1D2-2

RNNを利用したコンテンツ産業の消費トレンド予測

Contents Trend Prediction Method by RNN

野中　尚輝 *1　　中山　浩太郎 *1　　松尾　豊 *1
Naoki Nonaka　　Kotaro Nakayama　　Yutaka Matsuo

*1東京大学工学系研究科技術経営戦略学専攻
University of Tokyo Graduate School of Engineering

We propose a method to predict consumption trends from web data by Recurrent Neural Network(RNN). Recently, Japanese media contents (such as anime and manga) are gaining more and more attention internationally. But in terms of business it is not making successful results. The main reason is the difficulties of percieving local consumer consumption trends. In web mining field, many researches were conducted to predict trends using features extracted from web. Thus, in this research we aim to predict trend of Japanese media contents by using data extracted from web. Specifically, we calculate consumption trend based on sales ranking of online commerce site, and use features extracted from Twitter and Wikipedia. In this paper, we report result of trend prediction using several RNN models and compare its accucracy to SVR. Results shows that prediction with RNN is better accuracy compared to prediction with SVR.

1. はじめに

近年，日本のアニメ・マンガ作品が日本国内にとどまらず多くの国で人気を集めている．しかしながら，これらのコンテンツはビジネスの観点では十分な得られていないという点で海外展開は成功しているとは言えない．[杉山 06] では，このような状況に陥った原因の一つとして海外輸出をビジネスとして管理してこなかった歴史的背景を挙げている．これらの管理を行い，適切な意思決定を行うことが重要となる．意思決定材料として，自身の扱う製品の消費動向や製品の人気および潜在需要の変動といった消費トレンドは，それらを扱う企業や国にとって重要な要素である [Kuo 98], [保住 14].

このような状況の下ウェブマイニングの分野においては，映画を中心としてさまざまな商品やサービスの消費トレンドを予測する研究が行われてきた．特に映画の興行収入の予測に関連する研究は活発に行われており，レビューにおける評判をもとに映画の興行収入を予測する研究 [Yu 12] や Twitter におけるツイートをもとに映画の興行収入を予測する研究 [Asur 10] がこれまでに行われてきた．これらの研究において，トレンドの予測には各種の機械学習を用いるモデルが提案されており，特にサポートベクトル回帰を用いるモデル [保住 14] などが成果を出してきた．しかしながら，販売ランキングをはじめとするウェブから得られるデータは欠損値を含むことが多く，扱いが難しいという問題点があった．

近年，深層ニューラルネットワークの手法が様々なタスクにおいて成果をあげ，大きな注目を集めている．中でも系列データに対する適用では，再帰的ニューラルネットワーク（RNN）が大きな成果を出している．RNN は，過去の系列における入力も考慮できるため，RNN を用いることで欠損値に対して頑健なモデルを構築することが可能であると考えられる．

以上の背景から本研究では，ウェブマイニングにより得た素性に対して RNN を適用し，トレンド予測を行う手法を提案する．具体的には，幅広い作品を網羅する Wikipedia から得られる素性と Twitter から得られる素性を組み合わせ，オンラインショッピングサイトにおける日次の販売ランキングをもとにトレンドの予測を行う．

本論文は以下のように構成される．2 章にて関連研究に触れた後，3 章にて今回の提案手法について述べる．続いて 4 章にて実験概要とその結果について説明し，最後に 5 章にて結論を述べる．

連絡先: 野中尚輝，東京大学工学系研究科技術経営戦略学専攻
松尾研究室，nonaka@weblab.t.u-tokyo.ac.jp

2. 関連研究

この章では本研究に関連する研究について述べる．

2.1 ウェブマイニングによるトレンド・販売予測

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情報を発信することが可能になった．その結果，ウェブ上には多くの情報が存在するようになった．ウェブマイニングはそのような大量のウェブ上のデータから，データマイニングの技術を用いてそれまでに知られていない情報や知識を発見するプロセスを指す [Kosala 00]．ウェブマイニングの研究の一つとして，検索エンジンにおけるクエリの情報，Wikipedia，Twitter，ブログやそれらの組み合わせなどの素性を用いてトレンドの予測を行う研究が存在する．

Twitter からトレンドの予測を行う研究では，[Asur 10] が Twitter 上での上映一週間前における単位時間あたりの映画に対する言及数から映画の興行収入を予測している．レビューを用いた研究としては，レビューサイトから得られた文章に対して感情分析を行い映画の収益を予測している [Yu 12] が存在する．検索クエリを用いた研究には，Google における特定のクエリでの検索回数を素性として月次の自動車および自動車部品の販売数などを予測する研究 [Choi 12] が存在する．

