ユーザの行動履歴から学習したベクトル表現による コンテンツの人気予測

野中　尚輝† 　中山浩太郎† 　松尾　豊†

† 東京大学工学系研究科技術経営戦略学専攻　〒 113–8654 東京都文京区本郷 7 － 3 － 1
E-mail: †{nonaka,nakayama,matsuo}@weblab.t.u-tokyo.ac.jp

あらまし　コンテンツ産業は，日本の重要な産業の一つであり，近年海外においてもその人気が高まっている．このような背景の中で，コンテンツホルダーおよびコンテンツの二次利用を考える企業にとって，コンテンツの人気を予測することは重要な課題である．これまでの人気予測の研究では，消費者の嗜好をもとにしたジャンルや流行した年代といったコンテンツの多面的な情報が考慮されることは少なかった．本研究では，Wikipedia よりユーザの嗜好を抽出し，それに基づく商品の多面的な情報をもつベクトル表現を得ることを試みた．また，学習されたベクトル表現について定性的な分析を行うとともに，人気予測において精度を 2〜5%向上させることを示した．

キーワード　人気予測，Wikipedia，MLP

1. はじめに

コンテンツ産業は，日本の重要な産業の一つであり，近年海外でもその人気が高まっている．代表的なコンテンツであるアニメ産業の市場規模は，2015 年には前年比 12.0%増という大きな成長をしている [1]．また，海外におけるコンテンツ市場の規模も拡大が続くことが見込まれている [2]．

このような背景の中で，コンテンツホルダーおよびコンテンツの二次利用を考える企業にとって，コンテンツの人気予測は重要な課題である．また，海外の版権バイヤーに対して正確な人気の情報を伝えることでコンテンツ作品の海外進出が促進される点からもコンテンツの人気予測は重要である．商品の販売数や将来の人気を予測することは，マーケティングをはじめとする企業の意思決定において重要であり [13]，これまでに多くの研究が行われている [20], [30], [5]．代表的な研究としては，映画の売り上げを予測した研究 [20], [30] や将来の株価を予測した研究 [5] が挙げられる．コンテンツの人気予測の研究としては，検索クエリや Twitter，Wikipedia の情報を用いて予測を行なった [31] が存在する．これらの研究は近年急速に発達したソーシャルメディアの情報を用いることで，予測を行なっている．また将来の予測を行う上で，予測対象に対する消費者の口コミ情報を用いることで精度が向上することが知られている [9]．

商品の販売予測を行う上で，消費者の口コミ情報と同程度に，消費者の嗜好をもとにした商品の多面的な情報を考慮することが重要であると考えられる．商品の推薦において代表的な手法である協調フィルタリングは，商品とユーザのベクトル表現をもとに推薦を行うなど，ユーザの嗜好を反映したモデルである [23], [25]．このように推薦問題においては，消費者の嗜好を考慮するモデルが大きな成果をあげている．しかし，販売や人気を予測する研究では，消費者の嗜好が考慮された事例は少なかった．

そこで本研究では，Wikipedia よりユーザの嗜好を抽出し，それに基づくコンテンツのジャンルや流行した年代といった多面的な情報を得ることを試みる．Wikipedia は，幅広い内容を網羅しており，多くのユーザにより利用，編集されるソーシャルメディアの一つである．Wikipedia において，ユーザが自身の興味関心の高い項目に関するページを編集するため，各ユーザが編集する項目はそのユーザの嗜好を反映していると考えられる．したがって，編集履歴から得られる様々なユーザの嗜好を考慮することで，コンテンツ作品についての多面的な情報を得られると考えられる．

以上を整理すると本研究では，ウェブ上におけるコンテンツ作品に関するユーザの行動履歴の系列データから，消費者の嗜好に基づく商品の多面的な情報を学習し，それらを用いてコンテンツの人気予測を行う．より具体的には，コンテンツ作品についての Wikipedia の編集履歴に対し，自然言語処理の分野で大きな注目を集める Word2vec [18] を応用し，コンテンツ作品のベクトル表現を得たのち，Wikipedia の被リンク数を指標として人気予測を行う．被リンク数は，ブログ [28] やウェブページ [22] の人気や重要度の推定に用いられている．本研究では，Wikipedia におけるリンク構造から得た被リンク数を各コンテンツの人気指標とした．また，編集履歴より学習されたベクトル表現についての定性的分析も合わせて行なった．

本研究の貢献は以下である．

- ウェブ上でのユーザの行動履歴からコンテンツのベクトル表現を学習できることを示した．
- コンテンツの人気予測において，編集履歴から学習したベクトル表現を用いることで予測精度が向上することを示した．
- コンテンツについてクラスタリングを行うことにより，予測精度が向上する場合と低下する場合の違いを分析した．

本論文は以下のように構成される．2 章にて関連研究を示し，3 章で提案手法について説明する．4 章では提案手法の前提条件を検証する予備実験，5 章では提案手法の有効性を検証する

