オンラインショッピングにおける商品選択行動のモデル化

野崎　祐里†　　佐藤　哲司††

† 筑波大学図書館情報メディア研究科　〒 305–8550 茨城県つくば市春日 1-2
†† 筑波大学図書館情報メディア系　〒 305–8550 茨城県つくば市春日 1-2
E-mail: †{nozaki,satoh}@ce.slis.tsukuba.ac.jp

あらまし　Amazon や楽天市場など，オンラインショッピングサイトを利用する機会が増えてきている．このようなサイト上でユーザがどのような商品選択行動を取っているかを把握することは，検索支援を行う上で重要である．そこで本研究では，検索セッションのクエリ変更とページアクセスの推移を完了率推移曲線として表し，完了率が最大値に到達する早さを特徴量としてクラスタリングを行う．クラスタリングされた各クラスタをモデルとみなし，モデル間におけるクエリの生成確率や変更パターンを分析する．最後に，機械学習の手法を用いて，初期の検索行動からモデルの予測をすることで，検索行動に応じたオンライン推薦の可能性を検討する．実データに対し評価実験を行った結果をここに報告する．

キーワード　情報検索，EC サイト，クエリ遷移

1 はじめに

現在，Amazon や楽天市場，Yahoo!ショッピングなど，Web 上で買い物ができるサービスが数多く存在している．2016 年に行われた「ネットショッピングに関する実態調査」[1]では，調査対象者の約 9 割がネットショッピングを経験していることが示されている．また，ネットショッピングを利用している理由として，「安い商品が多い」，「24 時間いつでも購入できる」，「出かけなくてよい」，「品揃えが豊富」，「ポイントが貯まり，お得」などが上位に挙がっている．

オンラインショッピングサイトでは，入力フォームからクエリを入力して検索できるという特徴がある．そのため，以下のような検索行動が想定される．まず，ユーザは 1 個あるいは複数個のキーワードをクエリとして入力し，検索結果の画面を確認する．検索結果に期待した商品が含まれていると判断したとき，商品の詳細ページをアクセスする．また，クエリは適切と判断でき，更に検索結果を閲覧したい場合はクエリを変えずに別ページに移動する．入力されたクエリが適切ではないと判断したとき，クエリを変更し新たなクエリで検索結果を確認する．以上の行動を繰り返し，ユーザの要求が満たされたとき検索は終了する．

ユーザの検索行動の例を図 1 に示す．図中の数字は検索手順を示しており，手順 3 で入力したクエリ「Red Wine」は以降クエリを変更していない．このため，この地点でユーザは入力したクエリに満足したとみなすことできる．同様に，手順 7 の「Red wine 300ml」のページをアクセスした地点で，ユーザはアクセスしたページに満足したと考えることができる．

このようなことから，検索セッションにおけるクエリの変更とアクセスしたページの完了地点を分析することは，ユーザの商品選択行動を把握するのに重要である．本研究では，検索行動の完了を満足度とみなし，完了率が最大になる地点に基づきクラスタリングをすることで，オンラインショッピングにおけるユーザの商品選択行動のモデル化を行う．モデル化によって，ユーザの検索行動に応じた支援の実現が期待できる．

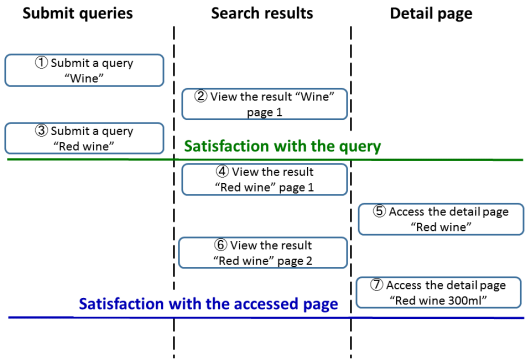

図 1　ユーザの検索行動の例

本論文の構成は以下の通りである．まず，第 2 章で関連研究を概観し，本研究の位置づけを示す．第 3 章では提案手法として，ログデータから完了率推移曲線を構築する手法と，構築した完了率推移曲線をクラスタリングする手法について述べる．第 4 章では実データを用いた評価実験を実施し，提案手法の有効性を検証する．第 5 章では，前章で得られた結果から，ユーザの検索行動について考察を行う．第 6 章で本論文のまとめと今後の課題を示す．

2 関連研究

Web 検索行動の分析に関する研究として，Jansen ら [1] は，Web 検索における 1 クエリあたりのキーワード数や，キーワードの追加・削除数，ページ閲覧数，検索演算子の使用数などの調査を行っている．検索行動の理解を目的とした研究とし

1：https://www.gmo.jp/news/article/5567/

て，富士谷ら [2] は複数のページ遷移から閲覧意図を分類する手法を提案している．また，クエリの変更意図を推定する研究 [3], [4], [5], [6] も数多くなされている．これらの研究は，クエリの変更行動をいくつかの推移タイプに分け，それぞれのタイプについてクエリ変更直前の行動との関係を分析している．