Wikipedia や検索クエリ，Twitter など複数の素性を組み合わせてトレンドを予測する研究としては [保住 14] がある．この研究では，アニメ・マンガなどのコンテンツ作品の消費トレンドを予測している．予測における正解データとしてオリコンによるマンガの販売部数を用い，学習を行っている．しかしながら，マンガの販売部数などは月次のデータのみしか得られないため，良い学習モデルを構築するための十分なラベルデータを入手することが難しいという課題があった．

2.2 RNN を用いた系列データの予測

深層ニューラルネットワークは，画像分類をはじめとするタスクにおいて目覚ましい成果を出しており，大きな注目を集めている．系列データに対しての適用には，再帰的ニューラルネットワークを用いる研究が優れた成果あげている．系列データの予測では，与えられた文章に対する応答を生成する Seq-to-seq モデルを用いた対話生成の研究 [Vinyals 15] が存在する．この研究では，映画での会話データをもとに対話を学習した後，与えられた会話文に対して適切な応答をするモデルが提案されている．モデルは，入力として与えられる単語についての系列データを学習し，出力として適切な単語の系列を予測し返す．また，ウェブ上のデータへの適用としては，レビュー文章から投稿者の意見を抽出するオピニオンマイニングにおいて，研究 [Irsoy 14] が存在する．この他にも，RNN の応用事例として機械翻訳 [Cho 14] や株価の予測研究 [Rather 15] が挙げられる．

3. RNN による消費トレンド予測モデル

本章では本論文において提案する，消費トレンドを予測する手法について述べる．

3.1 消費トレンドの元データ

消費トレンドは，商品の人気や潜在的な需要をあらわす指標である [保住 14]．本研究において予測対象である消費トレンドは，オンラインショッピングサイト「楽天」から得られた日次データをカテゴリーごとの販売ランキングをもとに作成した．カテゴリーとしては,「少年マンガ」,「少女マンガ」,「青年マンガ」,「DVD」,「Blu-ray」を対象としてデータの収集を行った．オンラインショッピングサイトでは，出版元やジャンルなどを問わず幅広い商品を取り扱っており，またカテゴリーごとの販売ランキングが日次で API 形式で提供されている．そのため，長期間に渡り安定的に販売ランキングを取得可能であり，このデータにおける販売ランキングの推移を消費トレンドと見なせるためである．

3.2 データ

提案手法において扱う対象とするタイトルの選別と用いる各素性の収集方法と計測方法について述べる．得られたデータは，[保住 14] に記された方法に従ってクエリの前処理を行った後，trie 木を用いた検索により Wikipedia のページタイトルにひも付けた．

Twitter から得る素性は，期間内に指定したクエリを含むツイートの数とした．Twitter の公式 API であるストリーミング API を用いて全ツイートの 10 %を取得し，取得したツイートに含まれる各作品への言及回数を素性とした．クエリは Wikipedia のページタイトルおよびそのエイリアスを用いた．

Wikipedia から得る素性としては，各ページの編集回数を用いた．Wikipedia のダンプデータに含まれる各ページの編集履歴とそのタイムスタンプをもとに，ページごとに日次の編集回数を算出した．

消費トレンドは，楽天より得た日次のカテゴリーランキング情報をもとに作成した．具体的には，商品カテゴリー，ランキング，商品タイトルを取得した後，商品タイトルの文章に対して Twitter のデータに対する手法と同じ手法を用いて，商品タイトルと Wikipedia のページタイトルをひも付けた．販売ランキングには，同一のタイトルが複数含まれる場合がある（例えば,「進撃の巨人 10 巻」が 1 位,「進撃の巨人 9 巻」が 5 位といった場合がある）．このような場合に下位のランキングをトレンドスコアに反映させるため，ページタイトルごとに，あらかじめ定めた定数からランキングを引いた値の合計を算出し，トレンドスコアとした．