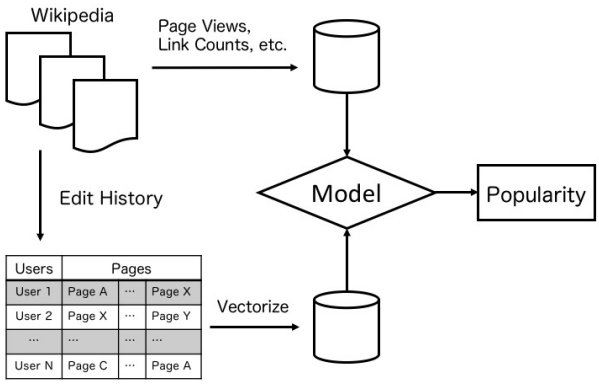

図 1　提案手法の全体像

実験について述べ，6 章にて実験の考察と手法の発展性について議論する．最後に結論を 7 章に記す．

2. 関 連 研 究

提案手法に関連する研究について述べる．これまでの人気予測についての研究について言及した後，Wikipedia，ベクトル表現の獲得に関わる関連研究について記す．最後に，ページに対する被リンク数からページの人気度を計測する手法について述べる．

2.1 販売数・人気の予測

商品やコンテンツの人気予測はマーケティング戦略を決定する上で重要な課題であり [13]，多くの研究が行われている．近年インターネットやスマートフォンの普及により，一般のユーザが情報を発信することが容易になった．その結果，ウェブ上には多くの情報が存在するようになり，それらを分析することで商品の流行や人気を予測することが先行研究にて行われている．特に，ソーシャルメディアやブログの普及により豊富なデータを取得することができるようになり，これらを用いた人気予測の研究が行われている [29], [3]．その他にも，検索クエリを用いた研究 [8] や各種ソーシャルメディアから得られた素性を組み合わせて用いる研究 [31] も存在している．

このような商品の販売や人気予測の問題において，口コミ情報を用いることで予測の精度を向上できることが示されている [9]．例えば, [16] では，口コミ情報の変遷とそれによる映画の収益の説明を試みており, [10] では新しい製品の成功の度合いを口コミ情報により予測している．また, [30] では，レビューサイトに投稿されたユーザの口コミ情報に対して，感情分析を行い，映画の収益を予測している．

商品の販売や人気を予測する上で，消費者の口コミ情報と同程度に，消費者の嗜好をもとにした商品の多面的な情報を考慮することが重要であると考えられる．商品の推薦において代表的な手法である協調フィルタリングは，商品とユーザのベクトル表現をもとに推薦を行うなど，ユーザの嗜好を反映したモデルである [23], [25]．また，商品の人気には自己強化的な側面があり，人気は消費者の意思決定に影響を与えることが知られている [24], [7]．このことから，消費者の嗜好と商品の人気の間には関連性が存在すると考えられる．しかしながら，これまで商品の販売や人気を予測する研究において，消費者の嗜好が考慮された事例は少なかった．

本研究は，人気予測において，予測対象についてのユーザの嗜好に基づくジャンル，流行した年代といった多面的な情報を考慮している点が新しい．

2.2 Wikipedia における編集行動

オンライン百科事典 Wikipedia は，幅広い内容を網羅しており，多くのユーザにより利用および編集されるソーシャルメディアであり，多くの研究がなされている．具体的には，大量の知識データおよびリンク構造によるそれらの関係性を知識ベースとして用いる研究 [19], [26] や Wikipedia に関わる社会的側面を分析した研究 [27], [6] などが存在する．

Wikipedia に関する研究の中でも，特に編集や編集行動に焦点を当てた研究も存在している．Wikipedia には，登録ユーザと未登録の一般ユーザが存在しており, [21] では，Wikipedia の登録ユーザに対し，アンケート調査を行い，彼らのモチベーションを調べている．また, [27] では，登録者の編集行動における社会的な役割を分析している．人気予測の文脈では，ページの編集回数を予測における素性として用いた, [31] が存在する．

2.3 多面的情報をもつベクトル表現の獲得

系列データを入力として，その各要素について単語における意味表現のような多面的な情報を取得する手法が注目されている．代表的な事例としては，単語の分散表現を獲得する Word2vec [18] およびその研究から派生し，文章の表現を獲得する [14] が存在し，近年自然言語処理の分野で大きな成果をあげている．Word2vec により学習されたベクトル表現は，単語の意味表現を保持したものとなったことが示されており，学習されたベクトル表現は様々なタスクに利用されている．

自然言語処理以外の分野においても，ベクトル表現を学習する手法が提案されている．一例として，グラフ構造を入力として与え，グラフ内の各ノードのベクトル表現を獲得する [11] らの Node2vec がある．また [4] では，ユーザの商品閲覧履歴から商品に関するベクトル表現を学習し，推薦問題に用いている．

2.4 リンク構造による指標

ウェブページのリンク構造から各ページの重要度，質，人気を推定する方法が提案されている．代表的な手法は，被リンク数の多いページまたは重要なページからリンクが貼られているページを重要なページであると考え，上位に位置づけるページランク [22] である．ページランクは，ウェブページの人気・注目度の指標として，Google の検索エンジンに導入され，大きな成果を上げた．

リンク構造を用いて人気・重要度の指標とする事例はウェブページ以外でも存在する．ブログにおけるリンク構造は，冪乗則に従う [15] ことが知られており, [28] では，リンク構造を用いてブログのランク付けを行なっている．また, [12] では Wikipedia における各ページの被リンク数を Wikipedia における人気の指標として扱っている．

3. 提 案 手 法

本章では，Wikipedia 上でのユーザの編集履歴をもとにコンテンツのベクトルを取得し，コンテンツの人気予測に用いる提案手法について説明する．

3.1 コンテンツベクトルの取得

Wikipedia の編集履歴からコンテンツベクトルを学習する手法について述べる．ユーザは自身の興味関心に基づいてページを編集すると考えられる．またユーザの編集系列において隣接するコンテンツは，ユーザの興味が似ているコンテンツであると考えられる．

