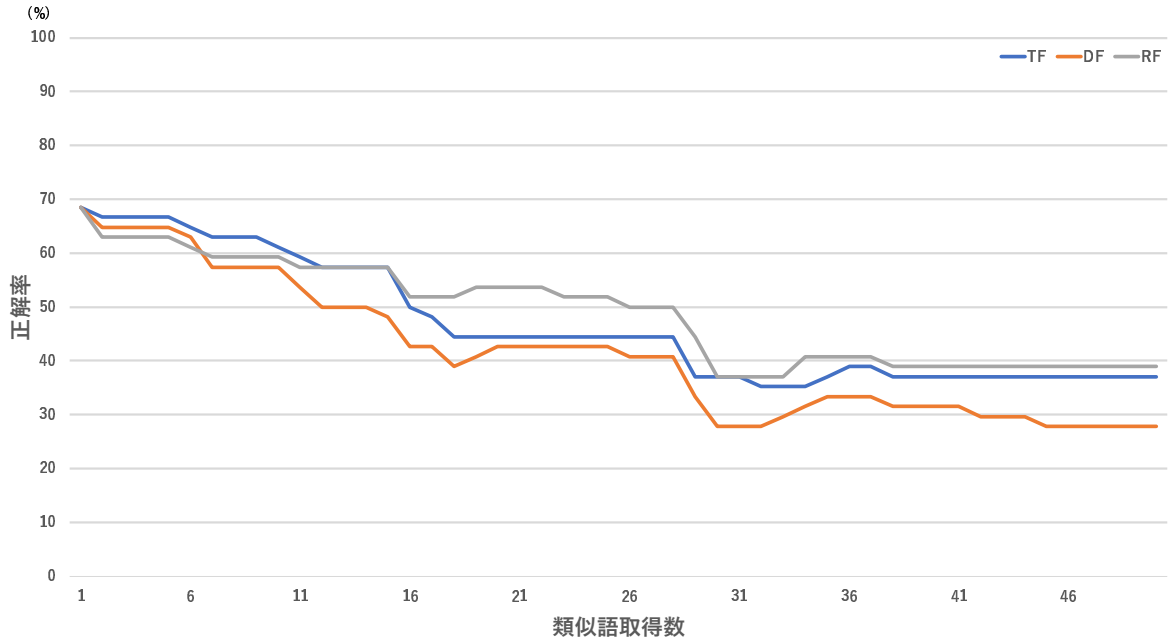


図 6.12　回答率 80% 以上における手法 $W5$ での類似語取得数に依る正解率変化 ($N = 54$)

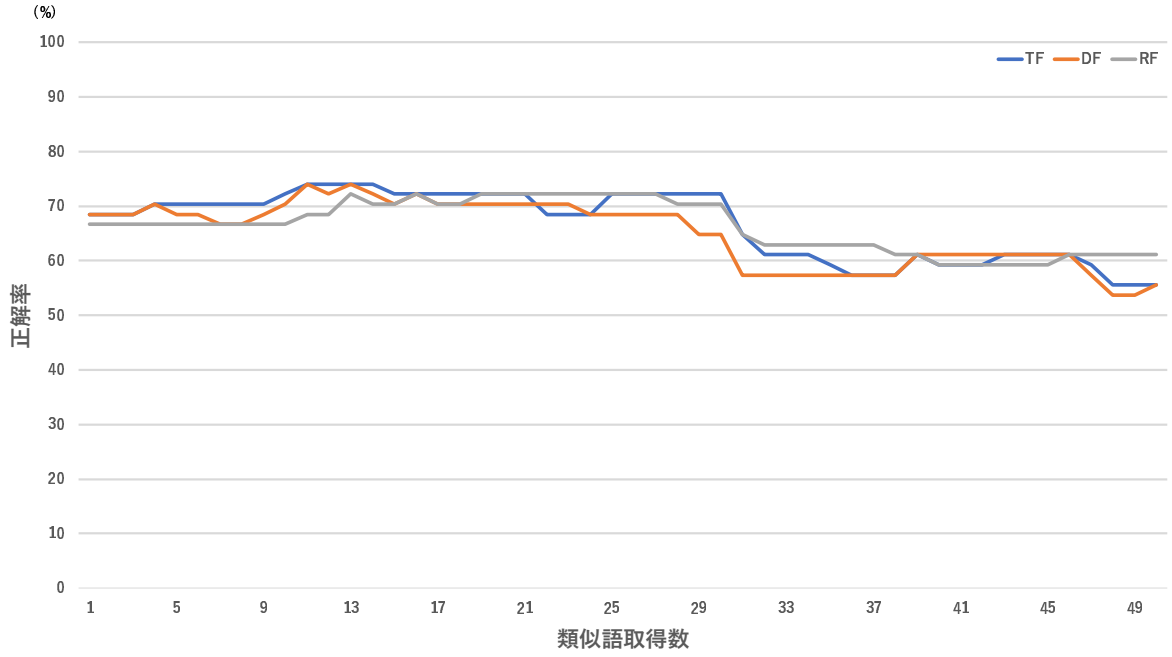


図 6.13　回答率 80% 以上における手法 $W6$ での類似語取得数に依る正解率変化 ($N = 54$)

