

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に関する研究

筑波大学

図書館情報メディア研究科

2019年3月

野崎 祐里

目次

第1章	序論	1
1.1	背景	1
1.2	本研究の目的	2
1.3	本論文の構成	2
第2章	関連研究	3
2.1	検索行動の分析に関する研究	3
2.2	検索ログを用いた推薦手法に関する研究	4
2.3	本研究の位置づけ	4
第3章	商品選択行動のモデル化手法の提案	5
3.1	セッションの抽出	5
3.2	検索行動のラベル付け	6
3.3	完了率の算出	8
3.4	完了率推移曲線のクラスタリング	9
第4章	実験評価	13
4.1	データセット	13
4.2	実験方法	14
4.2.1	出現キーワードの比較	14
4.2.2	クエリ変更行動の比較	15
4.2.3	クラスタの予測	15
4.3	実験結果	15
4.3.1	クラスタリング結果	15
4.3.2	出現キーワードの比較結果	17
4.3.3	クエリ変更行動の比較結果	17
4.3.4	クラスタの予測結果	18
第5章	考察	25
5.1	クラスタリングの結果について	25
5.2	クラスタの予測結果について	26
第6章	結論	27
6.1	本研究のまとめ	27

6.2 今後の課題	27
謝辞	28
参考文献	29
発表論文	32

目次

1.1 ユーザの検索行動の例	2
3.1 クエリ変更完了率の推移曲線	11
3.2 ページアクセス完了率の推移曲線	11
3.3 完了率推移曲線	12
4.1 セッション数の分布	14
4.2 クラスタ内誤差平方和	16
4.3 クエリ変更完了率の推移曲線 ($k = 2$)	17
4.4 ページアクセス完了率の推移曲線 ($k = 2$)	17
4.5 クエリ変更完了率の推移曲線 ($k = 3$)	18
4.6 ページアクセス完了率の推移曲線 ($k = 3$)	18
4.7 クエリ変更完了率の推移曲線 ($k = 4$)	18
4.8 ページアクセス完了率の推移曲線 ($k = 4$)	18
4.9 完了率推移曲線 ($k = 3$)	19
4.10 各クラスタにおける平均データ	20
4.11 キーワードの確率分布	21
4.12 キーワードの出現確率の累積分布	21
4.13 モデル別の分類正解率	23
4.14 特徴量別の分類正解率	24
4.15 パス別の分類正解率	24

表 目 次

3.1	セッション抽出例	5
3.2	検索行動ラベルの付与	7
3.3	クエリ変更完了率の処理手順	10
3.4	ページアクセス完了率の処理手順	10
4.1	出現キーワード上位 10 件	22
4.2	各クラスタの特徴語	22
4.3	クエリ変更行動の出現確率	23

第1章 序論

1.1 背景

現在, Amazon¹や楽天市場², Yahoo!ショッピング³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」⁴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を行うことができる. 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1.1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ができる. 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. 本研究では行動が完了するまでの推移を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実現する.

¹<https://www.amazon.co.jp/>

²<https://www.rakuten.co.jp/>

³<https://shopping.yahoo.co.jp/>

⁴<https://www.gmo.jp/news/article/5567/>

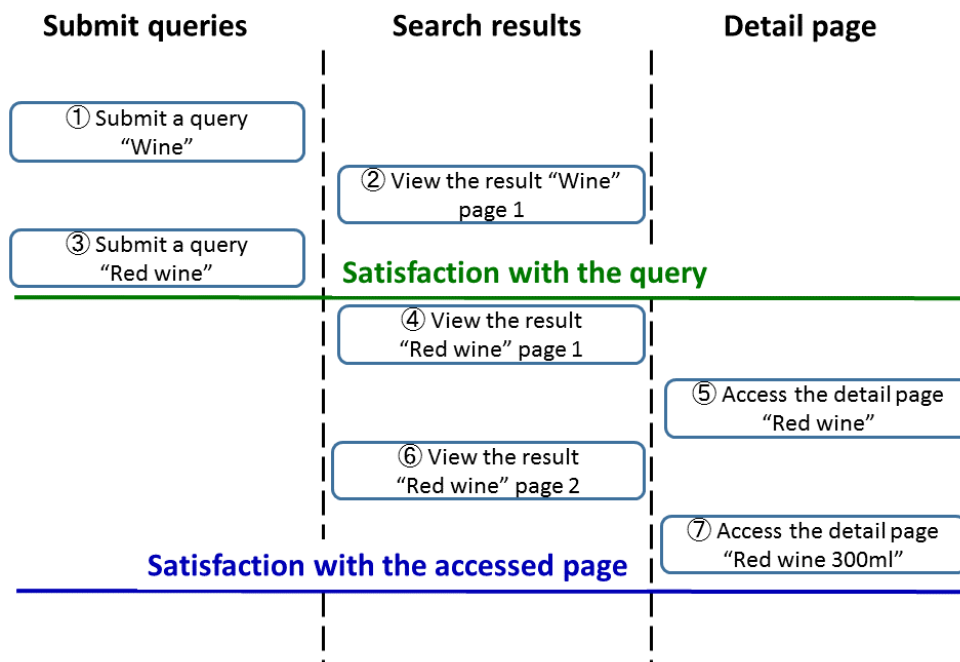


図 1.1: ユーザの検索行動の例

1.2 本研究の目的

本研究では、検索行動の完了を満足度とみなし、その推移パターンに基づいて、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をモデルする。具体的には、1セッションにおけるクエリ変更とページアクセスの完了率の推移曲線を構築し、クラスタリングの手法によって推移パターン进行分类する。そして、分類された各パターンにおいて、キーワードの生成確率やクエリ変更パターンの分析を行う。最後に機械学習の手法を用いて、初期の検索行動から推移パターンの予測をすることで、各商品選択行動モデルに適した検索支援をインタラクティブに行うことを試みる。

1.3 本論文の構成

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。

第2章 関連研究

Web 検索に関連する研究として、検索行動の分析に焦点を当てている研究と、検索ログを利用したクエリやアイテムの推薦を目的とした研究が存在する。

2.1 検索行動の分析に関する研究

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。Granka ら [2] や加藤ら [3] は、視線計測で Web コンテンツの注目している箇所を分析している。検索行動の理解を目的とした研究として、富士谷ら [4] 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 [5], [6], [7], [8] 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。

検索行動の評価に関する研究も提案されている。Buscher ら [9] は、クエリ、クリック、カーソル移動などの情報から、検索結果ページにおけるユーザーとタスクへの影響を分析している。Odijk ら [10] は、セッションで最後に入力されたクエリにおける検索行動に対して、成功した検索と失敗した検索の要因を分析している。Kim ら [11] や Borisov ら [12] は滞留時間を用いてアクセスしたページの満足度を評価するモデルを提案している。

オンラインショッピングサイトを対象にした研究もなされている。Moe [13] 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。Guo ら [14] は、広告のクリックスルー率を予測するために、オンラインショッピングサイトでユーザが商品を購入するかしないかの 2 値分類モデルを構築している。笹谷ら [15] 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。さらに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。Jing ら [16] や大河原ら [17] は、ショッピングサイトにおけるインターフェイスの研究を行っている。

2.2 検索ログを用いた推薦手法に関する研究

大量に蓄積されたクエリログデータを活用し、クエリを推薦する研究が数多く提案されている。関口ら [18] は、特定のクラスに対する属性語を取得するために、ログ中での出現頻度を正規化した平均情報量を用いることで属性語候補のスコア付けを行っている。鹿島ら [19] は、クエリの類似語を Word2Vec を利用して提示する手法を提案している。クエリに対して Web 検索を行い、検索結果に含まれる各単語から Word2Vec における分散表現を取得し、これらの平均ベクトルをクエリのベクトルとして類似単語を提示する。