そこでユーザごとに編集履歴を時系列に並べ，あるコンテンツの前後に出現するコンテンツが与えられた場合に元のコンテンツを予測するように学習を行う．より厳密には，あるコンテンツの系列 $c_1, c_2, c_3, ..., c_T$ が与えられた際に，c_t をその前後に存在するコンテンツ $c_{t-k}, ..., c_{t+k}$ により予測できるように各コンテンツ c のベクトル表現を学習する．ここで，各コンテンツ c を単一のベクトルにマッピングする行列を C とする．コンテンツの系列が与えられた時，以下の平均対数確率を最大化することでコンテンツベクトルを学習する．

$$\frac{1}{T}\sum_{t=k}^{T-k}\log p(c_t|c_{t-k}, ..., c_{t+k}) \tag{1}$$

予測タスクは通常，ソフトマックス関数に代表される多クラス分類によって行う．

$$p(c_t|c_{t-k}, ..., c_{t+k}) = \frac{\exp(y_{ct})}{\sum_i \exp(y_i)} \tag{2}$$

ここで y_i は，各コンテンツ c_i についての正規化されていない対数確率であり，

$$y = b + Uh(c_{t-k}, ..., c_{t+k}; C) \tag{3}$$

で算出される．U および b はソフトマックス関数のパラメータであり，h は C から得られるコンテンツのベクトル表現の平均値である．

3.2 人 気 予 測

人気予測は，Wikipedia における各コンテンツ作品ページの被リンク数を人気の指標として，Wikipedia から得られる素性およびコンテンツのベクトル表現をモデルに与えることで行う．Wikipedia から得られる素性として，ページ閲覧数，編集回数，被リンク数を用いる．ページ閲覧数および編集回数は日次のデータとして取得し，その平均値を月次データとする．一方，被リンク数は日次の変動が小さいため当該月の初日の値を記録し用いる．学習時には，コンテンツのベクトル表現と月次の素性からの特徴量を合わせて人気の予測を行う．

月次の素性およびコンテンツベクトルからの特徴量の抽出には，多層ニューラルネットワーク（MLP）を用いる．あるコンテンツ c についてのある月 t における月次の編集回数を e_c^t，ページ閲覧数を v_c^t，被リンク数を l_c^t とする．月次の素性 $\boldsymbol{X}_{c,M}^t$ を

$$\boldsymbol{X}_{c,M}^t = [e_c^t, v_c^t, l_c^t] \tag{4}$$

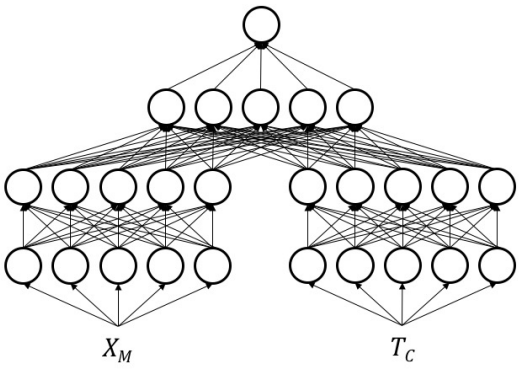

図 2　コンテンツ人気予測を行うニューラルネットワークの構造

とし図 2 に示すように MLP に与える．

また，コンテンツの系列を用いて学習された C から得られる，c についてのベクトル表現を T_c とする．月次の素性を与える MLP とは別に T_c を入力として与える MLP を作成する．それぞれの MLP から得られた出力を，新たな MLP への入力とし，最終的な出力を得る．

モデルに対して，$\boldsymbol{X}_{c,M}^t$ および T_c，正解ラベル y を与え，学習を行う．

4. 予 備 実 験

本章では，提案手法を用いる際に検証すべき点の前提条件についての分析を行う．実験は日本語版の Wikipedia を用いて行い，人気予測対象はアニメ，漫画，ゲームなどのコンテンツ作品を対象とした．提案モデルを用いて人気予測を行う際の前提としている，Wikipedia の編集履歴の系列においてユーザの嗜好が反映されていること，編集履歴を用いて学習したコンテンツベクトルがジャンルや流行した年代といった多面的な情報をもつことを検証した．合わせて，実験に用いる際に行なった前処理についても説明する．検証および人気予測には 2015 年 9 月から 2016 年 7 月末までの期間に得られたデータを用いた．

4.1 デ ー タ

Wikipedia のデータは，MediaWiki(注1) から取得した．

1 時間ごとのページ閲覧についてのダンプデータから日本語のページのページ閲覧数を取得した．取得した 1 時間ごとのデータから 24 時間分のページ閲覧数の合計値を算出し，日次のページ閲覧数のデータとした．また編集履歴情報から各ページが編集された回数を日次で算出した．

月次のデータとして，当該月の前月におけるページ閲覧数の平均値および平均の編集回数を算出した．加えて，毎月初日における各ページへの被リンク数を月次のデータとして算出した．すなわち 2015 年 10 月の各素性の値として，2015 年 9 月 1 日から 30 日までのページ閲覧数および編集回数の平均値から平均ページ閲覧数および平均編集回数を算出し，2015 年 10 月 1 日時点での被リンク数を算出した．