表 6.2　回答率 50% での 3 種類のアルゴリズムの手法 $W3$ から $W6$ における最も正解率の高い類似語取得数

手法	$W3$	$W4$	$W5$	$W6$
TF	1*	19	1*	11
DF	1*	19	1*	11
RF	1**	28	1*	16

*：類似語取得数 0 と正解率が同じ

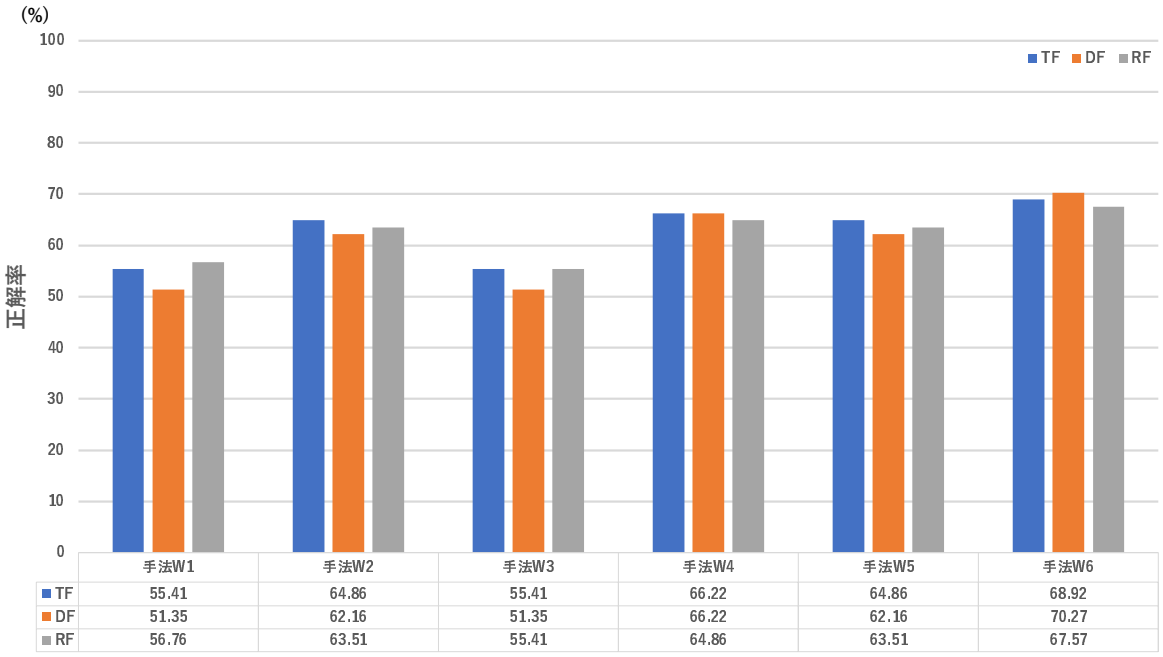
**：類似語取得数 0 より正解率が低い

表 6.3　回答率 80% での 3 種類のアルゴリズムの手法 $W3$ から $W6$ における最も正解率の高い類似語取得数

手法	$W3$	$W4$	$W5$	$W6$
TF	1*	25	1*	13
DF	1*	19	1*	13
RF	1**	28	1*	13

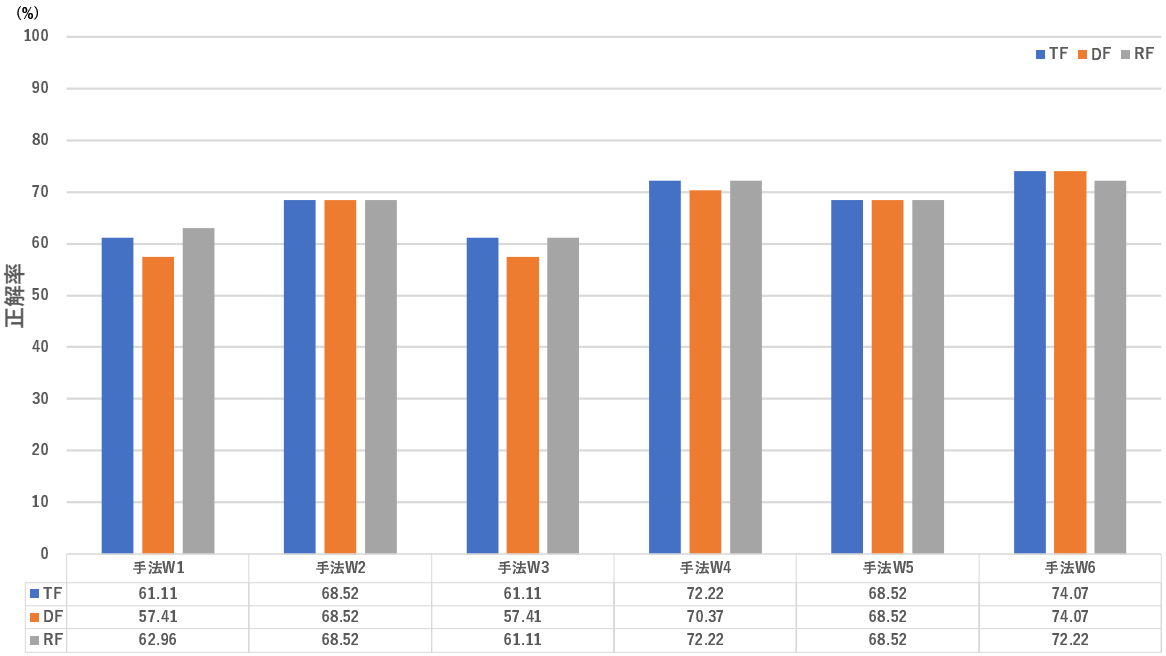
*：類似語取得数 0 と正解率が同じ

**：類似語取得数 0 より正解率が低い



	手法W1	手法W2	手法W3	手法W4	手法W5	手法W6
TF	55.41	64.86	55.41	66.22	64.86	68.92
DF	51.35	62.16	51.35	66.22	62.16	70.27
RF	56.76	63.51	55.41	64.86	63.51	67.57

図 6.14　回答率 50% 以上でのジャンル判定の比較結果 ($N = 74$)



	手法W1	手法W2	手法W3	手法W4	手法W5	手法W6
TF	61.11	68.52	61.11	72.22	68.52	74.07
