4Q-05

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

中本　昌吾†　　宮治　裕‡
青山学院大学　社会情報学部

1　はじめに

現代社会において，漫画や楽曲といったコンテンツ作品は人々に欠かせない重要な存在である．しかし，日本国内のコンテンツ産業の市場規模は，みずほ銀行[1]によると 2012 年時点で2007年のピーク時から約1.7兆円縮小しており、経済産業省[2]によるとその後2017年まで停滞している．みずほ銀行はコンテンツのデジタル化による選択的な消費を原因としている．一方で，日本動画協会[3]によると，産業内部のアニメ業界は 2013 年以降，売上増加が続いている．このことからコンテンツ産業内の活気がある産業の消費を他の産業に向けることができれば，産業全体の市場活性化が期待できる．

以上の背景から，日本コンテンツ産業の消費者に，従来の推薦よりも消費を促すことを本研究の目的とする．そのためには，コンテンツ産業内部の各々が相互に影響を与え合うことが有効である．

そこで本研究では，映画・漫画・アニメ・小説・楽曲・ドラマ・ゲーム作品のクロスドメイン推薦を実現し，それを評価するシステムを構築する．

2　関連研究

クロスドメイン推薦に関する研究には，以下のものが挙げられる．

富士谷ら[4]は，コンテンツの多様性を持つドメインとしてテレビ番組を対象に，放送ごとの番組に適した書籍推薦を研究した．テレビ番組の内容は多様であるため，有効な特徴量も一様ではない．富士谷らは多様性を考慮して，TF-IDF と LDA の特徴量を併用し，クロスドメイン推薦をおこなった．

また、コンテンツの特徴量を抽出している研究には以下が挙げられる．

山下ら[5]は，コミックの探索を支援するシステムを構築した．その際にコミックの画像から直接特徴量を抽出するのではなく，Web 上のレビューから TF-IDF と hLDA を用いてコミックの特徴量を抽出した．

3　提案手法

本研究では、山下らのように，作品リソースからではなく間接的に作品の特徴量を抽出しする．データには，Wikipedia の全文データから，扱うドメインに対応する記事カテゴリに属する作品記事を取得し， Paragrapth Vectorにより得られる，作品の分散表現のコサイン類似度を用いて推薦をおこなう．

まず，記事カテゴリを指定する．映画・漫画・漫画作品・アニメ・アニメ作品・小説・楽曲・ドラマ・ゲーム・ゲームソフトから始まる，または「漫画作品_(平仮名一字)」・「アニメ作品_(平仮名一字)」・「楽曲_(平仮名一文字)」のカテゴリを抽出し、属する記事を取得する．

さらに、取得した記事からさらに作品記事の抽出をおこなう．記事群から 20 または 19 から始まる，または年の・一覧・題材・作品」・賞を含む，または映画祭・映画・漫画・アニメ・小説・楽曲・ドラマ・ゲームで終わるタイトルの記事を削除する．これにより，映画 22917・漫画 15167・アニメ 7662・小説 8296・楽曲 31758・ドラマ 11077・ゲーム 15772・計 112649 件のデータを得た．

次に得られた記事本文に前処理をする．オープンソースの形態素解析エンジンである MeCab の辞書を mecab-ipadic-NEologd に設定し，単語分かち書きをおこなう．その際「名詞」・「動詞」・「形容詞」以外,または接頭語・接尾語・固有名詞・英単語・代名詞・数に該当する単語を削除する．

そして，Python のライブラリである gensim のうち Paragraph Vector に対応する doc2vec を用いて分散表現を獲得し，それを用いて作品間のコサイン類似度を算出する．

4　評価実験

本実験では実験用紙システムを構築し，新規性・妥当性・セレンディピティを観点として評価実験をおこなった．

4.1　実験方法

Recommendation system across multiple content domains with distributed representations of Wikipedia
† Shogo Nakamoto, Aoyama Gakuin University
‡ Yutaka Miyaji, Aoyama Gakuin University