オンラインショッピングサイトを対象にした行動分析の研究も存在する．Moe [7] は，オンラインショッピングサイトにおけるユーザの検索行動を knowledge building, hedonic browsing, directed buying, search/deliberation の 4 種類を仮定し，実データに対してクラスタリングを行うことでその有効性を検証した．笹谷ら [8] は検索ワードや注文カテゴリ，注文回数などを特徴量として，ロジスティック回帰の手法でショッピングサイトの顧客ランクを予測している．また，笹谷らはグレードの高い優良顧客のランクダウンを防止するために，離脱予測モデルの構築も行っている．

クエリ遷移の研究として，クエリ間の遷移関係をグラフとして表現する研究がある．Boldi ら [9] は，クエリ間の遷移確率や同一セッションに属する確率を用いて 2 クエリ間のエッジの重み付けを行い，ランダムウォークを利用したクエリ推薦を実施している．Bordino ら [10] は，Spectral Embedding でグラフを低次元に圧縮した上でクエリ間の類似度を計算する手法を提案している．

本研究は，オンラインショッピングサイトにおけるユーザの検索行動をモデル化することが目的である．そのため，同じくオンラインショッピングサイトの行動を分類している Moe [7] の研究と類似しているが，Moe の研究は「時代が合わない」，「変数がよくわからない」などの問題点が指摘されている [11]. また，我々の既存研究 [12] ではクエリ変更とページアクセスの推移を用いて商品カテゴリのグループ化を行ってきた．既存研究では行動パターンによって商品カテゴリを分類したが，本研究では行動自体をモデル化する点で目的が異なっている．

3 商品選択行動のモデル化手法の提案

3.1 セッションの抽出

本研究で提案する完了率推移曲線とは，1 セッションにおけるクエリ変更完了率とページアクセス完了率の推移を表現したデータである．そのため，ログデータからセッションの抽出を行う必要がある．

本研究ではクエリログとアクセスログを使用する．クエリログとは，検索結果のページを表示するときに記録されるログで，ユーザ ID，タイムスタンプ，クエリのフィールドがあることを想定する．アクセスログは，商品の詳細にアクセスしたときに記録されるログで，ユーザ ID，タイムスタンプ，アクセスに到ったクエリのフィールドが存在することを想定する．

セッションの抽出は以下の手順で行う．まず，クエリログとアクセスログを統合し，ユーザ ID，タイムスタンプの順に優先した昇順ソートを行う．次に，ユーザ ID に基づき同一のユーザごとにログを切り出す．切り出されたユーザごとのログを順に走査し，もし前後のログの時間差が 30 分より大きい場合別のセションとして切り出す．セッションの切り出しの時間を 30 分に設定したのは，[9], [10], [13] の研究を参考にした．

3.2 検索行動のラベル付け

抽出されたセッションにおける各ログに，検索行動を表すラベルを付与する．本研究では検索意図の研究をしている [3], [4], [5] を参考に 7 種類のラベル (R，M，A，D，C，S，P) を設定する．ラベル付けには，前後のログにおけるキーワードの関係を参照することがある．キーワードとは，クエリを半角または全角スペースで分割した際の各文字列である．ただし，半角あるいは全角スペースが出現しない場合，キーワードはクエリの文字列と一致することとする．クエリを Q，キーワードを w とし，前のクエリを $Q_{pre} = \{w_{pr_1}, w_{pr_2}, ...w_{pr_n}\}$，後のクエリを $Q_{post} = \{w_{po_1}, w_{po_2}, ...w_{po_m}\}$ と表す．各ラベルについて，以下で詳細に説明する．

全部置換 (R)

前後のクエリログ間で，共通するキーワードが 1 つも存在しない場合，全部置換を表すラベル R(Replacing keywords) を付与する．ゆえに前後のクエリ間で以下の関係が成り立つ．

$$Q_{pre} \wedge Q_{post} = \Phi \tag{1}$$

一部置換（M）

前後のクエリログ間で共通するキーワードが 1 つ以上存在し，なおかつ包含関係が生じない場合，一部置換を表すラベル M(Modifying keywords) を付与する．よって前後のクエリ間で以下の関係が成立する．

$$Q_{pre} \wedge Q_{post} \neq \Phi \tag{2}$$

$$Q_{pre} \neq Q_{post} \tag{3}$$

$$Q_{pre} \not\subset Q_{post} \tag{4}$$

$$Q_{pre} \not\supset Q_{post} \tag{5}$$

追加（A）

前クエリのキーワードが後クエリのキーワードの部分集合となる場合，追加を表すラベル A(Adding keywords) を付与する．ゆえに前後のクエリ間で以下の関係が成り立つ．

$$Q_{pre} \subset Q_{post} \tag{6}$$

削 除 (D)