(注1)：https://dumps.wikimedia.org/

編集者によるページの編集履歴は以下のように取得した．コンテンツベクトルの学習に耐えうる十分に長い系列長を確保するため，編集履歴は 2015 年 8 月末までにアニメ・漫画・ゲームカテゴリに属するページの編集回数が多く，Wikipedia の編集者として登録されているユーザによる編集系列を選択した．対象としてアニメ・漫画・ゲームカテゴリに属するページの編集者と各編集者の編集回数はジップの法則に従って分布していた．そこで，編集回数が全体の上位 5 %にあたる 2,500 人の編集者を選択し，解析の対象とした．選択されたユーザによる最低の編集回数は，86 回であった．

選択されたユーザの編集系列を用いてタイトルベクトルの学習を行った後，作成した日次および月次のデータを用いて，将来の人気予測を行った．

4.2 データに対する前処理

4.2.1 タイトルの統合

長期間連載されている作品の Wikipedia ページでは，メインページ以外にもキャラクターページや関連作品ページなどが存在する．元々は一つのページに記述されていた内容が複数のページに分散されて記載されるため，そのようなタイトルのメインページは，関連ページに情報が分散しページの閲覧数や被リンク数が減少する．そのため，見かけ上のページ閲覧数や被リンク数が少なくなってしまうという問題が生じる．同一タイトルに関するページであるのでそれらを統合することでより現実に近い素性となると考えられる．この問題を解決するため，Wikipedia の関連ページを大元のページに対する構造をユーザが記述する Path Navi を利用する．解析対象とした全ページの中で Path Navi が存在するページがある場合，Path Navi を解析しメインページとの対応を取得する．Path Navi が存在する場合，予測に用いる編集回数，ページ閲覧，被リンク数をメインページの値に加えることで，見かけ上の値が減少する問題に対処した（Path Navi が存在しない場合，統合はなされない）．

4.2.2 解析対象とする作品の選択

本研究では，アニメ・ゲーム・漫画を人気予測の対象とした．予測対象とするタイトルの取得は，二段階に分けて行った．

まず，Wikipedia のカテゴリー「アニメ」,「漫画」,「ゲーム」の下に存在するタイトルのうち「○○の一覧」といったページを除いて取得した．続いて，タイトルの統合を行い，各タイトルごとに分散しているページごとの指標を一つにまとめた後，対象とした期間の被リンク数の平均上位 2,000 件を取得した．その後，低次元にベクトル化でき，対象とした期間に編集が一回以上行われているという条件を満たすタイトルに絞り，さらに各月における被リンク数の変動が前月比 10%以内の作品を選択した．これは，被リンク数に急激な変動が起こる場合は，今回のモデルによる予測対象外としたためである．絞り込みの結果，解析対象となった作品は 1,547 件となった．

4.3 編集者系列の定性分析

本研究では，登録されている編集者のうち 2015 年 10 月以前の編集回数上位 2,500 人に着目した．ベクトル化を行う前提として，編集系列にユーザの嗜好が反映され，何らかの形で類似するタイトルが，編集系列上にて隣接することが求められる．そこで，今回対象とした編集者からランダムに 5 名を選択し，その編集系列の一部について定性的な分析を行った．

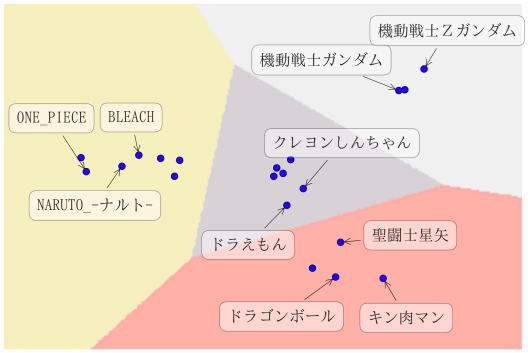

図 3　コンテンツベクトルのクラスタリング結果

ベクトル化の際には，あるタイトルのベクトル表現を得るために周辺の 10 のタイトルを用いている．そこで 10 の系列長の編集履歴の断片を選択し，分析を行った．表 1 に対象とした編集者の中からランダムに選択した編集者の編集系列からランダムに位置を指定し，その後編集された 10 のタイトルを選択した．

5 つの系列を選択した結果，系列 1 では,「ルパン三世」の関連作品および「金田一少年の事件簿」といった作品が並んだ．系列 2 では,「ワンパンマン」や「斉木楠雄のΨ難」など 2010 年代のギャグ漫画が並んだ．系列 3 では,「ジョジョの奇妙な冒険」関連作品とアメコミの関連のゲームが多く見られた．系列 4 では，手塚治虫の作品が並び，系列 5 では「コードギアス」シリーズの関連ページが並んでいた．

以上をまとめると，編集系列において隣接するタイトルは，ジャンル，作者，流行した年代，商品としてのカテゴリーなどが類似している傾向が見られ，ユーザの嗜好を反映していることが示唆された．この結果から，編集系列において隣接するタイトルのベクトル表現により対象とするタイトルのベクトルを表現できることが示唆された．

4.4 コンテンツベクトルの取得

Wikipedia の編集者による編集履歴から各コンテンツのベクトル表現を学習する方法について述べる．人気予測の対象としたカテゴリに属するページの編集回数が多い編集者の編集系列を用いた．編集回数が少ない編集者のデータからの学習は難しいため，全 50,000 人の上位 5 %に当たる編集回数上位 2,500 人の編集系列をベクトル化の対象とした．選択された編集者の編集回数は 86 回～20,793 回であった．