日次の形式で作成したデータを 7 日間分足し合わせることで，週次のデータとして予測実験に用いた．

3.3 モデルの学習

モデルの学習方法と比較手法の学習方法について述べる．

学習は，RNN により行った．RNN のユニットとしては，Long short-term memory（LSTM），Gated recurrent unit（GRU）および Fully-connected RNN（全結合 RNN）を用いた．ユニット数を 128 とし，損失関数は平均二乗誤差（MSE），活性関数は線形関数とした．学習のエポック数は 150 とした．ユニット数および学習のエポック数は事前の実験により決定した．

x_t を t 番目の入力，y_t を t 番目の出力とするとき，RNN は以下に定式化される．

$$\boldsymbol{a}^{(t)} = \boldsymbol{b} + \boldsymbol{W}h^{(t-1)} + \boldsymbol{U}^{(t)}\boldsymbol{x}^{(t)} \tag{1}$$

$$\boldsymbol{h}^{(t)} = tanh(\boldsymbol{a}^{(t)}) \tag{2}$$

$$\boldsymbol{o}^{(t)} = \boldsymbol{c} + \boldsymbol{V}\boldsymbol{h}^{(t)} \tag{3}$$

$$\boldsymbol{y}^{(t)} = \phi(\boldsymbol{o}^{(t)}) \tag{4}$$

ここで，ϕ は活性化関数，h_t は t 番目の隠れ層の出力であり，$\boldsymbol{W}$，$\boldsymbol{V}$，$\boldsymbol{U}$ はそれぞれ入力層から隠れ層，隠れ層から出力層，隠れ層から隠れ層への重み行列，$\boldsymbol{b}$ および $\boldsymbol{c}$ はバイアス項である．

データに含まれる Wikipedia のページタイトルのうち、7 割を学習セット、残りの 3 割をテストセットとした．データの節にて準備した素性を Wikipedia のページタイトルごとに時系列に並べ，N 週分のデータごとに区切り入力素性とした．すなわち Wikipedia のタイトル (e) についての入力素性として，Wikipedia の編集回数 $c_i^{(e)}$，ツイート数 $d_i^{(e)}$，トレンドスコア $s_i^{(e)}$ からなる 3 次元の週次データ $\boldsymbol{x}_i^{(e)}$ を N 週分並べた行列 $\boldsymbol{X}_i^{(e)}$ を一つの入力単位として準備した．出力のラベル $y_i^{(e)}$ としては，入力のデータにおける最新のデータである N 週目からさらに M 週後のトレンドスコア $s_{i+N+M}^{(e)}$ を用いた．

$$\boldsymbol{x}_i^{(e)} = [d_i^{(e)}, c_i^{(e)}, s_i^{(e)}] \tag{5}$$

$$\boldsymbol{X}_i^{(e)} = [x_i^{(e)}, \ldots, x_{i+N}^{(e)}] \tag{6}$$

$$y_i^{(e)} = s_{i+N+M}^{(e)} \tag{7}$$

学習セットに含まれるページタイトルについて，入力と出力のペア $(\boldsymbol{X}, y)$ を作成し、学習データとした．テストセットに含まれるページタイトルについても同様の操作を行い，テストデータとした．続いて，学習データを用いて学習したを行った後，学習されたモデルを用いてテストデータの入力素性から出力を得た．

出力として得られた値のうち，予測値が負の値となった場合には，その値を 0 として扱った．

3.4 モデルの検証

モデルの検証は，手法ごとに学習したモデルを用いて行った．

学習セットと同様にテストセットにおいても入力データを作成する．続いてページタイトルごとに，学習済みのモデルに与え，予測されるトレンドスコアを得る．実際のトレンドスコアと予測されたトレンドスコアを比較し，二乗平均誤差（MSE）

	Average MSE of test title (MSE $\times 10^2$)	Number of titles with smaller MSE than ν-SVR
全結合 RNN	**0.49**	**76**/110
LSTM RNN	0.54	65/110
GRU RNN	0.55	59/110
ν-SVR	0.53	-

表 1: 提案手法と SVR による予測精度の比較

によりタイトルごとに予測精度を算出する．その後，テストセットに含まれるページタイトルごとに手法ごとの予測精度を MSE にて比較し議論する．

4. 実験

MSE $\times 10^2$	全結合 RNN	ν-SVR
SLAM DUNK	**0.03**	0.16
となりのトトロ	**0.66**	0.73
ONE PIECE	3.03	**1.03**
宇宙兄弟	**0.87**	1.17
天元突破グレンラガン	**0.10**	0.25
BLEACH	0.85	**0.82**

表 2: タイトルごとの MSE 値の比較

この章では提案したモデルの有効性を比較実験の結果をもとに述べる．

4.1 予測対象としたデータ

実験は，2015 年 8 月から 2016 年 1 月末までの 6ヵ月間に得られたデータを用いて行った．対象としたタイトルは，Wikipedia においてマンガ・アニメ・ゲームに関連するカテゴリー（「漫画作品一覧」、「日本のアニメ作品一覧」など）に属するページタイトルを選択した．その後，対象とした期間内において販売ランキングにおける出現回数が少ないタイトルについては，正確な予測が難しいため予測の対象から外した．具体的には，予測対象期間におけるトレンドスコアの合計が 0.3 以上であるアニメ・マンガ・ゲームに関連する Wikipedia のページタイトルを予測の対象とした．また，Wikipedia の編集回数，ツイート数，トレンドスコアのそれぞれについて対数をとり正規化を行った．その後，対象としたタイトルについて学習データとテストデータに分割し，3 章の内容に従ってモデルを学習した．学習は，過去 4 週間分のデータを用いて，翌週のデータを予測するという条件（N = 4 および M = 1）にて，約 400 件のページタイトルについて行った．

