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
IN2019-65 (2020-01)

電子書籍検索のためのレビューを用いた漫画特徴タグのWeb抽出

村尾　和也†　荒澤　孔明†　服部　峻††

†,††室蘭工業大学 ウェブ知能時空間研究室　〒 050–8585 北海道室蘭市水元町 27–1
E-mail: †{16024151,18096001}@mmm.muroran-it.ac.jp, ††hattori@csse.muroran-it.ac.jp

あらまし　近年，書籍の電子化が進み，電子書籍の漫画を販売するサイトが増えており，また，スマートフォンの普及などの要因から電子書籍を購入する人が増加している．電子書籍の販売サイトの中には検索補助や内容を示すためタグやキーワードを付与しているサイトが存在するが，ジャンルやターゲット年齢層，メディア化の有無など簡易的なものが多く，漫画の内容を十分に把握することは難しい．そこで，本稿では，漫画作品の内容を表す単語やその単語の重要語，ジャンルの割合など，その作品について知らないユーザにもより解り易い漫画特徴タグを，Web 上の大量の漫画レビューを収集・分析し，単語出現数や単語重要度を測ることで自動的に生成し，各電子書籍の漫画と一緒に提供することを可能にする，より解り易い電子書籍検索システムの開発を目指す．
キーワード　タギング，キーワード抽出，特徴抽出，レビュー分析，電子書籍

Manga Feature Tag Extraction from Reviews on the Web for e-Book Search

Kazuya MURAO†, Komei ARASAWA†, and Shun HATTORI††

†,†† Web Intelligence Time-Space (WITS) Laboratory, Muroran Institute of Technology
27–1 Mizumoto-cho, Muroran, Hokkaido 050–8585, Japan
E-mail: †{16024151,18096001}@mmm.muroran-it.ac.jp, ††hattori@csse.muroran-it.ac.jp

Abstract　In recent years, the digitization of books has progressed, and the number of Web sites for selling e-book comics has increased. And the number of people who purchase electronic books has increased due to the spread of smartphones and other factors. Some e-book sales sites assign each electronic book with tags and keywords for search assistance and indicating its content, but it is difficult for users to fully grasp its content because most of them are too simple ones such as its genres, target age groups, media presence, and so on. Therefore, this paper aims at developing a more comprehensible e-book search system that allows to offer each manga e-book with its more comprehensible Manga Feature Tags, such as terms representing its content, their importance, and its genres with their proportion, which are automatically generated by analyzing a lot of reviews on the Web by calculating Term Frequency and Term Importance.
Key words　Tagging, Keyword Extraction, Feature Extraction, Review Analysis, e-Book

1. 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つながるが推薦された漫画がどのような内容なのかは把握できない．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つながる可能性がある．

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見て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わからず，作品についてよく知らない人にとってはわかりづらいものとなっている．

そこで本稿では，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う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．

2. 提 案 手 法

本章では，まず，漫画の特徴タグを抽出するシステムの構成を図 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 2 に示す．図 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による補強によりジャンルタグを，特徴語抽出により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 2 に関しては，基本タグは作品に関する基本情報が表示されるタグである．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．

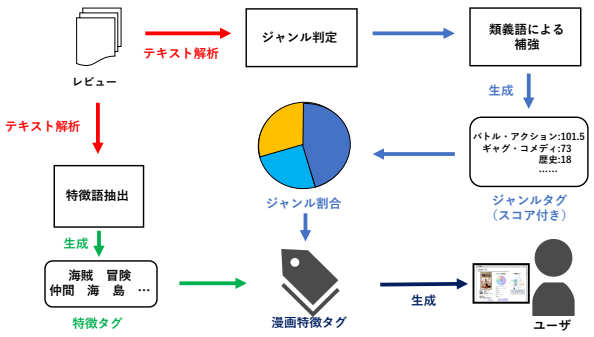

図 1　システム構成

図 2　システムイメージ

2.1　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」(注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，また固有名詞は漫画特有の造語や登場人物の名前も取得してしまうが，提案システムにおいてノイズとなるためである．