各コンテンツを単語と見なし，編集系列においてあるコンテンツの前後に現れるコンテンツを予測するように学習した．epoch 数は 50,000 回とし，window サイズは 10，編集系列における出現数が 10 回以上のコンテンツをベクトル化した．ベクトル化の際のモデルには Continuous Bag of Words（CBOW）[17] を用いた．

学習を行った後，得られたベクトルについて，定性的な分析

表 1　コンテンツ作品についての編集履歴（例）

	editor1	editor2	editor3
1	Lupin the Third	One-Punch Man	Marvel vs Capcom (Game)
2	The Kindaichi Case Files	Chagecha	JoJo's Bizarre Adventure (Game)
3	Lupin the Third and Detective Conan	The Disastrous Life of Saiki K.	JoJo's Bizarre Adventure (Manga)
4	Lupin the Third (Movie)	Chagecha	Deleted page
5	Lupin the Third and Detective Conan	The Disastrous Life of Saiki K.	Street Fighter: Sakura Ganbaru!
6	Lupin the Third (Movie)	Blue Exorcist	Oh! Edo Rocket
7	Lupin the Third and Detective Conan	The Disastrous Life of Saiki K.	Spider-Man
8	Lupin the Third (Movie)	Assassination Classroom	X-Men vs. Street Fighter
9	The Kindaichi Case Files	The Disastrous Life of Saiki K.	Marvel Super Heroes (video game)
10	Lupin the Third (Movie)	NARUTO (Movie)	X-Men: Children of the Atom (video game)

	editor4	editor5
1	Rainbow Parakeet	Code Geass (Character)
2	Neo Faust	Code Geass (Character)
3	Phoenix (manga)	Code Geass (Anime)
4	The Vampires (manga)	Code Geass (video game)
5	Pipi-chan	Code Geass (Character)
6	Ambassador Magma	Code Geass (Character)
7	Mako, Rumi and Chii	Code Geass (Manga and Novel)
8	Microsuperman	Code Geass (Character)
9	Dororo	Code Geass (Character)
10	Astro Boy	Code Geass (Character)

表 2　コンテンツベクトル（例）

original title	SLAM DUNK	Dragon Ball (Anime)	NARUTO (Manga)	Doraemon
1	Yu Yu Hakusho	Dragon Ball (Manga)	NARUTO (Computer game)	Doraemon (Movie)
2	Touch (manga)	Dragon Ball (Anime; Special)	NARUTO (Movie)	Crayon Shin-chan
3	Dr. Slump	Dragon Ball (Movie)	FAIRY TAIL	Doraemon (Movie)
4	Kimagure Orange Road	Dr. Slump	NARUTO (Computer game)	Doraemon (Movie)
5	I"s	Dragon Ball (Anime; Special)	D.Gray-man	21emon

を行った．学習されたベクトルの有効性を検証するため，コンテンツを 4 件選択し，それぞれのコンテンツと最も類似しているコンテンツ上位 5 件を取得した．上位 5 件には，クエリとしたコンテンツに関連する作品や同時期に放映・連載された作品，似た傾向にある作品が選ばれた．選択したコンテンツと類似のコンテンツをまとめたものを表 2 に示す．

また，図 3 に被リンク数上位 20 件のコンテンツに対してカーネル主成分分析を行い，二次元に射影した結果を示す．左の領域には，ポケットモンスター，ONE PIECE，NARUTO といった 10 代後半から 20 代にかけて人気のあるコンテンツが並んだ．中央の領域には，クレヨンしんちゃんやドラえもんといった子供向けのコンテンツが多く見られた．右上の領域には，機動戦士ガンダムおよび機動戦士 Z ガンダムとガンダムシリーズの作品が位置し，右下の領域には，ドラゴンボール，聖闘士星矢，Dr. スランプなど 80 年代から 90 年代にかけて流行したコンテンツが位置した．各領域ごとに対象年齢や放送していた年代といった性質が類似するコンテンツが置かれていた．この結果から学習されたベクトルは各作品のジャンルや流行した年代といった情報を内包していることがわかった．

5. 実　　験

本章では予備実験の結果，ジャンルや流行した年代といったコンテンツの多面的な情報を捉えていることが示唆されたコンテンツベクトルを用いて，コンテンツの人気予測を行う．また，ベースラインとしてコンテンツベクトルを用いない場合の精度を算出し，コンテンツベクトルを用いることにより予測精度が有意に向上することを示す．

5.1 人 気 予 測

人気予測は，Wikipedia における各コンテンツ作品ページの被リンク数を人気の指標として，Wikipedia から得られた素性とコンテンツベクトルを入力として与えたモデルにより行う．被リンク数は，ウェブ上のページの重要度の指標 [22] やブログの人気度の指標 [28] として用いられる．そこで本研究では, [12] をもとに，Wikipedia における各ページの被リンク数をコンテンツの人気の指標とした．

予測モデルには，図 2 に示すように多層ニューラルネットワーク (MLP) に月次データ $\boldsymbol{X}^t_{c,M}$ およびコンテンツベクトル T_c を入力として与えた．また，ベースラインとしてコンテンツベクトルを用いないモデルによる予測を置き，提案手法と比較

