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

Category Analysis Based on Query Change and Page Access in Online Shopping

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　Web ページや EC サイト，SNS などにおいて，ユーザは目的のページにアクセスするために，クエリを試行錯誤しながら探索行動を行っている．ユーザは入力したクエリが適切だと確信した際は，同一のクエリでページ遷移をしながら探索を行い，クエリが不適切だと判断した際はクエリを変更する．そして，目的に合うようなページを発見したらページにアクセスし，ユーザの目的が達成されれば探索が終了，達成されなければ探索を続行する．このような探索行動は，検索対象のページのカテゴリによって異なると考えられる．本稿では，EC サイトのログデータからセッション情報を抽出し，クエリ変更とページ遷移，ページアクセスの系列変化の分析を行う．各セッションに対して，クエリ変更完了率とページアクセス完了率をセッションの各時点ごとに算出し，それぞれの完了率の到達の早さから，カテゴリの特徴を分析する．分析の結果から，有用な知見が得られたのでここに報告する．

キーワード　情報検索，EC サイト，カテゴリ，セッション分析，クラスタリング

1. はじめに

インターネット技術の発達により，ユーザが Web 上で検索を行う機会が増えてきている．ユーザはクエリを入力して，欲しい情報のあるページを探し出すが，検索結果のページを見て入力したクエリが適切でないと判断した際はクエリの修正を行う．また, クエリが正しいと確信したときは，同一クエリのまま，次のページに移動し探索を続ける．クエリの修正とページの移動を組み合わせ, 目的に合うページを発見したらページへのアクセスを行う．アクセスしたページがユーザの要求を満たしたとき，そこで探索が終了し，満たさない場合は探索行動を続行する．

このような探索行動は，ユーザの探し出したいページのカテゴリに依存するのではないかと考える．クエリ変更がセッションの後半時点まで行われるようなカテゴリは，適切なクエリをなかなか見つけられないことから，クエリの想起が難しいカテゴリであるといえる．また，ページアクセスがセッションの後半に偏ることが多いカテゴリは，目的のページまでたどり着くのが難しいカテゴリとみなすことができる．

本稿は EC サイトのアクセスログを用いて，セッションを切り出しクエリ変更回数とページ遷移回数，ページアクセス回数をもとに，クエリ変更完了率とページアクセス完了率の系列データを算出する．クエリ変更完了率とページアクセス完了率それぞれのデータにクラスタリングを行い，クエリ変更完了率のクラスタとページアクセス完了率のクラスタの組み合わせから，商品カテゴリの特徴を分析する．探索行動の特徴をカテゴリごとに把握することで，カテゴリごとに適した探索支援の基盤を構築することができると考える．

本稿の構成を以下に示す．第 2 章では，本稿と関連性が高いクエリログやアクセスログを利用した研究を提示し，本稿の位置づけを示す．第 3 章では提案手法として，ログデータからセッションを抽出し，クエリ変更完了率とページアクセス完了率の系列データを算出する手法，セッションのクラスタリング手法について説明する．第 4 章で実データを用いた評価実験を行い結果を提示する．第 5 章では，前章の結果について考察を行う．そして，第 6 章で本稿のまとめと今後の課題を述べる．

2. 関連研究

情報検索分野でログデータを用いた分析は数多くなされている．クエリログを用いた研究として，山口ら [1] は同位語の発見手法を提案している．共起する語群との cos 類似度や出現順序，HITS アルゴリズムなど様々手法で，抽出される同位語候補の比較分析をしている．関口ら [2] は，シード語句に対して絞込に使用している語句群の類似度を計算することで，同一の属性語の抽出を行っている．深澤ら [3] は，レシピ検索サイトにおいて共起する食材の頻度を時系列に追い，その分散の変化を見ることで検索語の意味変化を分析している．関口ら [4] は機械学習を用いて，クエリ変更意図を絞込，汎化，関連，修正，新規の 5 つのクラス分類タスクを行っている．矢野ら [5] は，クエリの曖昧性をクリックページのばらつきで評価する際，ページ間のトピックの差を考慮した手法を提案している．Kobayashi ら [6] は，オンラインショッピングサイトのクエリ中キーワードに対するアクセスしたページの確率を行列分解することで，新たなカテゴリの生成およびカテゴリに対応する単語の推薦を

