2014 年度卒業研究概要

ベクトル空間モデルを用いたあらすじに基づくアニメの嗜好判定手法

大谷紀子研究室

0932088 郡司喬史

1.研究の背景・目的

現在、日本では独立 UHF 局で放送されるアニメや、ライトノベルが原作のアニメが増加したため、毎年 100 本以上のテレビアニメが放送されており、放送本数は年々増加する傾向にある。多数のアニメの中から自分の好みに合うものを見つけるには、最初の 2,3 話を視聴するか、各アニメの公式サイト等に記載されているあらすじの文章を読まなければならない。しかし、ほとんどのアニメは放送時間が 30 分であるため、前者の方法では 1 つの作品を視聴するのに 1 時間～1 時間半もの時間がかかり、100 本以上のアニメをチェックすることは困難である。また、後者の方法でも大量のアニメの中から自分の好みに合うものを見つけるのは手間がかかる。本研究は、ベクトル空間モデルを用い、あらすじの文章を元に好みに合うアニメを抽出する方法を提案し、自分の好みのアニメをより簡単に見つけることを目的とする。

2.研究の内容

本研究では、アニメの公式サイト等に記載されているあらすじの文章を用いて、自分の好みに合うアニメを判定するアニメ嗜好判定システムを構築する。判定にはベクトル空間モデルを用いる。ベクトル空間モデルとは、文章中の出現単語に基づいて文章を 1 つのベクトルで表現し、ベクトルの向きで内容を判断する情報検索技法である。あらすじの文章から名詞、形容詞、形容動詞の語幹を形態素解析で抽出し、単語間共起によって有効語を決定する。単語間共起に基づく方法では、特定のカテゴリに偏って出現する単語を有効語とみなす。各文書が L 個のカテゴリ $C_1, C_2, \cdots, C_L$ に分類されているとき、カテゴリ C_i 内の文書のうち単語 T を含む文書数を $dfreq(T, C_i)$ 、カテゴリ C_i の文書数を $dnum(C_i)$ とすると、単語 T の重要度 $w(T)$ は式(1)で求められる。ここで r は任意の定数とする。

$$w(T) = (1 + \sum_{i=1}^{L} \frac{dfreq(T, C_i)}{dnum(C_i)} \log_2 \frac{dfreq(T, C_i)}{dnum(C_i)}) \cdot \frac{1 - r^{\frac{dfreq(T)}{M}}}{1 + r^{\frac{dfreq(T)}{M}}} \quad (1)$$

重要度の高い N 個の単語 $V_1, V_2, \cdots, V_N$ を有効語とする。各有効語の意味的な類似性はそれぞれ異なっているため、有効語そのものを軸としてベクトルを生成するのは望ましくない。単語の意味的類似度をベクトルに反映させるために、有効語を共起係数に基づくベクトルで表現する。有効語 V_i を表すベクトル $\vec{v}_i$ の定義を式(2)に示す。ここで、有効語 V_i と有効語 V_j の両方を含む文書数は $dfreq(V_i, V_j)$、有効語 V_i を含む文書数は $dfreq(V_i)$ であり、有効語 V_i と有効語 V_j との共起係数を成分とする N 次元ベクトル $\vec{v}_i$ で有効語 V_i を表す。

$$\vec{v}_i = (\frac{dfreq(V_i, V_1)}{dfreq(V_i)}, \frac{dfreq(V_i, V_2)}{dfreq(V_i)}, \cdots, \frac{dfreq(V_i, V_N)}{dfreq(V_i)}) \quad (2)$$

文書 D を表すベクトル $\vec{d}$ は文書 D に含まれる有効語のベクトルの和で求められる。

$$\vec{d} = \sum_{i=1}^{N} tfreq(V_i, D) \cdot \vec{v}_i \tag{3}$$

有効語ベクトルを足し合わせるとき、文書内での出現位置や文章中における言語的役割などにより重みをつけることで、より内容を反映したベクトルを作成することができる。

文書の内容を反映したベクトル空間が生成されると、同一カテゴリに属する文書のベクトルは互いに近くに存在する。$dnum(C)$ 個の文書を含むカテゴリ C のベクトル $\vec{a}$ は、各文書のベクトル $\vec{d}_1, \vec{d}_2, \cdots \vec{d}_{dnum(C)}$ の平均で表される。

$$\vec{a} = \frac{1}{dnum(C)} \sum_{i=1}^{dnum(C)} \vec{d}_i \tag{4}$$

ある文書 D がカテゴリ C に属するか否かは、両者の類似度と閾値 α との大小関係により決定する。類似度 $sim(D,C)$ は、文書ベクトル $\vec{d}$ とカテゴリベクトル $\vec{a}$ とのなす角の余弦値で表す。

$$sim(D,C) = \frac{\vec{d} \cdot \vec{a}}{\|\vec{d}\| \cdot \|\vec{a}\|} \geq \alpha \Rightarrow D \in C \tag{5}$$

$$sim(D,C) = \frac{\vec{d} \cdot \vec{a}}{\|\vec{d}\| \cdot \|\vec{a}\|} < \alpha \Rightarrow D \notin C \tag{6}$$

3.評価実験

評価実験では、筆者が過去に視聴した様々なジャンルの約 250 本のアニメを用いて、クロスバリデーションにより本システムの有用性を確認する。各アニメは、筆者の好みに合うものと、好みに合わないものに分類されており、好みに合うアニメとして抽出すべきものが明確になっているものとする。すべてのアニメを無作為に 10 個のグループに分け、そのうちの 9 個のグループに所属するアニメを訓練データとして、好みのアニメのベクトルと好みでないアニメのベクトルを作る。残る 1 個のグループに所属する各アニメをテストデータとして、好みに合うアニメを判定し、再現率と適合率を求める。再現率と適合率は検索システムや分類システムの評価に用いられる指標で、再現率は検索漏れの少なさを表し、適合率は検索ノイズの少なさを表す。一般に、再現率と適合率は相反関係にあり、一方の値が増加するともう一方の値が減少する。

4.結果と考察

上記の実験の結果、好みに合うアニメ 180 本のうち 115 本が正しく判定され、64 本が誤った判定になった。また、1 本のアニメについては判定ができなかった。ここから再現率が 64.2%、適合率が 73.2%となった。本システムでは好みに合うアニメのうち約 3 割の判定が誤っていた上、好みに合わないアニメの半数以上が好みに合うと判定されたので、アニメのジャンルごとに分けて判定する、各アニメの好みの度合いを考慮して判定する等の改善が必要だと考える。

参考文献

[1] 大谷紀子, “情報検索におけるベクトル空間モデルの応用”, 武蔵工業大学環境情報学部紀要, vol5, pp99-109, 2004.