

G-32

レビューに基づく複数ジャンルを対象とした類似作品推定手法の検討 Methods for Estimating Similar Works in Multiple Genres based on Reviews

福本 虎太郎† 伊藤 淳子†
Kotarou Fukumoto Junko Itou

1. はじめに

ユーザは求める商品やサービス等を探す際、様々な推薦システムを利用する。推薦システムには大きく分けて協調フィルタリングと内容ベースフィルタリングの2種類が存在する。

協調フィルタリングはユーザの履歴等のプロフィールを用いて、ユーザの求める商品やサービス等を予測する手法である[1]。また、内容ベースフィルタリングは商品やサービス等に付けられた属性をもとにユーザの求める商品やサービスを予測する手法である[1]。既存のサービスで用いられている推薦システムの多くは協調フィルタリングであるが、映画や漫画、小説などの作品推薦においては、作品の内容までは考慮されていない。そのため、内容が類似していても購入実績のない作品は推薦されない問題が生じる[2]。また、他ユーザの購入履歴等をもとに推薦を行うため、続編や同じ作者の作品、人気のある作品等が推薦されやすくなる問題もある[2]。

この問題に対して、Wikipediaやレビューを用いた内容ベースフィルタリングによって、他ユーザの履歴に依存せず類似作品を推定する研究が行われている[3][4]。しかし、これらの研究で用いられる手法を映画や漫画、小説等複数のジャンルに横断した作品推薦に応用すると「演技」や「作画」等の主に特定のジャンルでしか用いられない単語の影響で、同じジャンルの作品が類似作品として提示されやすくなる可能性がある。

本研究では特定のジャンルに依存する単語の影響を受けない、複数ジャンルを対象とした類似作品の推定手法について検討する。

2. 関連研究

2.1 概要と読者レビューを用いた特徴語に関する研究

白石らは漫画推薦における協調フィルタリングの問題を解決するため、概要と読者レビューの両方を用いた特徴語の抽出を提案している[3]。Wikipediaから抽出した特徴語を概要キーワードと、読者レビューから抽出した評価に関する特徴語を評判キーワードと定義した。そして、これら2種類の特徴語を併用した。これにより、作品の特徴を直感的に理解可能とした。一方で、Wikipediaを用いる問題点として作品によって記述量が大きく異なる点が指摘されている。

2.2 レビューの類似度と作品間の類似度に関する研究

櫛見らはレビューの持つ多様な情報に着目し、レビュー間の類似度を作品間の類似度とする手法を提案している[4]。レビューをもとにDoc2Vecを用いて分散表現を獲得し、類似度の算出にはコサイン類似度を用いた。また、全体的観点と部分的観点に分けて類似度を算出し、その比率を調整しながら、ユーザの求める特徴を含んだ映画の提示を試みた。その結果、映画のジャンルや出演俳優、緊迫感、恋愛要素などに応じた映画が推薦された。しかし、対象は映画のみであり、複数のジャンルを対象とした場合に同様の手

法で同程度の推薦が可能かは不明である。

2.3 クロスドメイン推薦に関する研究

中本らは「映画・漫画・アニメ・小説・楽曲・ドラマ・ゲーム」を対象としたクロスドメイン推薦を行っている[5]。その際、作品の内容はWikipediaから取得し、Doc2Vecを用いて分散表現を獲得した。その分散表現を用いて作品間のコサイン類似度を算出し、類似作品を予測した。この手法では推薦される作品の妥当性は保証されているが、セレンディピティの観点で課題がある。

3. 提案手法

3.1 方針

2章で挙げた問題点に対し、本研究ではWikipediaではなく、レビューから作品の特徴を抽出する。類似作品の推定手法は2.2節で述べた既存研究と同じくレビュー間の類似度を作品間の類似度とする。一方で、類似度の計算については特徴語の共起関係を用いる。これは、ジャンルに依存する単語の影響を検討するためである。

また、2.3節で挙げた研究では対象ジャンルを「映画・漫画・アニメ・小説・楽曲・ドラマ・ゲーム」としていたが、本研究では類似作品を推定するため、楽曲を除いた「映画・漫画・アニメ・小説・ドラマ・ゲーム」の6種類とする。

3.2 特徴語の抽出

本研究では、解析対象として「作品データベース¹」に投稿されたレビューを利用した。このWebサイトを利用した理由は複数ジャンルの作品が投稿対象となっていたためである。対象とした作品数はジャンルごとに以下の通りである。

- ・映画：8414 作品
- ・漫画：9294 作品
- ・アニメ：6908 作品
- ・小説：4588 作品
- ・ドラマ：2772 作品
- ・ゲーム：6448 作品