表 1　セッション抽出例

ユーザ ID	タイムスタンプ	クエリ	キーワード 1	キーワード 2	カテゴリ	セッション	ラベル
1	2016-09-05 19:37:41	usb メモリ	usb メモリ			1	s
1	2016-09-05 19:37:48	usb メモリ	usb メモリ			1	t
1	2016-09-05 19:38:25	usb メモリ 送料無料	usb メモリ	送料無料		1	c
1	2016-09-05 19:38:34	usb メモリ 送料無料	usb メモリ	送料無料	computer	1	a
1	2016-09-05 19:41:44	花	花			2	s
1	2016-09-05 19:53:40	花 プレゼント	花	プレゼント		2	c
2	2016-09-05 19:55:11	花	花			3	s

行っている．

ログを系列データとして処理する研究も行われている．Boldi ら [7] はクエリログからクエリフローグラフの構築を行っている．これはクエリ間の遷移確率をグラフで表現したもので，クエリの系列からクエリ候補の推薦手法を提案している．また，Bordino ら [8] はクエリフローグラフを用いて類似クエリの抽出を行っている．早川ら [9] は，Web ページのアクセスログを系列データとしてグラフ化し，属性の付与や頻出パターンの抽出を行っている．

本研究は，アクセスログからセッションを抽出して分析を行うので，系列ログデータ分析の研究に近い．先行研究では，クエリ中のキーワードの内容に着目しているが，本研究はキーワードの内容までは扱わず，探索行動パターンのみに焦点を当てた分析を行う．

3. 提 案 手 法

3.1　セッションの抽出

検索ログと商品のアクセスログのデータからセッションの抽出を行う．検索ログとは，ユーザ ID，タイムスタンプ，検索クエリのフィールドを持ち，アクセスログは．ユーザ ID，タイムスタンプ，アクセスした際の検索クエリ，アクセスしたページの商品カテゴリを持つデータを使用する．以下，セッションの抽出手法について説明する．まず，ユーザ ID を用いて同一ユーザの検索ログ，アクセスログを 1 つにまとめる．次に集めたユーザのログ群をタイムスタンプを元にソートする．ソートされたログに対して，前後のクエリに共通するキーワードが 1 つ以上存在するとき，それらを同一のセッションとする．キーワードとは，クエリを全角または半角で分割した際の各文字列である．

抽出されたセッションの各ログに対して，探索行動情報のアノテーションを行う．アノテーション対象のログを直前に出現した検索ログと比較し，クエリが一致している場合に transition を意味する「t」の文字列を，クエリが一致しない場合に change を意味する「c」をアノテーションする．本研究では，ログの前後のキーワードが一致する場合，同一のクエリで別のページに遷移したとみなす．セッション最初のログは直前のクエリと比較ができないため，便宜的に start を意味する「s」をアノテーションする．また，アクセスログが出現した場合は access を意味する「a」を付与する．

セッション抽出例を表 1 に示す．ログの種類がアクセスログの場合のみ，カテゴリのフィールドに値が入る．セッション 1 は「usb メモリ」，セッション 2 は「花」が共通するキーワードとなっている．セッション 3 は，キーワード「花」がセッション 2 と共通しているが，ユーザが異なるため別のセッションとして切り出されている．

3.2　特徴量の算出

アノテーション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．算出方法は以下の通りである．

クエリ変更完了率の算出手法

- セッション最初のラベル「s」を初期値 0 とし，順に走査
- 「c」が出現したとき末尾の値に+1 した値を追加
- 「t」が出現したとき末尾の値をそのまま追加
- 走査終了後，数値列を最大値が 1 になるように正規化

ページアクセス完了率の算出手法

- セッション最初のラベル「s」を初期値 0 とし，順に走査
- 「a」が出現したとき，末尾にある値に+1
- 「t」または「c」が出現したとき末尾の値をそのまま追加
- 走査終了後，数値列を最大値が 1 になるように正規化

例としてセッションにアノテーションされたラベルが「s, c, c, t, t, a, t, a」であるとき，クエリ変更完了率算出の流れを表 2，ページアクセス完了率算出の流れを表 3 に示す．同じラベルでも，クエリ変更完了率とページアクセス完了率の算出結果は異なる．

2 種類の完了率は，系列データとしてグラフに表すことができる．先程の例で算出されたクエリ変更完了率は図 1，ページアクセス完了率は図 1 のようになる．グラフから，クエリ変更が早期に完了し，ページアクセスが後半に集中しているセッションであることがわかる．セッション中に出現する「s」「t」「c」のラベルの総和がセッションのパス長になる．