20 代の学生 28 人の実験協力者に以下の手順で実験をおこなった．

まず、作品消費頻度をドメインごとに回答してもらう．次にドメインから好きな 1 作品を選択してもらい，その作品との類似度が高い順に 3 作品，各ドメインから全 21 作品を推薦した．
表示される画面の例を図 1 に示す．各作品について新規性があるかを調査し，新規性が作品については推薦が妥当であるかを調査した．新規性があった作品については，仮想の消費活動として，作品名をクエリとした Web 検索に誘導しブラウジングをおこなってもらった．その後，作品が好みであったか，推薦により新たな出会いがもたらされたかを調査した．

4.2　実験結果及び考察

新規性について，全実験協力者に推薦された作品のうち，未知の作品の割合が 84.2%であり，既知の作品の割合が 15.8%であった．実験協力者ごとの作品消費頻度を考慮すると妥当な結果が得られた．

既知の作品群に対する妥当性調査の回答について，割合は「妥当である」が 48.2%，「どちらかといえば妥当である」が 27.7%，「どちらでもない」が 4.8%，「どちらかといえば妥当でない」が 4.8%，「妥当でない」が 14.5%となった．しかし既知の作品として，選択作品と同シリーズの作品が多数推薦されるといった問題が生じた．

未知の作品群に対するセレンディピティ調査の回答について，割合は「好みである」が 17.9%，「どちらかといえば好みである」が 22.3%，「どちらでもない」が 13.8%，「どちらかといえば好みでない」が 19.6%，「好みでない」が 26.3%となった．記述回答内容から好みでない作品が多数であった実験協力者の傾向として，作品の関心要因が内容とは無関係であることがわかった．

以上の結果から，提案した推薦は，新たな出会いを提供する機能や，実験協力者から信頼を得る妥当性があると判断できる．一方で本研究の目的達成に必要なセレンディピティをもたらせておらず，この点で改良の余地がある．

また，推薦をおこなう際にシリーズ全体の販売状況などを記載した作品でない記事を扱っていた場合もあり，データセットの整備も必要である．

セレンディピティの向上のためには，ドメインの特徴や関係の考慮や，分散表現の精度向上，ユーザの関心対象を考慮した推薦手法との併用が求められる．

5　おわりに

本研究では，コンテンツ産業の市場停滞を打破するために，作品のクロスドメイン推薦を提案した．

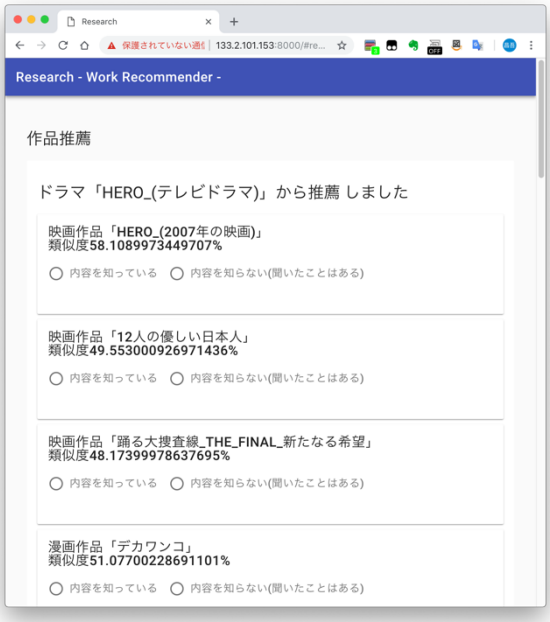

図 1 :推薦結果画面例

評価実験より，新たな出会いを提供する性能とユーザから信頼を得る妥当性に関して，推薦に必要な基本的な水準に達しているといえる．一方で，本研究の目的であるコンテンツ産業市場停滞の解決につながる消費意欲をもたらすことが十分にできたとはいえない．

今後の課題として，データセットの整備，ドメイの考慮，精度向上，他手法との併用が挙げられる．

参考文献

[1] みずほ銀行. コンテンツ産業の展望. みずほ産業調査. みずほ銀行産業調査部, 9 2014.
[2]経済産業省商務情報政策局コンテンツ産業課. コンテンツ産業政策について. 2017.
[3] 一般社団法人日本動画協会. アニメ産業レポート 2017. 1 2018.
[4] 富士谷康, 村尾和哉, 望月祐洋, 西尾信彦. コンテンツの多様性を考慮したクロスドメイン推薦. 情報処理学会論文誌, Vol. 57, No. 10, pp. 2210-2221, oct 2016.
[5] 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, pp. WII-D 1-11, 2017.