クリックと放棄に基づくモバイルバーティカルの順位付け

川崎　真未† Inho Kang†† 酒井　哲也†

† 早稲田大学基幹理工学研究科　〒 169–0072 新宿区大久保 3-4-1

†† Naver Corporation

E-mail: †marerhg@ruri.waseda.jp, ††once.ihkang@navercorp.com, †††tetsuyasakai@acm.org

あらまし　本研究は，与えられたモバイルクエリに対し，より適切なカード (従来研究におけるバーティカルと呼ばれるものに近い) の順位付けを目的とする．従来研究は，ウェブ検索において，クリックセッションのクリックデータを用いて URL を順位付けた．彼らの手法を本研究の目的に適用できると考えられるが，モバイル検索では，クリックせずにそれを閲覧するだけで所望の情報が得られる「良い放棄」が PC での検索よりも多く起こることが知られている．従って，モバイル検索においては，クリックデータだけでなく放棄セッションでのユーザの操作もカードの順位付けに役立つ可能性が高い．そこで本稿では，クリックセッションに加えて放棄セッションも用いた，カードの順位付け方法を提案する．評価には，韓国で最も普及している検索エンジン NAVER の実際のモバイルクエリログを使用した．評価データは，3 人の評価者により作成された，992 のユニーククエリを含む 2,472 組のカードの種類同士のプリファレンスデータである．放棄セッションを用いることにより，クリックセッションのみを用いた場合に比べ，5.0 ポイント高いプリファレンス精度を達成できた．

キーワード　クリック, 良い放棄, モバイル検索, バーティカル, ランキング

1. はじめに

スマートフォンによるモバイル検索は，ユーザがわからないことがある時にいつでも簡単に情報を得る手段として，現代では欠かせないものとなっている．例えば 2015 年に Google は以下のように述べている．“*more Google searches were completed on mobile devices than desktop computers.*”(注1) このため，モバイル検索の質は常に向上が求められている．本研究では，韓国で最も普及している検索エンジン NAVER(注2) のモバイルクエリログを題材に，モバイル検索の有効性向上を目的とする．NAVER のモバイル検索結果は，URL の順位付きリストではなく，カードと呼ばれる様々な情報のタイプ (例えば「店舗」「天気」「テレビ番組」「株価」など) の順位付きリストである．各カードは画像やテキストで構成され，その検索結果上に占める面積もそれぞれ異なる．そこで，上記目的達成のための一手段として，与えられたモバイルクエリに適したカードの種類の順位付けを研究課題とする．NAVER のカードは従来研究におけるバーティカル [2], [12], [16] やアンサー [3], [15] と呼ばれるものに近いが，詳細は 3 章で説明する．ウェブ検索の研究において，Agrawal ら [1] は，与えられたクエリに対し，各 URL に適合性ラベル (すなわち正解ラベル) を自動的に付与する取り組みについて報告している．これは，与えられたクエリに対するクエリログをもとに,「クリックされた URL は他の URL よりもユーザが好むものであった」という仮定のもと，URL をノードとする有向グラフを構築するものである．ここで，ノード間を結ぶ有効エッジは URL 間のプリファレンス (選好) を表す．Agrawal らの手法は，上記有向グラフをもとに URL 間に半順序関係を与えることができる．従って，URL に対するクリックの代わりに NAVER のカードに対するクリックを用いることにより，与えられたモバイルクエリに対し適切なカードの種類の順位付けを行う我々のタスクに，彼らの手法を応用できる可能性がある．

上記のアプローチにとどまらず，一般に検索エンジンの最適化はユーザのクリックに強く依存している (例えば [5], [14])．しかしながら，モバイル検索，特に NAVER のようにカードという形態で視覚的に情報を提示するモバイル検索においては，ユーザが検索結果を放棄 (ユーザがクリックをせずに検索セッションを終えること) するケースが少なくない．放棄には，ユーザが検索結果を閲覧するだけで所望の情報を得て満足する「良い放棄」と，ユーザが検索結果に失望しセッションを終了する「悪い放棄」があり，モバイル検索では，良い放棄が PC での検索よりも多く起こることが知られている [11]．本研究で扱う NAVER の検索エンジン結果は，URL の順位付きリストではなくカードの順位付きリストであるので，良い放棄が特に起こりやすいと考えられる．従って，クリックセッション (クリックがあったセッション) をもとにクリックされたカードに関するプリファレンスを取得できるのと同様に，放棄セッション (クリックがなかったセッション) においても，モバイル検索結果画面中でユーザが情報を得たと思われるカードを推定することにより，プリファレンスを取得できるのではないかと考えた．そこで本研究では，カードの順位付けのために，クリックセッションに加えて放棄セッションを用いることを提案する (注3)．

(注1)：https://techcrunch.com/2015/10/08/mobile-searches-surpass-desktop-searches-at-google-for-the-first-time/

(注2)：https://www.naver.com

(注3)：本研究は，ACM CIKM 2017 で発表した short paper [9] に追加実

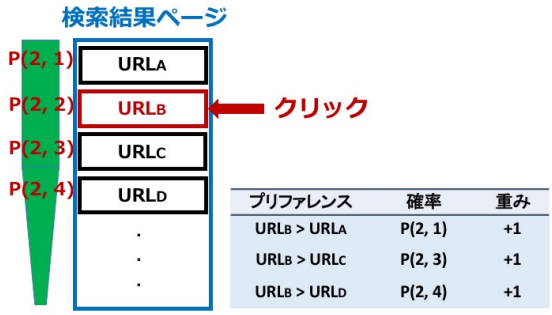

図 1　Agrawal らの確率的なプリファレンスルール

2. 関 連 研 究

以下，2.1 節で，検索エンジンの正解データ作成を目的とした，クリックセッションに基づく URL の順位付けに関する従来研究について説明する．本研究は，この手法を放棄セッションを使用できるよう拡張したものである．2.2 節では，放棄セッションが良い放棄であったか悪い放棄であったかを推定する従来研究について述べる．なお，後述するように，本研究では両者の区別は行っていない．2.3 節では，モバイル検索におけるビューポート (ページ全体のうち画面に表示される領域) とユーザの注視の相関を調査した従来研究について述べる．本研究では，ここでの知見を参考に，放棄セッションにおいてユーザが注視したカードを推定する．