比較手法としては，ν-サポートベクトル回帰（ν-SVR）を用いた．入力データとしては，学習データに含まれるページタイトルについて，Wikipedia の編集回数，ツイート数，トレンドスコアを時系列に並べた後，過去 4 週分からなる 12 次元のベクトルを用いた（N=4 として，3N 次元のベクトルを ν-SVR に与えた）．出力データとしては，ニューラルネットワークの場合と同様に，入力における最新のデータから 1 週後（M=1）のトレンドスコアを用いた．

4.2 トレンドの予測

トレンドの予測は RNN にて行い，予測されるトレンドスコアと実際のトレンドスコアを比較し，その精度を SVR と比較した．テストデータに含まれる 110 のタイトルについて，翌週のトレンドスコアを予測するという設定にて 26 週間分の予測を行った．それぞれの手法について，ページタイトルごとに予測されるトレンドスコアと実際のトレンドスコアとの MSE 値を算出し，ν-SVR における MSE 値と比較し手法の評価を行った．

各 RNN の手法について，比較手法である ν-SVR における MSE 値より小さいタイトル，すなわち予測精度が高くなっていたタイトルの数を算出した．表 1 に示すように，今回の実験条件ではいずれの RNN による手法も ν-SVR による予測精度を上回るタイトルが半数を超えた．RNN の手法の中では，全結合 RNN が最も良い結果となっていた．同様に，110 タイトルにおける MSE 値の平均値も全結合 RNN が最も良い結果を示し，ν-SVR を上回った．

予測された消費トレンドと実際の消費トレンドをプロットしたものを図 1 に示す（RNN の結果については，全結合 RNN のみを示す）．ここでは，110 のタイトルのうち代表的な 6 つのタイトルについてのみ結果を図示する．また，図 1 に示す各タイトルについて，全結合 RNN および ν-SVR における MSE 値を表 2 に示す.「SLAM DUNK」,「宇宙兄弟」,「天元突破グレンラガン」,「となりのトトロ」については RNN による予測精度が上回る一方,「ONE PIECE」および「BLEACH」については ν-SVR による予測の方が高くなっている．

5. 結論

本研究では，消費トレンド予測問題に対して LSTM や GRU を含む RNN のモデルを適用し，比較した．予測に用いる素性としては，ウェブから得た Twitter および Wikipedia の情報を用いた．また，消費トレンドをあらわす指標は，オンラインショッピングサイト「楽天」のカテゴリーごとの販売ランキングをもとに作成した．ν-SVR による予測結果と提案手法である RNN による予測結果を比較した結果，RNN による予測精度は ν-SVR による予測精度を上回っていた．消費トレンドを予測する際には，直近の変動のみならず変化の前触れをとられることが重要になる．このような前触れや長期の依存関係を捉えれることが，トレンド予測のモデルにおいては重要になる．

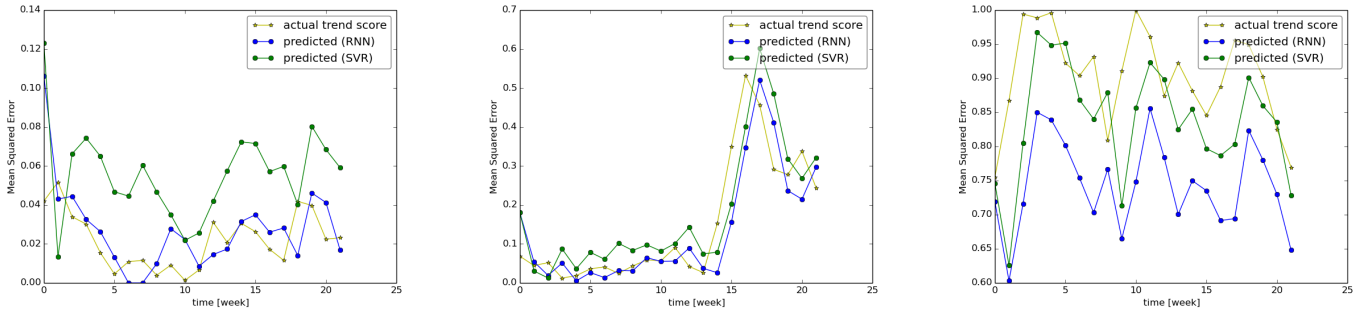
過去 4 週間のデータから翌週のトレンドスコアを予測するという今回の問題設定では，全結合の RNN を用いた場合の精度が最も良かった．より長期にわたる予測を行う場合には，長期の依存関係を捉えることのできる LSTM を用いたモデルの精度が全結合の RNN を上回る可能性がある．また RNN が ν-SVR の精度を上回った理由としては，RNN においては入力の素性が明確に区別されているが ν-SVR では 1 次元のデータとして与えている点が考えられる．

