DEIM2022 E43-4

インターリービングにおける正確性と効率性の理論的考察

飯塚洸二郎† 加藤 誠††

† 株式会社 Gunosy/筑波大学大学院 情報学学位プログラム
〒 150-6139 東京都渋谷区渋谷 2-24-12 渋谷スクランブルスクエア 39 階
†† 筑波大学/JST さきがけ 〒 305–8550 茨城県つくば市春日 1-2
E-mail: †kojiro.iizuka@gunosy.com, ††mpkato@acm.org

あらまし 本論文では，オンライン評価手法の 1 つであるインターリービングの正確性と効率性について理論的な考察を与える．インターリービングは，A/B テストに比べ評価の効率が良いことからオンライン評価に広く用いられている一方で，その効率の良さがどこからくるのかの議論は十分になされてこなかった．我々は，インターリービングの動作を説明するモデルを構築し，モデルの解析を行った．このモデルの解析を通し，正確性と効率性の面でインターリービングが A/B テストよりも同等以上に優れている場合の条件を明らかにした．最後に数値実験を行い，これらの考察と実験結果に整合性があることを確認した．

キーワード オンライン評価，インターリービング

1 はじめに

オンライン評価は，情報検索や情報推薦のアルゴリズムの評価を行うための重要な手法の 1 つである [1]．特に，A/B テストは実施の容易さからウェブサービスを運用する企業の間で日々数多く実施されている [2] [3]．この A/B テストは，実施が容易である一方で，多くのユーザーがテストに関わったり長期間に渡りテストをする場合に，テストを行う実サービスに大きな悪影響を及ぼすリスクがある．近年では，オンライン評価を効率化しユーザーやサービスへの影響を小さくするための手法が開発されており，インターリービングはそのために開発された手法の 1 つである．このインターリービングは，A/B テストよりも 10 倍から 100 倍程度効率がよいことが実験的に知られており，実サービス上での利用用途が模索されてきた [4] [5]．

しかしながら，現状インターリービングの利用範囲は限定的である．理由の 1 つとして，インターリービングの動作原理が理論的に解明されていないことがあげられる．そのため，実験的にインターリービングをサービス上で実施しない限り，インターリービングがその効果を発揮するかは明らかではなかった．もう 1 つの理由として，インターリービングは A/B テストに比べて評価できる指標が限られていることがあげられる．インターリービングは基本的にクリックに関する評価を行うための手法である．近年では，インターリービングを利用して，Post-Click 指標や因果的な効果を評価するための手法が開発されはじめている [6] [7]．このような中で，インターリービングの動作原理を解明することは，インターリービングの評価の適用範囲を広げるための手法の開発に役立つと考えられる．

インターリービングの動作原理は，直感的には以下の例から説明できる．ここでは，ランキング A がランキング B よりも相対的にユーザー満足度が高いランキングである場合を考える．インターリービングは，ランキング A とランキング B を混ぜたランキングをユーザーに提示する．混合ランキングにおいて，ユーザーはランキング A 由来のアイテムをクリックしたさいに，内容に満足してランキングから離脱する．この混合ランキングからの離脱に起因して，ランキング B はランキング B 単体をユーザーに提示する場合に比べてクリックされる機会が少なくなる．このように，相対的に満足度の高いランキングがもう片方のランキングのクリック機会を奪うようにして，クリック数の差がつきやすくなるのがインターリービングの作用である．

本研究では，この例に基づいて，インターリービングの正確性と効率性について，理論的な側面から解析を行う．まずはじめに，インターリービングの個々の手法について，ランキングの生成方法に着目して整理を行った．そこで，インターリービングによって生成される混合ランキングは，元の入力ランキングのアイテムがユーザーに等しい確率で表示されるように生成されている共通の性質があることを示した．次に，インターリービングの動作モデルを構築し，表示確率の等しさを利用してモデルの解析を行った．この動作モデルは，アイテムへのクリックを，走査確率とアイテムの適合度に分解し，確率的に定義したものである．我々は，このモデルの解析を通して，ユーザーがアイテムの適合度に依存してランキングから離脱するケースにおいて，インターリービングが A/B テストよりも誤差率が小さくなり，正確性と効率性が優れていることを示した．