DF	57.41	68.52	57.41	70.37	68.52	74.07
RF	62.96	68.52	61.11	72.22	68.52	72.22

図 6.15　回答率 80% 以上でのジャンル判定の比較結果 ($N = 54$)

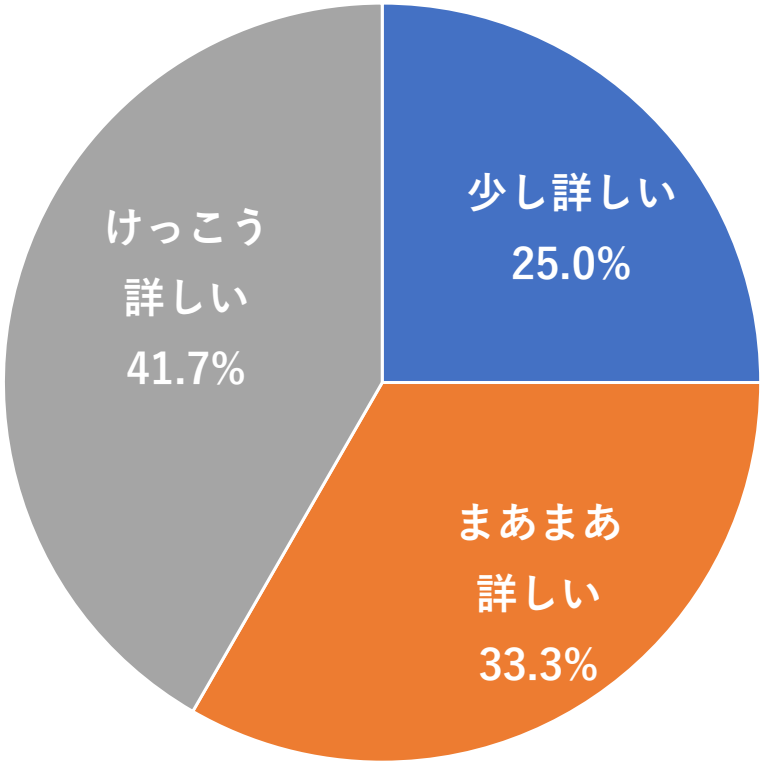


図 6.16　被験者の漫画の詳しさに対するアンケート結果

6.3　ジャンルタグの評価実験

6.3.1　実験概要

男性 12 名に対し，6.2.2 項における正解セット作成時，漫画に対して何らかジャンルが回答された回答率が 100% であった（「わからない」の回答が無かった）作品のうち 15 作品のジャンルの割合を示す円グラフをジャンルタグとして，ジャンル判定の精度評価において最も正解率の高かったスコアリングアルゴリズム DF を用いて，ベースラインの手法 $W1$ と，回答率 80% 以上の場合に最良である類似語取得数 13 語を用いた手法 $W6$ によってジャンルタグを作成し，どちらがより相応しいジャンルタグであるかと，どの程度漫画に詳しいかに関してアンケートを取った．

どれほど漫画に詳しいかは「あまり詳しくない」，「少し詳しい」，「まあまあ詳しい」，「けっこう詳しい」，「とても詳しい」の 5 段階に分けて質問を行っており，2 段階目以上で答えた対象者のアンケート結果を使用した．被験者の漫画の詳しさに対するアンケート結果の円グラフを図 6.16 に示す．しかし図 6.16 からわかるように，対象者全員が 2 段階目以上であったため 12 名全てのアンケート結果を用いている．