3.3　クラスタリング

算出したクエリ変更完了率，ページアクセス完了率それぞれのデータに対して，クラスタリングを行う．クラスタリングの手法は k-medoids を用いる．k-medoids はデータを k 個のクラスタに分類するために，クラスタ内の非類似度が最小となるようなデータ (medoids) を代表点として探し出す教師なし学習のアルゴリズムである．データを $D = \{d_1, d_2, \ldots, d_n\}$，クラスタを $D = \{C_1, C_2, \ldots, C_k\}$ とすると，medoids は以下の式 (1) で定義される．

$$min_{\in}(C_i - d) \sum_{d' \in (C_i - d)} dist(d, d') \quad (1)$$

k-medoids を行うためにはデータ間の距離行列を算出する必要があり，本稿ではユークリッド距離を利用した．クエリ変更完了率，ページアクセス完了率それぞれのクラスタの組み合わせの所属確率をカテゴリ間で比較を行い分析する．

4. 評価実験

4.1 実験準備

提案手法を実データに対して適用し，カテゴリ間の探索行動について分析を行う．データセットとして，株式会社リクルートテクノロジーズから提供されたポンパレモールのアクセスログデータを使用する．アクセスログデータには，ユーザ ID，タイムスタンプ，検索クエリが格納された検索キーワードログと，ユーザ ID，タイムスタンプ，アクセスした際の検索クエリ，商品の上位カテゴリ，商品の下位カテゴリの情報が格納されたアイテムクリックログがフィールドとして存在する．データセットは 2016 年 1 月〜2016 年 12 月の期間のデータをリサンプルしたもので，検索ログ，商品アクセスログを合わせるとデータの総数は 36,426,807 である．データからセッションを抽出した際の，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図 3〜5 に示す．

評価実験では，セッション中に出現するクリックログの上位カテゴリが多い順に並び替え，上位 3 つのカテゴリ「womens-fashion」「grocery」「daily-goods」のデータを使用する．womens-fashion はレディースファッション，grocery は食品，daily-goods は日用雑貨を表す．表 4 に上位カテゴリの出現セッション数上位 10 件を示す．実験用データとして，セッション内のアイテムクリックログ中に対象カテゴリが出現し，クエリ変更が 1 回以上行われているセッションから，パス長が 10 のセッションを各カテゴリ 400 件ずつ計 1,200 件のセッションを収集した．

表 2　クエリ変更完了率算出例

	ラベル	処理	結果
1	s	初期値 0 として走査開始	0
2	c	末尾の値に+1 した値を追加	0,1
3	c	末尾の値に+1 した値を追加	0,1,2
4	t	末尾の値をそのまま追加	0,1,2,2
5	t	末尾の値をそのまま追加	0,1,2,2,2
6	a	何もしない	0,1,2,2,2
7	t	末尾の値をそのまま追加	0,1,2,2,2,2
8	a	何もしない	0,1,2,2,2,2
		最大値 1 に正規化	0,0.5,1,1,1,1

表 3　ページアクセス完了率算出例

	ラベル	処理	結果
1	s	初期値 0 として走査開始	0
2	c	末尾の値をそのまま追加	0,0
3	c	末尾の値をそのまま追加	0,0,0
4	t	末尾の値をそのまま追加	0,0,0,0
5	t	末尾の値をそのまま追加	0,0,0,0,0
6	a	末尾にある値に+1	0,0,0,0,1
7	t	末尾の値をそのまま追加	0,0,0,0,1,1
8	a	末尾にある値に+1	0,0,0,0,1,2
		最大値 1 に正規化	0,0,0,0,0.5,1

