ECサイトにおける消費者の探索行動データを用いた購買生起と購買カテゴリの予測に関する研究
—再帰型ニューラルネットワークを用いたモデル構築—

経営システム工学専攻　橋本　鴻

1　研究背景と目的

近年，インターネットの急激な普及に伴い，オンラインショッピングの利用者数は増加の一途を辿っている．また，購買チャネルの変化により，消費者の購買行動も変化している．特にEC（Electronic Commerce）サイトでは，多種多様な消費者及び商品を有することから，すべての消費者購買行動を単純なパターンで表現することは適切ではない．即ち，ECサイトにおける消費者の購買行動は非常に複雑であるといえる．また，ECサイトでは消費者の行動履歴を取得し，それらのデータに基づいて顧客を管理するような仕組みを導入している企業が多い．このようなデータを用いたサイト上での消費者の探索行動に着目した研究が産学問わず注目を集めている [1, 2]．これらの研究により，消費者が商品を購買するまでの一連の流れを利用することが消費者の購買意思決定に影響を与えることが知られるようになった．中でもECサイトでの購買意思決定において重要なウェイトを占めているのが,「いつ」,「何を」買うかという購買生起，商品選択の2つの観点である．スーパーマーケットのような店舗では日常的に反復購買されるような消費財が主な商品であるが、ECサイトの場合，耐久品や専門品などの高額かつ購買頻度が低いカテゴリが存在する．したがって，どのタイミングで消費者が購買するかを予測することは難しい．また，当然のようにECサイト上では消費者の商品選択の様子を目で確認することができないため，購買が起きる前の「そろそろこの商品を買うのではないか」という消費者の状態を把握することは容易ではない．そのため，どの商品をどのくらい閲覧しているのかという消費者の探索行動を随時追わなければ，適切な消費者の状態を把握することが難しい．

本研究では，ECサイトのアクセスログデータを用いて，消費者が「いつ」,「何を」買うかという購買生起と商品選択の2つの観点から消費者行動の予測を試みる．購買商品カテゴリの予測モデルと購買生起の予測モデルという2つのモデルを提案することで，従来では困難であったECサイト上での消費者購買行動の状態を把握することが期待される．また，本研究では近年，文章や音声などの系列データに対して多くの成果を挙げているRNN (Recurrent Neural Networks) を用いて，消費者の一連の探索行動を学習する．消費者の探索行動を文脈と捉え，商品の閲覧を開始した時点から購買に至る一連の行動を時系列データとして扱い，RNNを用いた解析を行う．本研究では，RNNの複数のアルゴリズムを比較し，消費者行動の予測手法としての有効性を評価する．

2　本研究で用いる分析手法

RNN (Recurrent Neural Networks) は主に文章や音声といった系列データに対して適用される．内部に有向閉路を持つニューラルネットワークであり，内部の再起項によって，過去の情報を一時的に保持することが可能になる．図1にRNNの基本的なネットワーク構造を示す．隠れ層を展開すると2つの層が存在し，t 時点の情報だけでなく，$t-1$ 時点の情報を保持している．前時点の隠れ層を現時点の入力と合わせて学習に用いることで、時系列情報を考慮したネットワーク構造となっている．

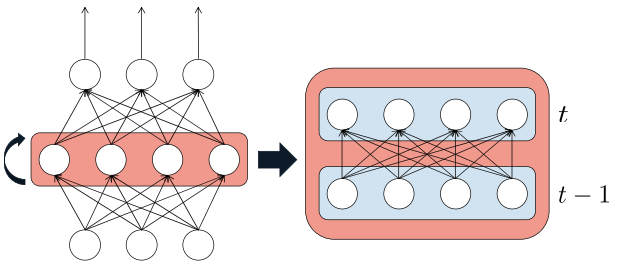

図 1: RNNの基本構造と隠れ層の内部構造

RNNは可変な長さ T の系列データ $(\mathbf{x}_1, \cdots, \mathbf{x}_T)$ を入力として受け取る．t 時点のネットワークへの入力データを $\mathbf{x}_t$ とする．隠れ層への入出力を