実験では，ユーザーが適合度に応じてどの程度離脱しやすいかをパラメータとし，インターリービングと A/B テストの性能を比較した．データセットは，先行研究 [8] [9] にならって，インターリービングの評価に利用されるデータセットを利用した．結果として，ユーザーが適合度に応じて離脱しやすい場合において，正確性と効率性が優れていることを確認した．また，ランキングのアイテムの重複が多いケースにおいて，よりインターリービングの効率がよいことを確認した．これらの結果は，インターリービングの動作モデルの考察の結果と一致することから，インターリービングの動作原理が理論的，実験的に確か

められたといえる．

本研究の貢献は以下のようにまとめられる．

- インターリービングの動作原理について理論的側面に着目し，議論を展開した．
- インターリービングの動作を説明する解釈が容易なモデルを設計した．
- インターリービングの動作モデルに基づいて，インターリービングが A/B テストに比べて正確性と効率性が優れている場合の前提条件を明らかにした．
- 数値実験の結果，理論的な考察が実験結果と合致することを確認した．

本論文の構成は下記のとおりである．第二章では関連研究について述べる．第三章では，記法の導入をはじめとした準備を行う．次に，第四章ではインターリービングの動作モデルを導入し，正確性と効率性に関する解析を行う．第五章では，数値実験の設定と実験結果を述べ，第四節で導入したインターリービングの動作モデルに関する考察を行う．最後に，第六章で本論文のまとめを行う．

2 関連研究

この節では，インターリービングに関連する２つの研究トピックを紹介する．１つはユーザーのクリック行動に関する関連研究で，もう１つはランキング評価手法である．

2.1 クリックに関する行動モデリング

ユーザーからのフィードバックの中でも，クリックに関する行動のモデリングは古くから研究がなされている．検索システムにおいて，クリックに関するユーザーの行動モデルはクリックモデルと呼ばれる [10] [11]．このクリックモデルは，実ユーザーがいない場合，あるいは実ユーザーを使った実験でユーザー体験に支障をきたす場合にユーザーの行動を分析することに役立つ．クリックモデルという言葉をはじめて使ったのは，Craswell らのカスケードクリックモデルが紹介されている論文である [12]．なお Craswell 以前にも現在ではクリックモデルとよばれるモデルがランキング評価のために導入されていた [13] [14]．Craswell らの発表後，基礎的なクリックモデル [10] [15] [16] が開発され，現在でも様々な改良がなされている．クリックモデルの開発は、検索システムの周辺技術の発展とも密接に関係している．例えば，マウスカーソルの動きの分析をクリックモデルに取り込む研究 [17] や，ユーザーの好みをクリックモデルに取り込む研究がある [18]．本研究では逆に，クリックモデル内部の走査確率や適合度の扱いをインターリービングのモデリングに活用し，インターリービングの動作原理の解明に役立てる．

2.2 オンライン評価

オンライン評価は，ランキング評価に利用される代表的な手法の 1 つである．その中でも，A/B テストは手軽に評価を行える手法として，広く利用されている．A/B テストは，ユーザーを群 A と群 B に分け，評価するランキング A とランキング B をそれぞれのユーザー群に提示する手法である．この A/B テストは，評価の効率に課題があり，評価に多くのユーザーが必要となる．この A/B テストの延長上にあり，評価の効率を改善するための手法も存在する．これらの手法は評価のさいの分散を減少させる手法が基本となっている [1] [19] [20]．オンライン評価のもう 1 つの手法に，インターリービングがある．インターリービングは，A/B テストに比べて評価の効率が 10 倍から 100 倍程度に及んで効率的であることが実験的に知られている [4] [5]．インターリービングは，ランキングの混ぜ方とクリックのスコアリングの仕方に応じて様々な種類が存在する [4] [21] [22]．インターリービングは２つのランキングを評価する手法であり，３つ以上のランキングを評価する手法はマルチリービングとよばれる [8] [9] [23]．このように，多くのインターリービング手法が開発されている一方で，どのような原理によってインターリービングの効力が発揮されるのかといった理論的な考察はなされてこなかった．本研究では，インターリービングの動作原理を説明するモデルを通して，インターリービングの正確性や効率性についての考察を与える．