2.1 クリックデータに基づく URL の適合性ラベルの付与

Joachims ら [7] はユーザのクリックに基づいて URL のプリファレンス対を予測し，Agrawal ら [1] は彼らのアプローチを拡張し，クリックデータに基づいて，与えられたクエリに対し，自動的に各 URL の適合性ラベルを付与する手法を提案した．Agrawal らの手法は，与えられた特定のクエリに対するクリックデータをもとに，以下の処理を行う．

ステップ 1　ノードを URL とし，エッジとその重みが URL 対のプリファレンスを表す有向グラフ (プリファレンスグラフ) を作成する．重みはクリックデータの集計により決定する．

ステップ 2　エッジの重みに基づいて，グラフのノード (URL) の順位付けを行う．

ステップ 3　順位付け結果をいくつかの領域に区切り，各 URL に適合性ラベルを付与する．

本研究の目的は，URL への適合性ラベル付与ではなく，モバイル検索におけるカードの種類を適切に順位付けすることである．そこで，提案手法では Agrawal らの手法を URL ではなくカードを有向グラフのノードとした上で拡張し，また上記ステップ 3 は行わない．以下，提案手法において拡張する上記ステップ 1 および 2 について詳述する．

ステップ 1 を説明する．基本的なアイデアは，「クリックされた URL はそれまでに見ていた URL よりもユーザが好むものであった」という仮定 [7] に基づき，各クリックセッションから URL 対のプリファレンス (どちらの URL が好まれたか) を得て，それらを集計することである．ある URL 対に対しプリファレンスが得られた場合，この URL 対に対しグラフ上でエッジを生成し，重みを 1 とする．以後，同じ URL 対に対し同じプリファレンスが得られた場合には，エッジの重みをインクリメントする．

例えば，図 1 のように，$\mathrm{URL}_A, \mathrm{URL}_B, \mathrm{URL}_C, \mathrm{URL}_D$ がこの順番で提示され，ユーザが URL_B のみをクリックしたセッションがあるとする．このとき，Agrawal らの手法では，URL_B をクリックしたユーザが，クリックをする前に他の各 URL を見たと思われる確率 (推定閲覧確率) を利用する．例えば，URL_B すぐ上およびすぐ下に位置する $\mathrm{URL}_A, \mathrm{URL}_C$ を見た確率はそれぞれ 1，それより下の URL_D 以下を見た確率は 1 未満で，同図の左側に視覚的に示したように，徐々に小さくなるよう定める．推定閲覧確率は，Joachims らのアイトラッキングを用いたユーザ実験結果を参考に定めらたものである．図中では，例えば検索結果中第 2 位の URL をクリックした場合の第 1 位の URL に対する推定閲覧確率を $P(2,1)$ のように表している．「クリックされた URL はそれまでに見ていた URL よりユーザが好むものであった」という仮定より，この例からは，以下のプリファレンスが考えらえる．$\mathrm{URL}_B > \mathrm{URL}_A$, $\mathrm{URL}_B > \mathrm{URL}_C$, $\mathrm{URL}_B > \mathrm{URL}_D$. そして，例えばプリファレンス $\mathrm{URL}_B > \mathrm{URL}_A$ は，確率 $P(2,1)$ で有向グラフに反映させ，$\mathrm{URL}_B > \mathrm{URL}_D$ は確率 $P(2,4)$(比較的小さい値) で反映させる．ここで，反映させるとは，ノード間に新たなエッジを設けて重みを 1 とするか，既存のエッジの重みをインクリメントすることを意味する．以上では，ひとつのセッションにおけるひとつのクリックから得たプリファレンスを有向グラフに反映させる手順について示したが，実際には与えられたクエリに対応する複数セッションの各クリックをもとに同様の処理を行い，そのクエリに対する有向グラフを完成させる．

上記の例では，例えばプリファレンス $\mathrm{URL}_B > \mathrm{URL}_A$ の確からしさを推定閲覧確率 $P(2,1)$ で，$\mathrm{URL}_B > \mathrm{URL}_D$ の確からしさを $P(2,4)$ で表しており，この確からしさを考慮した上で有向グラフを構築していることになる．前述の通り，Agrawal らはこの確からしさ，すなわち推定閲覧確率をユーザ実験を参考に定めている．一方，本研究では，ユーザがクリックを行った際にスマートフォン画面上で実際にどのカードが提示されていたかがモバイルクエリログから求められるため，このような推定値を用いる必要がない (4.1 節)．

Kadotami ら [8] は，ヤフーのモバイル検索におけるバーティカル (本研究におけるカードに相当) の順位付けのために Agrawal らのアルゴリズムを適用している．しかし，我々のモバイルクエリログとは異なり，彼らのモバイルクエリログにはどのバーティカルが画面に表示されていたかの情報が含まれていなかったため，Agrawal らと同様に推定閲覧確率に依存した処理を行っている．また，Kadotami らの研究では放棄セッションを扱っていない．

