1D2-3

ウェブサイトのユーザ遷移に基づいた コンバージョン促進要因の抽出

Extraction of Conversion Promotional Factor Based on User Transition on Webpage

龍野 翔 *1　飯塚 修平 *2　松尾 豊 *2
Sho Tatsuno　Shuhei Iitsuka　Yutaka Matsuo

*1東京大学大学院 情報理工学系研究科 創造情報学専攻
Department of Creative Informatics, The University of Tokyo

*2東京大学大学院 工学系研究科 技術経営戦略学専攻
Department of Technology Management for Innovation, The University of Tokyo

It is important for website development to improve their performance by analysis of data on the web. Users perceive information through webpages they visited. Therefore, the connection of pages that they have visited is very important to make strategy for improving their performance. However, many existing researches concentrated on only single page and few researches have taken the connection between webpages into consideration. In this research, we propose the method of extracting the property of user flows. Also, we visualize the property and show the promotional factor of conversion.

1. はじめに

近年，ウェブにおけるユーザの動きに着目する研究が増加している．その原因として，検索クエリの記録やウェブページに埋め込んだタグからウェブページにおけるユーザの流入や滞在時間，購入記録等のデータの記録ができるようになったという背景がある．これらのユーザのデータを分析することで，例えばコンバージョンに繋がるユーザを特定することや，ユーザの嗜好に合わせた効率的なリコメンドを行うことが可能となる．

一方で，ウェブに蓄積されたデータを利用し，より良いウェブサイトの構築を目指す取り組みが進められている．ウェブサイトの改善には Google Analytics*1 のようなウェブサイト分析ツールを用いて，ウェブページの改善のために最もコンバージョンが多い経路の特定を行い，その経路を強化するように経路上のウェブページの改善や，無駄なページの削減を行うといった施策がとられている．ここで述べるコンバージョンとは，サイトを訪問したユーザが広告のクリックや資料請求などサイト運営者が定めた目標に至ることを指す．

しかし，単純なコンバージョンの数値のみで経路が有用か否かということを判定することは困難である．なぜならば，ユーザがウェブページを巡回する経路によって同じコンバージョンであってもユーザがコンバージョンに結びついた要因が異なる可能性があるからである．例えば，1 つの資料請求ページに到達することをコンバージョンに置いたサイトでも「利用者数が多い」という場合と「24 時間サポートで安心」という 2 つの経路があったとする．各々のサイトを経由したユーザは同じコンバージョンであっても，コンバージョンの動機が経路によって異なると考えられる．つまり，コンバージョンしたユーザの知覚はコンバージョンという上では一定の傾向があるはずだが，閲覧したウェブサイトの経路により異なった知覚に基づいているはずである．そこで，ユーザが特定の経路を通りコンバージョンした場合，コンバージョンに辿り着いた原因を特定することができれば，どのような順番でウェブページを見せていけば良いか，あるいは今のウェブ上にどのような経路を追加すれば良いかといったことを考慮に入れたウェブサイトの設計が可能になる．

既存の研究では多くが単一ページに関する解析や最適化であった．そこで本研究では，経路によるユーザの知覚の可視化手法を提案する．Google Analytics で取得したデータとウェブページをクローリングして得られたデータを元に経路別で得られるユーザの知覚を可視化することで，単純なコンバージョンではなく，コンバージョンの原因となった知覚の特定やコンバージョンの改善に繋がる具体的な施策が打ちやすくなる可能性を示唆する．

連絡先: 龍野 翔，東京大学大学院情報理工学系研究科創造情報学専攻，東京都文京区弥生 2-11-16，03-5841-7672，sho_tatsuno@ipc.i.u-tokyo.ac.jp

*1 Google Analytics https://www.google.co.jp/analytics/

2. 関連研究

2.1 ユーザの嗜好

ウェブ上のユーザの挙動に着目した研究としては，行動ターゲティング [Yan 09] が挙げられる．行動ターゲティングとは，広告の対象となるユーザの購買や閲覧履歴を元に，そのユーザの嗜好に合ったリコメンドを行う研究である．ウェブサイトにおいては単純にクリックが多い商品のコンバージョンが高いわけではない (つまり，クリック率 $\neq$ コンバージョン率)[Pandey 11]．そのため，協調フィルタリング [Resnick 94] のようにユーザの過去の行動履歴と類似した行動を行った別のユーザの商品履歴に基づくリコメンドや別のカテゴリの商品を提示して，ユーザの好みを明確化するリコメンド [高玉 13] 方法やユーザの嗜好をネットワークとして可視化した研究 [飯塚 16] が存在する．