クエリ間の遷移をグラフとして表現し、クエリ推薦に利用する研究も行われている。Boldi ら [20] 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 [21] 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。

ユーザの検索行動から、クエリではなくアイテムを推薦する手法も提案されている。清水ら [22] は、ユーザが閲覧したレビューから重視している項目を推定し、アイテム推薦に適用する手法を提案している。西尾ら [23] は、レシピサイトにおいて、閲覧している時間帯とサイトに遷移してきた経路情報を利用して、レシピのレコメンド成功率を改善する手法を提案している。渡邊ら [24] は、ユーザが興味を持っているツイートの推薦手法を提案している。潜在的ディリクレ配分法 (LDA) に基づき、トピックごとの単語分布とユーザ毎のトピック分布を推定する手法に、協調フィルタリングの手法を組み合わせることで、生成確率が低かったトピックの評価値を改善している。

2.3 本研究の位置づけ

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理解するため、クエリの変更とページアクセスの満足度を完了率として定義し、その推移パターンを分析していくものである。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析している Moe [13] の研究と類似しているが、Moe の研究ではセッションのページ閲覧数や経過時間、閲覧したカテゴリ数などを特徴量としてクラスタリングをしているのに対して、我々の研究ではセッションにおける各時点のクエリの変更とページアクセスの完了率という時系列なデータの特徴量としている点で異なっている。

本研究では、モデル化した各行動の分析やモデルの予測の際に、クエリの変更意図の研究で用いられているクエリ推移タイプを利用する。モデルによってクエリの推移タイプや単語の生成確率、行動パターンが異なる場合、モデルに適したクエリやアイテムの推薦手法を適用することでより効果的な検索支援が期待できる。

第3章 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を持っているものとする。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を持つものとする。

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ッ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、[20], [21], [25]の研究を参考にした。

表3.1にセッションの抽出例を示す。Typeフィールドは、ログの種類がクエリログかアクセスログかを示している。セッション1とセッション2は時間差が30分より大きくなっているため分割されている。セッション2とセッション3は時間差は30分以内であるが、ユーザIDが異なるため別のセッションとして切り出されている。

表 3.1: セッション抽出例

User ID	Time stamp	Query	Type	Session
1	2016-09-05 19:37:41	usb	query	1
1	2016-09-05 19:37:48	usb	query	1
1	2016-09-05 19:38:25	usb 64gb	query	1
1	2016-09-05 19:38:34	usb 64gb	access	1
1	2016-09-05 20:11:44	flower	query	2
1	2016-09-05 20:13:40	flower present	query	2
2	2016-09-05 20:15:11	rose	query	3

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 [5], [6], [7] の研究を参考に 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 $Q_{pre} = \{w_{pr_1}, w_{pr_2}, \dots, w_{pr_n}\}$ 、後のクエリを $Q_{post} = \{w_{po_1}, w_{po_2}, \dots, w_{po_m}\}$ と表す。

各ラベルについて説明していく。

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。

$$Q_{pre} \cap Q_{post} = \Phi \quad (3.1)$$

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。

$$Q_{pre} \cap Q_{post} \neq \Phi \quad (3.2)$$

$$Q_{pre} \neq Q_{post} \quad (3.3)$$

$$Q_{pre} \not\subset Q_{post} \quad (3.4)$$

$$Q_{pre} \not\supset Q_{post} \quad (3.5)$$

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。

$$Q_{pre} \subset Q_{post} \quad (3.6)$$

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。

$$Q_{pre} \supset Q_{post} \quad (3.7)$$

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．

$$Q_{pre} = Q_{post} \quad (3.8)$$

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．ラベル P は商品の詳細ページに遷移する行動，P 以外のラベルは，商品の検索結果の一覧画面に遷移する行動である．そのため，パス長は検索結果のページの閲覧数と一致する．

表 3.2 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．

表 3.2: 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

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、系列データに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 $\{v_1, v_2, \dots, v_i, \dots, v_n\}$ とすると、正規化データ V' は以下ようになる。

$$V' = \left\{ \frac{v_1}{v_n}, \frac{v_2}{v_n}, \dots, \frac{v_i}{v_n}, \dots, \frac{v_n}{v_n} \right\} \quad (3.9)$$

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 $L = \{1, 2, \dots, i, \dots, n\}$ の正規化した値 L' は以下のように表せる。

$$L' = \left\{ \frac{1}{n}, \frac{2}{n}, \dots, \frac{i}{n}, \dots, \frac{n}{n} \right\} \quad (3.10)$$

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 $L' = \left\{ \frac{1}{n}, \frac{2}{n}, \dots, \frac{i}{n}, \dots, \frac{n}{n} \right\}$ 、各地点の正規化した完了率が $V' = \left\{ \frac{v_1}{v_n}, \frac{v_2}{v_n}, \dots, \frac{v_i}{v_n}, \dots, \frac{v_n}{v_n} \right\}$ のとき、あるパス長の地点 $x (0 \leq x \leq 1)$ における完了率 $rate(x)$ は以下の式で求めることができる。

$$rate(x) = \frac{\frac{v_{j+1}}{v_n} - \frac{v_j}{v_n}}{\frac{j+1}{n} - \frac{j}{n}} \times \left(x - \frac{j}{n} \right) + \frac{v_j}{v_n} \quad (3.11)$$

ただし、 $\frac{j}{n} \leq x \leq \frac{j+1}{n}$ 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 3.2 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 3.3 に、ページアクセス完了率を計算した例を表 3.4 に示す。また、クエリ変更完了率の推移曲線を図 3.1 に、ページアクセス完了率の推移曲線を図 3.2 に、完了率推移曲線を図 3.3 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。クラスタリングは以下の手順に沿って行われる。

1. ランダムに k 個データを選択しクラスタの重心とする。
2. 各データを k 個の重心の中から最も近い重心に割り振り、これを初期クラスタとする。

表 3.3: 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 3.4: 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

3. クラスタ内の全データに対し、重心を計算し、新たな重心に対して他のデータの所属クラスタを割り振り直す。

4. 3を繰り返す、重心が収束したら処理は終了する。

k-means 法はクラスタ数を人手で設定する必要がある。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。