後クエリのキーワードが前クエリのキーワードの部分集合となる場合，削除を表すラベル D(Deleting keywords) を付与する．よって前後のクエリ間で以下の関係が成立する．

$$Q_{pre} \supset Q_{post} \tag{7}$$

継続（C）

前後のクエリログのキーワードが一致するとき，継続を表すラベル C(Continuing same keywords) を付与する．つまり，前後のクエリにおいて以下の関係が成り立つ．

$$Q_{pre} = Q_{post} \tag{8}$$

ラベル C は，同一のクエリで検索結果の別ページに移動する行動である．

開 始（S）

セッションの最初のログは，前後の行動が比較できないため，便宜的にセッションの開始を表すラベル S(Starting session) を付与する．

ページアクセス（P）

ログの種類がアクセスログの場合は，ページアクセスを表すラベル P(Page access) を付与する．

セッションにおけるラベル R，M，A，D，C，S の出現数の総和をパス長，R，M，A，D の総和をクエリ変更回数，P の総和をページアクセス回数と定義する．パス長は検索結果のページの閲覧数を表す．

表 1 に検索行動のラベルを付与した例を示す．このセッションは，ラベル S が 1 つ，R が 1 つ，P が 2 つ，A が 1 つ，C が 1 つ，M が 1 つ，D が 1 つある．よって，セッションのパス長は 6，クエリ変更回数は 4，ページアクセス回数は 2 となる．

3.3 完了率の算出

セッションの各ログに付与されたラベルに基づいて，クエリ変更完了率とページアクセス完了率を算出する．どちらの完了率も，時系列順にセッションのログを走査し，各ラベルに応じた処理を実行していき，セッション内の全ログに対して処理が完了したとき走査は終了する．

クエリ変更完了率

クエリ変更完了率は，各パス地点において，クエリ変更がセッション全体の中でどれだけ完了しているかを表す系列データである．つまり，セッションにおけるユーザのクエリの満足度の推移を表現している．

クエリ変更完了率における各ラベルにおける処理は以下の通りである．クエリ変更のラベル R，M，A，D が出現した際，値が上昇する仕組みとなっている．

- S: 初期値 0 をセットする
- R，M，A，D: 末尾の値に+1 した値を追加する
- C: 末尾の値を追加する
- P: 処理なし

ページアクセス完了率

ページアクセス完了率は，各パス地点において，ページアクセスがセッション全体の中でどれだけ完了しているかを表す系列データである．つまり，セッションにおけるユーザのアクセスしたページの満足度の推移を表現している．

ページアクセスにおける各ラベルにおける処理は以下の通りである．ページアクセスのラベル P が出現した際，値が上昇する仕組みとなっている．

- S: 初期値 0 をセットする
- R，M，A，D，C: 末尾の値を追加する
- P: 末尾の値を+1 する

次に，それぞれの完了率およびパス長に対して正規化を行う．これは，異なるパス長に対して，同一の基準で分析を行えるようにするためである．

完了率の正規化は系列データの各値について，系列データの最大値を割ることで実現できる．系列データの最大値は，系列データの末尾に出現する値と一致する．そのため，完了率の系列データ V における各値を $\{v_1, v_2, ..., v_i...., v_n\}$ とすると，正規化データ V' は以下のようになる．

$$V' = \{\frac{v_1}{v_n}, \frac{v_2}{v_n}, ..., \frac{v_i}{v_n}...., \frac{v_n}{v_n}\} \tag{9}$$

その後，パス長を正規化する．パス長も完了率の正規化と同様の手法で正規化を行える．そのため，パス長 $L = \{1, 2, ..., i...., n\}$ の正規化した値 L' は以下のように表せる．

$$L' = \{\frac{1}{n}, \frac{2}{n}, ..., \frac{i}{n}...., \frac{n}{n}\} \tag{10}$$

次に，ある特定の地点における完了率を算出する手法について説明する．これにより，パス長に存在しない地点における完了率も算出することができるようになる．方針として，各完了率の間を直線で補正することによって，対応する値を取得できるようにする．つまり，正規化したパス長が $L' = \{\frac{1}{n}, \frac{2}{n}, ..., \frac{i}{n}...., \frac{n}{n}\}$，各地点の正規化した完了率が $V' = \{\frac{v_1}{v_n}, \frac{v_2}{v_n}, ..., \frac{v_i}{v_n}...., \frac{v_n}{v_n}\}$ のとき，あるパス長の地点 $x(0 \leqq x \leqq 1)$ における完了率 $completion\_rate(x)$ は以下の式で求めることができる．

$$completion\_rate(x) = \frac{\frac{v_{j+1}}{v_n} - \frac{v_j}{v_n}}{\frac{j+1}{n} - \frac{j}{n}} \times (x - \frac{j}{n}) + \frac{v_j}{v_n} \tag{11}$$

$$\text{ただし，} \quad \frac{j}{n} \leqq x \leqq \frac{j+1}{n} \quad \text{である．}$$

正規化したクエリ変更完了率を x 軸，ページアクセス完了率を y 軸，パス長を z 軸に取ることによって完了率推移曲線を描画する．表 1 の例では，ラベルが「S，R，P，A，C，M，P，D」の順に出現するが，このラベルにおいてクエリ変更完了率を計算した例を表 2 に，ページアクセス完了率を計算した例を表 3 に示す．例では，パスを 0 から 1 までの 0.1 の刻み幅 (11 次元) に設定してそれぞれの完了率を算出している．