ステップ 2 を説明する．本ステップではエッジの重みに基づいてグラフのノードの順位付けを行う．Agrawal らが試した手法のうち，最も単純かつ効果的であった Δ-*order* [1] を説明する．ステップ 1 により，例えばプリファレンス $\mathrm{URL}_B > \mathrm{URL}_A$ は，有向グラフ上では，URL_B のノードから URL_A のノード

験と考察を加えたものである．

への有向エッジおよびその重みにより表現されている．そこで，各ノードのスコアを，そのノードから出ていくエッジの重みの和から，そのノードに入ってくるエッジの重みの和を除くことにより算出する．そして，グラフ中の全ノードを上記スコアにより順位付ける．本研究では，ステップ 2 については Δ-order をそのまま採用しているため，具体例について 4.2 節で改めて説明する．

2.2 良い放棄と悪い放棄

良い放棄と悪い放棄を自動的に判別する問題に取り組んだ研究 [4], [13], [15] がある．この中で，Williams ら [15] の研究では，検索結果中最初の視覚的なアンサー (本研究のカードにほぼ相当) の推定閲覧時間，最初のアンサーのうち画面に表示されているピクセル数の割合などがユーザの検索に対する満足度と相関があると報告している．

また，Lagun ら [10] や，Guo と Song ら [6] は，モバイルユーザの満足度推定に，ビューポート時間 (検索されたアイテムが表示されていた時間) が有用であることを示した．

提案手法では，良い放棄と悪い放棄の判別は行なわないが，上記の研究はカードがユーザに与えた情報量を予測するのに役に立つと考えられる．

2.3 ビューポート情報とユーザの注視の関係

Lagun ら [10] は，セッション時間に対して，ユーザが検索結果ページ上のどの結果をどのくらい見たかについて，ビューポートデータから予測した場合と，アイトラッキングデータで決めた場合との相関を報告した．ビューポートデータにはビューポート時間，検索されたアイテムのうちどれくらいの面積がユーザに見えていたか，検索されたアイテムがビューポートの面積中どれくらいを占めていたかの 3 つがあり，それぞれを用いた場合より，3 つ全てを用いた場合の方が相関が高いと述べている．

提案手法は，カードがユーザに与えた情報量を予測するために上記 3 つの要素を使用する．

3. カードとは

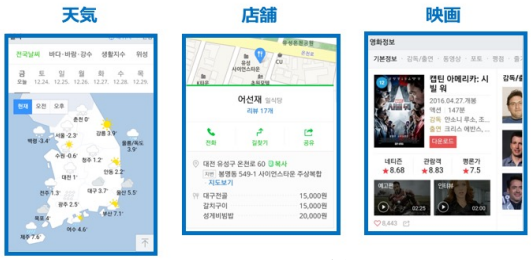

図 2 カードの例

本研究ではカードの順位付けを行なうため，カードの説明をする．カードは画像や文章で構成され，検索結果ページで直接情報を提供することを目的としている．図 8 にカードの例を示す．左のカードは地域の一週間の天気をアイコンを使って示し，選択した日の天気を地図上に示している．中央のカードは連絡先や地図などの店舗情報を示し，右のカードは映画の情報を示している．このように，カードは情報を簡潔に示し視覚的にも見やすいため，検索結果ページ上で URL よりも情報を得やすい，すなわち良い放棄を起こしやすいと考えられる．

4. 提案手法

本研究の提案手法は 2.1 で説明した手法を拡張したものであり，与えられたクエリに対し，モバイルクエリログを用いて以下の処理を行う．

ステップ 1　セッションデータ中のユーザの操作ログをもとに，与えられたクエリについて，どのカードがどのカードより好まれるか (プリファレンス) を推定し，これを集計した有効グラフ (プリファレンスグラフ) を作成する．

ステップ 2　プリファレンスグラフのエッジの重みに基づいて，グラフのノード (カード) の順位付けを行う．

4.1 ステップ 1: プリファレンスグラフの作成

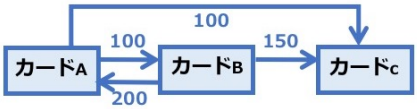

図 3 プリファレンスグラフ

図 3 に，ある与えられたクエリに対し作成されるプリファレンスグラフの簡単な例を示す．プリファレンスグラフのノードはカードであり，それぞれのエッジの向きはカード同士のプリファレンスを表し，エッジの重みはそのプリファレンスがセッションデータ中で何回推定されたかを表す．この例の 200 と書かれたエッジは，ある与えられたクエリに対して「カード$_B$ はカード$_A$ よりもユーザが好むものであった」とセッションデータ中のユーザの操作ログにおいて 200 回推定したことを示す．

提案手法は,「ユーザに多くの情報を与えたカードは，その他のユーザが閲覧したカードよりも好まれた」という仮定のもとプリファレンスを推定する．多くの情報を与えたカードとユーザが閲覧したカードの定義を以下で述べる．