まず、前処理として得られたレビューの全文に分ち書きを行う。この際、作品の特徴を表すと考えられる名詞と形容詞以外の単語を除去した。また、名詞の中でも代名詞や人名等は作品の特徴を表す単語としては不適であるためこれらについても除外した。他にも、単一作品のレビューでのみ用いられる単語はその作品固有の単語であり、他作品との関連には用いることができないため、これについても除外した。形態素解析エンジンにはJanome²を利用し、辞書にはmecab-ipadic-NEologd³の一部改変したものを利用した。

次に分ち書きによって得られた単語群を対象にTF-IDFを求める。ジャンル j の作品 d の単語 w のTF-IDFを求める

¹ 作品データベース, <https://sakuhibdb.com/>

² Janome, <https://mocabeta.github.io/janome/>

³ mecab-ipadic-NEologd, <https://github.com/neologd/mecab-ipadic-neologd/blob/master/README.ja.md>

†和歌山大学システム工学部,
Faculty of Systems Engineering, Wakayama University.

際の計算式を(式 1)に示す。また、単語の出現頻度を表す TF 値の計算式を(式 2)に、単語の逆文書頻度を表す IDF 値を(式 3)に示す。(式 3)について、計算はジャンルごとに分けて行う。これは特定のジャンルでのみ頻出する単語が作品の特徴語として出力されなくなることを期待したためである。その後、求められた TF-IDF の値の高い順に上位 50 語を各作品の特徴語として定義する。

$$tfidf(w, d, j) = tf(w, d) \times idf(w, j) \quad (式 1)$$

$$tf(w, d) = \frac{\text{単語 } w \text{ の出現回数}}{\text{作品 } d \text{ に含まれる全単語数}} \quad (式 2)$$

$$idf(w, j) = \log \frac{\text{ジャンル } j \text{ の全作品数} + 1}{\text{単語 } w \text{ が出現するジャンル } j \text{ の作品数} + 1} + 1 \quad (式 3)$$

3.2 類似度の計算

本稿では、特徴語の共起関係を用いて 3 種類の類似度計算手法を提案する。なお、以下で用いるスコアとは類似度の値を指す指標であり、スコアが大きいほど類似度が高いものと定義する。また、スコアは同一の類似度計算手法の中で類似度の大小を決定するための指標であり、異なる手法間でスコアを比較することはできない。

1 つ目の類似度計算手法は特徴語の共起数をそのままスコアとする手法である。計算式は特徴語の共起関係が 1 つ増加するたびにスコアに 1 を加算する。これを手法 A とする。2 つ目は共起関係にある特徴語の IDF 値の合計をスコアとする方法である。計算式を(式 4)に示す。関数 idf は(式 3)を参照するものとし、単語 W は作品間で共起関係にある単語とする。また、(式 3)の定義の通り、各特徴語の IDF 値はジャンルごとに異なる。そのため、本手法では作品間の関係を入力作品と出力作品とし、ジャンル J は入力作品のジャンルとする。これを手法 B とする。3 つ目は共起関係にある特徴語の IDF 値の二乗の合計をスコアとする方法である。計算式を(式 5)に示す。これを手法 C とする。この手法の入出力も手法 B と同様とする。

$$\sum idf(W, J) \quad (式 4)$$

$$\sum \{idf(W, J)\}^2 \quad (式 5)$$

以上の 3 つの類似度計算手法の提案意図は次のとおりである。手法 A は特徴語の共起数をそのままスコアとする単純な手法であるため、ジャンルに依存する特徴語の検討に適切であると考えた。一方で、作品の内容を反映している特徴語とそうでない特徴語を同じ値として計算する点が課題点として挙げられる。

手法 B はそのような課題を解決するための手法である。IDF 値に着目した理由は、(式 3)より、特定のジャンルで頻出する特徴語の IDF 値は低く算出され、他作品では出力されないような特徴語の IDF 値は高く算出されるためである。そのため、同じジャンルの作品が出力されにくくなり、かつ作品の内容を反映した特徴語が抽出されることを期待した。

手法 C は、スコア計算時の IDF 値の分散をより大きくしたものである。この手法では、共起数が少ない場合でも IDF 値が高ければスコアが高く算出される。これにより、手法 B 以上に複数ジャンルの作品が出力されやすくなると考えられる。