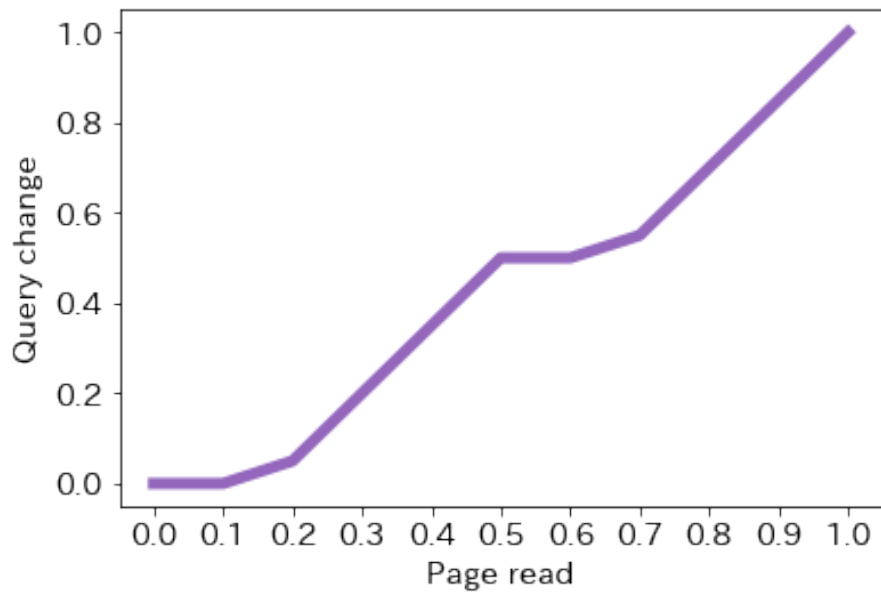


図 3.1: クエリ変更完了率の推移曲線

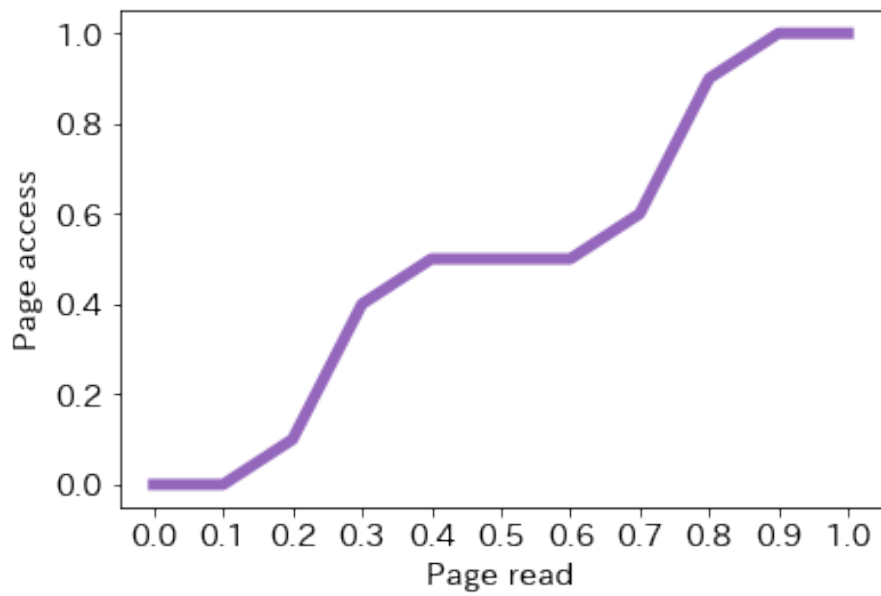


図 3.2: ページアクセス完了率の推移曲線

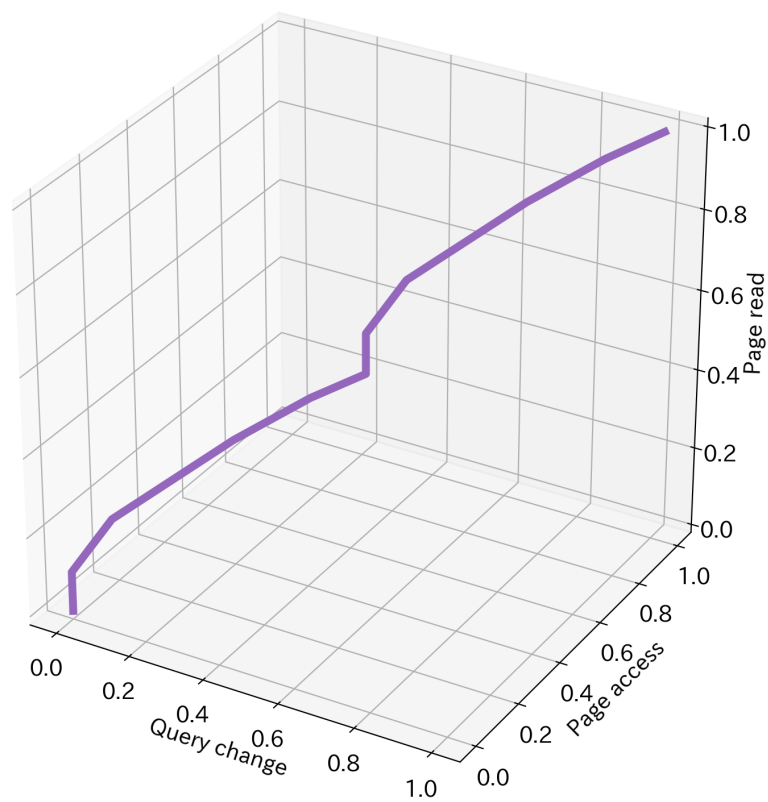


图 3.3: 完了率推移曲線

第4章 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¹のデータセットを利用する。ポンパレモールは、モール型のショッピングサイトで、同一の商品が複数の店舗によって異なる値段やサービスで販売されることがある。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。図4.1はデータセットから抽出されたセッションのパス長、クエリ変更回数、ページアクセス回数の頻度分布を示している。

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

クエリ変更とページアクセスの推移を分析するため、条件1と2を設定している。回数を3回以上にしたのは、完了率という相対的な指標を使っているからである。もし、セッション中にクエリ変更やページアクセスが1度しか出現しない場合、1回の変化で完了率が最大値まで到達してしまう。これを複数回の推移と同等に評価するのは望ましくないと考え、3回以上のクエリ変更とページアクセスを条件として設定した。

また、本研究の目的である、商品の選択行動を分析することを考えると、ある程度試行錯誤しているセッションを実験データとするのが適切であるといえる。一方で、探索行動が長すぎるのは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。ポンパレモールでは、商品は階層構造のカテゴリに分類される仕組みになっている。セッション内でアクセスしたページのカテゴリを同一のものに限定することで、複数の検索意図が混在するセッションを取り除くことができる。評価実験では階層構造の中で最上位のカテゴリである第一カテゴリを用いる。

クエリ変更、ページアクセス完了率におけるパス長の刻み幅は0.1に設定した。

¹<https://www.ponparemall.com/>

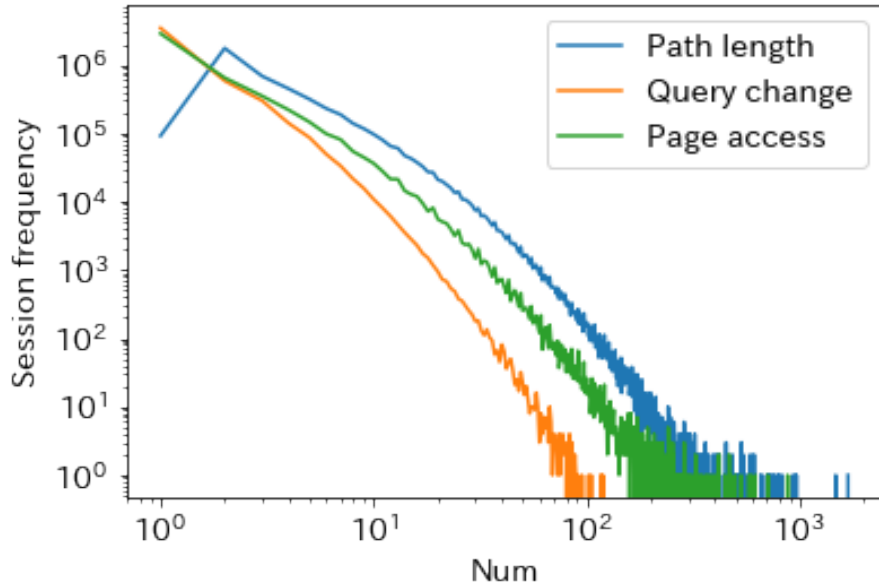


図 4.1: セッション数の分布

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。

$$P(w_n) = \frac{freq(w_n)}{\sum_{w_n \in C_n} freq(w_n)} \quad (4.1)$$

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。これは、同一のクエリで別の検索結果のページに遷移するラベル C に出現したキーワードは、出現頻度に数えないことを意味する。