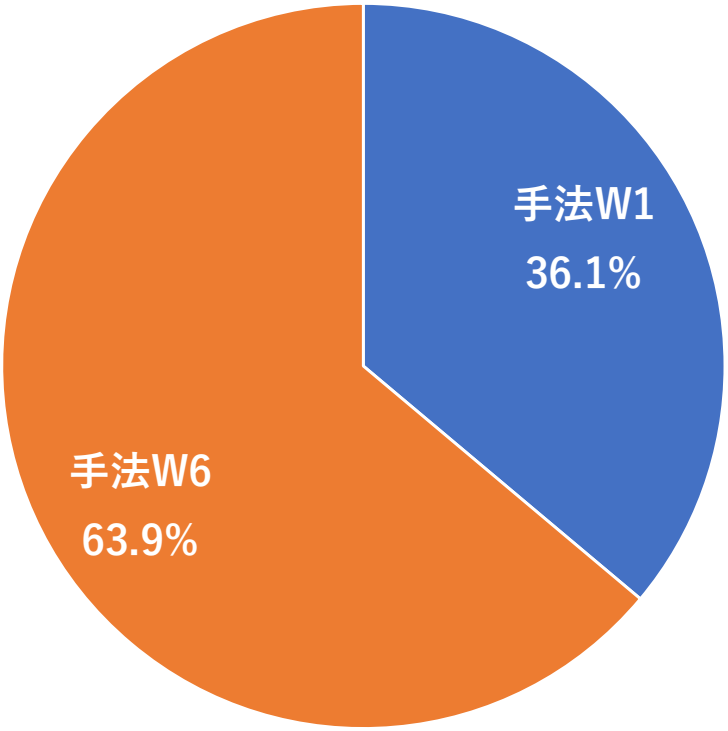


図 6.17　ジャンルタグの相応しさの比較に関するアンケート結果

6.3.2　実験結果

ジャンルタグの相応しさの比較に関するアンケートの結果を図 6.17 に示す．図 6.17 から，手法 $W1$ が 36.1%，手法 $W6$ が 63.9% であり，手法 $W1$ のようにただジャンルの単語だけでスコアを計算し，ジャンルタグを生成するよりも，手法 $W6$ のように類義語や類似語を用いてスコアを計算し，ジャンルタグを生成する方が漫画に相応しいジャンルタグを生成できるということがわかった．理由としては，ジャンル単語のみでスコアを算出する手法 $W1$ では取得できないジャンルが，手法 $W6$ では類義語，類似語により取得できるジャンルが増えたことにより，より漫画に相応しいジャンルタグを生成できたためと考えられる．またこのアンケート結果の有意確率が $p = 1.3608 \times 10^{-7}$ であることから，この結果は有意差があると言える．しかし，手法 $W1$ が 36.1% も割合を占めたことに関しては，手法 $W1$ と手法 $W6$ で取得できたジャンルタグが似たようなジャンルタグとなっている作品がいくつかあり，ユーザからの評価が分かれるジャンルタグがあったためと考えられる．

第 7 章

特徴語抽出の評価実験

本章では，漫画のレビューや漫画作品に関する Wikipedia を用いて，TF-IDF 法ベースの 4 種類のアルゴリズムで単語重要度を求めて特徴語抽出を行い，その精度を比較した実験，及び，提案手法によって抽出した特徴タグと実際にある電子書籍サイトのタグとの比較実験の 2 種類の評価実験を行った結果を示す．

7.1　特徴語抽出の精度評価

7.1.1　実験概要

最も精度の良い特徴語抽出アルゴリズムを求めるため，漫画 5 作品に関して，5 章で示した 4 種類の単語重要度を算出するアルゴリズム TF-IDF，DF-IDF，TF-IMF，DF-IMF に加えて，少ない作品数でしか登場しないような単語を除去するためのフィルターとして，漫画レビューにおいて少数の作品でしか登場しない単語を除去するためのフィルター ρ と，漫画作品 700 作品以上に関する Wikipedia から文書頻度を計算し，少数の作品でしか登場しないような単語を除去するためのフィルター ω も用意した．各アルゴリズムにおいて重要度が高い上位 10 個を取得し，正解セットと比較して平均 F 値を求める．4 種類のアルゴリズム TF-IDF，DF-IDF，TF-IMF，DF-IMF において両方のフィルター ρ と ω を用いて，2 種類のフィルターをパラメータとして 0 から 30 まで変動させ，平均 F 値を求めた．

フィルターを用意したのは，著者が特徴タグとして相応しいと考える単語が登場人物や漫画の造語を含まない一般的な単語であるからである．また特徴語を 10 語取得することについては，特徴タグとして舞台や登場人物の特徴，漫画の特色を表す特徴語が 10 個ほどあることで，漫画の内容をある程度解り易く表すことができると考えたためである．またジャンル単語は除去している．

7.1.2　特徴タグの正解セットの作成

著者を含む男性 3 名で，3 名全員の知る 5 作品の漫画に関して特徴タグとして相応しいと考えられる単語について協議し，10 単語ずつ正解セットとして用意した．ただし，ジャンル単語