表 1 「ノルウェイの森」の特徴語 50 語

['突撃', '恋愛', 'ノルウェイ', '小説', '自殺', '主人公', '描写', '青春', '官能', '強い', '哀しみ', '女性', '理解', 'エンターテインメント', '世界', '登場人物', '共感', '会話', '表現', '感情', '素晴らしい', '文章', '死生', '教え子', '精神', '過激', '心理', '上下', '大学生', '死者', 'テーマ', '海辺', '文学', '影響', '療養', '独特', '身体', '雰囲気', '一番', '上流', '作者', 'ヒロイン', 'エピソード', '世界観', 'ジョーク', '真理', '前者', '共通', '理想', '難しい']

表 2 各手法のジャンル別出力作品数

	手法 A	手法 B	手法 C
映画	0	0	6
漫画	5	5	1
アニメ	8	6	6
小説	7	9	5
ドラマ	0	0	0
ゲーム	0	0	2

4. 評価・考察

4.1 評価基準

3 章で提案した 3 つの手法を適用し、その結果を評価・考察する。評価基準は次の 2 点である、多様性として「ジャンルに依存する特徴語の影響を減少させ、多様なジャンルの作品が出力されたか」と定義し、また、類似度の指標として「提案手法による計算の結果得られたスコアは作品の類似度を示す指標として機能しているか」と定義する。

本稿では、提案手法による効果を比較するため村上春樹氏の「ノルウェイの森」という小説を入力作品とした場合を例に比較する。「ノルウェイの森」の特徴語 50 語を表 1 に示す。また、手法 A, B, C それぞれの手法で算出されたスコア上位 20 作品の内訳を表 2 に示す。

4.2 手法 A の評価及び考察

手法 A によって出力された作品のジャンルの多様性を評価する。手法 A によって出力された作品は「漫画・アニメ・小説」の 3 ジャンルに偏っていることが表 2 から分かる。その原因を考察する。手法 A によるスコア上位 20 作品中小説は 7 作品であり、その中で「小説」「文章」が含まれる作品は 6 作品であった。このことから、小説が多く出力された原因として「小説」や「文章」等、主に小説からのみ抽出される特徴語の影響が考えられる。

また、漫画、アニメの作品が多く出力された原因を考察する。手法 A によるスコア上位 20 作品のうち漫画、アニメは 13 作品であり、その中で「恋愛」「ヒロイン」が含まれる作品は 10 作品であった。このことから特定のジャンルのみではなく、いくつかのジャンルで多く抽出される傾向のある特徴語が存在する可能性がある。

次に、手法 A によって得られるスコアが作品の類似度を表す指標として機能しているかを評価する。手法 A によって計算されたスコア上位 5 作品の作品名と共起関係にある特徴語とスコアの組み合わせを表 3 に記載する。最もスコアが高く算出された作品は「灼眼のシャナ」という作品であるが、この作品はファンタジーを題材としたライトノベル作品であり、「ノルウェイの森」の内容を反映しているとは言えない。また、表 3 の 5 作品全体を通して「主人公」や「世界」等の特徴語が共通して共起関係にある。しかし、これらの特徴語が作品の内容を反映する特徴語であるとは言えない。よって、手法 A を用いたスコアは作品の類似度

を示すことが出来ていない。

4.3 手法 B の評価及び考察

手法 B によって出力された作品のジャンルの多様性を評価する。手法 B によって出力された作品は「漫画・アニメ・小説」の 3 ジャンルに偏っていることが表 2 から分かる。この偏りは、4.2 節で述べた理由と同様であると考えられる。手法 B によるスコア上位 20 作品のうち 15 作品が、手法 A で得られた作品と同一であった。同じ作品が抽出された原因は、共起関係にある特徴語の数に応じて加算されるスコアにある。手法 A の場合は 1 である一方、手法 B の場合は IDF 値である 1~8 程度の値となり、加算される値の差が小さかった。そのため、IDF 値によるスコア計算では手法 A で出力されたスコア上位 20 作品の内 6 作品の順位を変える程度の影響しか及ぼさなかった。

次に手法 B によって得られるスコアが作品の類似度を表す指標として機能しているかを評価する。手法 B によって計算されたスコア上位 5 作品の作品名と共起関係にある特徴語とスコアの組み合わせを表 4 に記載する。最もスコアが高く算出された作品は「海辺のカフカ」という作品であり、これは「ノルウェイの森」と同じ作者の作品である。この作品のスコアが最も高くなった原因は「ノルウェイ」や「海辺」等の特徴語の IDF 値が高く算出されたためであると考えられる。しかし、これらの特徴語は特定の作品名を示しているに過ぎず、作品の内容を反映しているとは言えない。また、「灼眼のシャナ」も 3 番目に高いスコアが算出されていることから、手法 A の問題点を解決できていない。よって、手法 B を用いたスコアは作品の類似度を示すことが出来ていない。