2.2 ウェブページ改善手法

ウェブページ最適化とはデザインや機能の異なるウェブページのパターンを複数用意し，各々の場合でのユーザ数やクリック数，滞在時間などの評価指標が最大となる条件を発見することである [飯塚 15]．ウェブページの最適化を行う場合は 2 つのサイトを同条件で比較する A/B テストや，2 つ以上のパターンを同時に試行する多変量テストが存在する．一般的に A/B テストや多変量テストはテストを試行したいパターンの数だけ期間が必要となるため，多くのパターンをテストする場

記号	パラメータの定義
entranceRate	流入率
exitRate	離脱率
log_avgTimeOnPage	平均滞在時間の対数
log_entrances	流入数の対数
log_exits	離脱数の対数
log_pageviews	ページビュー数の対数
log_pageviewsPerSession	ページ単位のセッション数の対数
log_timeOnPage	滞在時間の対数
log_uniquePageviews	ユニークなページビュー数の対数

表 1: 分析に用いたデータ

合には効率的ではない．そこで近年はテストと運用を同時に行うバンディットアルゴリズム [White 12] が提案されている．

3. 手法

本研究では，ウェブサイトから取得したデータを加工し，経路の可視化を行うために各ページにおける特徴ベクトルを抽出し，次元削減を行う．削減したページごとの特徴ベクトルを基に経路の定式化を行い，経路ごとの可視化を行う．

3.1 データの加工

用いるデータは株式会社 WACUL[*2] の扱うウェブサイトのうち，ある人材派遣会社のウェブサイトを取り扱う．人材派遣会社のウェブサイトは EC サイトと異なり，目的が人材派遣や資料請求といった目標が明確であり，経路が追跡しやすく，ユーザの利用数も年間 5 万人程度の規模であるために採用した．データの要素として，Google Analytics により得られた 2015 年 2 月 1 日 2016 年 2 月 1 日の期間のユーザのデータとクローリングにより得られたウェブサイト内の各ページ本文のテキストデータ用いる．Google Analytics により得られた 9 個の要素 (表 1) を用いる．平均滞在時間，流入数，離脱率，ページビュー数，ページ単位のセッション数，滞在時間，ユニークなページビュー数は指数分布に近かったため，対数変換を行った．一方，ホームページにより得られたテキストはタグを外した上で Mecab による形態素解析を行い，単語を分離した上で文章を Bag-of-Words に分解した．これらのパラメータを G を Google Analytics のデータ行列，B を Bag-of-Words の行列として $[G|B]$ のようにまとめ，データとして利用した．

3.2 特徴ベクトルの抽出

次に，図 1 のように特徴ベクトルの抽出を行う．特徴ベクトル抽出方法としては Sparse Non-negative Matrix Factorization(SNMF) [Kim 07] を用いる．NMF とは以下の式のように非負値行列 ($A \in \mathbf{R}^{m\times n}$) を 2 つの非負値行列 ($W \in \mathbf{R}^{m\times k}, H \in \mathbf{R}^{n\times k}$) に分解するアルゴリズムである．

$$A \approx WH^T \quad (W, H \geq 0) \tag{1}$$

この該当する W,H を抽出するのにフロベニウスノルム F を考え，以下の最小化問題を解く．

$$min \ \ \frac{1}{2}||A - WH^T||_F^2 \tag{2}$$

SNMF は最小二乗制約に基づく非負性を用いたスパース NMF である．本実験では文章を Bag-of-Words として取り出しており，比較的スパースな行列となっているため，スパースな行列に適した SNMF を圧縮方法として採用した．

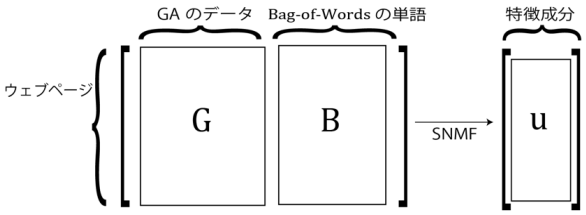

図 1: ページごとの素性を用いた特徴量抽出

$$min \ \ \frac{1}{2}\left[||A - WH^T||_F^2 + \eta||W||_F^2 + \beta\sum_{j=1}^{n}||H(j,:)||_1^2\right] \tag{3}$$