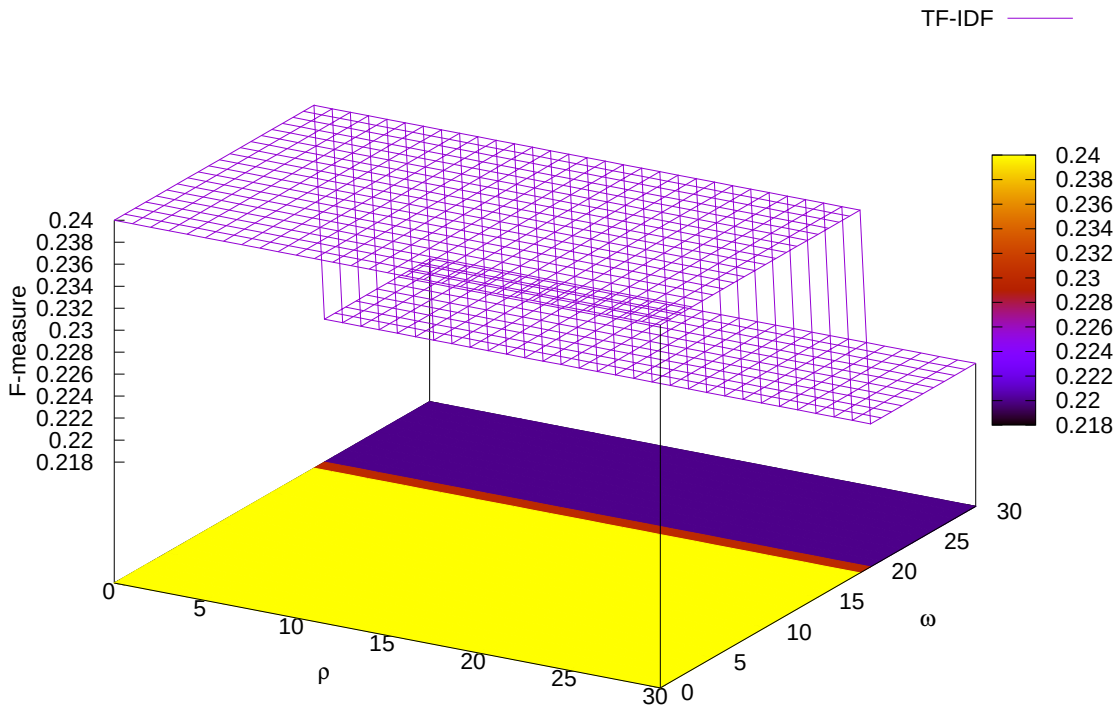


図 7.1　TF-IDF におけるパラメータに依る平均 F 値の比較結果

は正解の中に含めないようにしている．

7.1.3　実験結果

4 種類の特徴語抽出アルゴリズムそれぞれにおけるパラメータに依る平均 F 値の比較結果を示した図 7.1 から図 7.4 より，最も精度の良い結果は図 7.3 の TF-IMF における平均 F 値 0.40 となった．まず，TF-IDF，DF-IDF の結果では 2 種類のフィルターを設けても大きく平均 F 値が変化することは無かった．理由としては逆文書頻度である IDF を使用しているため，どの作品でも出現するような単語が重要度の上位に出現してしまい，作品数によるフィルターがほとんど機能していないためと考えられる．そして，TF-IMF，DF-IMF においては 2 種類のフィルターのパラメータが 1 から 10 に上がるまでに平均 F 値が上昇し，それ以降下降していくことから，特徴タグとして特徴語抽出する上で，ある程度の作品数で単語を除去することが効果的であると言える．また TF-IDF，DF-IDF に比べて，TF-IMF，DF-IMF の方が平均 F 値が高いことから，逆作品頻度 IMF を特徴語抽出に用いることは，有効であるということがこれらの結果から言える．有効である理由としては，どの作品でも出現するような，例えば「作品」,「漫画」などといった単語が重要度の上位に出現せず，特徴的な単語をその作品の重要度の高い単語として抽出することができたためであると考えられる．そして TF-IMF の方が DF-IMF より精度が良い理由としては，大量にレビューを書くユーザと少量しかレビューを書かないユーザが存在するため，DF-IMF では上位に現れるような単語も TF-IMF

