EC サイトにおけるページ遷移順序を考慮した購買行動分析

1X14C115-6　保戸田 未桜
指導教員　後藤 正幸

1　研究背景・目的

近年，情報技術の発展に伴い，EC サイトを通じた商品の購買が盛んになっており，Web マーケティング施策の重要性が高まっている．一方，EC サイトにおいて，ユーザの閲覧開始から終了までの 1 セッションの間に購買に至る割合は高々数%である．このため，サイトにアクセスしたユーザのうち，購買に至るユーザの割合（以下，CVR）を向上させるための施策が求められている．一般にユーザはセッション中に EC サイトの各 Web ページ（以下，ページ）に対して検索や商品，レビューなどの閲覧や商品の購買を行う．このとき，購買に至るユーザと至らないユーザのセッション内には，異なるページ遷移（以下，閲覧パス）が存在すると考えられる．そこで本研究では，EC サイトに蓄積されたユーザの閲覧履歴データを活用し，閲覧パスを素性として，ユーザが商品を購買するか否かの 2 値分類モデルを構築することを考える．これにより，分類境界近くの購買に至らないユーザに対して適切な施策を打つことで，CVR の向上につながると考えられる．また，分類に寄与する素性を抽出し購買に影響する閲覧パス（以下，影響パス）を特定することで，具体的な施策の一助となることが期待される．

本研究では，個々のユーザの閲覧パスと購買・非購買の関係性をモデル化するために適した素性として，N-gram[1] を導入し，かつ相互情報量 [1] を用いて適切な素性を選択する方法を提案する．

提案手法の有効性を示すため，株式会社ヴァリューズ提供の実閲覧履歴データに提案手法を適用し，購買に至るか否かの 2 値分類を行う．またこの分類結果を用いて CVR の向上の対象となるセッションや素性を抽出し考察を行う．

2　準備

2.1　EC サイトにおける閲覧履歴

EC サイトにおいて，一般にユーザは 1 セッションの間に商品の検索，閲覧などを行う．そして，興味のある商品をカートに入れ，注文操作をした後に購買に至る．その際に，EC サイト上の様々なタイプのページを遷移する．例として，EC サイトのトップページから商品の検索を行い，ある特定の商品の詳細ページを閲覧する，といった閲覧行動が考えられる．このとき,「トップページ→検索→商品詳細」というログが閲覧履歴データとして蓄積される．ただし、どの程度異なるページタイプを同じページとみなすか（以下，ページ粒度）は，閲覧履歴の中から個々のページの内容や特徴を活かすことができるように設定する必要がある．

2.2　N-gram[1]

N-gram とは，単語や文字を 1 単位として，任意の系列中に出現する N 個の連続した部分系列を抽出し，1 つの素性とするモデルである．N を大きくすることでモデルのパラメータ数が指数的に増加するが，モデルの精度は向上しないことがあるため，N-gram においてパラメータ N の設定は大きな問題の 1 つである．

1 ページのみを素性とする Bag of Words（以下，BoW）に対して，本研究では，ページの閲覧順序を考慮するために閲覧履歴データに N-gram[1] を適用し，N 個のページが含まれた閲覧パスを 1 つの素性とみなす．

3　提案手法

ユーザはセッション内で複数のページを遷移するが，遷移先のページは閲覧中のページに対するユーザの興味によって変化すると考えられる．したがって，ページ遷移の順序に着目することで，ユーザの興味の移り変わりを把握することができる．そこで本研究では，N-gram を用いて，ページ遷移の順序から閲覧パスを素性として抽出し，閲覧履歴データのベクトル化を行う．

このとき，N として取り得る値はデータに依存し，また複数の N を素性として組み合わせることも考えられる．しかし，N の設定方法により，素性の種類数も指数的に増加するという問題がある．そのため，購買に至るセッションと至らないセッションの分類に寄与する素性を相互情報量 [1] により選択する．本提案手法により，性能の良い 2 値分類器を学習することを目的とする．

3.1　有効な N の範囲の決定

前述のように，N-gram は N の値を大きくしても精度が向上するとは限らない．したがって，本研究が扱う EC サイトの閲覧履歴データに有効な N の範囲を探索的に決定する．