3 準　　備

3.1 記　　法

まず，ユーザーのクリックに関する記法を導入する．クリックの有無を表す確率変数を $Y = \{0, 1\}$，ユーザーの走査を表す確率変数を $O = \{0, 1\}$，アイテムの適合度を表す確率変数を $R = \{0, 1\}$ で表す．$O = 1$ は，ユーザーがアイテムを実際に目にすることを意味する．本研究では，$Y = O \cdot R$ と，O と R の独立を仮定する．

次に，ランキングに関する記法を導入する．本研究では，ランキング A とランキング B の 2 つのランキングの評価を考える．$P_A(X)$ をランキング A の確率変数 X に関する確率，$P_{AB,A}(X)$ をランキング A の A/B テストにおける確率変数 X に関する確率，$P_{I,A}(X)$ をランキング A のインターリービングにおける確率変数 X に関する確率と表す．同様に，$E_A(X)$ をランキング A の確率変数 X に関する期待値と表す．また $P_*(S = A)$ を A/B テストまたはインターリービングにおいて，ランキング A 由来のアイテムが選択される確率と表す．

本研究では，議論を容易にするため，ランキングにおける各アイテムごとの統計量ではなくアイテム全体の統計量を考える．なお期待値や分散の和の性質より，一般性を失わずに各アイテムごとの議論も同様に行える．

3.2 インターリービング

ここでは，代表的なインターリービング手法の整理を行う．

3.2.1 Balanced Interleaving (BI) [21]

BI は，ランキングのトップ k 番目までの位置において，ランキング A 由来のアイテムの個数とランキング B 由来のアイテムの個数の差が 1 以下になるように混合ランキングを生成する手法である．具体的には，トップ k までに含まれるアイテムの数が少ない方のランキングのアイテムを貪欲にランキングへ

追加を行う．

3.2.2 Team Draft Interleaving (TDI) [4]

TDI は，偏りのないコインを投げ，コインの表裏に応じて確率的にアイテムを追加していく手法である．例えば，表が出たらランキング A，ランキング B 由来のアイテムの順に混合ランキングに追加する．裏が出たら，ランキング B，ランキング A のアイテムの順に混合ランキングに追加を行う．それぞれのランキングからのアイテムの選出は，まだ混合ランキングで選ばれていないアイテムの中から選ぶ．

3.2.3 Probabilistic Interleaving (PI) [22]

PI は，できるだけ入力ランキングの順位を維持しつつも，相対的に低い確率で任意の順序で入れ替えを許容して混合ランキングを生成する手法である．TDI では，混合ランキングに追加するアイテムは，入力ランキングの上位から順に選択していた一方で，PI ではランキングの上位のアイテムのほうが下位のアイテムよりも確率的に選択されやすいよう設計されている．期待値を考えると，入力ランキング A，B それぞれのアイテムの個数が等しくなることには留意する．

3.2.4 Optimized Interleaving (OI) [24]

OI は，いくつかの制約のもとで感度の期待値を最大化するように確率的に混合ランキングを提示する手法である．OI では，混合ランキングは 1 つではなく，事前に複数生成し，それらを確率的に提示する．この確率は，ユーザーに出力する混合ランキングのアイテムの表示回数が各入力ランキングに対して偏らないという制約のもとで，ランキングの感度が最大になるように決定される．

一　般　化

これらのインターリービングの各種法について，共通するのはユーザーに表示されるアイテムの回数が，各入力ランキングに対して偏らないように設計されている点である．言い換えると，ランキング全体を考えたときに，各入力ランキングに対する走査の期待値は互いに等しくなるように，つまり $E_{I,A}(O) = E_{I,B}(O)$ となるように混合ランキングは生成されている．次の節では，このインターリービングの基本的な性質を活用して議論を行う．