$\mathbf{u}_t, \mathbf{z}_t$ とし，隠れ層の状態を $\mathbf{h}_t$ とする．このとき，RNN の順伝播計算は入力データ $\mathbf{x}_t$ と $t-1$ 時点の隠れ層の状態 $\mathbf{h}_{t-1}$ のセルから次のように計算することができる．

$$\mathbf{u}_t = \mathbf{W}_{input}\mathbf{x}_t + \mathbf{W}_{hidden}\mathbf{h}_{t-1} + \mathbf{b} \quad (1)$$

$$\mathbf{h}_t = \sigma_{hidden}(\mathbf{u}_t) \quad (2)$$

$$\mathbf{z}_t = \mathbf{W}_{out}\mathbf{h}_t + \mathbf{b}_{out} \quad (3)$$

$$\mathbf{y}_t = \sigma_{out}(\mathbf{z}_t) \quad (4)$$

$\mathbf{W}_{input}$ と $\mathbf{W}_{hidden}$，$\mathbf{W}_{out}$ はそれぞれ入力層と隠れ層，出力層における学習重み行列であり，$\mathbf{b}$ と $\mathbf{b}_{out}$ は学習バイアスである．また，σ_{hidden} と σ_{out} は隠れ層と出力層における活性化関数である．隠れ層の状態 $\mathbf{h}_t$ は入力された系列データ $(\mathbf{x}_1, \cdots, \mathbf{x}_T)$ から，現時点 t までの情報を得る．これにより，学習開始初期の入力からの情報を時間経過とともに保持することができる．

一方で，RNN には勾配消失問題が生じることが知られている．この問題により，長期の系列データに対しては学習が困難になる可能性が高い．そこで，勾配消失問題を抑制する 2 つの RNN アーキテクチャを消費者行動の学習に適用する．一つは Hochreiter らが提案した LSTM (Long Short-Term Memory) [3]，二つ目は Cho らが提案した GRU (Gated Recurrent Unit) [4] である．両者は長期にわたる記憶を実現できるようにする方法として提案された．本研究では，これら 2 つの手法と隠れ層の活性化関数に双曲線正接関数を適用した RNN のそれぞれを用いた商品カテゴリと購買生起の予測モデルを作成し，その有効性について比較する．

3　対象データ

本研究では，ゴルフポータルサイトのデータを用いて分析を行う．データは 2014/01/01~2016/11/30 の期間におけるショッピングページの購買履歴データとアクセスログデータである．

商品カテゴリの選定では，購買履歴データ内のクラブ，用品・小物，ウェアのといった商品大分類（商品クラス）の内，購買割合が 5%以上の商品カテゴリに限定した．具体的には，クラブからは 6 カテゴリ，用品・小物からは 6 カテゴリ，ウェアからは 7 カテゴリを対象とした．また，対象ユーザの選定では，購買履歴データとアクセスログデータから，ユーザの選定を行った．その結果，選定した 92,748 人を本研究で対象とした．

4　分析概要

本研究では，消費者が探索行動を経て，購買に至る一連のデータを学習することで，購買商品カテゴリと購買生起をセッション単位で予測する．以下では，購買商品カテゴリの予測モデルを model-A，購買生起の予測モデルを model-B とする．

4.1　商品カテゴリの予測モデルにおける入力変数

本モデルでは，ユーザがセッション時点において，どの商品カテゴリを選択するかを予測する．model-A の入力変数を表 1 に示す．表 1 の上 3 つの入力変数は対象カテゴリ数に合わせて設定する．ユーザのデモグラフィック属性については性別と年代に応じたユニットを与えている．性別は男性，女性の 2 つのユニットを与え，年代は 10 代以下，20 代，30 代，40 代，50 代，60 代以上の 6 つのユニットを与えた計 8 ユニットとしている．なお，出力層のユニット数は対象カテゴリ数に合わせ，19 個とし，活性化関数にロジスティックシグモイド関数を使用する．

表 1: 商品カテゴリの予測モデルにおける入力変数一覧

変数名称	ユニット数
各カテゴリのセッションにおける閲覧（購買）回数	19
各カテゴリの累積購買回数	19
各カテゴリの累積閲覧回数	19
ユーザのデモグラフィック属性	8