2.2　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．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．

2.3　類義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の単語の類義語，ジャンルの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，ジャンルの単語を Weblio 辞書 [2] 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．

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

表 1　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
類似語	ラブストーリー		スポーツ漫画
	青春		競技
	三角関係		スポ根
	恋愛漫画		野球
	ラブ		サッカー
	恋愛模様		ポーツ
	恋愛関係		アメフト
	コメディー		剣道
	ハーレム		メフト
	恋		卓球

表 1 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「ポーツ」や「メフト」といったノイズが混ざってしまっている．

2.4　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．

3.　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．全レビュー集合を R，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 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 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．

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

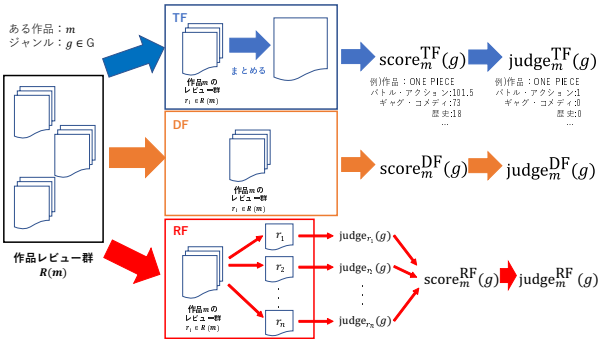

図 3　3 種類のジャンル判定アルゴリズム

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

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 3 に示す．

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

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 4 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 $W4$：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 $W3$ の Wikipedia のモデルは公開されている Wikipedia のモデル [3] を使用し，手法 $W4$ の漫画のレビューのモデルは 2.1 節で収集したレビューから作成したモデルである．

4.　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．

$$\text{TF-IDF}_m(w) = \text{TF}_m(w) \cdot \text{IDF}(w)$$
$$\text{DF-IDF}_m(w) = \text{DF}_m(w) \cdot \text{IDF}(w)$$
$$\text{TF-IMF}_m(w) = \text{TF}_m(w) \cdot \text{IMF}(w)$$
$$\text{DF-IMF}_m(w) = \text{DF}_m(w) \cdot \text{IMF}(w)$$

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

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 $\text{TF}_{r_i}(w)$ を算出するだけであるが，本稿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 $\text{TF}_m(w)$ を求めている．また全レビュー集合における文書頻度 $\text{DF}_R(w)$ ではなく，全作品集合における作品頻度 $\text{MF}(w)$ を用いる狙いとして，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て来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．

5. 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 4 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の単語の類義語，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．

5.1 電子書籍サイトによる正解セットを用いた精度評価

5.1.1 実験概要

3 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 4 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 $W3$，$W4$ 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，手法 $W4$ で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同数で，さらに最小の類似語取得数であった 16 語の類似語取得数を用いて各手法を比較する．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．

5.1.2 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 箇所から漫画 100 作品のタグ，キーワードを取得し，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．

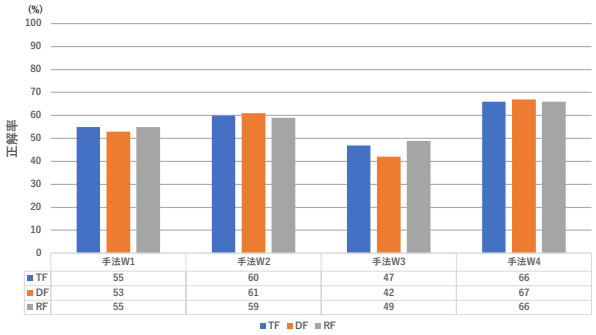

図 4 ジャンル判定の比較結果 ($N = 100$)