4　理論的考察

この章では，インターリービングの正確性と効率性について，理論的な考察を行う．

4.1　正　確　性

下記の条件が成り立つときに，A/B テストにおけるランキングの優劣と，インターリービングにおけるランキングの優劣が，入力ランキングの優劣と一致することを示す．本研究ではこの入力ランキングとの優劣の一致を正確性と呼ぶ．

条件 1. $E_A(R) > E_B(R)$*:* ランキング A のアイテムの適合度の期待値がランキング B のアイテムの適合度の期待値よりも大きい．

条件 2. $E_{I,A}(O) = E_{I,B}(O)$*:* 混合ランキングの各ランキングに対する走査の期待値は互いに等しい．

定理 1. 条件 *1,2* が成り立つとき，$E_A(Y) > E_B(Y) \rightarrow E_{AB,A}(Y) > E_{AB,B}(Y) \wedge E_{I,A}(Y) > E_{I,B}(Y)$ となる．言い換えると，ランキング A のクリックの期待値がランキング B よりも大きいならば，A/B テストによって得られる評価の優劣とインターリービングによって得られる評価の優劣が一致する．

Proof. まず，$E_A(Y) > E_B(Y) \rightarrow E_{AB,A}(Y) > E_{AB,B}(Y)$ を示す．A/B テストにおいて，ランキング A とランキング B は同一の確率で選択されるため，$P_{AB,A}(O=1) = P_{AB}(S=A)P_A(O=1) = P_A(O=1)/2$ となる．

$$
\begin{aligned}
E_{AB,A}(Y) &= E_{AB,A}(O) \cdot E_{AB,A}(R) \\
&= P_{AB}(S=A)P_A(O=1) \cdot E_{AB,A}(R) \\
&= \frac{P_A(O=1)}{2} \cdot E_{AB,A}(R) \\
&= \frac{E_A(O)}{2} \cdot E_A(R) \\
&= \frac{E_A(Y)}{2} \\
&> \frac{E_B(Y)}{2} \\
&= E_{AB,B}(Y)
\end{aligned}
$$

よって，$E_A(Y) > E_B(Y) \rightarrow E_{AB,A}(Y) > E_{AB,B}(Y)$ となる．

次に，$E_{I,A}(Y) > E_{I,B}(Y)$ を示す．

$$
\begin{aligned}
E_{I,A}(Y) &= E_{I,A}(O) \cdot E_{I,A}(R) \\
&= E_{I,B}(O) \cdot E_{I,A}(R) \\
&> E_{I,B}(O) \cdot E_{I,B}(R) \\
&= E_{I,B}(Y)
\end{aligned}
$$

よって，$E_{I,A}(Y) > E_{I,B}(Y)$ となる．

以上から，条件 1,2 がなりたつとき，$E_A(Y) > E_B(Y) \rightarrow E_{AB,A}(Y) > E_{AB,B}(Y) \wedge E_{I,A}(Y) > E_{I,B}(Y)$ となる．

□

4.2　効　率　性

次に，A/B テストとインターリービングの効率について議論する．本研究では，効率性を一定のランキングの表示回数における評価の誤差確率によって定義する．

4.2.1 誤差の定義

$$
\mathrm{Error} = (\mathrm{sign}(E_A(Y) - E_B(Y)) \neq \mathrm{sign}(E_{*,A}(\bar{Y}) - E_{*,B}(\bar{Y})))
$$

ここで，sign は，値が 0 以上であれば 1 を返し，値が 0 未満であれば 0 を返す．また $\neq$ は，左辺と右辺の値が異なるときに 1 を返し，同じ値であるときに 0 を返す．このように誤差の値は 0 か 1 の値をとる．