4.2　購買生起の予測モデルにおける入力変数

本モデルでは，ユーザがセッション内で購買するか，しないかという購買生起を予測する．model-B の入力変数を表 2 に示す．入力変数の「購買セッション」は学習時のレコードが購買セッションのレコードである場合は 1，非購買セッションの場合は 0 が入っている．非購買セッションは単にユーザの閲覧行動データであることを示している．また，入力変数の「model-A の出力値」は model-A から出力される値を model-B の入力変数として使用することを意味している．なお，出力層のユニット

表 2: 購買生起の予測モデルにおける入力変数一覧

変数名称	ユニット数
購買セッション { 購買または非購買 }	2
model-A の出力値	19

数は 2 個とし，活性化関数にソフトマックス関数を使用する．

4.3 ネットワークモデルの概要と学習方法

本研究ではで述べた 3 つの RNN の手法を使用し，その比較を行う．本研究では，model-A，model-B において同じ RNN モデルを使用した場合の比較を行う．作成したネットワークの設定を表 3 に示す．ミニバッチ学習では，バッチサイズを 1000 とする．なお，本研究で分析対象ユーザ 92,748 人を訓練データ:検証データ:テストデータ = 8:1:1 の割合になるように分割した．また，本分析では，隠れ層のレイヤー数と Adam のパラメータ α，隠れ層のドロップアウト率について複数のケースを設け，学習を行う．隠れ層のレイヤー数については [1,2,3] のケースについて検討する．また，最適化手法 Adam のパラメータ α は [0.001, 0.01, 0.2, 0.5] のケースについて検討する．加えて，隠れ層のドロップアウト率は [0.0, 0.2, 0.5] とする．これらのパラメータを用いた組み合わせ $3 \times 4 \times 3 = 36$ 通りのケースで学習を行う．

表 3: ネットワークモデルの概要

	ネットワークの設定
隠れユニット数	100
損失関数	交差エントロピー誤差
最適化手法	Adam
学習方法	ミニバッチ学習
エポック数	20

5 分析結果と考察

5.1 作成したモデルの比較評価

パラメータの組み合わせ 36 通りで学習した場合の model-A の検証データに対する正答率を比較する．その際，勾配消失が起きたケースを除き，LSTM，GRU，RNN の各手法において検証データに対する正答率が他の 2 手法よりも高かった場合の数を集計する．その後，隠れ層のレイヤー数が 1 層の場合のみテストデータに対して正答率を算出し，同様にテストデータに対する正答率が他の 2 手法よりも高かった場合の数を集計し，表 4 にまとめる．

表 4: model-A の正答率が他の 2 手法よりも高かったケース数

	LSTM	GRU	RNN
検証データ	16	13	4
テストデータ	8	3	0

表 4 より，model-A において，LSTM，GRU，RNN の各手法を用いて検証データ，テストデータに対する正答率が他の 2 手法よりも高かった場合の数を集計すると LSTM が安定して高い精度を得ている．LSTM は内部にメモリセルと呼ばれる長期記憶のためのセルを有しており，消費者の探索行動によって与えられた過去の情報を保持しやすい仕組みとなっている．このような仕組みを持ったネットワーク手法で探索行動を学習することが，商品カテゴリの予測を行う上で有効であると考えられる．なお，LSTM で隠れ層のレイヤー数が 1，Adam のパラメータ α が 0.01，ドロップアウト率が 0.5 の場合，検証データに対する正答率が 61.9%，テストデータに対する正答率が 65.3%であった．

5.2 2 つのモデルを併用した消費者行動の予測精度の評価

以下の 4 つのシチュエーションを考え，model-A と model-B の 2 つを併用した際の評価を行う．

1. 商品カテゴリの予測モデル（model-A）と購買生起の予測モデル（model-B）が共に正解した場合（$S1$）
2. 商品カテゴリの予測モデル（model-A）は不正解だったが，購買生起の予測モデル（model-B）は正解だった場合（$S2$）
3. 商品カテゴリの予測モデル（model-A）は正解したが，購買生起の予測モデル（model-B）は不正解だった場合（$S3$）
4. どちらのモデルも不正解だった場合（$S4$）