4.4 手法 C の評価及び考察

手法 C によって出力された作品のジャンルの多様性を評価する。表 5 に示すように、手法 C において出力された作品は「映画・漫画・アニメ・小説・ゲーム」の 5 ジャンルであり、手法 A、B と比較して増加したことが分かる。このような結果になった要因が、手法 A、B で課題として挙げた特定ジャンルからのみ抽出される傾向のある特徴語の影響を回避できたためであるのかを考察する。

まず、手法 C によるスコアの上位 20 作品中小説は 5 作品であり、その中で「小説」「文章」が含まれるのは 3 作品であった。次に、手法 C によるスコアの上位 20 作品中漫画、アニメは 7 作品であり、その中で「恋愛」「ヒロイン」が含まれるのは 4 作品であった。手法 A と比較して、割合が減少しており、また小説、漫画、アニメジャンルの出力された作品数の合計も 20 作品から 12 作品に減少した。これはスコア計算時に「小説」「文章」「恋愛」「ヒロイン」等の IDF 値の低い特徴語ではなく、他の IDF 値の高い特徴語の影響を強く受けるためである。よって、今回の例では、手法 C は特定ジャンルからのみ抽出される傾向のある特徴語の影響を減少させ、出力される作品のジャンルの多様性を向上させるのに有効であった。

一方で、入力された作品のジャンルでは稀少度が高いが、他の特定ジャンルにおいては頻出する特徴語が存在する場合、問題が生じる。小説家をテーマにした映画を入力作品とした場合を例に挙げる。小説家をテーマにした映画の特徴語として「小説」が抽出された場合、映画ジャンルにおいて「小説」という特徴語の IDF 値は高く算出される。一方で、「小説」という特徴語は小説ジャンルでは、多く抽出される傾向のある特徴語である。3.2 節で述べたように

表 3 手法 A で出力されたスコア上位 5 作品の作品名、共起関係にある特徴語、スコア

作品名	共起関係にある特徴語	スコア
灼眼のシャナ(小説)	['恋愛', '小説', '主人公', '描写', '強い', '理解', '世界', '表現', '文章', '心理', '作者', '世界観', '難しい']	13
雲のむこう、約束の場所 (アニメ)	['恋愛', '主人公', '描写', '青春', '理解', '世界', '表現', '感情', '独特', '雰囲気', 'ヒロイン', '世界観']	12
海辺のカフカ(小説)	['ノルウェイ', '小説', '主人公', '描写', '理解', '世界', '会話', '表現', '精神', '海辺', '作者', '世界観']	12
羅生門(小説)	['小説', '主人公', '描写', '世界', '表現', '文章', '心理', 'テーマ', '文学', '作者', '難しい']	11
true tears (アニメ)	['恋愛', '主人公', '描写', '青春', '登場人物', '表現', '感情', '心理', 'テーマ', '雰囲気', 'ヒロイン']	11

表 4 手法 B で出力されたスコア上位 5 作品の作品名、共起関係にある特徴語、スコア

作品名	共起関係にある特徴語	スコア
海辺のカフカ(小説)	['ノルウェイ', '小説', '主人公', '描写', '理解', '世界', '会話', '表現', '精神', '海辺', '作者', '世界観']	39.5
雲のむこう、約束の場所 (アニメ)	['恋愛', '主人公', '描写', '青春', '理解', '世界', '表現', '感情', '独特', '雰囲気', 'ヒロイン', '世界観']	34.8
灼眼のシャナ(小説)	['恋愛', '小説', '主人公', '描写', '強い', '理解', '世界', '表現', '文章', '心理', '作者', '世界観', '難しい']	33.1
アダルト・エデュケーション(小説)	['小説', '描写', '官能', '女性', '登場人物', '共感', '表現', '心理', 'テーマ', '一番', '作者']	32.4
true tears (アニメ)	['恋愛', '主人公', '描写', '青春', '登場人物', '表現', '感情', '心理', 'テーマ', '雰囲気', 'ヒロイン']	31.8

スコアの計算は入力された作品のジャンルを基準とした IDF 値で計算されるため、この場合、出力される作品のジャンルは小説に偏る可能性がある。つまり、手法 C は IDF 値が高く、かつ特定のジャンルからのみ抽出される傾向のある特徴語が存在する場合において出力される作品のジャンルが偏る可能性がある。

