4A1-4in

レビュー文を用いたコミックの内容判別手法の検討

Basic study on a content discrimination method for comics using their reviews

山下 諒[*1] 松下 光範[*2]
Yamashita Ryo Matsushita Mitsunori

[*1]関西大学大学院 総合情報学研究科 Graduate School of Informatics, Kansai University
[*2]関西大学 総合情報学部 Kansai University, Faculty of Informatics

The objective of our research is to develop a search system that supports comic exploration. To achieve this objective, it is necessary to acquire content information from comics, such as the world setting and characters. In this paper, we propose a method to acquire content information about comics from their reviews on the Web. The proposed method determines a set of keywords that reflects the content of target comics appropriately and visualizes a relationship between the comics by connecting their common keywords. To determine these keywords, the term frequency - inverted document frequency (TF-IDF) and latent Dirichlet allocation (LDA) algorithms are used. TF-IDF is used to determine keyword sets that reflect on *explicit* information, which appear in the reviews, and LDA is used to determine keywords that reflect on *implicit* topics underlying the comics. We conducted two analyses to confirm the performance of these algorithms. Thus far, we determined that TF-IDF provides meaningful keyword sets, but LDA provides relatively unclear keyword sets.

1. はじめに

「2012 年版出版指標年報」(出版科学研究所刊) によると 2011 年度に発刊されたコミックの新刊は 12, 356 点にのぼる. これらのコミックに関する情報は，Twitter[*1]をはじめとする SNS や，インターネットに記載された広告などから得ることができる．しかし，これらの媒体から発信される情報の多くは，最新のコミックや注目を集めているコミックである．そのため，このような状況では，必ずしも読者が自らの嗜好に合致したコミックに出会えるとは限らない．現状でも，コミックを検索するサービスはあるものの，コミックの内容にまで踏み込んだ検索を行うことは難しい．

そこで本研究では，読者が大量のコミックの内容を直感的かつ効率的に理解でき，読者の情報要求が曖昧な状態でも，各々の嗜好に沿ったコミックの探索を容易にするシステムの実現を目指す．本稿では，システムを実現するための基礎検討として，コミックの内容情報 (e.g.，ストーリー情報，登場キャラクタ情報) を Web 上から獲得し，それらの情報を可視化することでコミックの内容を容易に把握できるようにする手法を提案する．

2. コミック検索における課題と着眼点

2.1 コミック検索における課題

コミックは絵とテキストで構成されたクロスモーダルなコンテンツであり，そこから直接コミックの内容情報を抽出することは至難である [2]．そのため，現状のコミック検索サービスでジャンル検索を行う際は，コミックに付与されたジャンル情報 (e.g.,「少年漫画」,「冒険」) に基づいて行われることが多い．これらのジャンル情報は全て人手で付与したものであるため，各コミックに付与される情報は少なくなる傾向にあり，コミック同士の類似関係を測るのは難しい．例えば,「ラブストーリー」というキーワードで検索すると 6628 件ものコミックが提示される[*2]．

人はコミックに限らず，自分の趣味や嗜好について明確に認識していることはまれである [1]．つまり，読者が能動的に嗜好に合う対象の特徴を列挙することは困難であるといえる．このことから，読者の嗜好に合ったコミックを見つけるためには，コミックの内容情報を提示する必要がある．

このような背景の下，本稿では，コミックから内容情報を獲得するのではなく，他の情報源から内容情報の獲得を目指す．

2.2 抽出対象となる情報源

1. 章で述べたようにコミックは増加し続けている．従って，各コミックの内容情報を自動で獲得できる方が望ましい．そこで，本稿では，Web 上に記載されている情報から抽出することとした．

抽出する際に，着目すべき観点を 2 つ挙げる．1 つ目は，本研究が扱う情報の単位がコミック単位であるため，獲得する情報もコミック単位であるという点である．2 つ目は，1 つのコミックに対して，複数の観点の情報を得ることができるという点である．コミックは，世界設定や登場キャラクタ，アイテムなど，複数の要素から構成されており，それぞれの要素が読者の嗜好の対象になりうる．そのため，読者によって嗜好に合うと考える観点は異なると思われる．これは，多くの読者の幅広い嗜好を満たすためには，各々のコミックに対して複数の観点の情報が必要であることを意味する．