テストデータの購買セッションにおけるすべてのテストデータに対して予測を行った際の 4 つのシ

チュエーションの割合を算出する．なお，model-Aでは予測値の上位 3 つを予測リストとして使用する．model-B では F 値などを参考に購買ユニットの確率が 0.2 を超えた場合は購買と予測し，0.2 未満の場合は非購買と予測する．また，各ユーザが購買を行った購買セッション数は 85,204 であった．表 5 に 4 つのシチュエーション，$S1 \sim S4$ の割合を示す．

表 5: $S1 \sim S4$ の割合

シチュエーション	割合（%）
$S1$	31.85
$S2$	0.26
$S3$	65.81
$S4$	2.08

表 5 より，本研究で目的としている「商品カテゴリの予測モデル（model-A）と購買生起の予測モデル（model-B）が共に正解した場合（$S1$）」は 31.85%と購買セッションにおいて約 3 分の 1 を占めるシチュエーションであることがいえる．また，$S2$，$S4$ の割合が低く，$S3$ の割合が高いことから，商品カテゴリの予測モデル（model-A）は非常に高い精度で予測できていることがいえる．

また，図 2 は $S1$ のシチュエーションの割合が大きいカテゴリの順番で並び替えたグラフである．この図より，$S1$ の割合が大きいカテゴリは“tee”や“socks”，“gloves”など，用品・小物やウェアといったクラスが多いことがわかる．これらのカテゴリは比較的安価な商品であることがいえる．一方で，$S1$ の割合が小さいカテゴリとしては“driver”や“ iron ”，“ putter ”などゴルフクラブ類を中心としたカテゴリであることがわかる．ゴルフクラブなどの商品カテゴリは比較的高価な商品が多く，購買に踏み切るまでに閲覧時間やセッション数が多くなるため，model-B が購買と予測するセッションも多くなる．そのため，S3 の割合が大きくなり，S1 の割合は小さくなることが考えられる．また，ゴルフクラブなどは計画購買が多いため，他サイトとの比較を行ったり，サイト内で購買するかを吟味したりすることが考えられる．したがって，サイト内でのクラブ類に対する商品閲覧を含めた探索行動は多くなるが，当該 EC サイト上のみのデータであることや複数のセッションをまたいで購買に至るため，購買生起を予測することが難しいことが考えられる．

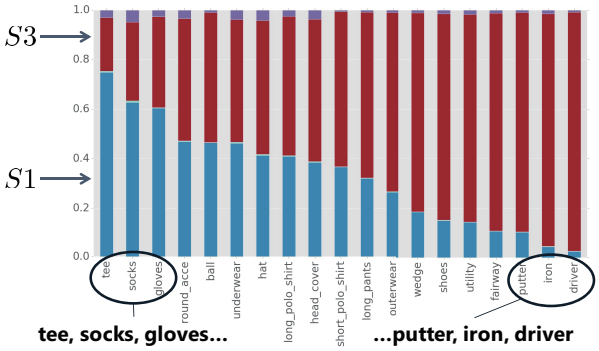

図 2: 各商品カテゴリにおけるシチュエーション別の帯グラフ

6 まとめと今後の課題

消費者が「いつ」,「何を」買うかという購買生起と商品選択の 2 つの観点から消費者行動の予測を試みた．消費者行動の予測には，RNN を用いてモデルの作成を行った．その結果，約 3 割程度の精度で，購買する商品カテゴリ並びに購買生起を正しく予測することが出来た．

購買生起の予測モデルにおいて，購買確率はカテゴリごとに異なる可能性がある．したがって，商品カテゴリごとに購買確率の閾値を設け，評価を行う必要があると考える．

参考文献

[1] 久松俊道, 外川隆司, 朝日弓未, 生田目崇, “EC サイトにおける購買予兆発見モデルの提案,” オペレーションズ・リサーチ，Vol. 58, No. 2, pp. 93–100, (2013).

[2] 石井久治, 市川祐介, 佐藤宏之, 小林透, “Web アクセスログからのパターンマイニングによる購買行動の推定,” 電子情報通信学会技術研究報告. LOIS, ライフインテリジェンスとオフィス情報システム, 109(272), pp. 89–94, (2009).

[3] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Computation 9(8), 1735-1780, (1997).

[4] K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” arXiv:1406.1078, 15 pages, (2014).