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。

$$\frac{P(w_n)}{\sum_{i=1}^k P(w_i)} - \frac{1}{k} \geq \theta \quad (4.2)$$

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。

$$P(l_n) = \frac{freq(l_n)}{\sum_{l_n \in C_n} freq(l_n)} \quad (4.3)$$

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リアルタイムに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、SVM, ランダムフォレスト, ロジスティック回帰の 3 種類を使用し、5 分割交差検定を行った際の分類正解率を比較する。SVM のカーネルは線形カーネルを利用する。また、他のパラメータにおいては、Python の機械学習ライブラリである scikit-learn² のバージョン 0.19.0 のデフォルト値を用いる。

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を図 4.2 に示す。 $k=2\sim 4$ においてクラスタ内誤差平方和の減少値は大きいですが、明確にエルボーとなるような k の値を発見することができなかった。このため、クラスタ数を 2~4 に絞り、実際のクラスタリング結果を確認することで k の値を決定する。図 4.3~4.8 は $k=2\sim 4$ におけるクエリ変更完了率、ページアクセス完了率の推移曲線である。

それぞれのクラスタの特徴を取り上げる。 $k=2$ のとき、クラスタ 1(青線)はクエリ変更完了率には特徴は見られないが、ページアクセス完了率は早期に完了しない。クラスタ 2(黄線)は、クエリ変更完了は早期に完了しない。また、ページアクセス完了率は早期から上昇し始める。

$k=3$ のとき、クラスタ 1(青線)は、クエリ変更完了率が早期から上昇し始める。ページアクセス完了率は晩期になってから上昇し始めない。クラスタ 2(黄線)は、クエリ変更完了率がセッション後半に完了する。ページアクセス完了率は前半から上昇し始める。ク

²<https://scikit-learn.org/>

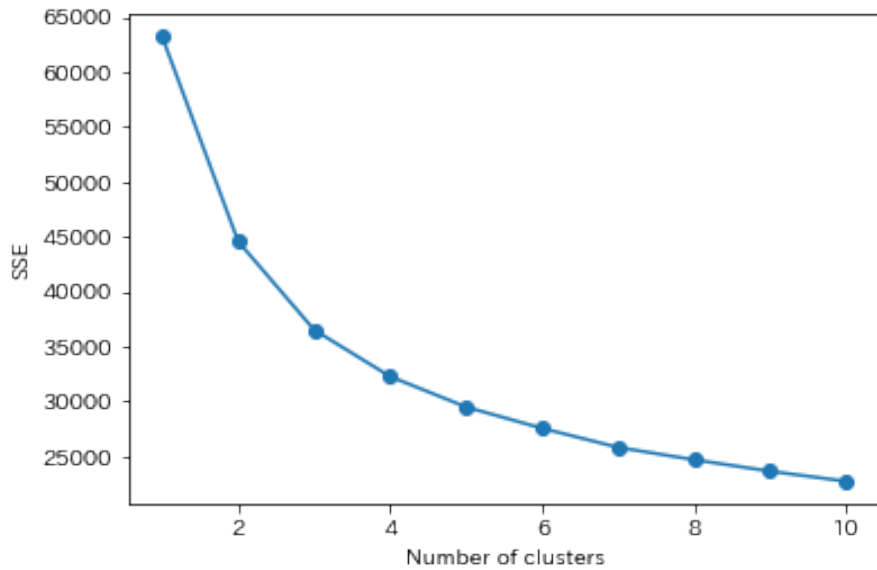


図 4.2: クラスタ内誤差平方和

クラスタ 3(緑線) は、クエリ変更完了率が早期に完了はせず、ページアクセス完了率が晩期に完了するのが特徴的である。

$k = 4$ のとき、クラスタ 1(青線) は、クエリ変更完了率が早期から上昇し始める。ページアクセス完了率は前半から上昇するが、完了は早期に行われない。クラスタ 2(黄線) は、クエリ変更完了率が晩期に完了する。ページアクセス完了率は前半から上昇するが、完了は後半になる。クラスタ 3(緑線) はクエリ変更完了率が後半に完了し、ページアクセス完了率が早期から上昇する。クラスタ 4(赤線) は、クエリ変更完了率が早期完了はせず、ページアクセス完了率は晩期に完了する。

$k = 2$ のときはクラスタ 1 のクエリ変更完了率に特徴が見られなかったため、分析をするには望ましくない。また、 $k = 4$ のときはクラスタ 2 とクラスタ 4 の差が小さいため、解釈を行うことは難しい。よって、本研究では $k = 3$ に設定して分析を行っていく。 $k = 3$ における 3 次元の完了率推移曲線を図 4.9 に、各クラスタの平均データ(セントロイド)を図 4.10 に示す。クラスタ 1 のデータ数は 23,052 件、クラスタ 2 は 31,446 件、クラスタ 3 は 19,719 件である。

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分布を図 4.11 に、出現確率の累積分布を図 4.12 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 4.1 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 1 は出現頻度のランキングが上位のキーワードの出現確率が高い。一方で、クラスタ 2 とクラスタ 3 はランキングが下位のキーワードの出現確率が高くなっている。

また、各クラスタの特徴語を評価する式 4.2 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 は、「レディース」「水着」「ドレス」などの衣料品や、「xperia」「iphone7」「スマホカバー」などのスマートフォン関連商品、「時計」「ショルダーバッグ」「バッグ」などの服飾品に関するキーワードが特徴語として抽出されている。クラスタ 2 は、特徴語の数が少なく、共通点も見られなかった。クラスタ 3 は、「送料無料」「お試し」「メール便」などのサービスに関するキーワード、「米」「水」「スイーツ」などの飲食物に関するキーワード、「キャンメイク」「化粧品」「洗顔」などの化粧品に関するキーワードが特徴語として抽出された。

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 4.3 のような結果になった。クラスタ 1 は、ラベル R の値が低いが、ラベル A の値がやや高くなっている。クラスタ 2 は、反対にラベル R の値が高いが、ラベル A の値が低い。クラスタ 3 はラベル R の値が高くなっている。

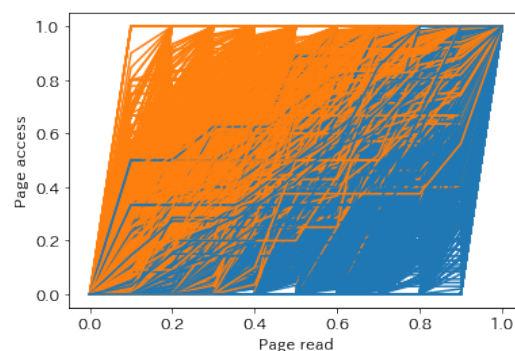
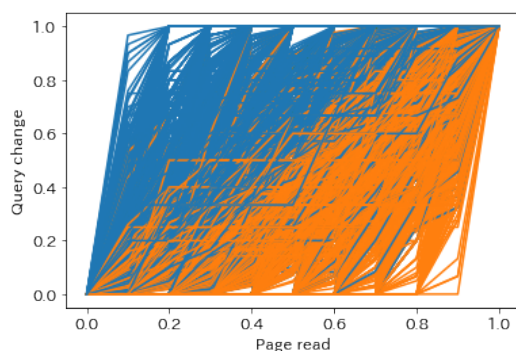


図 4.3: クエリ変更完了率の推移曲線 ($k = 2$) 図 4.4: ページアクセス完了率の推移曲線 ($k = 2$)

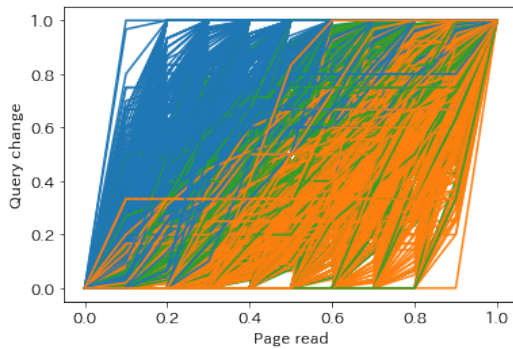


図 4.5: クエリ変更完了率の推移曲線 ($k = 3$)