上記の 2 つの観点を勘案し，本稿では，Web 上のレビューサイトに記載されているレビュー文に着目した．レビューサイトでは，各コミックに対してレビューを付与することができるため，コミック単位で情報を獲得することが可能である．加えて，1 つのコミックに対して複数人がレビューを記述するため，複数の観点の情報を獲得することができると期待される．

レビュー文をコミックの内容情報を獲得する対象とするにあたり，どのような内容情報を獲得できるのか分析を行う必要がある．レビュー文から内容情報を抽出する際，複数の観点が混在するため，表層的な情報のみでは，記述されたユーザの意見を正しく汲み取れない可能性がある．この問題を解決するため

連絡先: 山下 諒，関西大学大学院総合情報学研究科，大阪府高槻市霊仙寺町 2-1-1，k809231@kansai-u.ac.jp

*1 http://www.twitter.com

*2 http://www.cmoa.jp/ (2014 年 2 月 27 日時点)

に，テキストの表層情報だけでなく潜在情報を用いる手法に着目した. 次章でこの 2 つの情報を用いた分析手法について述べる.

3. 分析手法

3.1 分析における着眼点

上述したように，本研究ではそのコミックのレビュー文から表層情報と潜在情報を得ることで，コミックの内容情報を把握する.

表層情報とは，文章中に含まれる各々の単語から判断する情報のことを指す. 例えば，ある文章にサッカーという単語が出現していれば，その文章はサッカーに関する文章であると推定することができる. また，潜在情報とは，そのテキストに出現していなくても，そのテキストに現れうる情報のことを指す. 例えば，サッカーという単語が出現していなくとも，文章中に “ワールドカップ” や “ゴールキーパー” などの単語が出現していれば，その文章はサッカーに関する情報を表していると推測することが可能である.

原島らは，この表層情報と潜在情報の 2 種類の観点の情報を用いることで，対象となる文章に含まれる情報の意味合いをより正確に汲み取ることが可能であるとしている [5]. 先行研究にならい，本稿でも表層情報と潜在情報を用いて分析を行うこととした.

3.2 分析対象

本稿では，漫画レビュー.com[*3] から収集したレビュー文を分析対象とした. 漫画レビュー.com では，コミック単位でレビューを付与するため，コミックごとの情報を獲得することができる. 漫画レビュー.com では，レビューを記入する際，そのコミックを 10 段階で評価する. 今回は，その評価の上位 150 作品を分析の対象とし，2014 年 2 月 25 日時点の各コミックのレビュー文を収集した. なお，収集した各コミックに付与されているレビュー数は異なり，最多レビュー数が「SLAMDUNK」の 259 件であり，最小レビュー数が「がんばれ元気」他 6 タイトルの 10 件であった.

3.3 表層情報の分析

レビュー文に含まれる表層情報の分析には，TF-IDF を用いた. TF-IDF とは，文書に含まれる単語の相対的な重要性を表す指標であり，単語の出現頻度を表す TF (Term Frequency) 値と文書頻度を表す DF (Document Frequency) 値の逆数の積で表される. TF-IDF を用いることで，各々のレビュー文に記載されている特徴的な情報を確認することができると期待される. 本稿で分析の対象としたコミックのレビュー数は作品ごとに異なるため，情報量に偏りが生じてしまい分析結果に影響を及ぼす可能性がある. そこで，本稿では，各コミックの TF 値を各コミックのレビュー数で除算し正規化した.