η, β はそれぞれ分解した行列の調整やスパース性と正確さに関するバランス項である． NMF の初期化には同様に要素のスパース性を考慮し，Nonnegative Double Singular Value Decomposition(NNDSVD) [Boutsidis 08] を用いた．各ウェブページに対して SNMF を用いることで，各軸がその要素の特徴を示す特徴量空間にウェブページを特徴ベクトル u として落とし込むことが可能となる．

3.3 経路の定式化

抽出した個々のページの特徴ベクトルを元に経路の定式化を行う．今回は経路の定式化を重み付け線形和により定義した．各経路の重み付けを以下のように，流入したユーザが該当する経路を通った割合を重みとして設定する．

$$w_i = \frac{PV_{pt}}{PV_t} \quad (i = 1, 2, ..N) \tag{4}$$

(w:重み， N: 経路内のページの数，PV_{pt}: 経路の前のページから各ページに移ってきたページビューの数，PV_t: 各ページのページビューの数)

この重みの設定により，単純なユーザの流入が多いページの影響を削減し，ユーザの流出経路としての重要度を表現した．加えて，全体での正規化を行った．これにより経路の大きさではなく，経路の特徴の傾向を表現することが可能となる．

よって，以上の式をまとめると，経路の定式化は以下のように表される．

$$V_j = \frac{\sum_i^N w_i \ u_i}{\sum_i^N w_i \ |u_i|} \ (j = 1, 2, ..., N') \tag{5}$$

(w:重み， u: ページの特徴ベクトル，N': 経路の総数， V: 経路の特徴ベクトル)

4. 実験結果

4.1 実験概要

本実験では，前節に基づき加工したデータのうちでコンバージョンに結びついた経路のみの取得を行った．Google Analytics ではコンバージョンの 3 つ前のウェブページからしか経路を取得できないため，コンバージョンしたページから 3 つ前のページ，合計 4 ページ分を一つの経路とみなした．アクセス可能であった全 170 ページ，212 経路全体で 2 つの目標を選定した．各々の目標に至った経路を SNMF により特徴ベクトル

*2 株式会社 WACUL https://wacul.co.jp/

に落とし込み，可視化する．この時，特徴ベクトルの次元は 3 とした．目標の選定基準としては，コンバージョンに結びつく経路が複数あるもの，全体のコンバージョンが多いもの，また URL に到達するものを目標に設定しているものを採用した．

得られた 3 次元の特徴ベクトルを 2 次元の特徴空間に落とし込み，2 つの目標について特徴量 1–3 をそれぞれ軸とした 3 つの図を生成する．この時，各目標でコンバージョンが多かった上位 3 経路も同時にプロットする．また SNMF を行う部分 (図 1) を特異値分解 (SVD) に変えたものも比較対象として検証する．得られた図から目標の比較などの考察を行う．

4.2 経路ごとの特徴の抽出と可視化

2 つの目標 (target1，target2) とコンバージョン上位 3 経路 (highCV1，highCV2) について，3 次元の特徴空間に落とし込んだものを可視化したものは図 2–7 となった．図 2–4 が SVD に関する可視化，図 5–7 が SNMF に関する可視化である．図 2,5 は第 1 特徴量と第 2 特徴量，図 3,6 は第 1 特徴量と第 3 特徴量，図 4,7 は第 2 特徴量と第 3 特徴量を可視化した．また，SVD の変数寄与度は，[0.431, 0.239, 0.051] となった．

SVD による特徴量抽出では，ほとんどの点が図 2 上の半径約 1 の円周上に分布していることが分かる．また，図 3 ではグラフの右側に寄っているが目標ごとの傾向の差は見受けられず，図 4 に関しても第 1 目標は直線上にいくらか乗っているが，第 2 目標については大きな傾向は見受けられない．SNMF については図 5 に着目すると，第 1 目標は散らばりが見られるものの，原点中心とした大きさ 0.6 の円周上に密集していることが見て取れる．また，第 2 目標は横軸上 0.9 近辺に密集しているものの，他の要素も比較的横に繋がっていることが確認できる．図 6 に着目すると，第 1 目標は大まかに直線上に存在する傾向が見て取れる．また，第 2 目標は曲線上の傾向や繋がった傾向が見て取れる．図 7 に着目すると，第 1 目標は主に曲線上に分布している．第 2 目標は横軸上 0 付近か 1 付近に多く密集している．

4.3 SVD と SNMF の比較

SVD により抽出したベクトルについては第 1 特徴量と第 2 特徴量が第 3 特徴量より遥かに強く働く結果となり，図 2–4 についても傾向が掴みずらい結果となった．これは，SVD の寄与度は第 1，第 2 成分が第 3 成分よりも大きいため，第 3 成分以下の寄与度が非常に小さい (0.05 以下) ことが図 2 に表れていると考えられる．