まず，多くの情報を与えたカードを定義する．クリックセッションの場合は，クリックされたカードを多くの情報を与えたカードとする．放棄セッションの場合は，ユーザが最も注視したカードを多くの情報を与えたカードとする．具体的には，ある放棄セッションにおいて，画面にその一部もしくは全体が表示されたカードの集合 C があるとき，各カード $c(\in C)$ について以下の方法により $cardscore(c)$ を計算し，最も値の高いカードを多くの情報を与えたカードとする．各セッションデータは，図 4 に示すように，ユーザのスクロールにより定義される一般に複数の「画面」により構成される．そこで，あるセッションを構成する画面の集合を S とするとき，$cardscore(c)$ を以下のように算出する．

$$
\begin{aligned}
cardscore(c) &= \sum_{s\in S} score(c,s)\ , && (1)\\
score(c,s) &= time(c,s)\times dominance(c,s)\\
&\quad \times completeness(c,s)\ ,\\
time(c,s) &= \frac{c\text{ を }s\text{ に表示した時間}}{\text{セッションの時間}}\ ,\\
dominance(c,s) &= \frac{c\text{ の }s\text{ に表示されている部分の高さ}}{s\text{ の高さ}}\ ,\\
completeness(c,s) &= \frac{c\text{ の }s\text{ に表示されている部分の高さ}}{c\text{ の高さ}}\ .
\end{aligned}
$$

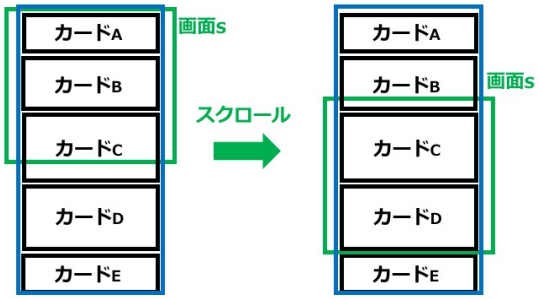

図 4　セッション中のスクロール

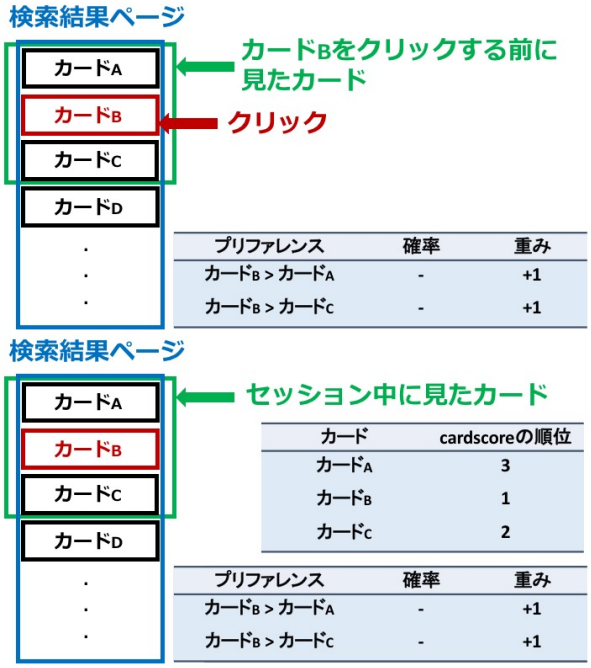

図 5　クリックセッション (上) と放棄セッション (下) でのプリファレンスの推定方法

次に，ユーザが閲覧したカードであるが，クリックセッションにおいては，各クリック以前に画面上に一部でも表示されたカードと定義する．一方，放棄セッションにおいては，セッション中に一部でも画面に表示されたカード，すなわち前述の $c \in C$ と定義する．提案手法ではこのように，スマートフォンの小さい画面に表示されたカードのみをプリファレンスの対象としているため，Agrawal ら，Kadotami らのような推定閲覧分布を考慮せずとも信頼性の高いプリファレンスが収集できると考えられる．

図 5 の上半分は，クリックセッションからプリファレンスを推定する様子を表している．このように，クリックセッションにおいては，クリックされた各カード (一般にはセッション中に複数存在する) を多くの情報を与えたカードと見なし，各クリック以前に閲覧されたカードに対するプリファレンスを推定する．一方，同図の下半分は，放棄セッションからプリファレンスが生成される様子を表している．この場合は，このセッションに対し $cardscore(c)$ が最大のカードを選定し，これをユーザに多くの情報を与えたカードとみなす．そして，同セッション中に画面に表示された全カードを対象にプリファレンスを推定する．なお，$cardscore(c)$ が最大のカードが複数枚ある場合は，それらのカードそれぞれについて，同様にプリファレンスを推定する．

4.2　ステップ 2: プリファレンスグラフのノードの順位付け

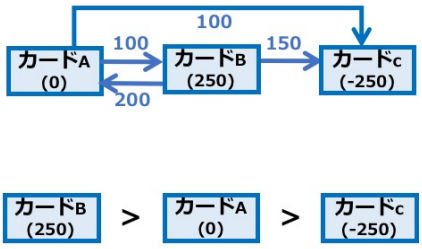

図 6　ノードの順位付け

各クエリに対して構築したプリファレンスグラフからカード間の半順序関係を得るために，2.1 で説明した Δ-$order$ を使用する．図 6 は，図 3 のプリファレンスグラフが与えられた場合に，ノードを順位付ける様子を表している．図の上半分には，カードごとにスコアを記載している．例えば カード_A のについては，ノードから出ていくエッジの重みの和が $100 + 100$ で 200，ノードに入ってくるエッジの重みの和が 100 であるので，スコアは 0 である．同様に カード_B，カード_C について計算を行なうと，同図の下半分のように，スコアに基づいたカードの順位付け結果が得られる．

5.　プリファレンスによる評価

提案手法の有効性を確認するため，NAVER のモバイルクエリログを用いて正解つきデータセットを作成し，評価実験を行った．

5.1　使用するモバイルクエリログ

評価実験に使用したデータは，2016 年 12 月 12 日から 18 日までの韓国の検索エンジン NAVER のモバイルクエリログの一部である．NAVER のモバイル検索結果ページにはカードが並んでいる．モバイルクエリログの形式は，USER_ID, SERP_ID, QUERY, UNIXTIME, INTERACTION_TYPE, VISIBLE_ITEMS, CLICKED_CARD このようになっており，VISIBLE_ITEMS には，その時に画面に表示されているカードの種類や，カードごとに，画面に表示されている部分の高さが記録されている．