3.4 潜在情報の分析

レビュー文に含まれる潜在情報の分析には，潜在的ディリクレ配分法 (Latent Dirichlet Allocation 以下 LDA と記す) を用いた [6]. LDA は文章の生成モデルの 1 つであり，文章の表層に現れない情報も考慮して分析を行うことができる. なお本稿では，分類されるトピック数 K を 30 に，確率分布を定めるパラメータである α 値を 50/K (K は推定する際に与えるトピック数) ，単語の事前ディリクレ分布のパラメータである β 値を 0.01 と定めて分析を行った. また，本稿では，各トピックの推定に Gibbs サンプリング法を採用し，試行数を 100 とした [7].

4. 結果と考察

4.1 表層情報

TF-IDF の分析結果の一部を表 1 に示す. 表 1 から各々のコミックに特徴的な情報を確認することができる. 例えば,「MOONLIGHT MILE」というコミックは,「宇宙」「開発」「月」「ロボット」「未来」などといった単語が特徴的な情報である. 今回は，TF-IDF 値上位 50 語を各コミックの表層情報とした. ただし，今回表層情報と定義した情報の中にコミックの内容に関係しない単語も含まれていたため，これらの単語に関しては手動で除外した. 各々のコミックの内容に関係しない単語として除外した 3 種類を下記に記す.

1. 登場キャラクタ名に関する単語

   「ハチ」などである. 未読のコミックに登場するキャラクタ名が提示されても，そこからそのコミックの内容を推測することはできない.

2. コミックの内容を推測することができない単語

   「〜」や「過渡期」,「哲学的」などである. このような情報が提案システムに含まれていた場合，ユーザがシステムを用いてコミックを探す際，各々のコミックの内容を推測する妨げになりうる.

3. 書誌情報に関する単語

   「デビュー作」や「打ち切り」などである. 本研究の対象はコミックの内容情報であるため，これらの情報は表層情報から除外することが望ましい.

上記に除外した単語以外をグラフ表現を用いて可視化したものを図 1 に示す. 今回は，表層情報の中に「SF」という単語が含まれているうちの 5 タイトルのコミック (「7SEEDS」,「MOONLIGHT MILE」「SF 全短篇」「プラネテス」「銃夢-GUNNM-」) を可視化の対象とした. 図 1 から，各々のコミックに紐付いている表層情報を用いてコミックの内容を推測できる可能性が示唆された. 例えば,「7SEEDS」は，表層情報から「荒廃して滅亡した地球で生き残りを懸けたサバイバルゲーム」のような内容であると推測できる.

また，このように各コミックの内容が推測できることによって，他のコミックの内容との比較が容易になった. 例えば,「MOONLIGHT MILE」は,「アメリカと中国が月を巡って争う話」であると推測できる. この推測結果と上記の「7SEEDS」の推測結果とを比較すると「MOONLIGHT MILE」の方が近未来な内容であると捉えることができる.

既存のコミック検索サービスでは,「SF」というタグ情報を用いてジャンル検索を行うと上記のコミック同士は類似しているとみなされ，同じ検索結果に現れる. しかし，今回の結果からレビュー文に記載されている表層情報から各コミックの内容が大きく異なることが確認された.

以上を踏まえ，表層情報をタグ情報に用いることで，より詳細にジャンルを分解でき，既存のコミック検索サービスではできないような，コミックの内容を直感的かつ効率的に理解できる可能性が示唆された.

*3 http://www.mangareview.com/

表 1: 各コミックの特徴量上位 10 語と特徴量

7SEEDS		MOONLIGHT MILE		SF 全短篇		プラネテス		銃夢-GUNNM-	
名詞	特徴量	名詞	特徴量	名詞	特徴量	名詞	特徴量	名詞	特徴量
サバイバル	2.916	宇宙	3.639	島	1.944	宇宙	1.624	ガリィ	1.594
キリギリス	2.119	開発	1.354	皮肉	1.639	ハチ	1.203	サイボーグ	1.066
夏	1.889	中国	0.903	PERFECT 版	1.541	マキ	0.978	クズ	0.711
チーム	1.594	月	0.804	絶滅	1.504	愛	0.389	LO	0.683
少女	1.056	天空	0.733	侵略者	1.328	フィー	0.326	鉄町	0.683
アリ	0.779	征服	0.733	オジサン	1.328	デビュー作	0.300	SF	0.570
冬	0.664	～	0.733	モンスター	1.238	木星	0.260	一昔	0.533
春	0.664	SF	0.671	怪物語	1.156	哲学的	0.228	サイバーパンク	0.455
未来	0.639	ロボット	0.541	ガク	1.156	ロック	0.221	ザレム	0.455
花	0.624	過渡期	0.539	日本語訳	1.156	未来	0.216	賞金	0.455