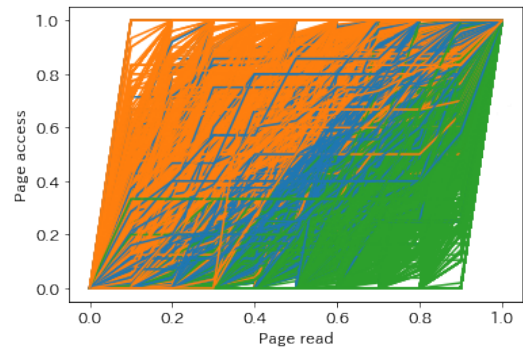


図 4.6: ページアクセス完了率の推移曲線 ($k = 3$)

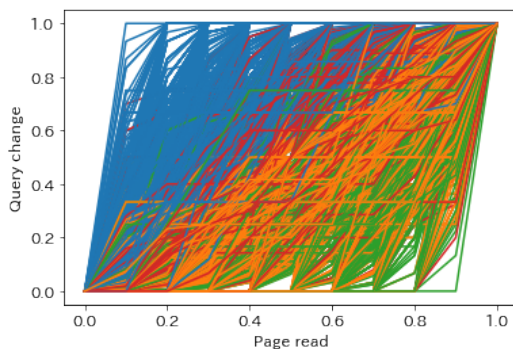


図 4.7: クエリ変更完了率の推移曲線 ($k = 4$)

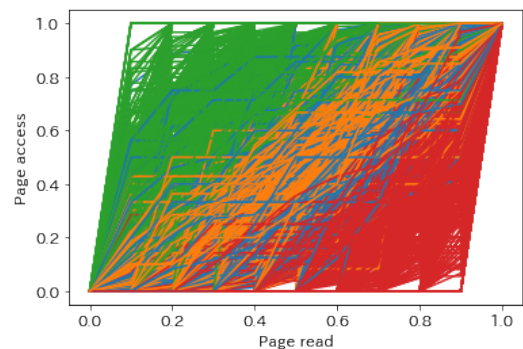


図 4.8: ページアクセス完了率の推移曲線 ($k = 4$)

4.3.4 クラスタの予測結果

図 4.13 はページ遷移行動とページアクセス回数の特徴量とした際の、各モデルの分類正解率である。SVM が学習するパス数が 2 のとき分類正解率が他の 2 つのモデルよりも低くなっているが、それ以外はモデル間に大きな差は見られなかった。

ランダムフォレストにおいて、特徴量別に学習・分類をした結果を図 4.14 に示す。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正解率を出している。個々の特徴量では、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が高いが、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。

さらにランダムフォレストにおいて、予測データのパス長別に算出した正解率を図 4.15 に示す。長いパスになるほどクラスタの分類正解率が低下している。学習パス数が長いほど分類正解率は高くなるが、予測対象のパス長が 20 を超えると、学習パス数が 2 と 3 の正解率にほとんど差はなくなっている。

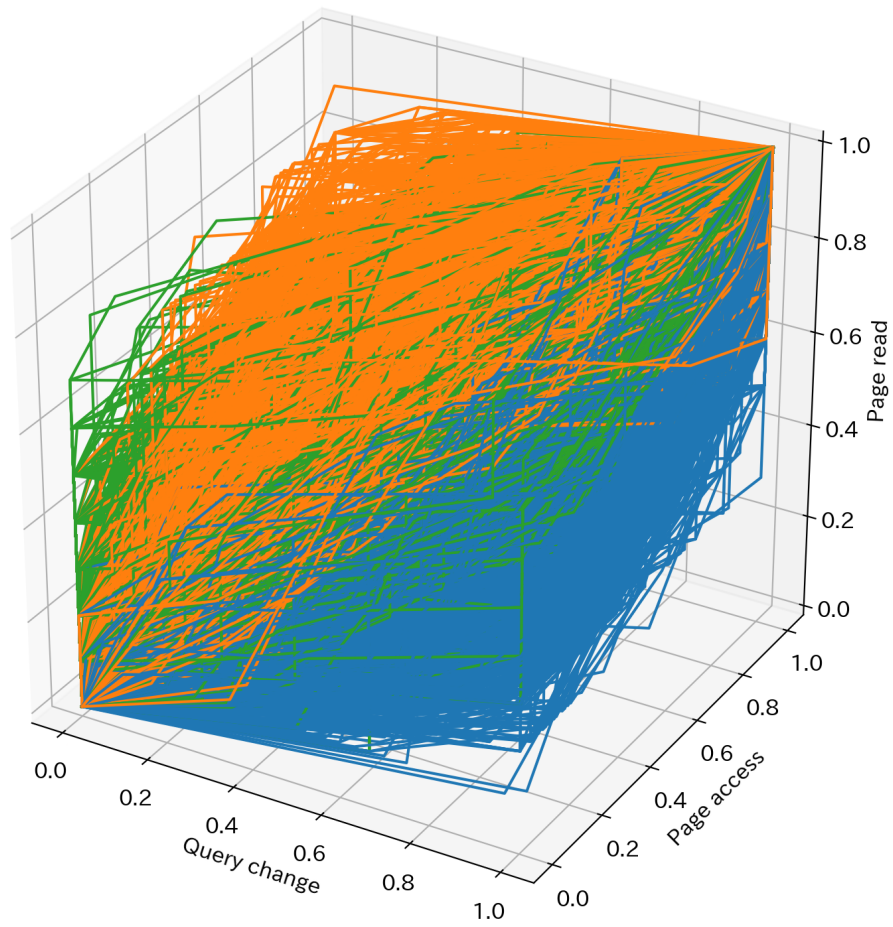


图 4.9: 完了率推移曲線 ($k = 3$)