本研究で提案したモデルを拡張することで，今後より予測精度を高め，また長期にわたる予測を行っていきたいと考えている．加えて，*nu*-SVR と RNN の精度に関してより多くの条件にて実験を行い，各々の手法がどのような条件において効果的に用いることができるかを明らかにしたいと考えている．

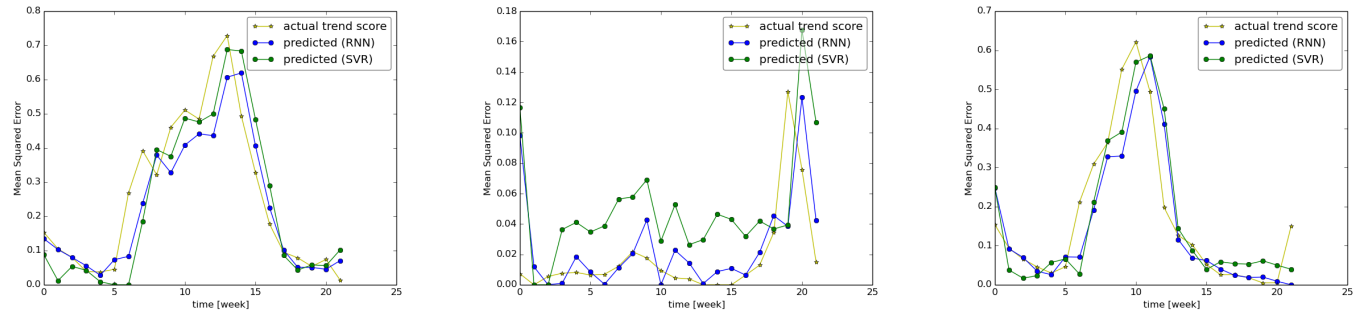
参考文献

[杉山 06] 杉山知之: クール・ジャパン 世界が買いたがる日本. 祥伝社 (2006)

[保住 14] 保住 純, 飯塚 修平, 中山 浩太郎, 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



(a) SLAM DUNK (b) となりのトトロ (c) ONE PIECE



(d) 宇宙兄弟 (e) 天元突破グレンラガン (f) BLEACH

図 1: RNN と SVR による予測の可視化結果

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. 449459 (2014)

[Kosala 00] KOSALA, Raymond; BLOCKEEL, Hendrik. Web mining research: A survey. ACM Sigkdd Explorations Newsletter, 2000, 2.1: 1-15.

[Asur 10] ASUR, Sitaram; HUBERMAN, Bernardo A. Predicting the future with social media. In: Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. IEEE, 2010. p. 492-499.

[Choi 12] CHOI, Hyunyoung; VARIAN, Hal. Predicting the present with Google Trends. Economic Record, 2012, 88.s1: 2-9.

[Kuo 98] KUO, Ren J.; XUE, K. C. A decision support system for sales forecasting through fuzzy neural networks with asymmetric fuzzy weights. Decision Support Systems, 1998, 24.2: 105-126.

[Vinyals 15] VINYALS, Oriol; LE, Quoc. A neural conversational model. arXiv preprint arXiv:1506.05869, 2015.

[Yu 12] YU, Xiaohui, et al. Mining online reviews for predicting sales performance: a case study in the movie domain. Knowledge and Data Engineering, IEEE Transactions on, 2012, 24.4: 720-734.

[Irsoy 14] IRSOY, Ozan; CARDIE, Claire. Opinion Mining with Deep Recurrent Neural Networks. In: EMNLP. 2014. p. 720-728.

[Cho 14] CHO, Kyunghyun, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078, 2014.

[Rather 15] RATHER, Akhter Mohiuddin; AGARWAL, Arun; SASTRY, V. N. Recurrent neural network and a hybrid model for prediction of stock returns. Expert Systems with Applications, 2015, 42.6: 3234-3241.