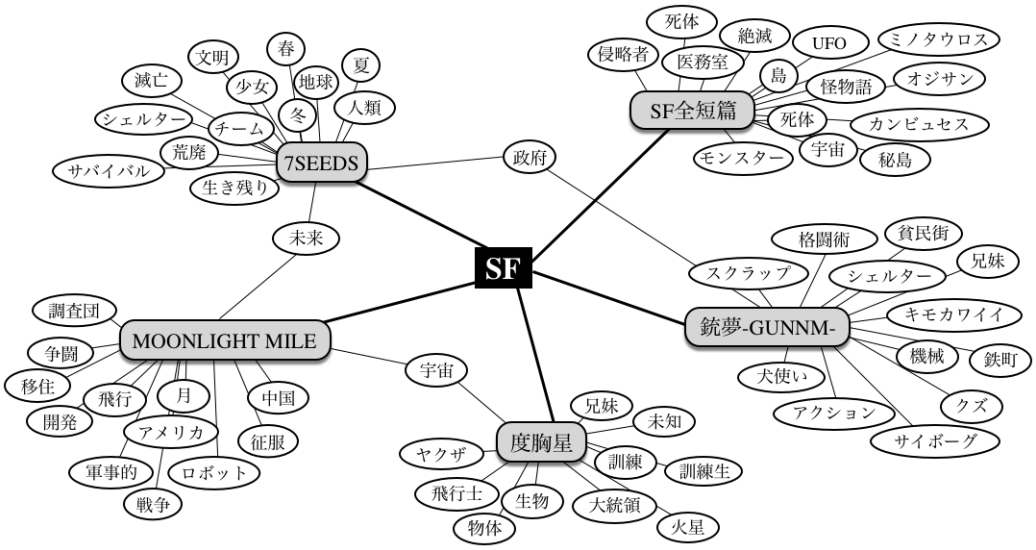

図 1: 表層情報を各コミックに紐づけたグラフ

4.2 潜在情報

LDA を用いた分析結果の一例を表 2 に示す．LDA は教師なし学習であるため，推定結果の各トピックに割り当てられた単語から，そのトピックが何を表しているのかを主観に基づいて判断する必要がある．しかし，今回の結果では，各トピックを推定することは困難な状況であった．各トピックの語群を確認すると，複数のトピックに出現している単語や，他の単語と組み合わせてもコミックの内容を推測することは難しい単語が存在している．これらの単語が各トピックの生成確率上位に出現しているため，各トピックを判断しづらい結果になったと考えられる．

次に， 1 つのコミックを対象に分析を行い 1 つのコミックのレビュー文にどのようなトピックが含まれているのかを分析した．今回は,「SLAMDUNK」というバスケットボールがテーマのコミックを対象に分析を行った．分析結果を表 3 に示す．今回の結果を見ると例えば，トピック 15 では,「最後」や「ラスト」という単語と「青春」,「感動」,「友情」などの単語から,「クライマックスに感動的なシーンが描かれている作品」であると推定することができる．しかし，他のトピックに類似した単語や同じ単語が出現しているため，上記の結果と同様に，主観に基づいてトピックを判断することは困難な結果である．