一方，SNMF により抽出したベクトルに基づく経路は図 5–7 についても各々，目標ごとの傾向が見て取れる．単純な半径 1 の円周上に乗っている訳ではなく曲線状や直線状の分布が観測されており，これは目標による経路ごとの知覚の傾向が表れていると考えられる．また，コンバージョン上位 3 項目について見てみると，それぞれの直線や曲線上の分布に乗っており，これらの経路はこのコンバージョンの高い方向，つまり highCV 方向へと要素を変化させていくことでより良いコンバージョンの経路に改善させていくことが可能になると考えられる．

こうした傾向から，コンバージョンに繋がると考えられる傾向のうち，例えば図 5 の直線・曲線上の延長など，コンバージョンの高い経路の方向への延長として考えられる部分が存在する．こうした第 1，第 2 特徴量上の点に対して，正規化の条件から第 3 特徴量の成分が一意に定まり，式 (5) を逆に解くことで，コンバージョンに近い可能性があるがまだコンバージョンできていない経路を示唆できる可能性が存在する．

5. まとめ

本研究では，ウェブサイトの最適化手法として既存の単一ページのみに着目した最適化手法と異なり，コンバージョンに結びついた経路全体を要素として取得し，次元削減を行い特徴ベクトル化して可視化を行った．この時，Google Analytics により取得したページごとのユーザのデータのみならず，ウェブページに内包されている html の文章データも加味したページごとの特徴ベクトルを抽出した．そして，得られた特徴ベクトルと前後の経路を加味した重み付けによる線形和をユーザ遷移に基づいた定式化手法として提案した．この時，SVD よりも SNMF を用いた方がより明確な経路の傾向の取得が可能になったことを検証した．特に単一ページではなく，ページの経路に基づき定式化して可視化を行ったことは本研究の新規性と言える．

次に行うべきこととして，単一ページと経路の比較が考えられる．単一ページの中からユーザ数が多いものを仮想的な経路として選出することで，経路の知覚を行った場合との比較を行うことが可能となる．

今後の展望として，各特徴量の軸の成分のサイト上の解釈を行うことが考えられる．軸の意味づけにより，より意識的な知覚の把握に繋がると考えられる．現状は BoW により特徴量の抽出を行っているが，word2vec や tf–idf の利用により，個々のウェブページに特有な文章ベクトルの抽出が可能となる可能性がある．また評価法として，一部のテストデータを基に提案した経路でのコンバージョンがあるか，また提案した経路を強化し実際に A/B テストを行い，コンバージョンが発生するかを検証するといったことが挙げられる．こうした経路別の特徴の抽出により，ウェブサイトの作成者にとって経路による違いを考慮しながら明確なウェブサイトの構築が可能になると考えられる．

謝辞

本研究は，大津裕史氏をはじめとする株式会社 WACUL よりデータの提供と多大なアドバイスを頂きました．この場を借りて，心より御礼申し上げます．

参考文献

[Boutsidis 08] Boutsidis, C. and Gallopoulos, E.: SVD based initialization: A head start for nonnegative matrix factorization, *Journal of Pattern Recognition*, Vol 41, Issue 4, pp. 1350–1362 (2007)

[Kim 07] Kim, H. and Park, H.: Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis, *Journal of Bioinformatics*, Vol 23, Issue 12, pp. 1495–1502 (2007)

[Pandey 11] Pandey, S., Aly, M., Bagherjeiran, A., Hatch, A., Ciccolo, P., Ratnaparkhi, A. and Zinkevich, M.:Learning to target: what works for behavioral targeting, *in Proceeding of CIKM 2011*, pp. 1805–1814 (2011)

[Resnick 94] Resnick, P., Iacovou, N., Suchak, M. and Bergstrom, P.: GroupLens: an open architecture for collaborative filtering of netnews, *in Proceeding of CSCW 1994*, pp. 175–186 (1994)

[Yan 09] Yan, J., Liu, N., Wang, G., Zhang, W., Jiang and Y., Chen, A.: How much can Behavioral Targeting Help Online Advertising?, *in Proceeding of WWW 2009*, pp. 261–270 (2009)

[White 12] White, J.: *Bandit Algorithms for Website Optimization*, O'Reilly (2012)

[飯塚 15] 飯塚 修平, 松尾 豊：高速なウェブサイト最適化のための KPI 設計手法の提案, 人工知能学会全国大会 (2015)