3.2　相互情報量による素性の削除

N-gram では，N をある程度大きく取ることで，重要な素性の抽出が期待される反面，膨大な数の重要ではない素性も抽出されてしまうという問題がある．そこで，各素性の分類に対する寄与度を相互情報量によって定量化し，相互情報量の大きい素性のみを学習させることで 2 値分類器の性能を向上させる．

相互情報量とは，2 つの確率変数間の相互依存の強さを表す尺度である．相互情報量が大きいほど 2 つの確率変数に強い相関がある．本研究では，ある閲覧パスの出現確率と購買に至る確率の相関を調べるために用いる．X は $X=1$ を,「全セッション内で対象の閲覧パスが出現する事象」，$X=0$ を,「出現しない事象」とする．また Y は $Y=1$ を,「購買に至る事象」，$Y=0$ を,「購買に至らない事象」としたとき，相互情報量は以下の式で定義される．

$$I(X;Y)=\sum_{y\in\{0,1\}}\sum_{x\in\{0,1\}}p(X=x,Y=y)\log\frac{p(X=x,Y=y)}{p(X=x)p(Y=y)} \quad (1)$$

4　実験

提案手法による，未分類のセッションが購買に至るか否かの 2 値分類の精度を検証するため，実データを用いた実験を行う．

4.1　実験データ

株式会社ヴァリューズ提供の楽天市場サイトの閲覧履歴データを用いた．データ収集期間は 2017 年 8 月 1 日から 2017 年 10 月 31 日，総ユーザ数 1,000 人，総セッション数 32,803 件，総購買数 1,046 件（CVR3.2%）である．

4.2　実験方法

Step1）ページ粒度の決定

本実験では，全ページを 23 種類に振り分け，以下のようにページ粒度を決定した．

表 1．ページ粒度（抜粋）

ページ内容	ページ内容	ページ内容
トップページ	カートページ	会員ページ
検索	注文操作	カテゴリーページ
商品詳細	ショップページ	レビューページ

Step2）素性の抽出

全セッションから N-gram を用いて素性を抽出する．本研究では閲覧履歴中の連続した N 個のページを持つ閲覧パスから，各閲覧パスの出現回数をカウントする．その後，各閲覧パスの中で出現回数が 2 回以上のものを素性として抽出する．得られた素性の種類数を以下の表 2 に示す．

表 2．素性の種類数

N	1	2	3	4	5
閲覧パス数	23	251	1,070	2,625	4,603

Step3）セッションのベクトル化

抽出した素性を基に，各セッションを各閲覧パスの出現回数を要素としたベクトルに変換する．

Step4）2 値分類器作成

Step3 で作成したベクトルを用いて，2 値分類する際の線形識別関数を構成する手法である Support Vector Machine[2]（以下，SVM）によって 2 値分類器を学習し，未分類のセッションが購買に至るか否かの予測を行う．

4.3　実験条件

本実験では，購買に至ったセッション数が購買に至らなかったセッション数に比べて少ないことから，購買に至ったセッションをオーバーサンプリング（購買に至ったセッションのデータをランダムに複製）し，10,000 件に増やすことでデータ数の偏りを解消する．また本実験では，最低 3 ページ遷移しなければ購買に至らない点を考慮し，閲覧ページ数が 2 以下のセッションは除外して実験を行う．

そして得られた結果に対して，5 分割交差検証法を用いて正解率を求め，評価指標とした．

4.4　実験結果

[有効な N の範囲の決定]

有効な N の範囲の決定するために，N を 1 から 10 まで変化させて実験を行った．実験より，N が 5 から 6 になる際に急激に正解率が低下した．したがって，1〜5 を有効な N の範囲とした．

[相互情報量による素性の削除]

有効な N の範囲 N=1〜5 の素性のうち，相互情報量の大きい上位 I 件の閲覧パスを影響パスと定義し，影響パスのみを基に各セッションをベクトル化し，実験を行った．このとき，I を 200 から 8,000 まで 600 刻みとし，I=8,572（全ての素性を用いる場合）と正解率を比較した結果を以下の図 1 に示す．