3.4 完了率推移曲線のクラスタリング

構築した完了率推移曲線をクラスタリングすることで，検索行動をいくつかのモデルに分類する．本研究では，クラスタリングの手法として k-means 法を利用する．k-means はデータを k 個のクラスタに分割する非階層型クラスタリングである．クラスタ数 k を人手で設定する必要があるが，本研究ではエルボー法に基づきクラスタ数を判断する．エルボー法とは，k の値を増やしていく中でクラスタ内誤差平方和 (SSE) が大きく低下し，飽和する点を最適なクラスタ数とみなす手法である．

表 1　検索行動ラベルの付与

User ID	Time stamp	Query	Type	Label
1	2016-09-05 10:07:11	water	query	S
1	2016-09-05 10:07:23	tea	query	R
1	2016-09-05 10:08:94	tea	access	P
1	2016-09-05 10:08:47	tea 500ml	query	A
1	2016-09-05 10:11:61	tea 500ml	query	C
1	2016-09-05 10:13:77	green tea	query	M
1	2016-09-05 10:15:39	green tea	access	P
1	2016-09-05 10:18:36	tea	query	D

表 2　クエリ変更完了率の処理手順

処理	結果
S:初期値 0 をセット	0
R:末尾の値に+1 した値を追加	0,1
P:処理なし	0,1
A:末尾の値に+1 した値を追加	0,1,2
C:末尾の値を追加	0,1,2,2
M:末尾の値に+1 した値を追加	0,1,2,2,3
P:処理なし	0,1,2,2,3
D:末尾の値に+1 した値を追加	0,1,2,2,3,4
完了率の正規化	0,0.25,0.5,0.5,0.75,1.0
11 次元に変換	0, 0.0, 0.05, 0.2, 0.35, 0.5, 0.5, 0.55, 0.7, 0.85, 1.0

表 3　ページアクセス完了率の処理手順

処理	結果
S:初期値 0 をセット	0
R:末尾の値を追加	0,0
P:末尾の値を+1	0,1
A:末尾の値を追加	0,1,1
C:末尾の値を追加	0,1,1,1
M:末尾の値を追加	0,1,1,1,1
P:末尾の値を+1	0,1,1,1,2
D:末尾の値を追加	0,1,1,1,2,2
完了率の正規化	0,0.5,0.5,0.5,1.0,1.0
11 次元に変換	0, 0.0, 0.1, 0.4, 0.5, 0.5, 0.5, 0.6, 0.9, 1.0, 1.0

4　実 験 評 価

4.1　データセット

提案手法の有効性を確かめるため，実データを用いた評価実験を実施する．本実験では，株式会社リクルートテクノロジーズが提供するポンパレモール [2]のデータセットを利用する．ログの期間は 2016 年 6 月から 2017 年 12 月までであり，クエリログは 24,582,912 件，アクセスログは 8,564,511 件である．本研究では，以下の条件を満たした 74,208 件のセッションを実験データとして用いる．

（1） クエリの変更が 3 回以上ある
（2） ページアクセスが 3 回以上ある
（3） パス長が 50 以下である
（4） アクセスしたページの第一カテゴリが同一である

条件 1 と 2 を設定したのは，完了率という相対的な指標を使うため，ある程度の回数のクエリ変更とページアクセスを要するからである．一方で，長すぎるセッションは bot の可能性があるため，条件 3 でパスの長さを制限した．条件 4 は同一のテーマに関する検索セッションを抽出するために設定した．

クラスタリングの特徴量として，各パスにおける完了率の値ではなく，完了率が最大値 1 に最初に到達した際のパスの値を用いる．実験におけるクエリ変更，ページアクセス完了率のパス長の刻み幅は 0.1 に設定した．

4.2　実 験 方 法

4.2.1　出現キーワードの比較

クラスタリングによって分類された各クラスタにおいて，キーワードの出現確率を評価する．n をクラスタ番号とし，クラスタ C_n におけるキーワード w_n の出現頻度を $freq(w_n)$ とすると，出現確率 $P(w_n)$ は以下の式の通りになる．

2 : https://www.ponparemall.com/

$$P(w_n) = \frac{freq(w_n)}{\sum_{w_n \in C_n} freq(w_n)} \quad (12)$$

キーワードの出現頻度 $freq(w_n)$ を数える対象ログは，ラベルが S，R，M，A，D のみとする．

次にクラスタごとに特徴的なキーワードを抽出する．方針として，他のクラスタより相対的に出現率が高いキーワードを抽出する．ゆえに，以下の式を満たすキーワードをクラスタ C_n における特徴的なキーワードとする．

$$\frac{P(w_n)}{\sum_{i=1}^{k} P(w_i)} - \frac{1}{k} \geqq \theta \quad (13)$$

k はクラスタ数を表す．パラメータ θ の値が大きいほどより特徴的なキーワードが抽出でき，小さくなるほど多くのキーワードを抽出できるようになる．実験ではパラメータ θ の値を 0.1 に設定した．