表 3　コンテンツ人気予測の結果

# Seed	Proposed (MAPE × 100)	Baseline (MAPE × 100)
1	0.148	0.150
2	0.148	0.154
3	0.149	0.150
4	0.148	0.151
5	0.147	0.151
6	0.149	0.157
7	0.150	0.150
8	0.147	0.150
9	0.149	0.153
10	0.150	0.153
Ave.	0.148	0.152

した．

月次の素性を扱う MLP は，各層のユニット数を 8 とし，2 層の構造とした．また，コンテンツベクトルを扱う MLP は，各層のユニット数を 32 とし全 2 層の構造で，各層間にドロップアウト層を挿入した．両 MLP の出力を結合し，ユニット数 24 の MLP 層に入力として与え，最終的な出力を得た．最終層を除く層の活性化関数には ReLU を用い，最終層では線形関数を用いた．なお，コンテンツベクトルを用いない場合には，上記のモデルからコンテンツベクトルを扱う部分を除いた MLP を用いた．

学習は誤差逆伝播法にて行い，最適化は RMSprop を用いた．学習率は 0.0001 とし，エポック数は 800 として，早期終了を用いて実験を行った．

解析の対象としたコンテンツについて，2015 年 10 月から 2016 年 3 月までの 6ヶ月分のデータを学習データ，2016 年 4 月から 7 月までの 3ヶ月分のデータをテストデータとした．学習データを用いてモデルの学習を行った．テストデータをモデルに与え，コンテンツごとに 3ヶ月分の予測を行い，実際の値からのずれを平均絶対パーセント誤差（MAPE）にて評価した．代表的なコンテンツ作品であるアニメの放映は，3ヶ月を 1 クールとしているため，本研究では予測期間を 3ヶ月と設定した．

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_{true} - y_{pred}}{y_{true}}\right| \tag{5}$$

コンテンツベクトルを用いた場合と用いない場合について，全コンテンツについての MAPE の値の平均値を 3 に示す．ランダムシードを固定し，10 回の実験を行った．コンテンツベクトルを用いた場合，用いない場合と比較して平均の MAPE の値は 2.3 %低下していた．また，t 検定による p 値は 0.0014 であり，コンテンツベクトルを用いることで有意に MAPE の値が低下していた．

5.2　予測精度の精度の高いクラスタ

MSE を算出した結果，コンテンツベクトルを用いない場合と比較して精度が上がるコンテンツと上がらないコンテンツが存在した．両者の違いを分析するため，コンテンツベクトルによるクラスタリングを行い，各クラスタごとの平均の MAPE 値を算出し，コンテンツベクトルを用いる場合と用いない場合の変化を比較した．表 4 に，各クラスタに含まれるコンテンツの数とコンテンツベクトルを用いない場合に対する平均の MAPE の改善を百分率で示した．クラスタリングは KMeans にて行い，セントロイドの数は全コンテンツ数 1,547 に対して 15 とした．

表 4　クラスタごとの MAPE

cluster id	ratio	number of contents	cluster id	ratio	number of contents
1	19.73	61	9	4.36	31
2	5.27	318	10	-36.09	22
3	1.00	119	11	3.29	183
4	17.72	22	12	-2.07	31
5	-5.64	120	13	11.96	31
6	18.92	1	14	1.36	246
7	2.66	132	15	4.81	23
8	3.20	192			

クラスタリングを行った結果，各クラスタに含まれるコンテンツ数はクラスタ 6 を除いて 22 から 318 の間であった．クラスタに含まれるコンテンツ数が 50 以上であり，コンテンツベクトルを用いた場合と用いない場合の MAPE の平均値の改善が大きい 5 つのクラスタ 1，2，5，8，11 に着目し，それぞれに含まれる被リンク数の多いコンテンツ 20 件を調べた．

クラスタ 1 には「機動戦士ガンダム」シリーズの作品のみが含まれており，コンテンツベクトルを用いることで 19.73 %精度が向上していた．クラスタ 2 では,「ひぐらしのなく頃に」や「エウレカセブン」といったコンテンツが含まれており，MAPE の改善は 5.27 %であった．クラスタ 8 では,「とある魔術の禁書目録」や「鋼の錬金術師」などのコンテンツが含まれ，改善は 3.20 %であった．クラスタ 11 では,「ゲゲゲの鬼太郎」,「銀河鉄道 999」,「サイボーグ 009」を始めとするコンテンツが含まれ，3.29 %向上していた．最後にクラスタ 5 では,「進撃の巨人」や「妖怪ウォッチ」といった作品が含まれ，5.64 %予測精度が低下していた．

6. 考　　察

この章では，編集者の編集系列，コンテンツのベクトル化，人気予測およびクラスタごとの人気予測の精度の変化について，得られた実験結果から考察する．

6.1　編集履歴の系列

コンテンツをベクトル化する際の前提として，入力される系列において隣接するコンテンツは類似していることが求められる．Wikipedia の編集者は，自身が詳しいページについて編集を行うため，編集者の系列において隣接するコンテンツは何らかの形で類似していると考えられた．解析の対象とした編集者の編集履歴系列からランダムに一部を取り出し，実際に隣接するコンテンツが類似しているか定性的に分析した．

その結果，隣接するコンテンツは同一コンテンツの関連ペー

ジである場合やジャンルが同一であるコンテンツ，作者が同一であるコンテンツなど関連があることが明らかとなった．この編集系列を用いることにより学習されるコンテンツのベクトル表現は，ジャンルや作者についての情報，流行した年代といった情報を保持するように学習されていた．