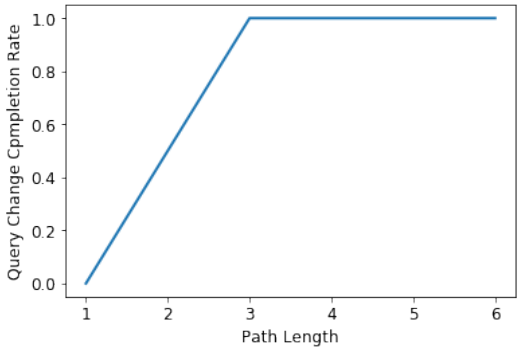

図 1　クエリ変更完了率の出力例

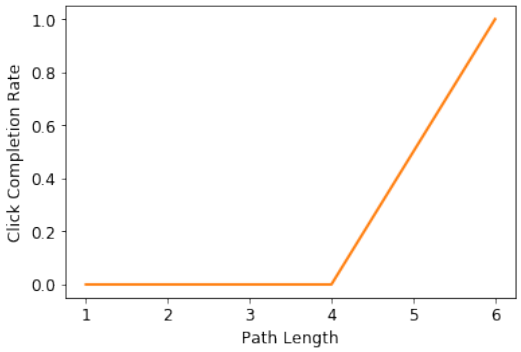

図 2　ページアクセス完了率の出力例

表 4　上位カテゴリの出現セッション数上位 10 件

上位カテゴリ名	出現セッション数
womens-fashion	237011
grocery	214737
daily-goods	165596
sports	117130
home-appliances	115607
cosmetics	109675
shoes	105969
sweets-snacks-cakes	104715
books	99888
bags	95273

k-medoids でクラスタリングを行う際，クラスタ数 k を人手で定める必要があるが，クラスタ数が多くなると各クラスタの所属確率が低くなってしまい組み合わせの解釈が難しくなるため，クエリ変更完了率，ページアクセス完了率それぞれ k=3 に設定する．k-medoids は初期値に選ぶ medoids によって結果が変わるため，クラスタリングを 10 回繰り返し，クラスタ内二乗誤差の総和が最小となる結果を採用する．

4.2 実験結果

収集したデータのクエリ変更完了率，ページアクセス完了率を図 6〜7 に示す．また，クラスタリングの結果を図 8〜9 に示す．これは k-medoids における各クラスタの代表データを

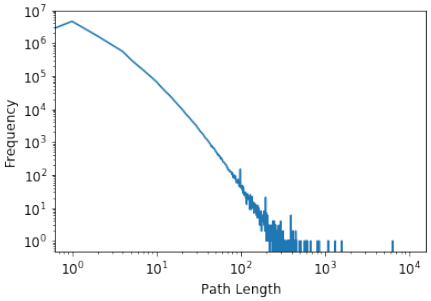

図 3　パス長の頻度分布

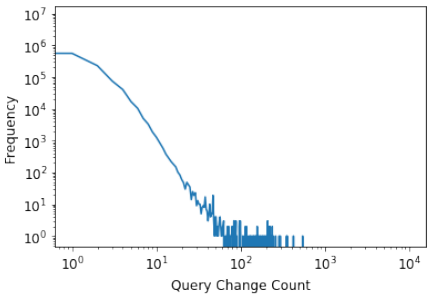

図 4　ページ変更回数の頻度分布

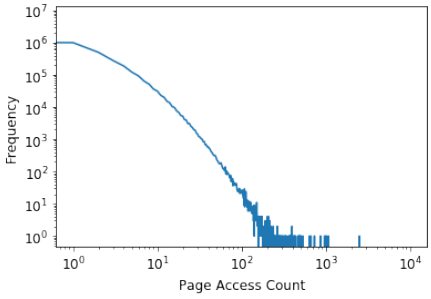

図 5　ページアクセス回数の頻度分布

プロットしたものである．各クラスタのデータ件数および代表データの各パス地点における具体的な値を表 5〜6 に示す．クラスタ a はクエリ変更完了率の中で最もデータ数が多いクラスタで，クエリ変更完了率が徐々に上昇している．クラスタ b はクエリ変更完了率の到達が早く，クラスタ c は到達が遅いクラスタである．クラスタ d はページアクセス完了率の到達が遅く，クラスタ e は到達が早くなっている．クラスタ f はページアクセス完了率の中で最もデータが多いクラスタで，完了率が徐々に上昇している．

各商品カテゴリのクラスタ所属確率の結果を表 7 に示す．セル内で上段の値が womens-fashion，中段が grocery，下段が daily-goods のクラスタ所属確率を表している．また，クラスタの項目は，アルファベット順ではなく，完了が早く行われる順に並び替えてある．

他の商品カテゴリよりも 0.05 以上所属確率に差があるクラスタの組み合わせを特徴的な行動とすると，womens-fashion は (b，e) のクラスタの組み合わせの割合が他のカテゴリより高い．grocery は (a，d)(b，d) が高く，(a，f)(b，f) が他よりも低い．daily-goods は (b, e) クラスタ以外は，womens-fashion と特徴的な差が見られなかった．