NAVER のカードの種類は数百にものぼるが，正解つき評価用データ作成のためには，カードの種類とクエリ数をある程度限定する必要がある．今回は，カードの種類を クリック率 > 0.8 であるカード，$0.8 \geqq$ クリック率 $\geqq 0.2$ であるカード，クリック率 < 0.2 であるカード，それぞれ 10 種類を選定し，合計 30 種類を対象とすることとした．

図 7 に，対象のカードの例を示す．

次に，評価データに含めるクエリは，上記 30 種類のカードを対象とすることを前提として以下のように選定した．

- 信頼性の高いプリファレンスグラフの作成には，与えられたクエリに対するある程度のセッション数が必要であるため，セッション数の合計で上位 70%に入るヘッドクエリとする．

図 7　本実験で対象とする 30 種類カードの例

- 提案手法は検索結果ページに現れるカードを順位付るため，1 つの検索結果ページに対象のカードが 2 つ以上現れるセッションを 1 つ以上持ち，対象のカードが 1 つ以上現れるセッションを複数持つクエリとする．
- 実験の目的が放棄を考慮したプリファレンスルールの有用性を示すことであるため，クリック率が 50 %以下のクエリとする．

上記の通りクエリを絞り，最終的に 992 のユニーククエリで合計 720,764 セッションが得られ，そのうち 315,759 が放棄セッションであった．このデータセットを使って，プリファレンスグラフを作成した．

5.2　正解データ

上述の 992 件のクエリに対する正解データを以下のように作成した (注4)．本研究は日本で行われたが，実験で用いた NAVER のカードデータは韓国語で書かれているため，クラウドソーシングサイト Lancers (注5) を通して 2 名の**ネイティブ**の韓国人と 1 名の韓国語が話せる日本人を雇い，各カード対に対し判定者 3 名によるプリファレンスを独立に収集した．ここで，個々のプリファレンスは，与えられたクエリに対し 2 種類のカードのうちいずれのカードがより適しているかの判断結果を表す．以上により合計 7,616 件のプリファレンスを収集し，各クエリについて 3 名のプリファレンスの多数決をとることにより最終的に 2,472 件の正解プリファレンスを得た．

図 8　正解データを作成するための画面．表示されている 2 種類のカードの内容は上のクエリと関係がない．

図 8 に判定者に提供した正解データ作成用画面のスクリーンショットを示す．クエリ (“香港為替”) が一番上に表示され，2 種類の異なるカードが左右に表示されている．本研究の目的は，個々のカードの事例の間のプリファレンスではなく，異なるカードの種類の間のプリファレンスを推定することであるため，図に示したように，クエリとは内容的に直接関係がないカードを左右に並べて提示している．また，カードの種類が同じであっても，実際の表示形式が若干異なる場合があるため，左右それぞれ複数事例がある場合にはスクロールによりこれらを閲覧できるようにしている．判定者への指示は日本語で書いたドキュメントにより与えた．この中で，“提示された「検索ワード」に対して，左と右のどちらのカードの形式が検索結果として適切かを判定してください.” と記載した．また，判定者はブラウザ上で上記の画面を使用してリモートで判定作業を行った．

表 1 に 3 名の判定者による判定ラベルの統計を示す．1 列目についてだが，**N1** ，**N2** はどちらも韓国語がネイティブである判定者 2 名を示し，**NN** は韓国語がネイティブでない判定者 1 名を示す．また，例えば L-L-R は，**N1** が左のカード，**N2** が左のカード，**NN** が右のカードを判定画面で選んだことを示す．3 列目の “正解カード” は上述の通り多数決で選んだ．2 列目の判定数の太字は，**NN** がネイティブの 2 名の判定と大きく異なることを示しており，このことから，**NN** の判定内容が信頼できない可能性があると言える．表 2 に判定者間の一致度を Fleiss’ κ と Cohen’s κ [17] により示す．Fleiss’ κ で 3 名の判定者間の一致度を測り，Cohen’s κ で 2 名ずつ判定者間の一致度を測る．95%信頼区間も算出する．

上記のように，ネイティブでない判定者とネイティブの判定者のラベルの一致度が低いことがわかったため，ネイティブ 2 名の判定結果のみを用いた第 2 の正解データを作成した．こちらでは多数決が適用できず，ネイティブ 2 名の判定結果が一致したデータのみを使用したため，正解プリファレンス数は 2,023

(注4)：実際に作成したプリファレンスデータは 1000 のユニーククエリに対する合計 7,476 件のプリファレンスデータであったが，エクセルでデータを整理した際に，8 のユニーククエリを変換 (例えば「34-1」というクエリを「Jan-34」と変換した) してしまい，そのまま判定者に表示していたことがわかったため，20 件のデータ，すなわち合計 60 件のプリファレンスデータを取り除いた．
(注5)：www.lancers.jp

件となった．

表 1　正解データラベルの内訳

N1-N2-NN	判定数	正解カード	判定数
L-L-L	993 (40.17%)	L	1,742 (70.47%)
L-L-R	**506** (**20.47%**)		
L-R-L	105 (4.25%)		
R-L-L	138 (5.58%)		
R-R-R	347 (14.04%)	R	730 (29.53%)
R-R-L	**177** (**7.16%**)		
R-L-R	92 (3.72%)		
L-R-R	114 (4.61%)		
合計	2,472 (100%)	total	2,472 (100%)