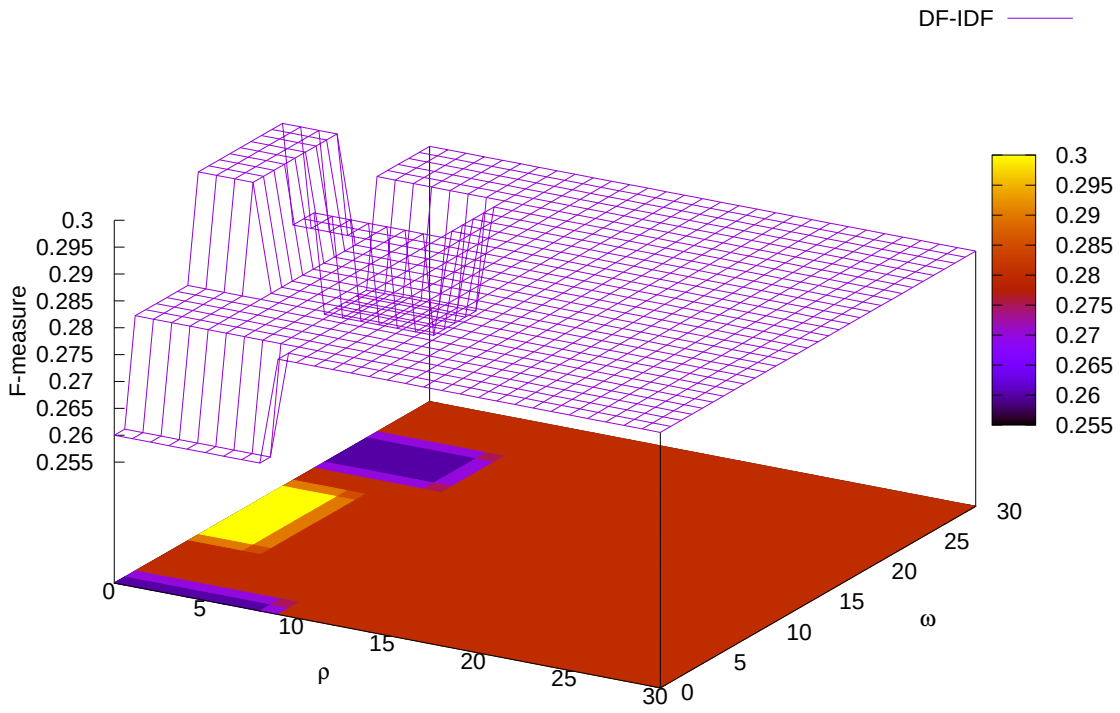


図 7.2　DF-IDF におけるパラメータに依る平均 F 値の比較結果

表 7.1　各電子書籍販売サイトのタグと 7.1.2 項で用意した特徴語を比較した結果の平均 F 値

電子書籍販売サイト	サイト A	サイト B	サイト C	サイト D	サイト E	サイト F
平均適合率	0.09	0.00	0.17	0.11	**0.20**	0.00
平均再現率	0.04	0.00	0.06	0.06	**0.10**	0.00
平均 F 値	0.06	0.00	0.09	0.08	**0.13**	0.00

で全ての単語をカウントすることにより，特徴語として相応しい単語が作品における重要度の高い単語として算出することが可能となったと考える．

7.2　特徴タグの評価実験

7.2.1　実験概要

まず，電子書籍販売サイト 6 箇所において 7.1 節と同様の 5 作品に関するジャンルを表すタグ以外のタグを取得し，「・」で分けられた単語に関しては分けた単語を 1 つ 1 つのタグとして，そのタグの中で 7.1.2 項で作成した正解セットと比較し，平均 F 値を算出した．各サイトの平均 F 値を表 7.1 に示す．表 7.1 から，最も平均 F 値の高かった電子書籍販売サイトのサイト E のタグと，7.1.3 項において最も平均 F 値が高くフィルターの値が小さかった TF-IMF の $\rho = 0$，$\omega = 10$ で取得した重要度の高い上位 10 単語とが，漫画の電子書籍のタグとして付

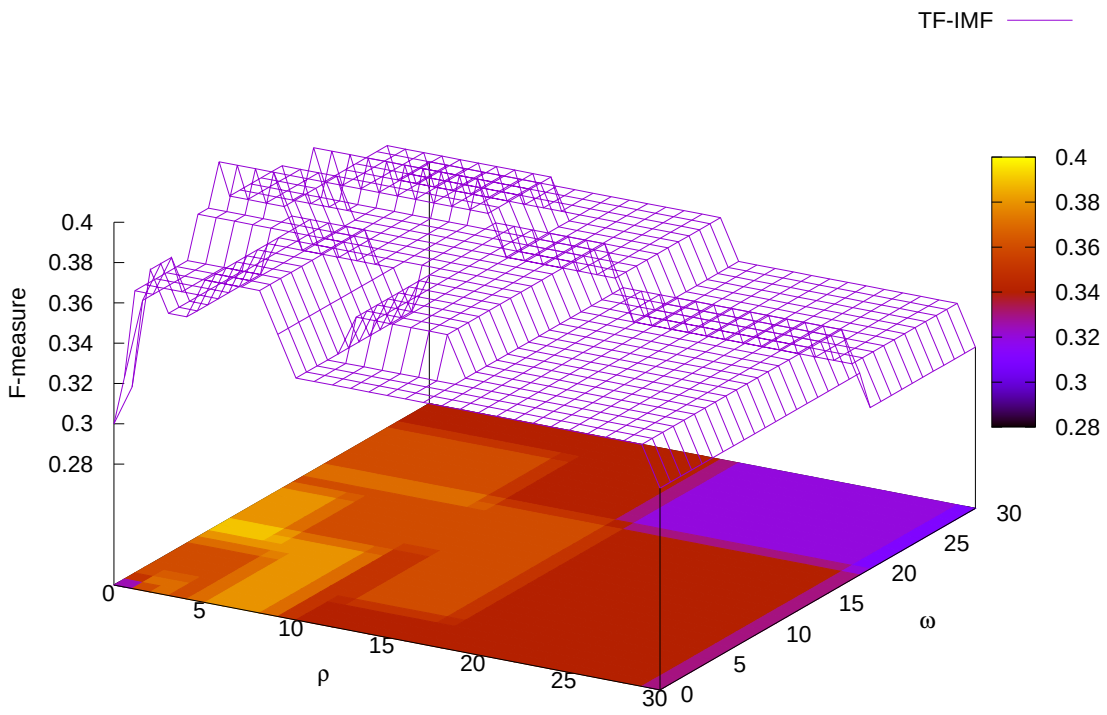


図 7.3　TF-IMF におけるパラメータに依る平均 F 値の比較結果

与されていた場合，どちらのタグの方が漫画の内容が解り易いと感じるかを選択してもらうアンケートを男性 12 名に対して行った．

7.2.2　実験結果

特徴タグの作品内容の解り易さの比較に関するアンケートの結果を図 7.5 に示す．図 7.5 から，電子書籍販売サイトのタグであるサイト E のタグでは 15.0%，提案手法で生成した特徴タグは 85.0% となった．この結果から，従来の電子書籍タグに比べて提案手法の特徴タグの方が漫画の内容が解り易いという点で非常に勝っているということがわかった．またこのアンケート結果の有意確率が $p = 1.75415 \times 10^{-14}$ であることから，この結果は有意差があると言える．図 7.5 のようになった理由としては，やはり従来の電子書籍販売サイトのタグでは「メディア化の有無」や「ターゲット年齢層」に関するタグが多く，提案手法の特徴タグは漫画の内容を表すような単語をタグとして使用しているためであり，提案手法により従来以上に漫画の内容が解り易いタグが生成できるようになったためであると言える．