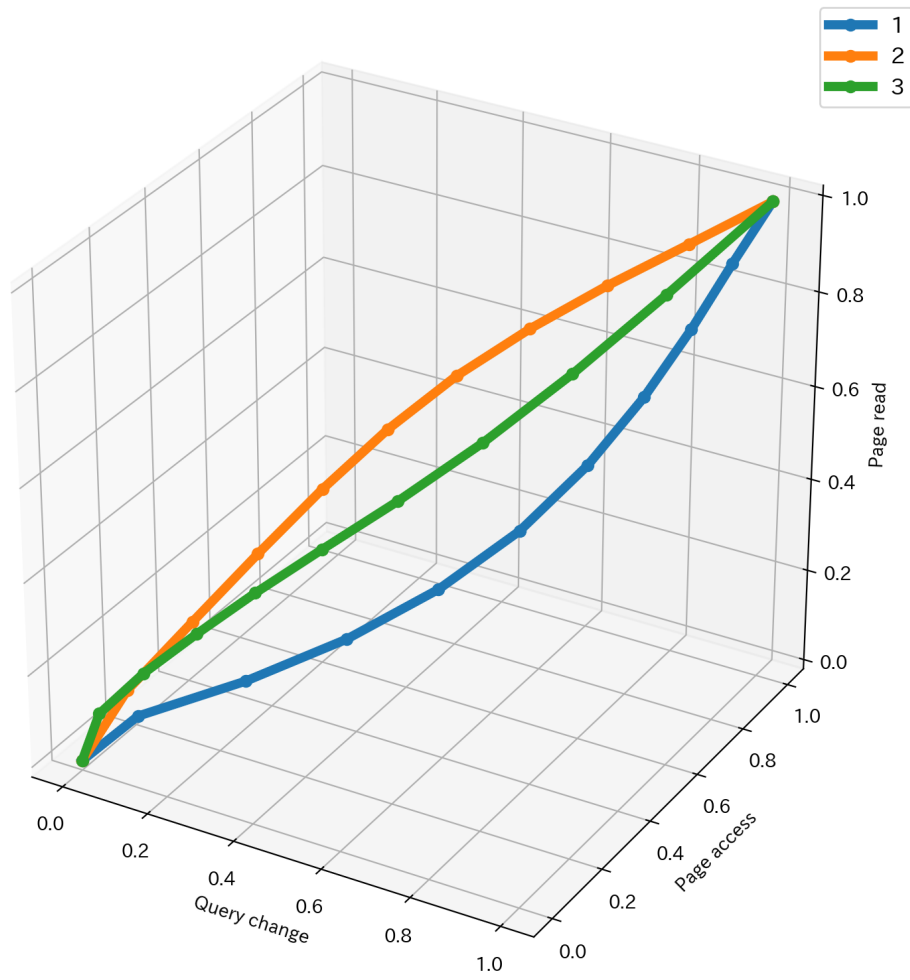


図 4.10: 各クラスにおける平均データ

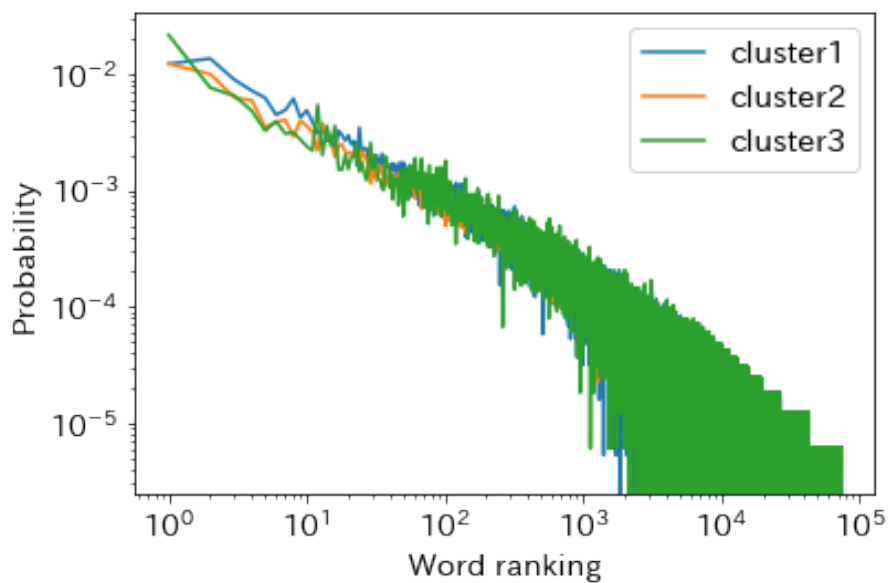


図 4.11: キーワードの確率分布

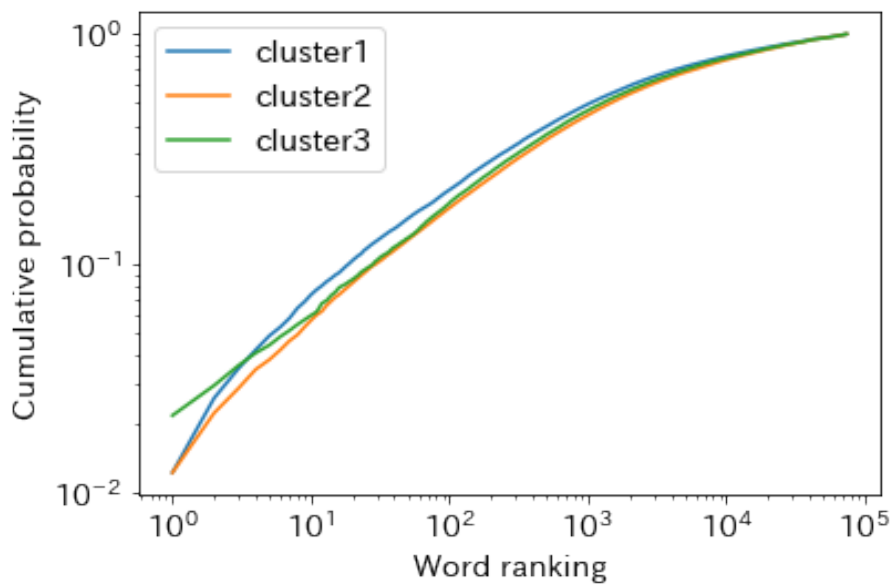


図 4.12: キーワードの出現確率の累積分布

表 4.1: 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

表 4.2: 各クラスタの特徴語

クラスタ	特徴語
1	'レディース', 'スマホケース', '水着', '大きいサイズ', '手帳', 'xperia', 'iphone7', 'スーツケース', 'ドレス', 'ハロウィン', 'シューズ', '時計', 'ショルダーバッグ', 'ラッシュガード', 'フォーマル', '傘', 'スマホカバー', 'スーツ', 'バック', '携帯ケース', 'エクスペリア', 'ベッド', 'クロックス', 'スリッポン', 'デニム', 'リング', 'ブラジャー', '子供服', 'テント', 'チュニック', 'ソファ', '牛肉', 'ロングカーディガン'
2	'掃除機', 'タンクトップ', '爽快ドラッグ', '進撃の巨人'
3	'送料無料', '米', 'コンタクトレンズ', '靴下', '水', 'ビール', 'コーヒー', 'お試し', 'キャンメイク', 'お米', '水筒', '炭酸水', '食品', 'メール便', '500 円', '無洗米', '化粧品', '本', 'お茶', '100 円', '洗剤', '柔軟剤', 'ダイエット', 'ポイント消化', '日焼け止め', 'スイーツ', 'クッキー', '商品券', '絵本', 'クリスタルガイザー', '洗顔', 'おむつ', 'スリッパ', 'ふりかけ', 'ナッツ', 'パン', 'うどん', 'sk-ii', 'プレスレット', '明太子', 'お取り寄せ', 'おしりふき', 'カレー', 'ワンデーアキュビューモイスト', 'ミネラルウォーター', '入浴剤', 'ゆめぴりか', 'くるみ', 'つや姫', 'ジュース', 'チーズケーキ', 'アクアエステソニック 2'

表 4.3: クエリ変更行動の出現確率

クラス	R	M	A	D
1	0.580	0.111	0.203	0.103
2	0.637	0.104	0.161	0.096
3	0.634	0.094	0.180	0.090