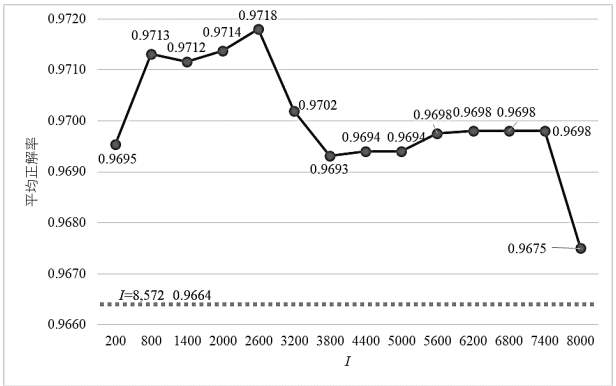

図 1．各 I の平均正解率

4.5　考察

図 1 より相互情報量で素性を絞ることで正解率が向上していることがわかる．また，I=2,600 のとき，正解率が最も高いことがわかる．これは，抽出した素性の 3 割程度であり，分類において寄与しない，精度低下を招く素性が多く含まれることを意味している．

この結果から，相互情報量により分類に寄与する閲覧パスのみを抽出することができ，提案手法の有効性を示すことができた．

N の範囲 1〜5 の素性のうち，相互情報量が高く，選択された上位の閲覧パスとその相互情報量を以下の表 3 に示す．

表 3．相互情報量上位閲覧パス

順位	N	素性	相互情報量
1	1	注文操作	0.1431
2	2	カートページ→注文操作	0.1336
3	1	カートページ	0.1006
4	2	注文操作→注文操作	0.0763
5	3	商品詳細→カートページ→注文詳細	0.0751
⋮	⋮	⋮	⋮
8	4	商品詳細→カートページ→注文詳細→注文詳細	0.0437
⋮	⋮	⋮	⋮
12	5	商品詳細→カートページ→注文詳細→注文詳細→注文詳細	0.0247

表 3 より，相互情報量が上位の閲覧パスは素性の種類数から考えると相対的に N が小さいものが多いことがわかる．これは，N が大きくなると極端に出現頻度が少なくなるためだと考えられる．また，相互情報量が上位の閲覧パスを見ると，購買に至る際に必要となる注文操作ページやカートページが購買に強く影響することがわかる．N=2 では，商品詳細ページからカートに入れるという閲覧パスよりも，カートに入れてから注文操作への閲覧パスの方が相互情報量が大きい．これは，カートに入れたとしても購買に至らないケースが多いためであると考えられる．

図 1 より，影響パスが I=2600 のとき最も分類精度が高くなる理由として 2 点挙げられる．まず I が小さいときは，相互情報量がある程度大きく，購買に影響のある閲覧パスを分類に利用することができない点が考えられる．また I が大きいとき，相互情報量が小さい閲覧パスは購買に影響があまりないため，そのような閲覧パスが多く素性に含まれたため精度低下を招いていると考えられる．

また，購買に至っていないにも関わらず，購買に至ったと誤分類したセッションには，大きく 2 つの傾向が見られる．1 つは，注文操作を繰り返した後，購買に至らずにセッション自体を終了するものである．この原因として注文操作が煩雑であることが挙げられる．そのため，注文操作の簡略化が望まれる．もう 1 つには，注文操作後に購買に至らずに別のページに遷移するものである．このようなセッションに対しては，ページの遷移検出時にリアルタイムにクーポンを発行するといった施策が有効と考えられる．

5　まとめと今後の課題

本研究では，EC サイトにおける閲覧履歴系列のページ遷移を考慮することで，購買に至るか否かを分類する手法を提案した．提案手法を実データに適用することで，提案手法の有用性を示した．

また，実験結果より分類には注文操作が大きく影響していることが分かった．したがって，今後の課題として，注文操作ページの扱いをより詳細にモデル化することが挙げられる．

参考文献

[1] 北研二,“確率的言語モデル”, 東京大学出版会,1999.

[2] Ioannis Tsochantaridis,Thorsten Joachims,Thomas Hofmann,Yasemin Altun,“Large Margin Methods for Structured and Interdependent Output Variables,” *Journal of Machine Learning Research 6*, pp.1453–1484, 2005.