表 2　判定者間の一致度．3 名の場合は Fleiss' κ を，2 名の場合は Cohen's κ を使用した

	κ	95% CI
3 名全員	0.562	[0.539, 0.584]
N1 と **N2**	0.570	[0.551, 0.588]
N1 と **NN**	**0.199**	[0.178, 0.219]
N2 と **NN**	**0.245**	[0.224, 0.265]

5.3 評価尺度

提案手法の評価指標として，以下のように正解率と精度を定義する．

$$\text{プリファレンス正解率} = \frac{n}{N} \quad \text{プリファレンス精度} = \frac{n}{N'}$$

N: 用意した正解プリファレンスの総数．

$N'(\leqq N)$: 上記のうち，各手法が作ったランキングにより推定できたプリファレンス総数．提案手法では画面に表示されたカード同士のみについてプリファレンスを推定するため，正解プリファレンスが提案手法によるカードのランキングに含まれない場合がある．また，プリファレンスグラフに Δ-*order* を適用した結果，2 つのカードの順位が同じになってしまう場合もある．以上により，一般に N' は N より小さくなる．

$n(\leqq N')$: 推定プリファレンスのうち，正解プリファレンスと一致するものの個数．

5.4 比較する手法

実験で比較する手法は以下の通りである．なお，C (Click) はクリックセッション，A (Abandonment) は放棄セッションを示し，score は多く情報を与えたカードを提案手法の cardscore により選択することを，random は多く情報を与えたカードを画面に表示したカードの中からランダムに選択することを示す．

C+A(score)　[提案手法]

クリックセッションと放棄セッションを使用し，放棄セッションで多く情報を与えたカードは提案手法の cardscore により選択する．

C+A(random)

クリックセッションと放棄セッションを使用し，放棄セッションで多く情報を与えたカードはランダムに選択する．

C

クリックセッションのみを使用する．

A(score)

放棄セッションのみを使用し，放棄セッションで多く情報を与えたカードは提案手法の cardscore により選択する．

A(random)

放棄セッションのみを使用し，放棄セッションで多く情報を与えたカードはランダムに選択する．

C(score)+A(score)

クリックセッションと放棄セッションを使用し，クリックセッションの場合も放棄セッションの場合と同じように，多く情報を与えたカードを提案手法の cardscore により選択する．

5.5 実験結果

表 3 と表 4 に，各手法のプリファレンス精度と正解率を示す．表 3 は判定者 3 名の判定結果から正解データを作った場合，表 4 はネイティブ 2 名の判定結果から正解データを作った場合の結果である．また，表 5 に，判定者 3 名による正解データを使用した場合の，各手法のプリファレンス精度の差について Tukey HSD 検定を行なった結果を示す．以降，各手法により推定されたプリファレンスの性質の違いについて考察を行うため，主としてプリファレンス精度に着目する．

表 3　各手法のプリファレンス精度と正解率 (判定者 3 名)

手法	精度	正解率	n	N'	N
C+A(score)	0.6932	0.5138	1270	1832	2472
C+A(random)	0.6045	0.4482	1108	1833	2472
C	0.6431	0.3354	829	1289	2472
A(score)	0.6772	0.4490	1110	1639	2472
A(random)	0.4865	0.3135	775	1593	2472
C(score)+A(score)	0.6678	0.5489	1357	2032	2472

表 4　各手法のプリファレンス精度と正解率 (判定者が 2 名)

手法	精度	正解率	n	N'	N
C+A(score)	0.7175	0.5348	1082	1508	2023
C+A(random)	0.6287	0.4696	950	1511	2023
C	0.6820	0.3584	725	1063	2023
A(score)	0.6884	0.4652	941	1367	2023
A(random)	0.4789	0.3134	634	1324	2023
C(score)+A(score)	0.6842	0.5645	1142	1669	2023

5.6 議　　論

以下，各評価結果について議論する．まず，表 3 より C の精度は 0.6431，A(score) の精度は 0.6772 であり，表 5 よりこれらの間に統計的有意差はない．このことから，提案した *cardscore*(c) に基づく放棄セッションからのプリファレンス推定により，クリックセッションからのプリファレンス推定結果と遜色のない結果が得られていることがわかる。

次に，表 3 と表 4 において，C+A(score) は C より精度が高く，推定できたプリファレンス数 N' もより多くなっている．さらに，表 3 における両者の精度の差は表 5 が示すように統計的に有意である ($p = 0.0454$)．よって，プリファレンスグラフ

表 5　各手法のプリファレンス精度の差について Tukey HSD 検定を行なった結果 (判定者 3 名).
有意水準 $\alpha = 0.05$ にて有意な p 値を太字で示す

手法		精度の差	同時信頼区間			p 値
C+A(score)	A(random)	0.2067	[0.1601	, 0.2534	]	**0.0000**
A(score)	A(random)	0.1907	[0.1428	, 0.2386	]	**0.0000**
C(score)+A(score)	A(random)	0.1813	[0.1357	, 0.2269	]	**0.0000**
C	A(random)	0.1566	[0.1056	, 0.2076	]	**0.0000**
C+A(random)	A(random)	0.1180	[0.0713	, 0.1646	]	**0.0000**
C+A(score)	C+A(random)	0.0888	[0.0438	, 0.1337	]	**0.0000**
A(score)	C+A(random)	0.0728	[0.0265	, 0.1191	]	**0.0001**
C(score)+A(score)	C+A(random)	0.0633	[0.0195	, 0.1072	]	**0.0006**
C+A(score)	C	0.0501	[0.0006	, 0.0996	]	**0.0454**
C	C+A(random)	0.0387	[−0.0108	, 0.0882	]	0.2256
A(score)	C	0.0341	[−0.0166	, 0.0848	]	0.3911
C+A(score)	C(score)+A(score)	0.0254	[−0.0185	, 0.0693	]	0.5643
C(score)+A(score)	C	0.0247	[−0.0238	, 0.0732	]	0.6956
C+A(score)	A(score)	0.0160	[−0.0303	, 0.0623	]	0.9231
A(score)	C(score)+A(score)	0.0094	[−0.0358	, 0.0546	]	0.9914