6.2 ベクトル化

編集者履歴から得られたコンテンツベクトルについて，定性的な分析を行った．表 2 に示すように，学習されたベクトル空間において，あるコンテンツの周囲に位置するコンテンツは，何らかの関わりがあった．最も多い関係性が，対象としたコンテンツとその関連作品という関係性であった．また，作者が同一人物である場合もベクトル空間上の近い位置に存在した（Dragon Ball と Dr. スランプ，ドラえもんと 21 エモン）．加えて，同時代に連載・放映されたコンテンツが近くに位置していた（SLAM DUNK と幽遊白書）．その他にも，ジャンル・対象年齢が近いコンテンツが選択される傾向にあった．

また，図 3 に対象としたコンテンツのうち，被リンク数上位 20 件のコンテンツのベクトルを 2 次元に射影した図を示した．各コンテンツについてカーネル主成分分析により得られた第一成分および第二成分の値を用いて KMeans クラスタリングを行った．その結果，一つ目のクラスタにはポケットモンスターや NARUTO，ONE PIECE といった 10 代後半から 20 代男性を中心に幅広い層に人気のある作品が位置した．二つ目のクラスタは，クレヨンしんちゃんやドラえもんといった広い年齢層に親しまれ，比較的古くから存在し近年にもアニメが継続的に放映されているまたは近年映画化した作品が位置した．3 つ目のクラスタには，機動戦士ガンダムに関連するシリーズが位置した．ガンダムシリーズは，他のクラスタからの距離も大きく，ユーザ層が異なることが示唆された．実際にガンダムシリーズは，中高年の男性からの支持が大きいことが知られている．最後のクラスタには，ドラゴンボールや聖闘士星矢など 80 年代から 90 年代初頭にかけて流行したコンテンツが位置した．各クラスタごとにそれぞれ流行した時代，主な支持層による違いが顕著に現れていると考えられる．以上の結果から，編集履歴を用いて学習したコンテンツベクトルは，ユーザの嗜好に基づくコンテンツの多面的な情報を表現していると考えられる．

6.3 人 気 予 測

解析対象としたコンテンツについて，3ヵ月後の Wikipedia の被リンク数を予測した．モデルには月次の入力素性とともに，編集者履歴から学習したコンテンツベクトルを入力として与えた．その結果，コンテンツベクトルを使用しない場合と比較して，予測精度は 2.3 %向上した．これは，対象としたコンテンツに関する情報をコンテンツベクトルとして入力することで，対象のコンテンツに対するユーザの嗜好がモデルに組み込まれたからであると考えられる．コンテンツベクトル以外のモデルに与えるページ閲覧数の推移や編集回数，当月の被リンク数といった特徴量は，それぞれ短期的な変動を捉えていると考えられる．しかしながら，対象としたコンテンツを支持するユーザ層や長期にわたり形成されるコンテンツに対するユーザの嗜好を考慮することができていない．コンテンツベクトルは Wikipedia の編集履歴から学習されており，長期にわたり形成されてきたコンテンツに対するユーザの嗜好や対象のコンテンツを支持するユーザ層の情報をモデルに与えていると考えられる．その結果，将来の人気予測においてコンテンツベクトルをモデルに与えることで予測精度が改善したと考えられる．

6.4 予測に寄与する素性

コンテンツをクラスタリングし，予測精度が上がるクラスタと低下するクラスタを調べた．精度の変化が大きいクラスタのうち 50 以上のコンテンツを含むクラスタに着目し，各クラスタに含まれる被リンク数の多い作品上位 20 件を確認することで分析を行った．精度が最も向上したクラスタ 1 に含まれる作品上位 20 件すべてが「機動戦士ガンダム」に関連する作品であった．ガンダムシリーズは，40 年代の男性を中心に非常に強い人気がある．シリーズ作品が一つのクラスタを形成しており，またコンテンツベクトルを与えることで予測精度が大幅に向上したことから，このクラスタには固有の傾向があると考えられる．クラスタ 2 および 8 には，少年漫画および少年向けアニメの人気作品が数多く含まれていた．これらの作品は，10 代後半から 20 代の男性を中心としたユーザ層に支持されていると考えらる作品である．コンテンツベクトルを用いることで予測精度は，両クラスタそれぞれ 5.27%および 3.20%向上していた．クラスタ内に存在する作品は，クラスタ内の作品すべてがガンダムシリーズであったクラスタ 1 ほど傾向が似ていないため予測精度の向上幅が小さくなったのではないかと考えられる．またクラスタ 11 には「サイボーグ 009」や「銀河鉄道 999」といった昭和期の作品が主に含まれていた．これらの作品は，中高年からの支持を得ていると考えられ，コンテンツベクトルによりそのような情報を与えられたことで精度が向上したと考えられる．最後に，精度が大幅に低下したクラスタ 5 について述べる．クラスタ 5 に含まれる作品の多くは，テレビアニメ，映画，ゲーム，漫画と幅広く様々なメディアに展開された作品であった．このため，ゲーム，映画など各メディアにおける支持層が広く存在し，結果として予測が難しくなったのではないかと考えられる．

以上をまとめると，コンテンツベクトルを用いる場合，作品のジャンルや支持するユーザ層がコンテンツベクトルに強く含まれるほど予測精度は向上する．逆に，複数のメディアにて展開され，支持するユーザ層が幅広くなった場合には予測精度が低下すると考えられる．