4.2.2 クエリ変更行動の比較

クラスタ間におけるクエリ変更行動を比較する．クエリ変更行動とは R，M，A，D のラベルを表す．クラスタ C_n におけるクエリ変更ラベル l_n の出現頻度を $freq(l_n)$ とすると，出現確率 $P(l_n)$ は以下の式で表せる．

$$P(l_n) = \frac{freq(l_n)}{\sum_{l_n \in C_n} freq(l_n)} \quad (14)$$

4.2.3 クラスタの予測

初期の検索行動からクラスタを予測できれば，クラスタに応じた検索支援をオンラインで行うことが可能になる．そのため，機械学習を用いたクラスタの予測問題に取り組む．

特徴量として，ページ遷移行動とページアクセス回数を用いる．ページ遷移行動とは検索結果のページを移動する行動で，R，M，A，D，C の 5 つのラベルを表す．各パスごとにページ遷移行動のカテゴリ変数を用意することで特徴量として表現する．

ページアクセス回数は，ページ遷移行動間にアクセスしたページ数を表す．つまり，ラベル R，M，A，D，C が出現するまでに出現したラベル P の回数が特徴量となる．

機械学習のアルゴリズムはランダムフォレストを使用し，5 分割交差検定を行った際の分類正解率を比較する．

4.3 実験結果

4.3.1 クラスタリング結果

k-means のクラスタ数 k を 1〜10 にした際のクラスタ内誤差平方和の変化を図 2 に示す．k =3 のとき，SSE が大きく低下し，これ以降値の減少値が停滞するため，k =3 をクラスタ数として採用する．完了率推移曲線のクラスタリング結果を図 3 に，各クラスタのクエリ変更完了率，ページアクセス完了率の推移曲線を図 4〜9 に示す．

それぞれのクラスタの特徴を取り上げる．クラスタ 1 はクエリ変更がセッションの後半に完了し，ページアクセスが前半に完了する検索行動である．クラスタ 2 はクエリ変更，ページアクセスともに晩期に完了するセッションである．クラスタ 3 は

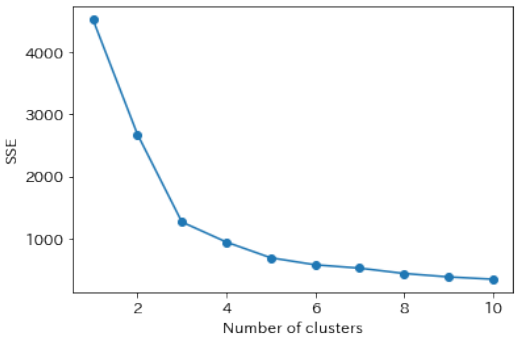

図 2　クラスタ内誤差平方和

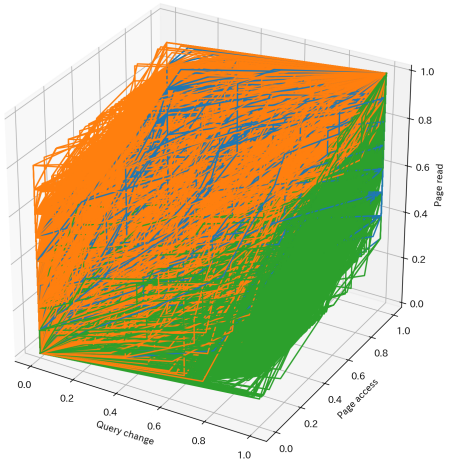

図 3　クラスタリング結果 ($k = 3$)

クエリ変更がセッションの前半に完了する一方で，ページアクセスは後半に完了する検索行動である．クラスタ 1 のデータ数は 116,12 件，クラスタ 2 は 47,074 件，クラスタ 3 は 15,522 件になった．

4.3.2 出現キーワードの比較結果

各クラスタにおけるキーワードの出現確率の累積分布を図 10 に示す．横軸は，実験データに出現するキーワードを頻度の多い順に並び替えたカテゴリーデータである．実験データにおけるキーワードの出現数上位 10 件を表 4 に示す．第 1 位はサービスを表す「送料無料」だが，第 2 位から第 10 位までは，カテゴリを表すようなキーワードになっている．

累積確率分布から，クラスタ 3 はよく使われるキーワードの出現確率が高い．

また，各クラスタの特徴語を評価する式 13 を満たすキーワードを表に示す．対象となるキーワードは，出現したセッション数がセッション全体の 0.1%を上回るキーワードとした．クラスタ 1 においては，特徴語の傾向を見出すことができなかったが，クラスタ 2 では,「カラコン」「キャンメイク」などのコスメに関するキーワードが多く抽出された．クラスタ 3 においては,「スマホケース」「xperia」などスマートフォンに関するキーワード,「長財布」「腕時計」などの服飾品に関するキーワード,「ドレス」「子供服」など衣料に関するキーワードが特徴語とし