5. 考　　察

表 7 より，各カテゴリにおける特徴的なクラスタについて考察する．womens-fashion は（b, e）クラスタが他のカテゴリよりも高かったが，このクラスタはクエリの変更，ページアクセスともに早期に完了しているにもかかわらずページの移動を続けている行動を表している．クエリの変更を行わず探索を続けているため，最適なクエリは見つかっているものの商品を探すのに手間取っていることが考えられる．そのため，womens-fashion カテゴリには，関連する商品を推薦することが有効な検索支援の手法であるといえる．

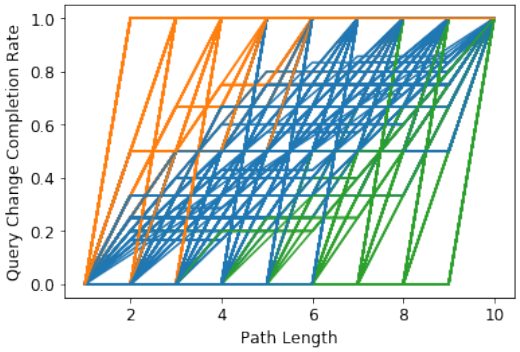

図 6　クエリ変更完了率

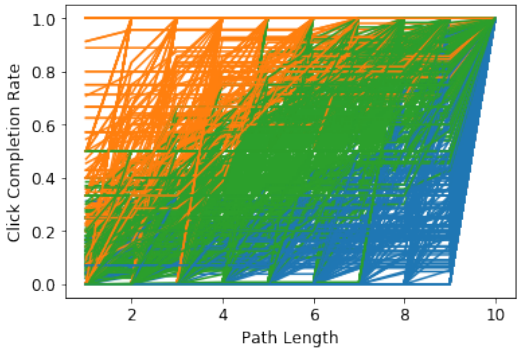

図 7　ページアクセス完了率

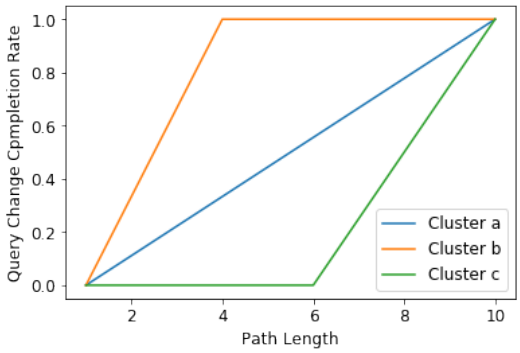

図 8　クエリ変更完了率のクラスタ代表値

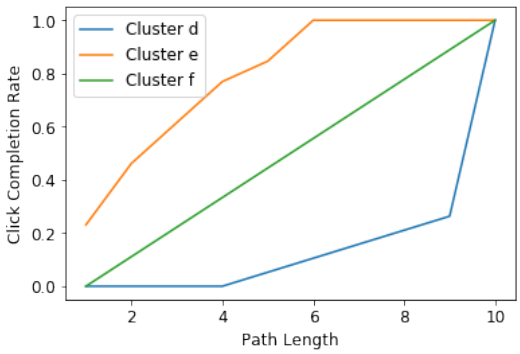

図 9　ページアクセス完了率のクラスタ代表値

表 5　クエリ変更完了率のクラスタ代表値

クラスタ名	件数	1	2	3	4	5	6	7	8	9	10
a	603	0	0.11	0.22	0.33	0.44	0.55	0.66	0.77	0.88	1.0
b	359	0	0.33	0.66	1.0	1.0	1.0	1.0	1.0	1.0	1.0
c	238	0	0	0	0	0	0	0.25	0.5	0.75	1.0

表 6　ページアクセス完了率のクラスタ代表値

クラスタ名	件数	1	2	3	4	5	6	7	8	9	10
d	368	0	0	0	0	0.05	0.1	0.15	0.21	0.26	1.0
e	260	0.23	0.46	0.61	0.76	0.84	1.0	1.0	1.0	1.0	1.0
f	572	0	0.11	0.22	0.33	0.44	0.55	0.66	0.77	0.88	1.0