7. お わ り に

本研究ではウェブ上でのユーザの行動履歴を用いて，ユーザの嗜好に基づくジャンルをはじめとするコンテンツの多面的な情報を学習し，それらを用いてコンテンツの人気予測を行った．具体的には，Wikipedia におけるユーザによる編集履歴に対して，Word2vec を応用した低次元ベクトル表現の獲得手法を適用した．その後，新たなニューラルネットワークに学習したベクトル表現を与えることで，人気予測の精度が向上することを示した．今回の実験では，ウェブ上でのユーザの行動履歴として Wikipedia での編集履歴を用いた．しかし，提案手法はユー

ザが興味を持っていると考えられる対象物についての系列データを入手することができれば適用可能である．したがって編集履歴に限らず，レビューサイトにおけるレビュー文章の投稿履歴やオンラインショッピングサイトにおけるページの閲覧履歴などの幅広いユーザの行動履歴に対して適用可能であると考えられる．学習された結果は，対象物に対して興味を抱いているユーザ群の推定や類似・競合商品の特定にも用いることが可能であると考えられる．また，取得する系列データをユーザ群により分割することで，ユーザ群それぞれについてのベクトル表現を学習し，ベクトル群間における同一商品の位置付けの違いを分析することも可能であると考えられる．

文　献

[1] アニメ産業レポート 2016 サマリー（日本語版）1.1. http://aja.gr.jp/?wpdmdl=991.

[2] コンテンツ産業の現状と今後の発展の向性. http://www.meti.go.jp/policy/mono_info_service/contents/downloadfiles/shokanjikou.pdf.

[3] Fabian Abel, Ernesto Diaz-Aviles, Nicola Henze, Daniel Krause, and Patrick Siehndel. Analyzing the blogosphere for predicting the success of music and movie products. In *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, pp. 276–280. IEEE, 2010.

[4] Oren Barkan and Noam Koenigstein. Item2vec: Neural item embedding for collaborative filtering. *arXiv preprint arXiv:1603.04259*, 2016.

[5] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, Vol. 2, No. 1, pp. 1–8, 2011.

[6] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1101–1110. ACM, 2008.

[7] Yubo Chen, Qi Wang, and Jinhong Xie. Online social interactions: A natural experiment on word of mouth versus observational learning. *Journal of marketing research*, Vol. 48, No. 2, pp. 238–254, 2011.

[8] Hyunyoung Choi and Hal Varian. Predicting the present with google trends. *Economic Record*, Vol. 88, No. s1, pp. 2–9, 2012.

[9] Chrysanthos Dellarocas, Xiaoquan Michael Zhang, and Neveen F Awad. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive marketing*, Vol. 21, No. 4, pp. 23–45, 2007.

[10] Tal Garber, Jacob Goldenberg, Barak Libai, and Eitan Muller. From density to destiny: Using spatial dimension of sales data for early prediction of new product success. *Marketing Science*, Vol. 23, No. 3, pp. 419–428, 2004.

[11] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[12] Tomáš Kliegr, Vojtech Svátek, Krishna Chandramouli, Jan Nemrava, and Ebroul Izquierdo. Wikipedia as the premiere source for targeted hypernym discovery. *Wikis, Blogs, Bookmarking Tools Mining the Web 2.0*, p. 38, 2008.

[13] Ren J Kuo and KC Xue. A decision support system for sales forecasting through fuzzy neural networks with asymmetric fuzzy weights. *Decision Support Systems*, Vol. 24, No. 2, pp. 105–126, 1998.

[14] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, Vol. 14, pp. 1188–1196, 2014.

[15] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM international conference on data mining*, pp. 551–556. SIAM, 2007.

[16] Yong Liu. Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing*, Vol. 70, No. 3, pp. 74–89, 2006.

[17] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[19] David Milne and Ian H Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 509–518. ACM, 2008.

[20] Gilad Mishne, Natalie S Glance, et al. Predicting movie sales from blogger sentiment. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 155–158, 2006.

[21] Oded Nov. What motivates wikipedians? *Communications of the ACM*, Vol. 50, No. 11, pp. 60–64, 2007.

[22] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[23] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pp. 175–186. ACM, 1994.

[24] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, Vol. 311, No. 5762, pp. 854–856, 2006.

[25] Upendra Shardanand and Pattie Maes. Social information filtering: algorithms for automating "word of mouth". In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 210–217. ACM Press/Addison-Wesley Publishing Co., 1995.

[26] Michael Strube and Simone Paolo Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI*, Vol. 6, pp. 1419–1424, 2006.

[27] Howard T Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, pp. 122–129. ACM, 2011.

[28] Yi Wu and Belle L Tseng. Important weblog identification and hot story summarization. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 221–227, 2006.

[29] Sheng Yu and Subhash Kak. A survey of prediction using social media. *arXiv preprint arXiv:1203.1647*, 2012.

[30] Xiaohui Yu, Yang Liu, Xiangji Huang, and Aijun An. Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Transactions on Knowledge and Data engineering*, Vol. 24, No. 4, pp. 720–734, 2012.

[31] 保住純, 飯塚修平, 中山浩太朗, 高須正和, 嶋田絵理子, 須賀千鶴, 西山圭太, 松尾豊. Web マイニングを用いたコンテンツ消費トレンド予測システム. 人工知能学会論文誌, Vol. 29, No. 5, pp. 449–459, 2014.