この問題を解決する手法の 1 つとして本稿では， term-score に着目した [8]. term-score とは， TF-IDF の単語の重みを取り入れた指標であり，各トピックの生成確率を TF 値，文書 (トピック) 頻度を DF 値とみなして，特徴量を算出するスコアである．この指標を用いることで複数のトピックに出現している単語の重みが下がり，上記の分析結果よりも各トピックを推定しやすくなると期待される．term-score を用いて対象コミックを分析した結果を表 4 に示す．

term-score を用いたことで例えば，トピック 16 の「作品」や，トピック 21 の「漫画」(10 位以下) などの単語の重みが下がっていることが確認された．しかし，各トピックに特徴的な情報を確認しても各トピックの内容を推定することは困難な状況であった．これは，term-score を用いて，各トピックに割り当てられる単語の重みを下げてもコミックの内容が推測することが難しい単語 (e.g.,「雰囲気」,「言葉」) や，自分のことを表す単語 (「個人的」,「私」) などの単語が多くを占めていたからであると推測される．今後は，コミックの内容に関する情報とはどのような情報であるのかを再考し，それ以外の情報を stopword として扱う必要があると考えている．

また，今回は，トピック数を 30 と定めて分析を行ったが，これは各分析対象に対して最適なトピック数ではない可能性がある．これを解決するために芹澤らは，term-score を用いて各トピックに生成されうる単語の特徴量を算出した後，cosine 類似度を用いて各トピックの類似度を算出している [9]．算出した類似度に対して閾値を設け，ある一定以上の類似度関係にあるトピックを関連トピックとしてまとめている．これを行うことで余分なトピックを除外することができる．今後 LDA を

表 2: 対象コミックの推定結果 (生成確率上位 10 語)

トピック	生成単語
トピック 0	絵，漫画，人，確か，私，為，レベル，一言，涙，手
トピック 1	最後，漫画，好き，作品，アニメ，話，世界観，展開，巻，評価
トピック 2	他，マンガ，個人的，意味，部，人物，話，オススメ，素直，男
トピック 3	方，先生，感じ，描写，題材，以外，世界，何，迫力，移入
トピック 5	漫画，事，点，頃，今，現実，作品，バトル，好み，絵柄
トピック 6	点，ギャグ，名作，ラスト，回，漫画，最後，話，絵柄，自分

表 3: 「SLAMDUNK」の推定結果 (生成確率上位 10 語)

トピック	生成単語
トピック 6	名作，他，キャラクター，全て，存在，言葉，中学，小学生，評価，満点
トピック 15	キャラ，時，最後，青春，上，秒，野球，ラスト，感動，友情
トピック 19	最高，事，涙，最終，人物，力，無理，過去，理解，勝手
トピック 20	漫画，描写，人物，本，最近，影響，天才，演出，賛否，初心者
トピック 22	漫画，ギャグ，レベル，読者，文句，全国，アニメ，ボール，人間，能力
トピック 25	作品，巻，成長，セリフ，巻数，地味，気分，メイン，ケガ，金字塔

表 4: term-score を用いた推定結果 (生成確率上位 10 語)

トピック	生成単語
トピック 12	ストーリー，個人的，一番，人生，内容，魅力，主人公，秀逸，最高，作品
トピック 16	巻，絵，セリフ，宇宙，後半，作品，雰囲気，テーマ，ネタ，完全
トピック 21	人物，キャラ，展開，当時，視点，結局，感想，感動，能力，冊
トピック 24	表現，私，作者，子供，画力，言葉，主人公，天才，とら，ジャンル
トピック 26	感じ，人間，敵，自分，上，愛，最近，世界観，好き，絶対
トピック 27	漫画，人間，主人公，好き，舞台，人，頭，エピソード，巻，少女

用いる際，この手法を用いてより適切な条件の下で分析を行うことで，各トピックが推定しやすくなると期待される．

5. 関連研究と本研究の位置づけ

本章では，新たなコミック探索の実現に向けた研究について概観し，本研究との差異を明らかにする．

岩間らは，DBpedia[*4]に作成されているコミックに関する情報を用いて，コミックへのアクセスを支援するシステムを提案している [3]. このシステムでは，コミックのタイトルを入力することで，そのコミックに関する情報がリンクとして表示される. さらに，リンクを選択することで，そのリンクの情報と同じ情報が紐付いているコミックが表示される. これを繰り返すことで，コミック間の関連に沿った探索を行うことができる．