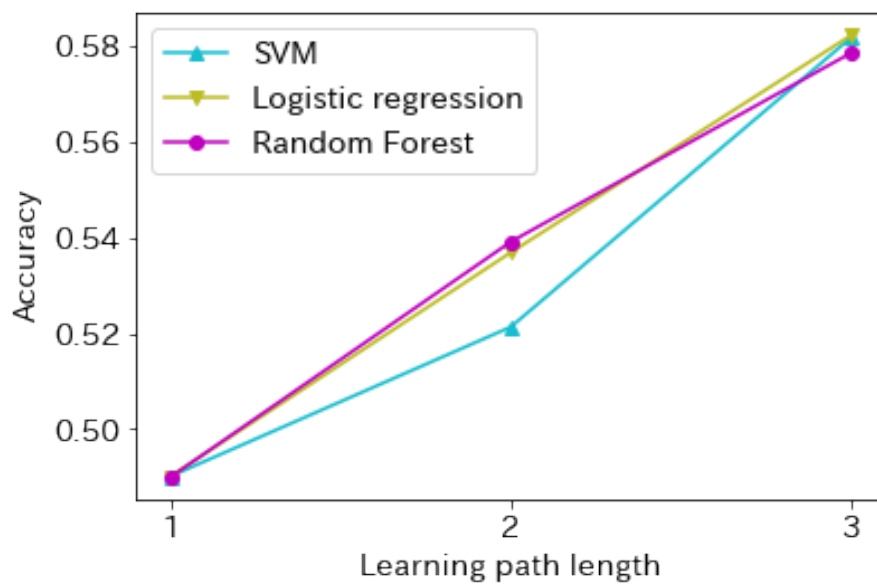


図 4.13: モデル別の分類正解率

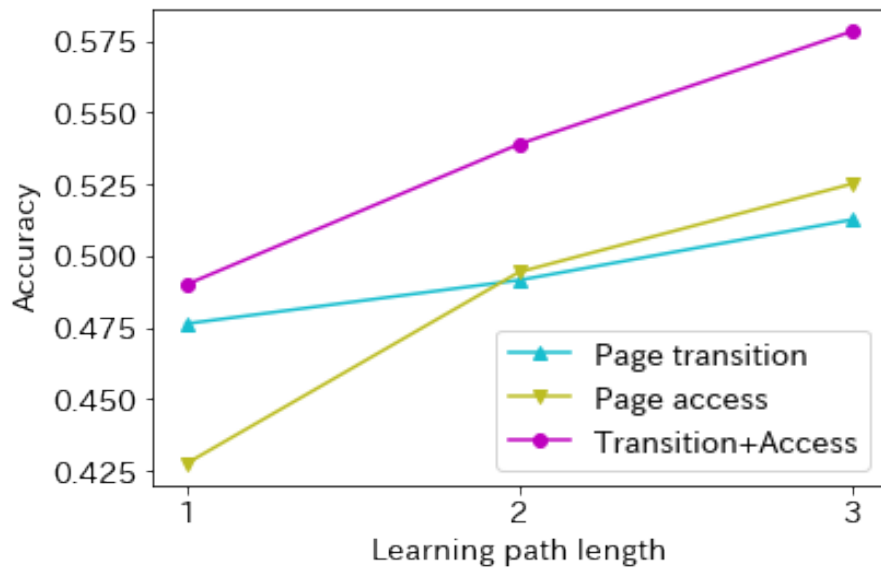


図 4.14: 特徴量別の分類正解率

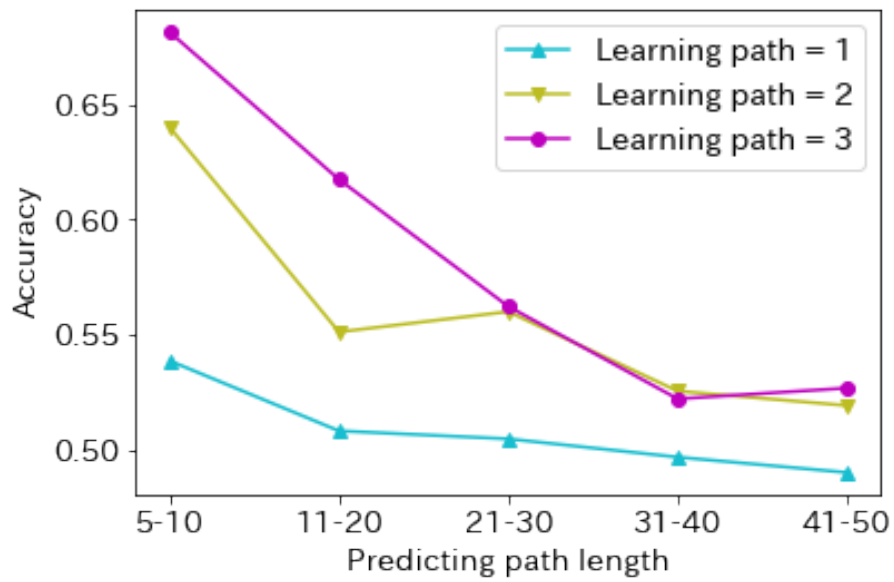


図 4.15: パス別の分類正解率

第5章 考察

5.1 クラスタリングの結果について

評価実験では k-means の k の値を 2~4 にして実施し、各クラスタの特徴を観察した。その結果、 $k = 3$ のときクラスタ間の特徴に明確な差が見られた。各クラスタの特徴を取り上げ、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早期に行われ、ページアクセスはセッションの晩期にし始めることはないセッション群である。特徴としては、頻度の高いキーワードが多く使われ、衣料品、スマートフォン、服飾品に関するキーワードが他のクラスタと比べて特徴的に出現する。また、クエリ変更パターンは全部置換 (R) の頻度が少なく、追加 (A) の頻度が他クラスタよりも高い。

以上のことから、クラスタ 1 では曖昧なクエリから始まり、キーワードを追加して商品を絞り込むような探索をするというモデルが考えられる。

よって、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多く推薦することが効果的な検索支援であることが示唆される。

クラスタ 2

クラスタ 2 はクエリ変更がセッションの晩期で完了し、ページアクセスはセッションの早期から行われるセッション群である。特徴として、頻度の低いキーワードの出現確率が高く、クエリ変更では全部置換 (R) の頻度が高く、追加 (A) の頻度が低い。

クエリ変更がセッションの晩期まで行われる一方で、ページアクセスは早期から行われることから、クラスタ 2 ではユーザは興味のあるページにアクセスをしながら知識を蓄え、クエリを洗練させていくという検索モデルが考えられる。

ゆえに、クラスタ 2 では現在のクエリに対して絞込を行うようなクエリを提示するというよりも、関連するキーワードを幅広く推薦する方が有効な検索支援であることが示唆される。

クラスタ 3

クラスタ 3 はクエリ変更が早期に完了せず、ページアクセスが晩期に完了するようなセッション群である。特徴語としてサービスや、飲食物、化粧品に関するキーワードが抽

出されており，全部置換 (R) のクエリ変更パターンの頻度が高い。

クエリ変更が早期に完了せず，ページアクセスが晩期に完了することから，クラスタ 3 ではクエリ変更やページ遷移を行いながら，満足がいくまで商品にアクセスをして比較していることが考えられる。

クラスタ 3 ではサービスに関するキーワードが特徴的だったことから，お得な商品を探していることが考えられ，値引きされている商品やポイントが多く付与されるような商品を推薦することが有効な検索支援であるといえる。

5.2 クラスタの予測結果について

図 4.13 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，どの機械学習のモデルでも学習パス数が 3 では 0.58 前後の分類正解率，学習パス数が 1 だけでも 0.49 前後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を手がかりに，ある程度の精度で行動が予測できると考えられる。

また，図 4.14 より，ページ遷移行動とページアクセス回数どちらも 0.33 を超える正解率で分類ができているため，クラスタの予測に有効な特徴量であるといえる。

図 4.15 の結果では，予測対象のセッションのパスが長いほど，クラスタの分類正解率が低下した。これは本実験では学習パス数を 1~3 で固定したため，予測データのパス長が長くなるほど相対的に学習データ量の割合が低下することが原因しているのではないかと考えられる。

第6章 結論

6.1 本研究のまとめ

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クラスタリングし、各クラスタにおける特徴的な行動推移、キーワード、クエリ変更パターンを分析することで、ユーザの商品選択行動のモデル化を行った。

評価実験において、k-means クラスタリングを行ったところ、 $k=3$ のとき絞り込みの検索をしているクラスタ、知識を拡充していくクラスタ、商品の比較をしているクラスタの検索モデルに分類することができた。また、各クラスタに対して適切な検索支援方法が示唆された。

最後にインタラクティブ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0.58前後の分類正解率、学習パス数が1だけでも0.49前後の分類正解率を算出した。