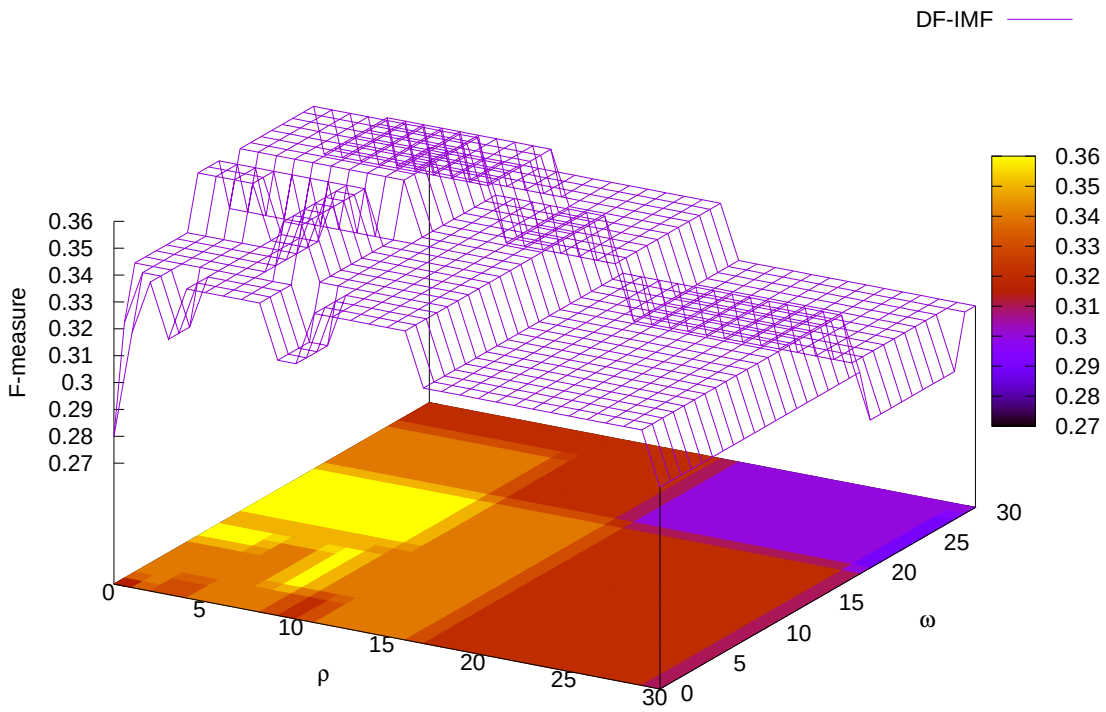


図 7.4　DF-IMF におけるパラメータに依る平均 F 値の比較結果

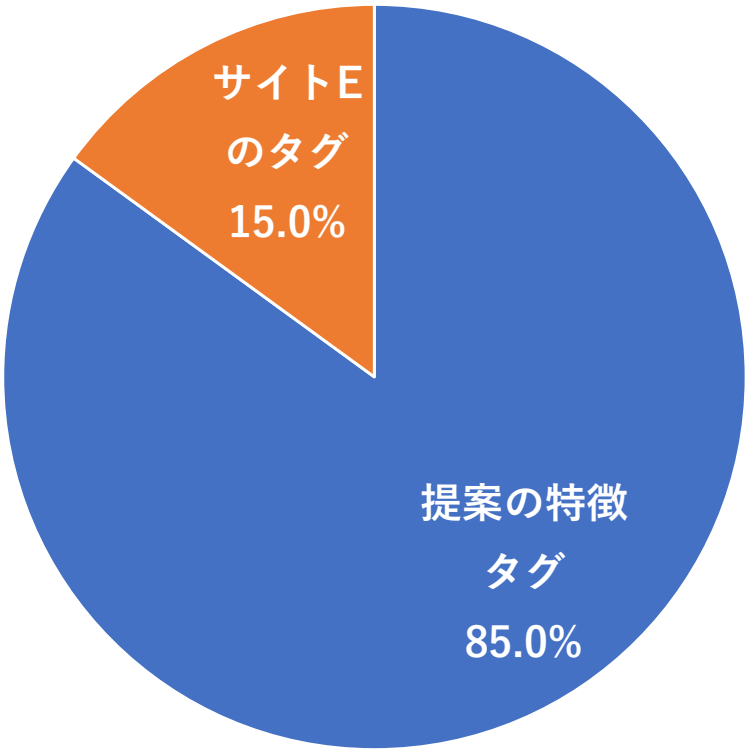


図 7.5　特徴タグの作品内容の解り易さの比較に関するアンケート結果

第 8 章

まとめと今後の研究課題

本研究では，漫画特徴タグ（ジャンルタグや特徴タグ）の生成方法について提案した．ジャンルタグについては，評価実験によって，レビューを基にジャンルを判定したり，ジャンルの割合からジャンルタグを生成したりする上で，ジャンル単語だけでなくその類義語やレビューから作成したモデルを使用した Word2Vec より得られた類似語を使用することが効果的であることがわかった．特徴タグについては，評価実験によって，特徴語抽出する上での逆作品頻度 IMF を用いることの有用性や，従来の電子書籍販売サイトのタグに比べ，提案手法の特徴タグの方が漫画の内容が解り易いタグを生成する上で，効果的であるということがわかった．

本研究の技術的な貢献としては，漫画のジャンルの判定に，レビューを用いてジャンル単語や類義語，類似語からスコアリングすることで精度の高いジャンルの判定を行えるようになっている．また，特徴語抽出において，基となった単語重要度を求める TF-IDF 法での逆文書頻度 IDF を用いるのではなく逆作品頻度 IMF を用いることで，ただ TF-IDF 法を用いるよりも作品に合った特徴語を抽出することに成功している．

また，社会的な貢献としては，現状の段階であっても作品のジャンルの割合を生成することで，作品のジャンルの構成が視覚的に解り易くなり，作品の特徴的な単語も知ることができるといったような効果があり，これだけでもある程度漫画作品の理解を助けることができる．

そして完成した場合のシステムでは，ジャンル割合のタグにより作品のジャンルが視覚的に解り易くなるだけではなく，似たようなジャンルの割合を持つ漫画が検索できる．また，特徴タグがあることで，作品の特徴が解り易くなり，漫画に詳しくない人でもその漫画がどのような漫画であるか理解でき，そして，特徴タグで検索することで，同じ特徴を持つ漫画を検索できるようなシステムとなる．このシステム，研究は従来以上に漫画について解り易いタグを生成することができるため，作品の理解の手助けや，電子書籍をより個人の好みに合わせて検索することができ，購入を促進できるものになると考えられる．

今後の研究課題として，ジャンルタグにおいては，3 種類のスコアリングアルゴリズム TF，DF，RF に大きな差が無く，ジャンルの割合としても大差が生まれなかった．そのため，3 種類のスコアリングアルゴルズムがジャンル判定に適しているかどうかの見直しや，ジャンル単語 17 語がジャンル判定において相応しいかどうか，また，類似語の *cos* 類似度であったり，単語の重要度による指標をスコアリングに用いることが必要である可能性が考えられる．そし