[飯塚 16] 飯塚 修平, 濱野 将司, 川上 和也, 松尾 豊, 萩原 静厳, 川上 登福, 浜田 貴之：閲覧・購買行動に着目した結婚情報サイトにおける商品間の勝敗関係の分析手法, 電子情報通信学会論文誌, pp.109-118 (2016)

[高玉 13] 高玉圭樹, 佐藤史盟, 大谷雅之, 服部聖彦：別カテゴリ商品提示による好みの明確化を促す推薦システム, 人工知能学会論文誌, pp. 261–270 (2013)

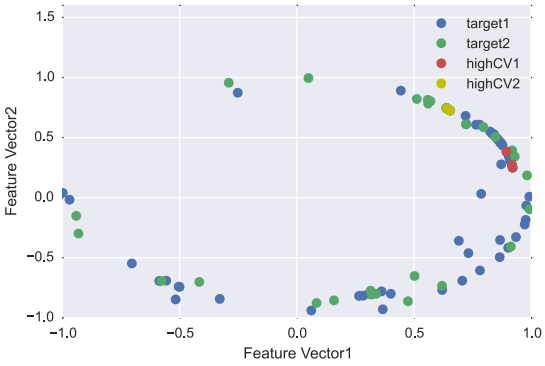

図 2: SVD：第 1 特徴量と第 2 特徴量の関係

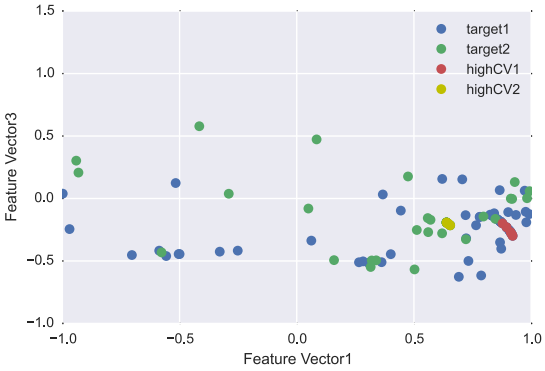

図 3: SVD：第 1 特徴量と第 3 特徴量の関係

図 4: SVD：第 2 特徴量と第 3 特徴量の関係

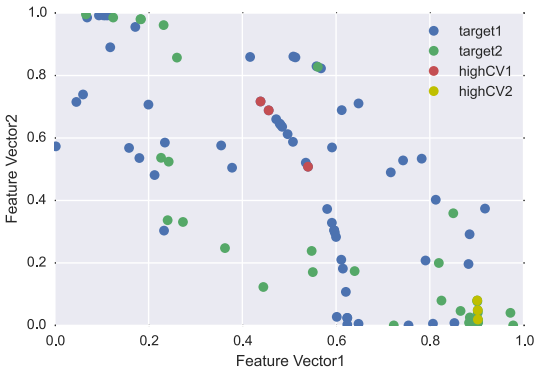

図 5: SNMF：第 1 特徴量と第 2 特徴量の関係

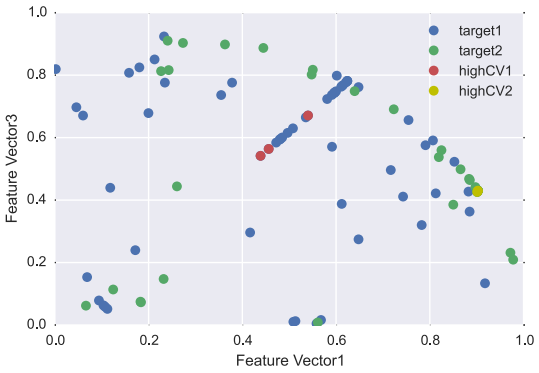

図 6: SNMF：第 1 特徴量と第 3 特徴量の関係

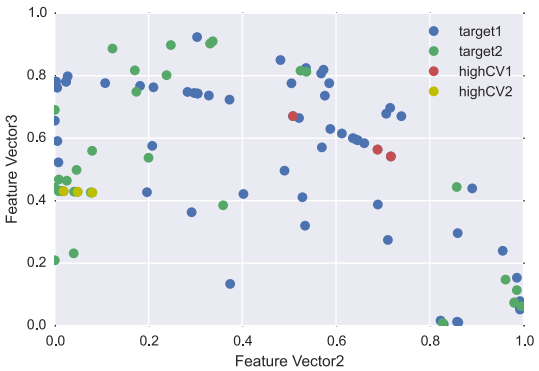

図 7: SNMF：第 2 特徴量と第 3 特徴量の関係