表 7　商品カテゴリごとのクラスタ所属確率
上段：womens-fashion
中段：grocery
下段：daily-goods

		ページアクセス完了率			
		クラスタ e	クラスタ f	クラスタ d	計
クエリ変更完了率	クラスタ b	0.10	0.16	0.07	0.33
		0.04	0.11	0.13	0.28
		0.05	0.19	0.06	0.30
	クラスタ a	0.09	0.26	0.11	0.46
		0.11	0.19	0.23	0.53
		0.11	0.27	0.15	0.53
	クラスタ c	0.07	0.10	0.05	0.22
		0.05	0.08	0.07	0.20
		0.04	0.09	0.06	0.19
	計	0.26	0.52	0.23	
		0.19	0.38	0.43	
		0.20	0.55	0.27	

grocery は (a，d)(b，d) が高く，(a，f)(b，f) が低いという結果になった．(a，d) はクエリの変更を行い続けているが，ページアクセスがセッションの後半に集中する行動を表しており，有効なクエリがなかなか見つからないことが考えられる．(b，d) はクエリの変更の完了は早いが，ページアクセスが後半に集中する行動である．これは，検索結果のページを移動していかなければ見つからない商品を探していると考えられる．ページを移動していかなければ商品が見つかってないため，そもそもクエリが最適でない可能性も考えられる．(a，f) はクエリの変更を行いながら，ページアクセスを行うような行動であり，ページにアクセスするようなクエリが見つかっていることがわかる．(b, f) はクエリが早期に定まり，アクセスを行いながらページを移動していく行動を表し，クエリの想起に苦労していないことがうかがえる．以上の点を踏まえると，grocery はクリックに至るクエリがなかなか見つからず，クエリ推薦がカテゴリに有効な検索支援であるといえる．

daily-goods は，(b, e) クラスタ以外は womens-fashion と大きな差は見られなかった．このため，daily-goods は関連商品の推薦が有効な検索支援の手法であるといえる．

grocery が他の 2 つのカテゴリよりクエリの想起が難しい理由として，同じ商品でも様々な売り方をしているため，商品の絞り込みが必要なことが考えられる．例として，分量や個数の指定，送料無料，訳あり品などサービスの指定が挙げられる．

6. おわりに

本稿は，EC サイトにおける商品カテゴリの探索行動をクエリ変更完了率とページアクセス完了率の到達の早さに着目して分析を行った．検索ログデータとページアクセスログデータからクエリ間のキーワードをもとにセッションを抽出し，クエリ変更完了率とページアクセス完了率の系列データを算出した．算出した特徴量でセッションのクラスタリングを行い，各商品カテゴリのクラスタ所属確率を出した結果，カテゴリ間で異なる特徴を示した．分析の結果から，womens-fashion，daily-goods は関連商品の推薦を，grocery はクエリ推薦が有効な検索支援の手法であることが示唆された．

今後の課題として，提案手法を他のカテゴリにも適用してみることが挙げられる．本稿では，評価データをパス長が 10 のセッションのみに固定したが，それ以外のパス長に対しても分析を行い，データ全体の傾向を把握する必要があるといえる．

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，研究を遂行するにあたり，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データを利用している．ここに記して謝意を示す．

文　献

[1] 山口雅史, 大島裕明, 小山聡, 田中克己. サーチエンジンのクエリログを利用した同位語の発見. 日本データベース学会 Letters, Vol. 5, No. 2, pp. 17–20, 2006.
[2] 関口裕一郎, 田中智博, 内山匡, 藤村滋, 望月崇由. 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.
[3] 深澤祐援, 原島純. 料理レシピサービスにおける検索語の意味変化に関する分析. 研究報告自然言語処理（NL）, Vol. 2016-NL-228, No. 2, pp. 1–8, 2016.
[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] 矢野友貴, 田頭幸浩, 田島玲. クリック文書の分散表現を用いたクエリ曖昧性の評価. 第 8 回データ工学と情報マネジメントに

関するフォーラム (DEIM2016), 2016.

[6] Kouga Kobayashi, Yuri Nozaki, Takayasu Fushimi, and Tetsuji Satoh. Category reformation using purchase logs. In *Proceedings of 19th International Conference on Information Integration and Web-based Applications & Services(iiWAS2017)*, 2017.

[7] 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.

[8] 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.

[9] 早川潤一, 中野智文, 犬塚信博. 頻出系列パターンマイニング手法を用いた web 利用パターン発見. 第 3 回人工知能学会データマイニングと統計数理研究会 (SIG-DMSM), pp. 89–96, 2007.