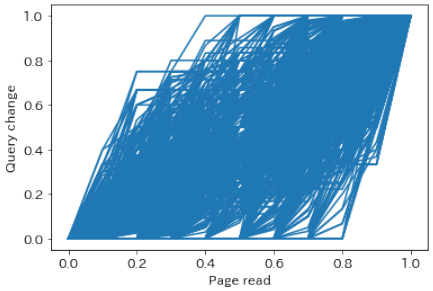

図 4　クエリ変更完了率
(クラスタ 1)

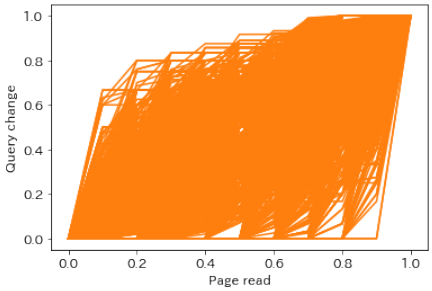

図 5　クエリ変更完了率
(クラスタ 2)

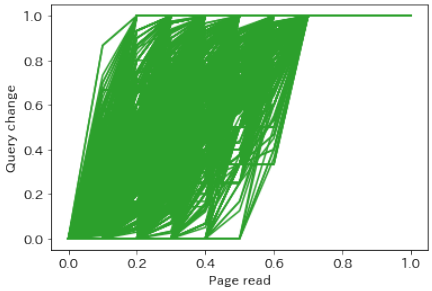

図 6　クエリ変更完了率
(クラスタ 3)

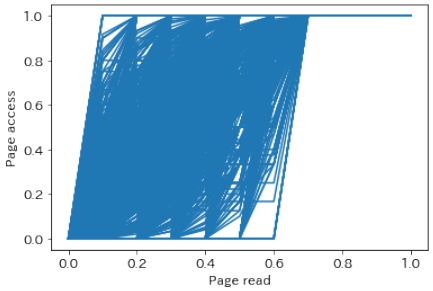

図 7　ページアクセス完了率
(クラスタ 1)

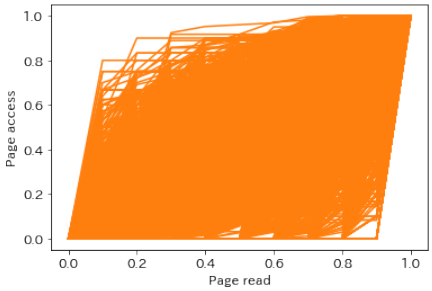

図 8　ページアクセス完了率
(クラスタ 2)

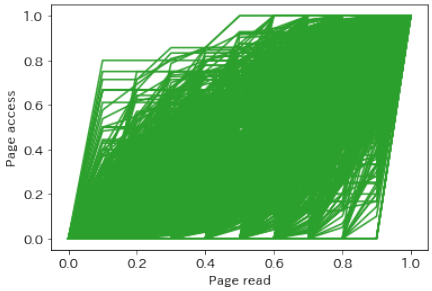

図 9　ページアクセス完了率
(クラスタ 3)

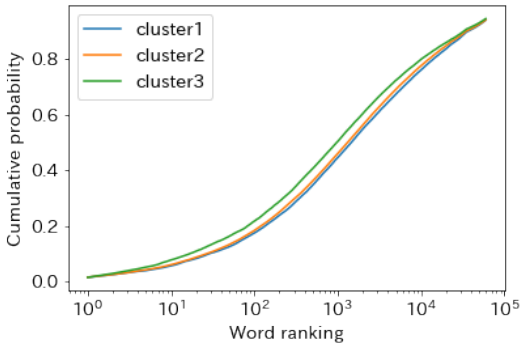

図 10　キーワードの出現確率の累積分布

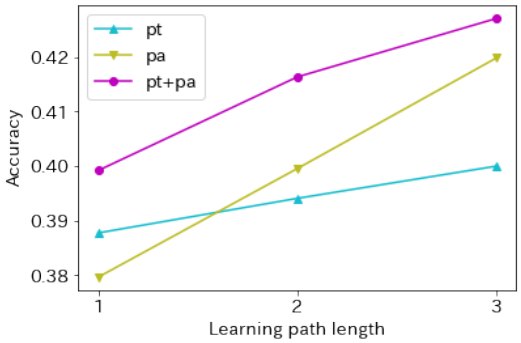

図 11　特徴量別の分類正解率

表 4　出現キーワード上位 10 件

順位	キーワード	出現頻度
1	送料無料	9160
2	レディース	6501
3	メンズ	4439
4	ケース	3741
5	手帳型	2647
6	セット	2525
7	財布	2511
8	スマホケース	2463
9	ワンピース	2307
10	水着	2270