5.1.3 実験結果

図 4 より手法 $W4$ 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．Wikipedia のモデルを使用した手法 $W3$ が他の手法に比べ正解率が低いことからその裏付けになっていると言える．しかし，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ってしまい，また正解率も 70% に届かない結果となってしまった．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，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が挙げられる．そこで次節で正解セットを新たに用意し，再度実験を行った．

5.2 人による正解セットを用いた精度評価

5.2.1 実験概要

正解セットの中で，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 5.1 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 5.1 節と同様に，事前実験として手法 $W3$，$W4$ 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，回答率が 50% 以上の場合では手法 $W4$ で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同数で，さらに最小の類似語取得数が存在しなかったため，3 種類のアルゴリズムにおいて，最も正解率の高い中で最小の TF では 11 語，DF では 11 語，RF では 16 語の類似語取得数を，回答率が 80% 以上の場合では手法 $W4$ で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同数で，さらに最小の類似語取得数であった 13 語の類似語取得数を用いて各手法を比較する．

5.2.2 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 個の中からを 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

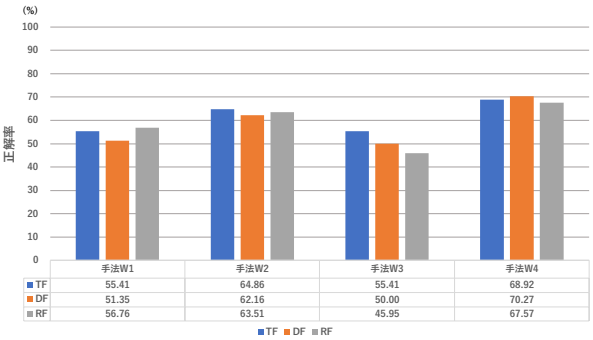

図 5　回答率 50% 以上でのジャンル判定の比較結果 ($N = 74$)

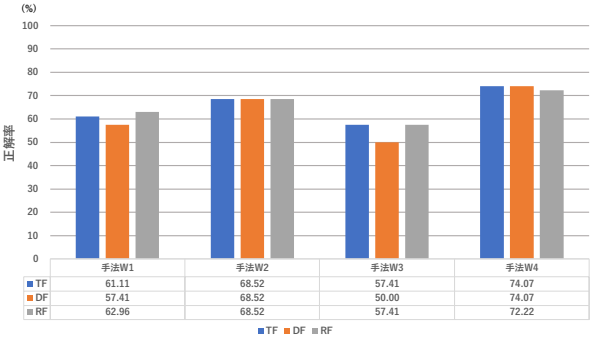

図 6　回答率 80% 以上でのジャンル判定の比較結果 ($N = 54$)

ルを各漫画の正解とした．また 5.1.2 節と同様に，「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．

5.2.3　実 験 結 果

図 5，図 6 の比較結果から最も正解率が高い手法は 5.1 節での実験と同様に手法 $W4$ 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．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，5.1 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．

5.3　ジャンルタグの評価実験

5.3.1　実 験 概 要

男性 12 名に対し，5.2.2 節における正解セット作成時，回答率が 100% であった作品のうち 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 $W4$ 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，どれほど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．しかし対象者全員が 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．

5.3.2　実 験 結 果

アンケートの結果から，手法 $W1$ が 36.1%，手法 $W4$ 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 $W4$ 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 $W4$ 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．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 $W4$ 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．

6.　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．

6.1　特徴語抽出の精度評価

6.1.1　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，4 章で示した 4 つ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，つまり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，少数の作品数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．TF-IDF，DF-IDF ではフィルター ρ による変化はなかったため，フィルター ω のみを用い，TF-IMF，DF-IMF では両方のフィルター ρ と ω を用いて平均 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．

6.1.2　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，ジャンル単語は正解の中に含めないようにしている．

