ISS-P-69

# 語句の出現頻度に着目したアニメの自動推薦

小柴　健太郎†　　延澤　志保†
† 東京都市大学知識工学部情報科学科

## 1. はじめに

日本で作られるアニメ作品は年々作品数を伸ばしており，2012 年までに作られたアニメ作品は 3013 作品にもなる．アニメファンがおすすめのアニメを紹介する Web サイトやブログが存在するが，あくまでも紹介者の好みのアニメであり，個々のユーザの好みに合うとは限らない．

本研究ではアニメの自動推薦を実現するため，アニメの紹介記事を基にアニメの類似度をジャンルを考慮して推定する手法を提案する．

## 2. 手法

### 2.1 アニメの特徴を得るための入力データ

アニメごとに Wikipedia のテキストをコーパスとして準備した．各アニメの概要，あらすじ，登場人物，世界観，オリジナル用語の説明をコーパスとして利用した．

### 2.2 類似度計算のためのキーワードの抽出

ある語 $t$ の重要度 $w(t,d)$ を，アニメ $d$ のコーパスに出現する頻度(TF 値) $tf(t,d)$，コーパスセット中のアニメの総数 D と語 $t$ を含むアニメの数(DF値) $df(t)$ とを用いて，一般的な TF IDF 法[1]をアニメ $d$ に出現する語の総数で正規化した式(2.1)で算出する．

$$w(t,d)=\frac{tf(t,d)}{\sum_t tf(t,d)}\times\left(\log\frac{D}{df(t)}+1\right) \tag{2.1}$$

各アニメについて名詞を対象に(1)DF 値が 2 以上かつコーパス中のアニメの総数 $D$ 未満，(2)重要度 $w(t,d)>0.0001$，(3)記号や数字を含まないとの条件を満たすものを抽出し，この単語群を類似度計算のためのキーワード群 $V$ とする．

### 2.3 ジャンルごとの特徴語を考慮した類似度推定

ジャンルの特徴語の条件としてそのジャンル内での DF 値の高さを仮定する．さらにジャンル内での出現頻度が高い語が特徴的である．しかし，コーパス全体の DF 値が高ければ日常で使用する一般的語句であり，特徴的とは言えなくなる．これらから，ジャンルごとの特徴語を考慮して，あるキーワード $t$ のあるジャンル $g$ に属するあるアニメ $d$ での評価値 $key(t,d,g)$ を，語 $t$ のアニメ $d$ での重要度 $w(t,d)$，語 $t$ のジャンル $g$ 内での出現文書数 $df_g(t)$ を基に算出する(式(2.2))．$key(t,d,g)$ の値が大きいほど，そのジャンルの特徴語の可能性が高い．

$$key(t,d,g)=w(t,d)\times df_g(t) \tag{2.2}$$

評価値 $key(t,d,g)$ を基に，アニメ $d_i$ とアニメ $d_j$ の類似度 $sim(d_i,d_j)$ をコサイン類似度[2]を使用して求める(式(2.3))．

$$sim(d_i,d_j)=\sum_{t\in V}(key(t,d_i,g_i)\times key(t,d_j,g_j)) \tag{2.3}$$

## 3. 実験結果

表 1 に提案手法を用いた推薦評価結果を示す．対象としたアニメは 7 ジャンル計 28 個であり，1 ジャンル平均 3～5 個のアニメを含む．表 1 は同じジャンル，異なるジャンル 2 区分について，各アニメ同士の類似度推定を行った結果である．それぞれのアニメについて類似度の平均の最高値より

表 1　提案手法による推薦評価結果

| | 強く推薦 | 推薦する | 推薦しない |
|---|---|---|---|
| 同じジャンル | 47.7% | 31.8% | 20.5% |
| 異なるジャンル | 2.7% | 21.7% | 75.6% |

高い値の個数(強く推薦)，それぞれのアニメの類似度の平均よりも高い値の個数(推薦)と類似度の平均よりも低い値の個数(推薦しない)の 3 段階のアニメ数の割合を示す．表 1 に示すとおり，同じジャンルのアニメについては 79.5%の正解率を示す一方，ジャンルが異なるアニメでは誤って推薦されたアニメは 23.4%にとどまった．同じジャンルのアニメの評価数は計 88，異なるジャンルのアニメの評価数は計 668 個であることを考えると，これは十分に低い値である．また，ジャンルの特徴語を考慮せずアニメごとの TFIDF 値を基に評価を行った結果(表 2)と比べても正解率は 20.4%向上しており，これらから提案手法は有効と結論付ける．

表2　ジャンル特徴語を考慮しない推薦評価結果

| | 強く推薦 | 推薦する | 推薦しない |
|---|---|---|---|
| 同じジャンル | 34.1% | 25.0% | 40.9% |
| 異なるジャンル | 1.8% | 7.6% | 90.6% |

## 4. まとめ

本研究では，アニメのジャンルの特徴語を考慮してアニメの同士の類似度を推定する手法を提案した．7 ジャンル 28 アニメを対象とした類似度推定実験の結果 79.5%の正解率を得ることができた．この結果はジャンルを考慮せずアニメごとの TFIDF 値のみを用いた方法の 59.1%という結果に対して 20%以上向上しており，本手法の有効性を確認することができた．

## 参考文献

[1]　奥村学，自然言語処理の基礎，コロナ社，2010.
[2]　久米雄介，打矢隆弘，内匠逸，“興味領域を考慮した Twitter ユーザ推薦手法の提案と評価”，情処研報，2015-ICS-179(1)，pp.1-8，2015.