また，書店に並ぶコミックは，掲載雑誌やジャンル，作者ごとに陳列されている. これに着目し，コミック探索の効率を高めるために，上記のような種類ごとを群として提示する取り組みもされている [4]. この研究では，Wikipedia[*5]に記載されているマンガ記事の分類に利用できるカテゴリ (e.g. 漫画家，連載雑誌) を群とみなし，各々の情報を DBpedia を用いて紐付けている．

関連研究も，本研究と同様に各コミックに関する情報を用いたコミックの横断を目的としている. しかし，各システムが提示すべき情報は異なる. 我々はコミックの内容を考慮した探索的検索の実現を目指している. そのため，提示すべき情報は，コミックの内容に関する情報である. 一方，関連研究では，あらゆるコミックの情報に基づいた横断を可能にすることを目指しているため，内容情報だけでなく，書誌情報 (e.g.，著者情報，掲載雑誌情報) も必要となる．

∗4 http://www.dbpedia.org/

∗5 http://www.wikipedia.org/

関連研究では，DBpedia に記載されているコミックに関する情報を紐づけているが，これらの情報の多くは，書誌情報であるため，コミックの内容に直接には関係せず，コミック同士を差別化することが難しい. この点から，本稿の提案手法のほうが関連研究と比較してコミックの内容情報をより適切に獲得できたといえる．

6. おわりに

本稿では，コミックの内容情報に基づいた探索的検索システムを実現するための基礎検討として，Web 上に記載されたレビュー文からコミックの内容情報の獲得を行った. TF-IDF を用いた表層情報の分析では，各コミックに特徴的な情報を確認することができ，それらの情報からコミックの内容を推測することは可能であった. 一方，LDA を用いた潜在情報の分析では，term-score を用いて各トピックに生成されうる単語の特徴量を算出し，各トピックに特徴的な単語を確認したものの，それらが何に基づいて分類されているのかを推定することはできなかった. 今後は，レビュー文にどのようなコミックの内容情報が含まれているのかを分析すると共に，更に潜在情報の分析を行っていきたいと考えている．

参考文献

[1] 土方嘉徳:推薦システムにおけるインタラクション研究へのいざない, ヒューマンインターフェース学会誌, Vol. 15, No. 2, pp. 131-134 (2012).

[2] 松下光範:コミック工学の可能性, 第 2 回 ARG WEB インテリジェンスとインタラクション研究会予稿集 pp. 63-68 (2013).

[3] 岩間勇介, 三原鉄也, 永森光晴, 杉本重雄: Linked Opend Data を利用したマンガへのアクセス支援 —メタデータによるマンガ情報の可視化— , 情報処理学会第 75 回全国大会講演論文集 (分冊 1), pp. 633-634 (2013).

[4] 岩間勇, 三原鉄也, 小平優衣, 永森光晴, 杉本重雄: リソース間関係のメタデータを利用したマンガコレクションブラウザ, HCG シンポジウム 2013 論文集, pp. 212-216 (2013).

[5] 原島純, 黒橋禎夫: テキストの表層情報と潜在情報を利用した適合性フィードバック, 言語処理学会誌, Vol. 19, No. 3, pp. 121-142 (2012).

[6] Blei, D. M. and Ng, A. Y. and Jordan, M. L.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022 (2003).

[7] Griffiths, T. L. and Steyvers, M.: Finding scienific topics, *Proceedings of the National Academy of Sciences*, Vol. 101, pp. 5228-5235 (2004).

[8] Blei, D. M. and Lafferty, J. D.: TOPIC MODELES, In A. Sricastaca and M. Sahami, editors, *Text Mining: Theory and Applications* (2009).

[9] 芹澤翠, 小林一郎: 潜在的ディリクレ配分法に基づくトピック類似度を考慮したトピック追跡, 第 4 回 DEIM フォーラム論文集 (2011).