て抽出されている．

4.3.3　クエリ変更行動の比較結果

各クラスタにおけるクエリ変更行動のラベルの出現確率は表 6 のような結果になった．クラスタ 1 は，ラベル R の値が高いが，ラベル A の値が低い．クラスタ 3 は反対にラベル R の値が低いが，ラベル A の値が高い．クラスタ 2 は，ラベル R と A がクラスタ 1 と 3 の中間の大きさになっている．

4.3.4　クラスタの予測結果

ランダムフォレストで特徴量別に学習・分類をした結果を図 11 に示す．データ数はクラスタ 2 が多く，クラスタ 1 と 3 は少ないため，評価実験ではアンダーサンプリングを実施した．凡例 pt はページ遷移行動を，pa はページアクセス回数を表している．分類正解率は，学習するパス数が増加すると上昇することが確認できる．特徴量別に比較をすると，ページ遷移行動とページアクセス回数両方を特徴量とした方が高い分類性能を示している．また，学習パス数が少ないときページ遷移行動の方が分類正解率は高いが，学習パス数が増加するとページアクセス回数の方が高い分類正解率になる．

表 5　各クラスタの特徴語

クラスタ	特徴語
1	’チョコレート’, ’ポーチ’, ’おもちゃ’, ’キーケース’, ’布団’, ’商品券’, ’トイレットペーパー’, ’マタニティ’, ’ワイン’, ’スリッパ’, ’爽快ドラッグ’, ’韓国’, ’mac’, ’ハンカチ’, ’入浴剤’, ’進撃の巨人’, ’チーズ’, ’ティッシュ’, ’RMK’, ’アクアエステソニック 2’
2	’カラコン’, ’ミルボン’, ’キャンメイク’, ’コート’, ’ロイヤルカナン’, ’シャチハタ’, ’柔軟剤’, ’チーク’, ’肉’, ’口紅’, ’日焼け止め’, ’rmk’, ’ハンドクリーム’, ’海苔’, ’ps4’, ’ルナソル’, ’クッキー’, ’クリスタルガイザー’, ’アーモンド’, ’キュレル’, ’アイブロウ’, ’ゴルフボール’, ’ふりかけ’, ’コスメデコルテ’, ’入浴剤’, ’ゆめぴりか’, ’くるみ’, ’進撃の巨人’, ’エチュードハウス’, ’イブサンローラン’
3	’スマホケース’, ’大きいサイズ’, ’手帳’, ’xperia’, ’長財布’, ’トートバッグ’, ’スーツケース’, ’ドレス’, ’腕時計’, ’水筒’, ’ブーツ’, ’時計’, ’ショルダーバッグ’, ’フォーマル’, ’スマホカバー’, ’携帯ケース’, ’エクスペリア’, ’クロックス’, ’スリッポン’, ’ブラジャー’, ’子供服’, ’アネロ’, ’キャリーバッグ’, ’アウター’, ’扇風機’, ’ワンデーアキュビュー’, ’ケイトスペード’, ’座椅子’, ’ガウチョパンツ’, ’チュニック’, ’ソファー’, ’アイス’, ’牛肉’

表 6　クエリ変更行動の出現確率

クラスタ	R	M	A	D
1	0.648	0.098	0.159	0.093
2	0.623	0.106	0.173	0.096
3	0.579	0.098	0.218	0.103

5　考　　察

5.1　クラスタリングの結果について

評価実験で k-means の k の値をエルボー法によって求めた結果，$k = 3$ のときエルボーが明確に見られた．各クラスタの特徴を取り上げることで，検索モデルの解釈と有効な検索支援手法について考察を行う．

クラスタ 1

クラスタ 1 はクエリ変更がセッションの後半に完了し，ページアクセスがセッションの前半に完了する検索行動である．ページアクセスがセッションの後半に行われないため，後半に入力したクエリが良くないことが考えられる．クエリ変更パターンは全部置換 (R) の頻度が高く，追加 (A) の頻度が他クラスタよりも低くなっていることからも，キーワードを追加して商品を絞り込む段階にまで到っていないことが考えられる．

このようなことから，クラスタ 1 に対する検索支援として絞り込みのクエリを推薦するよりも，幅広くクエリを推薦することが重要だといえる．

クラスタ 2

クラスタ 2 は，クエリ変更とページアクセスがセッションの晩期に完了する検索行動である．クエリ変更，ページアクセスともにセッションの最後まで行われるので，苦労して満足するページへのクエリを見つけ出せたことが示唆される．

クラスタ 2 に対しては，検索が完了するまでのステップ数を減らすことが課題である．クエリ変更，ページアクセスの完了の早さに偏りはないため，クエリとページアクセスをバランスよく推薦することが重要な検索支援だといえる．