を作成してカードを順位づける際，クリックセッションに加えて放棄セッションを用いることは有効といえる．一方，表 5 によれば，C+A(score) と A(score) の精度の差は統計的に有意ではない．以上を総合すると，C+A(score) の高い精度に大きく貢献しているのは放棄セッションでのプリファレンス推定のほうであると考えられる．

表 3 と表 4 において C+A(score) は C(score)+A(score) よりも精度が高い．このことは，クリックセッションでは，提案した *cardscore*(c) を利用するよりも，実際にクリックしたカードに基づきプリファレンスを作成したほうがよいことを示唆する．ただし，表 5 によれば両者の差は統計的に有意ではないので ($p = 0.5643$)，今回の結果から直ちに上記を結論づけることは出来ない．

表 3 と表 4 において C+A(score) は C+A(random) より精度が高く，表 5 より $p \approx 0.0000$ で有意差がある．また，同様に A(score) は A(random) より精度が高く，表 5 より $p \approx 0.0000$ で有意差がある．このことから，放棄セッションにおけるユーザーに多く情報を与えたカードの選ぶ際に，提案した *cardscore*(c) を用いることは有効と言える．

表 3 と表 4 を比較してみると，A(random) を除いて，全般的に表 4 におけるプリファレンス精度および正解率のほうが高いことがわかる．このことと，表 2 の判定者間一致度の結果を総合すると，今回実験した手法は韓国語ネイティブの判断に比較的近い結果になっていると思われる．

5.7　放棄セッションにおけるユーザが多くの情報を得たカードの選定方法

前節では *cardscore* に基づきユーザが多くの情報を得たカードを選定し，プリファレンスを推定する提案手法の有効性を示した．本節では，*cardscore* を構成する特徴量である time, dominance, completeness (4.1 節参照) の効果について考察する．表 6 に，*cardscore* で使用する特徴量のうち 1 つもしくは 2 つを用いない場合の手法 A(score) のプリファレンス精度を，通常の *cardscore* による手法 A(score) のそれと比較した結果を示す．判定者 3 名による正解データを使用した場合である．なお，表の t，d，c はそれぞれ time，dominance，completeness を示しており，例えば td は time と dominance を用いたことを示す．また，この結果に対応する Tukey HSD 検定の結果を表 7 に示す．第一列において，例えば “c t” は completeness のみを用いた場合と time のみを用いた場合の差を意味する．

表 6 より，completeness のみを使った場合が一番精度が高く，time のみを使った場合が一番精度が低く，表 7 よりこれらの間 (“c t”) には統計的有意差がある ($p = 0.0006$)．すなわち，カードが一部分でも映った時間の合計よりも，カード全体のうちどれくらいの部分が何度閲覧されたかの方が放棄セッションにおけるプリファレンス推定のために有効であると言える．さらに，time と completeness を使った場合 (tc) と全ての特徴量を用いた場合 (tdc) は，共に time のみを使った場合 (t) よりも精度が高く，有意差もある (それぞれ $p = 0.0195, p = 0.0211$)．このことも，completeness の有用性を示している．

表 6　*cardscore* で使用する特徴量変えた場合の A(score) のプリファレンス精度 (判定者 3 名)

手法	time	dominance	completeness	プリファレンス精度
t	使用	-	-	0.6236
d	-	使用	-	0.6486
c	-	-	使用	0.6928
td	使用	使用	-	0.6571
dc	-	使用	使用	0.6687
tc	使用	-	使用	0.6777
tdc	使用	使用	使用	0.6772

6.　結論と今後の課題

本稿では，与えられたモバイルクエリに対しより適切なカードの順位付けを目的として，放棄セッションからもプリファレンスが得られる可能性に注目し，放棄セッションでユーザによ

表 7　A(score) の *cardscore* に使用する特徴量を変えた場合のプリファレンス精度の差について Tukey HSD 検定を行なった結果 (判定者 3 名)．有意水準 $\alpha = 0.05$ にて有意な p 値を太字で示す

手法		精度の差	同時信頼区間				p 値
c	t	0.0692	[	0.0202	, 0.1183	]	**0.0006**
tc	t	0.0541	[	0.0051	, 0.1031	]	**0.0195**
tdc	t	0.0537	[	0.0047	, 0.1026	]	**0.0211**
dc	t	0.0451	[	−0.0038	, 0.0941	]	0.0940
c	d	0.0442	[	−0.0044	, 0.0929	]	0.1034
c	td	0.0358	[	−0.0130	, 0.0845	]	0.3157
td	t	0.0335	[	−0.0155	, 0.0825	]	0.4054
tc	d	0.0291	[	−0.0196	, 0.0777	]	0.5733
tdc	d	0.0286	[	−0.0200	, 0.0772	]	0.5905
d	t	0.0250	[	−0.0239	, 0.0740	]	0.7409
c	dc	0.0241	[	−0.0245	, 0.0728	]	0.7674
tc	td	0.0206	[	−0.0281	, 0.0693	]	0.8754
tdc	td	0.0202	[	−0.0285	, 0.0688	]	0.8859
dc	d	0.0201	[	−0.0285	, 0.0687	]	0.8867
c	tdc	0.0156	[	−0.0331	, 0.0643	]	0.9654
c	tc	0.0152	[	−0.0336	, 0.0639	]	0.9700
dc	td	0.0116	[	−0.0370	, 0.0603	]	0.9924
tc	dc	0.0090	[	−0.0397	, 0.0576	]	0.9982
tdc	dc	0.0085	[	−0.0401	, 0.0572	]	0.9986
td	d	0.0085	[	−0.0402	, 0.0571	]	0.9987
tc	tdc	0.0004	[	−0.0482	, 0.0491	]	1.0000