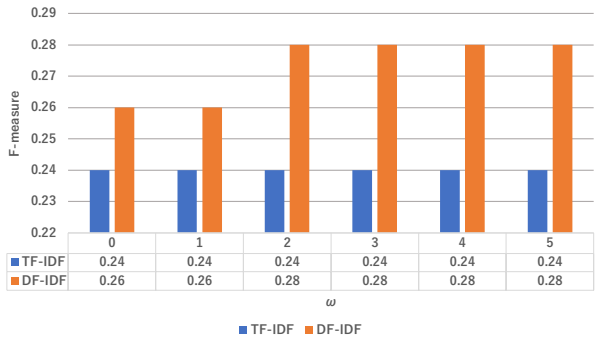

図 7　TF-IDF と DF-IDF の平均 F 値の比較結果

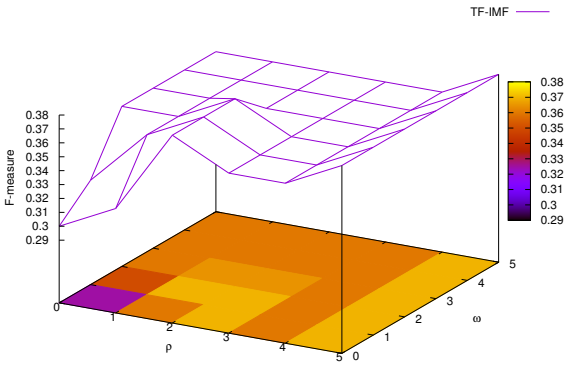

図 8　TF-IMF の平均 F 値の比較結果

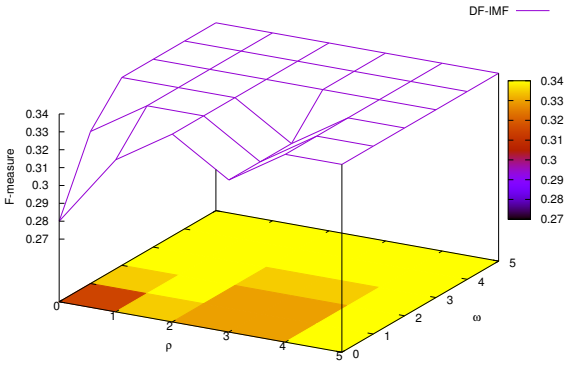

図 9　DF-IMF の平均 F 値の比較結果

6.1.3　実 験 結 果

図 7 から図 9 の比較結果から，最も精度の良い結果は図 8 の TF-IMF における平均 F 値 0.38 となった．いずれの結果においてもフィルターで一切除去しない精度が最も低いことから漫画レビューの作品数や Wikipedia での出現数でフィルターを設けることは有効であるということがわかる．また TF-IDF，DF-IDF に比べて，TF-IMF，DF-IMF の方が平均 F 値が高いことからレビューを用いてほとんどの作品で出現しないような単語を除去することは著者の目指す特徴タグにおいて一定の効果があることがわかる．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 では上位に来ないような単語も TF-IMF 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．

6.2　特徴タグの評価実験

6.2.1　実 験 概 要

まず，電子書籍販売サイト 6 箇所において 6.1 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 6.1.2 節で作成した正解セットと比較し，平均 F 値を算出した．その中で最も平均 F 値の高かった電子書籍販売サイト 1 箇所のタグと，6.1.3 節において最も平均 F 値が高くフィルターの値が小さかった TF-IMF の $\rho = 5$，$\omega = 0$ 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．

6.2.2　実 験 結 果

アンケートの結果から，従来の電子書籍タグでは 15.0%，提案手法で抽出した特徴タグは 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．理由としては，やはり従来のタグでは「メディア化の有無」や「ターゲット層」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と言える．

7.　まとめと今後の研究課題

本稿では，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，レビューから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類似語を使う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，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．しかし，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．今後はレビュー以外に Twitter などの他の Web 上の情報資源を使うことで精度が変わる可能性があるため，さらなる検討が必要である．また，特徴タグにおいては，本稿では上位 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，作品に対するタグの個数に関して検討する必要がある．さらに，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，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．

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