次に、手法 C によって得られるスコアが作品の類似度を表す指標として機能しているかを評価する。手法 C によって計算されたスコア上位 5 作品の作品名と共起関係にある単語とスコアの組み合わせを表 5 に記載する。最もスコア

が高く算出された作品は「海辺のカフカ」であり、手法Bと同じ結果となった。4.3節で述べた通り、特定の作品を示す特徴語の IDF 値は高く算出されるが、作品の内容を反映しているとは言えない。手法Cでは IDF 値の分散を大きくした影響で、手法B以上に「海辺のカフカ」のスコアが突出して高く算出された。そのため、「ノルウェイ」「海辺」等の IDF 値が高く、かつ作品の内容を反映しない特徴語が存在する場合、手法Cによるスコアは作品の類似度を表すことが出来ていない。

一方で、4.2節及び4.3節で述べた「灼眼のシャナ」等の内容に一切の関連が見られない作品については上位20作品から除去された。そのため、手法Cは作品の類似度を表す指標として手法A、Bよりも適切であると言える。

5. おわりに

本稿では、特定のジャンルに依存する特徴語の影響を受けない、類似作品の推定を実現するため、レビューから取得した特徴語の共起関係による類似度計算手法を3種類提案した。また、各提案手法を用いて、算出するスコアを作品の類似度を表す指標と定義した。

特徴語の共起数をそのまま類似度とする手法では、特定のジャンルからのみ抽出される傾向のある特徴語の影響で出力される作品のジャンルに偏りが生じた。また、作品の内容を反映しない特徴語の影響により、スコアが作品間の類似度を表せていないことが分かった。

これらの問題を解決するため、IDF 値に着目した。しかし、共起関係にある特徴語の IDF 値の合計をスコアとする手法では、IDF 値以上に特徴語の共起数に影響を受けた。その結果、出力される作品のジャンルの偏り及びスコアが作品間の類似度を表せていない問題を解決できなかった。これは IDF 値が共起数の影響を回避できるほどの大きさでないことが原因だった。

この問題を解決するため、共起関係にある特徴語の IDF 値の二乗の合計をスコアとする手法を提案、適用した。これにより、IDF 値の分散が大きくなるため、出力される作品のジャンルの多様性を向上させることができた。また、この手法によるスコアは今回提案した他の手法と比較して、類似度を表す指標として適切であることが分かった。一方で、この手法では、IDF 値が高く算出され、かつジャンルに依存する特徴語が存在した場合、出力される作品のジャンルに偏りが生じる可能性がある。また、IDF 値が高く算出され、かつ作品の内容を反映しない特徴語が存在した場合、この手法によるスコアが作品の類似度を表せないことが分かった。これらの問題の解決が今後の課題である。

参考文献

- [1] 寺野隆雄：Web上の情報推薦システム，IPSJ Magazine, Vol.44, No.7, pp.696-701(2003)
- [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，第11回情報科学技術フォーラム第4分冊，pp.319-325(2012)
- [3] 白石 絵里奈，田村 亮介，浅沼 爽汰，白井 聡一，藤田 和成，町田 翔，延澤 志保：概要と読者レビューに基づく漫画特徴抽出，情報処理学会第81回全国大会講演論文集，Vol.2019, No.1, pp.443-444(2019)

表5 手法Cで出力されたスコア上位5作品の作品名、共起関係にある特徴語、スコア

作品名	共起関係にある特徴語	スコア
海辺のカフカ(小説)	['ノルウェイ', '小説', '主人公', '描写', '理解', '世界', '会話', '表現', '精神', '海辺', '作者', '世界観']	171.9
かまいたちの夜2-監獄島のわらべ唄-(ゲーム)	['恋愛', '主人公', '青春', '官能', '登場人物', '真理']	120.7
おみおくりの作法(海外映画)	['主人公', '死生', '死者', 'ジョーク']	115.8
アダルト・エデュケーション(小説)	['小説', '描写', '官能', '女性', '登場人物', '共感', '表現', '心理', 'テーマ', '一番', '作者']	111.9
スプートニクの恋人(小説)	['主人公', '描写', '強い', '理解', '世界', '登場人物', '共感', '教え子', '心理', 'ヒロイン']	109.2

[4] 櫛見 圭司，北山 大輔：ユーザレビューを用いた全体的・部分的観点の類似に基づく映画推薦，第9回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2017），No.B5-4，pp.1-6 (2017)

[5] 中本昌吾，宮治裕：自然言語処理を用いたコンテンツ作品のクロスドメイン推薦，情報処理学会第81回全国大会講演論文集，Vol.2019, No.1, pp.441-442(2019)