り多く情報を与えたカードを定義し，クエリセッションと放棄セッションの両方からプリファレンスを得て，プリファレンスグラフを作成した．プリファレンス精度による評価を行なったところ，以下の知見が得られた．

- クリックセッションにおいてクリックに基づきプリファレンスを推定するのと同様に，放棄セッションにおいてはユーザ注視したカードを推定することにより，プリファレンスを高精度に推定することが可能である．
- プリファレンスグラフを用いてカードを順位付けする際に，クリックセッションに加えて放棄セッションを用いることは有効である．
- 放棄セッションにおけるユーザに多く情報を与えたカードを選ぶ際に，提案した time, dominance, completeness に基づく *cardscore* を用いることは有効である．
- 放棄セッションで多く情報を与えたカードを選ぶ際は，カードが一部分でも映った時間の合計 (time) よりも，カード全体のうちどれくらいの部分が何度閲覧されたか (completeness) の方が有効な手がかりとなる．

今後の課題を述べる．今回は，放棄セッション利用の有効性を検証するために意図的に放棄率の高いクエリを選んだので，実際のモバイルクエリデータからの代表的サンプルに対する実験には必ずしもなっていない．従って，今回得られた効果の実用上での影響について検証する必要がある．また，今回のモバイルクエリログ収集時点での元の検索結果と，本研究における新たなカードの順付けを適用した検索結果を，実運用において直接比較評価することを検討したい．

文　献

[1] Rakesh Agrawal, Alan Halverson, Krishnaram Kenthapadi, Nina Mishra, and Panayiotis Tsaparas. Generating labels from clicks. In *Proceedings of ACM WSDM 2009*, pp. 172–181, 2009.

[2] Jaime Arguello, Fernando Diaz, and Jamie Callan. Learning to aggregate vertical results into web search results. In *Proceedings of ACM CIKM 2011*, pp. 201–210, 2011.

[3] Lydia B. Chilton and Jaime Teevan. Addressing people's information needs directly in a web search result page. In *Proceedings of WWW 2011*, pp. 27–36, 2011.

[4] Aleksandr Chuklin and Pavel Serdyukov. Potential good abandonment prediction. In *WWW 2012 Companion*, pp. 485–486, 2012.

[5] Nick Craswell and Martin Szummer. Random walks on the click graph. In *Proceedings of ACM SIGIR 2007*, pp. 239–246, 2007.

[6] Qi Guo and Yang Song. Large-scale analysis of viewing behavior: Towards measuring satisfaction with mobile proactive systems. In *CIKM 2016*, 2016.

[7] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM TOIS*, Vol. 25, No. 2, 2007.

[8] Yuta Kadotami, Yasuaki Yoshida, Sumio Fujita, and Tetsuya Sakai. Mobile vertical ranking based on preference graphs. In *Proceedings of ACM ICTIR 2017*, 2017.

[9] Mami Kawasaki, Inho Kang, and Tetsuya Sakai. Ranking rich mobile verticals based on clicks and abandonment. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 2127–2130, 2017.

[10] Dmitry Lagun, Chih-Hung Hsieh, and Dale Webster Vidhya Navalpakkam. Towards better measurement of attention and satisfaction in mobile search. In *Proceedings of SIGIR 2014*, pp. 113–122, 2014.

[11] Jane Li, Scott Huffman, and Akihito Tokuda. Good abandonment in mobile and pc internet search. In *Proceedings of ACM SIGIR 2009*, pp. 43–50, 2009.

[12] Ashok Kumar Ponnuswami, Kumaresh Pattabiraman, Qiang Wu, Ran Gilad-Bachrach, and Tapas Kanungo. On composition of federated web search result page: Using online users to provide pairwise preference for heterogeneous verticals. In *Proceedings of ACM WSDM 2011*, pp. 715–724, 2011.

[13] Yang Song, Xiaolin Shi, Ryen White, and Ahmed Hassan Awadallah. Context-aware web search abandonment prediction. In *Proceedings of ACM SIGIR 2014*, pp. 93–102, 2014.

[14] Kuansan Wang, Toby Walker, and Zijian Zheng. PSkip: Estimating relevance ranking quality from web search clickthrough data. In *Proceedings of ACM KDD 2009*, pp. 1355–1364, 2009.

[15] Kyle Williams, Julia Kiseleva, Aidan C. Crook, Imed Zitouni, Ahmed Hassan Awadallah, and Madian Khabsa. Detecting good abandonment in mobile search. In *Proceedings of WWW 2016*, pp. 495–505, 2016.

[16] Ke Zhou, Thomas Demeester, Dong Nguyen, Djoerd Hiemstra, and Dolf Trieschnigg. Aligning vertical collection relevance with user intent. In *Proceedings of ACM CIKM 2014*, pp. 1915–1918, 2014.

[17] 酒井哲也. 情報アクセス評価方法論: 検索エンジンの進歩のために. コロナ社, 2015.