6.2 今後の課題

今後の課題として、第一に完了率推移曲線の構築手法を改善することである。本研究では、クエリ変更完了率やページアクセス完了率を算出する際、すべての行動を同じ重みとして扱っていた。しかし、実際にクエリの遷移やアクセスしたページ間の関係に着目してみると、内容が大きく異なっている関係や反対にかなり近い関係などが存在する。前後の行動間の大きさを考慮して完了率推移曲線を構築することで、より正確なユーザ行動の把握が可能になることが期待できる。

第二の課題として、クラスタの予測タスクにおいて分類正解率を上昇させることである。本研究では特徴量として、ページ遷移行動とページアクセス回数のみを利用したが、クエリ情報や滞留時間などを特徴量に加えることでさらなる分類正解率の向上を目指す。

謝辞

本研究は、筆者が筑波大学図書館情報メディア研究科博士前期課程に在学中の研究成果をまとめたものである。主研究指導教員の佐藤哲司教授には、学部4年次から手厚いご指導をいただきました。また、副研究指導教員の若林啓助教からは、院生合宿や学会などを通じて的確なアドバイスをいただきました。同研究室の院生である神澤義人さん、稲福和史さんからは、プログラムの実装や手法のアイデアの面で議論することができました。学部生の後輩である兼信光さん、高田未央さん、高橋武さん、野澤優佳さんの懸命に努力する姿にこちらも励まされました。ここに感謝の意を表します。

参考文献

- [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] Laura A. Granka, Thorsten Joachims, and Geri Gay. Eye-tracking analysis of user behavior in www search. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pp. 478–479, New York, NY, USA, 2004. ACM.
- [3] 加藤勇太, 岩本健嗣, 松本三千人. タッチ操作ログを用いた web コンテンツ閲覧時における興味度合い推定の研究. *情報処理学会論文誌*, Vol. 59, No. 2, pp. 508–518, feb 2018.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] Georg Buscher, Ryen W. White, Susan Dumais, and Jeff Huang. Large-scale analysis of individual and task differences in search result page examination strategies. In

Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12, pp. 373–382, New York, NY, USA, 2012. ACM.

- [10] Daan Odijk, Ryen W. White, Ahmed Hassan Awadallah, and Susan T. Dumais. Struggling and success in web search. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM '15*, pp. 1551–1560, New York, NY, USA, 2015. ACM.
- [11] Youngho Kim, Ahmed Hassan Awadallah, Ryen W. White, and Imed Zitouni. Modeling dwell time to predict click-level satisfaction. In *The 7th Annual International ACM Conference on Web Search and Data Mining (WSDM 2014)*. ACM, February 2014.
- [12] Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. A context-aware time model for web search. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16*, pp. 205–214, New York, NY, USA, 2016. ACM.
- [13] 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.
- [14] Qi Guo and Eugene Agichtein. Ready to buy or just browsing?: Detecting web searcher goals from interaction data. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10*, pp. 130–137, New York, NY, USA, 2010. ACM.
- [15] 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.
- [16] Jing Zhu, Xue Han, Rong Ma, Xiaowei Li, Tong Cao, Shuting Sun, and Bin Hu. Exploring user mobile shopping activities based on characteristic of eye-tracking. In Qiaohong Zu and Bo Hu, editors, *Human Centered Computing*, pp. 556–566, Cham, 2016. Springer International Publishing.
- [17] 大河原一輝, 平野廣美, 益子宗, 星野准一. ショッピングモール型 ec サイトのための店舗情報視覚化システム. 情報処理学会論文誌, Vol. 56, No. 3, pp. 847–855, mar 2015.
- [18] 関口裕一郎, 田中智博, 内山匡, 藤村滋, 望月崇由, 鈴木智也. 検索クエリログのセッション情報を利用した属性語句抽出. 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM2010), 2010.
- [19] 鹿島好央, 北山大輔. Word2vec と web 検索を用いた検索クエリ置換手法. 第 9 回データ工学と情報マネジメントに関するフォーラム (DEIM2017), 2017.

- [20] 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.
- [21] 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.
- [22] 清水涼人, 服部俊一, 高間康史. レビュー閲覧履歴からの価値観モデリングを用いた情報推薦システム. 第5回 Web インテリジェンスとインタラクション研究会, 2014.
- [23] 西尾義英, 藤井絵美子, 安松健. Web サイト閲覧行動のモデリングと評価. 人工知能学会全国大会論文集, Vol. JSAI2014, pp. 1L3OS17b1–1L3OS17b1, 2014.
- [24] 渡邊恵太, 加藤昇平. トピックモデルと協調フィルタリングに基づくユーザ興味を反映した情報推薦システム. 人工知能学会全国大会論文集, Vol. JSAI2014, pp. 2M34–2M34, 2014.
- [25] 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.

発表論文

学術雑誌論文

- Shuhei Yamamoto, Kei Wakabayashi, Tetsuji Satoh, Yuri Nozaki, Noriko Kando. Twitter user growth analysis based on diversities in posting activities. International Journal of Web Information Systems, Vol. 13, No. 4 pp. 370-386, 2017.

国際会議論文

- Yuri Nozaki, Tetsuji Satoh. Public/Private Tweet Classification Method to Prevent Posting to Wrong Accounts. Proceedings of 19th International Conference on Information Integration and Web-based Applications & Services, iiWAS2017, pp. 151-155, 2017.
- Kouga Kobayashi, Yuri Nozaki, Takayasu Fushimi, Tetsuji Satoh. Category Reformation using Purchase Logs. Proceedings of 19th International Conference on Information Integration and Web-based Applications & Services, iiWAS2017, pp. 146-150, 2017.
- Yuri Nozaki, Tetsuji Satoh. Category Classification Methods Reflecting Item's Search Behaviors in Online Shopping Sites. 9th International Conference on E-Service and Knowledge Management (ESKM 2018) in conjunction with 7th International Congress on Advanced Applied Informatics, pp. 32 - 37, 2018.
- Yuri Nozaki, Tetsuji Satoh. Search Log Analysis Method of Online Shopping Sites for Navigating Item Categories. Proceedings of 20th International Conference on Information Integration and Web-based Applications & Services, iiWAS2018, pp. 85-93, 2018.
- Yuri Nozaki, Fumiya Watanabe, Tetsuji Satoh. Analysis of Item Selection Behavior in Online Shopping. Proceedings of 20th International Conference on Information Integration and Web-based Applications & Services, iiWAS2018, pp. 41-45, 2018.

国内会議論文

- 野崎祐里, 佐藤哲司. Twitter の複数アカウント所持者を対象とした投稿アカウント推定手法の提案. 第9回データ工学と情報マネジメントに関するフォーラム DEIM2017, D1-1, 2017.
- 野崎祐里, 佐藤 哲司. 複数アカウント所持者のための企業・私的ツイート判別手法の提案. 情報処理学会, マルチメディア, 分散, 協調とモバイル (DICOMO2017) シンポジウム論文集, pp. 995 - 1001, 2017.
- 小林滉河, 野崎祐里, 伏見卓恭, 佐藤哲司. 購買検索ログを使用した注釈付きカテゴリの提案. WebDB フォーラム 2017, 電子情報通信学会技術研究報告 DE2017-18, pp. 69-74, 2017.,
- 野崎祐里, 佐藤哲司. クエリ変更とページアクセスに基づく EC サイトの商品カテゴリ分析手法の提案. 第10回データ工学と情報マネジメントに関するフォーラム DEIM2018, D4-2, 2018.
- 渡辺郁弥, 野崎祐里, 佐藤哲司. オンラインショッピングにおける商品選択プロセスのモデル化手法の提案. WebDB フォーラム 2018, 電子情報通信学会技術研究報告 DE2018-16, pp. 43-48, 2018.