クラスタ 3

クラスタ 3 は，クエリ変更がセッションの前半に完了し，ページアクセスが後半に完了する検索行動である．クエリ変更行動では追加の出現頻度が高く全部置換の出現頻度が低いことや，よく使われるキーワードの出現頻度が高いことから，キーワードを組み合わせてうまく検索結果を絞り込み，複数のページにアクセスをして商品を比較していることが考えられる．

よってクラスタ 3 ではクエリを推薦するよりも，アクセスしたページ情報を元に，よく比較される商品や，その分野で人気の商品を推薦することが有効な支援になるといえる．

5.2　クラスタの予測結果について

図 11 より，ページ遷移行動とページアクセス回数を特徴量として学習，分類を実施すると，学習パス数が 3 では 0.43 の分類正解率，学習パス数が 1 だけでは 0.40 の分類正解率となった．本実験は 3 クラスのマルチクラス分類であり，ランダムに予測した分類正解率が 0.33 になることを考えると，初期のページ遷移行動とページアクセス回数は特徴量として一定の有効性はあると考えることができる．しかし，分類正解率は十分な精度とはいえず，特徴量の種類や学習パス数を追加する必要性があるといえる．

6　お わ り に

本研究では，検索セッションにおけるクエリの変更とページアクセスの満足の度合いを完了率で表現し，クエリ変更完了率，ページアクセス完了率，正規化したパス長の 3 次元の完了率推移曲線を構築する手法を提案した．構築した完了率推移曲線を完了率が最大になるまでの早さでクラスタリングすることで，ユーザの商品選択行動のモデル化を行った．評価実験の結果，k-means のクラスタ数は 3 に定まり，各クラスタに対して有効な検索支援方法が示唆された．最後にオンラインな検索支援を

目指すために，初期の検索行動からクラスタを機械学習で予測する実験を行ったところ，学習パス数が 3 のとき 0.43 の分類正解率，学習パス数が 1 だけでは 0.40 の分類正解率を算出した．

今後の課題として，特徴量を完了率が最大になるときのパスだけにせず，25%，50%，75%完了地点などを追加することで，詳細な完了率の推移を分析することが期待できる．

また，クラスタの予測タスクにおいてクエリを特徴量に加えることも課題である

謝　　辞

本研究は JSPS 科研費 JP16H02904 の助成を受けたものである．また，DBSJ Data Challenge プログラムに参加し，株式会社リクルートテクノロジーズから提供を受けたポンパレモールのデータを利用している．ここに記して謝意を示す．

文　　献

[1] Bernard J. Jansen, Amanda Spink, and Tefko Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management*, Vol. 36, No. 2, pp. 207 – 227, 2000.

[2] 富士谷康, 吉田拓磨, 中村明順, 安積卓也, 望月祐洋, 西尾信彦. 閲覧履歴の関連性を考慮した閲覧意図の階層的分類手法. 情報処理学会論文誌, Vol. 56, No. 1, pp. 295–305, jan 2015.

[3] 梅本和俊, 中村聡史, 山本岳洋, 田中克己. 検索意図の遷移検出に基づく動的なクエリ推薦に向けた行動ログデータの分析. 研究報告 ヒューマンコンピュータインタラクション（HCI）, Vol. 2012, No. 24, pp. 1–8, may 2012.

[4] 関口裕一郎, 杉崎正之, 内山匡, 藤村滋, 望月崇由. 検索クエリログを用いたクエリ変更意図の自動推定. 第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM2011), 2011.

[5] Jeff Huang and Efthimis N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pp. 77–86, New York, NY, USA, 2009. ACM.

[6] Jiepu Jiang and Chaoqun Ni. What affects word changes in query reformulation during a task-based search session? In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16, pp. 111–120, New York, NY, USA, 2016. ACM.

[7] Wendy W. Moe. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*, Vol. 13, No. 1, pp. 29 – 39, 2003. Consumers in Cyberspace.

[8] 笹谷奈翁美, 坪内孝太, 田代昭悟, 鍛治伸裕, 清水伸幸. ショッピングサイトにおける優良顧客の離脱抑止施策について. 人工知能学会全国大会論文集, Vol. JSAI2016, pp. 2E34in2–2E34in2, 2016.

[9] Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, Aristides Gionis, and Sebastiano Vigna. The query-flow graph: Model and applications. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pp. 609–618, 2008.

[10] Ilaria Bordino, Carlos Castillo, Debora Donato, and Aristides Gionis. Query similarity by projecting the query-flow graph. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pp. 515–522, 2010.

[11] 三富悠紀. Ec サイトにおける消費者の閲覧行動をどう分類するのか? 赤門マネジメント・レビュー, Vol. 16, No. 6, pp. 261–272, 2017.

[12] Yuri Nozaki and Tetsuji Satoh. Search log analysis method of online shopping sites for navigating item categories. In *Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services, iiWAS 2018, Yogyakarta, Indonesia, November 19-21, 2018*, pp. 87–95, 2018.

[13] Lara D. Catledge and James E. Pitkow. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN Systems*, Vol. 27, No. 6, pp. 1065 – 1073, 1995. Proceedings of the Third International World-Wide Web Conference.