て，レビューでは人気の漫画ではレビューが多く書き込まれるが，マイナーな漫画はレビューが少なく，マイナーな漫画のタグの生成が難しいという問題もまた存在する．今後はレビュー以上に，よりマイナーな漫画についても情報のある可能性の高い Twitter などの他の Web 上の情報資源を使用することでマイナーな漫画にも対応できる可能性があり，レビュー以外の情報資源の活用に関して検討が必要である．

また，特徴タグにおいては，本研究では特徴タグに用いる単語を全ての作品で上位 10 単語としたが，作品によって増減するものであるとも考えられるため，作品に対する特徴タグの個数の相応しさに関して検討する必要がある．そして，本研究ではジャンル判定や特徴語抽出においては名詞のみを用いたが，作品のジャンルや特徴を表す動詞や形容詞なども存在すると考えられるので，動詞や形容詞もタグとして利用できる可能性がある．それらを用いることで作品に抱く感情などがわかり，作品に対する感じ方で検索することも可能になるのではないかと考えられる．さらに，図 3.2 のシステムイメージのように重要度により大きさが変化するタグの実装にまでは至れなかったため，重要度を正規化してタグの大きさの変化に利用するなど，解り易いデザインにするための検討などの課題が残されている．

謝辞

本研究に際して，様々なご指導を頂きました服部峻助教に厚く御礼申し上げます．また，日常の議論を通じて多くの知識や示唆を頂いた服部研究室の皆様，実験に協力して頂いた被験者の皆様にも深く感謝の意を表します．

参考文献

[1] 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, p.WII-D_1-11 (2017).

[2] 村瀬 尊好, 柊 和佑, 安藤 友晴, “マンガの概要に基づく作品推薦システム,” 情報科学技術フォーラム講演論文集, Vol.11, No.4, pp.319–325 (2012).

[3] 類語辞典・シソーラス・対義語-Weblio 辞書, https://thesaurus.weblio.jp/ (2019).

[4] word2vec の学習済み日本語モデルを公開します, http://aial.shiroyagi.co.jp/2017/02/japanese-word2vec-model-builder/ (2019).

[5] DMM 電子書籍, https://book.dmm.com/ (2019).

[6] ebookjapan, https://ebookjapan.yahoo.co.jp/ (2019).

[7] BookLive, https://booklive.jp/ (2019).

[8] BOOK WALKER, https://bookwalker.jp/ (2019).

[9] まんが王国, https://comic.k-manga.jp/ (2019).

[10] Renta!, https://renta.papy.co.jp/ (2019).