4.2.2 誤差確率

まず，A/B テストの誤差確率を与える．$E_A(Y)-E_B(Y)>0$ のとき，$Y\sim\mathcal{N}(E[Y],V[Y])$ と仮定すると，

$$\begin{aligned}
&P(sign(E_A(Y),E_B(Y))\neq sign(E_{AB,A}(\bar{Y}),E_{AB,B}(\bar{Y})))\\
&=P(P_{AB,A}(\bar{Y})-P_{AB,B}(\bar{Y})\leqq 0)\\
&=\int_{-\infty}^{0}\mathcal{N}(x|E_{AB,A}(\bar{Y})-E_{AB,A}(\bar{Y}),V_{AB,A}(\bar{Y})+V_{AB,A}(\bar{Y}))dx\\
&=\int_{-\infty}^{-(E_{AB,A}(\bar{Y})-E_{AB,A}(\bar{Y}))}\mathcal{N}(x|0,V_{AB,A}(\bar{Y})+V_{AB,A}(\bar{Y}))dx
\end{aligned}\tag{1}$$

$E_A(Y)-E_B(Y)<0$ のときも同様．インターリービングの誤差確率も下記のように与えられる．

$$\begin{aligned}
&P(sign(E_A(Y),E_B(Y))\neq sign(E_{I,A}(\bar{Y}),E_{I,B}(\bar{Y})))\\
&=P(P_{I,A}(\bar{Y})-P_{I,B}(\bar{Y})\leqq 0)\\
&=\int_{-\infty}^{0}\mathcal{N}(x|E_{I,A}(\bar{Y})-E_{I,A}(\bar{Y}),V_{I,A}(\bar{Y})+V_{I,A}(\bar{Y}))dx\\
&=\int_{-\infty}^{-(E_{I,A}(\bar{Y})-E_{I,A}(\bar{Y}))}\mathcal{N}(x|0,V_{I,A}(\bar{Y})+V_{I,A}(\bar{Y}))dx
\end{aligned}\tag{2}$$

以上から，A/B テストとインターリービングのエラー確率は，分散の和とクリック期待値の差に依存することがわかる．特に，分散の和が小さいほどエラー確率も小さく，クリック期待値の差が大きいほどエラー確率が小さくなる．以下では，この分散とクリック期待値の差について，走査確率が静的な場合と動的な場合に分けて大小関係を比較する．

4.2.3 走査確率が静的な場合

走査確率が静的な場合とは，走査確率が適合度に依存しない場合を指す．例えば，走査確率がランキングの位置のみに依存するポジションベースのクリックモデルは走査確率が静的な場合である．

以下では，走査確率が一定という条件の元，インターリービングの効率が A/B テストと同等以上によいことを示す．

条件 3. $P_{*,A}(O=1)=P_{*,B}(O=1)=w$*:* 走査確率は一定である．

定理 2. 条件がなりたつとき，$P(sign(E_A(Y),E_B(Y))\neq sign(E_{AB,A}(\bar{Y}),E_{AB,B}(\bar{Y})))\geqq P(sign(E_A(Y),E_B(Y))\neq sign(E_{I,A}(\bar{Y}),E_{I,B}(\bar{Y})))$ となる．

Proof.

$$\begin{aligned}
&P(sign(E_A(Y),E_B(Y))\neq sign(E_{AB,A}(\bar{Y}),E_{AB,B}(\bar{Y})))\\
&=\int_{-\infty}^{-(E_{AB,A}(\bar{Y})-E_{AB,B}(\bar{Y}))}\mathcal{N}(x|0,V_{AB,A}(\bar{Y})+V_{AB,A}(\bar{Y}))dx\\
&=\int_{-\infty}^{-(E_{I,A}(\bar{Y})-E_{I,B}(\bar{Y}))}\mathcal{N}(x|0,V_{AB,A}(\bar{Y})+V_{AB,A}(\bar{Y}))dx\\
&\geqq\int_{-\infty}^{-(E_{I,A}(\bar{Y})-E_{I,B}(\bar{Y}))}\mathcal{N}(x|0,V_{I,A}(\bar{Y})+V_{I,A}(\bar{Y}))dx\\
&=P(sign(E_A(Y),E_B(Y))\neq sign(E_{I,A}(\bar{Y}),E_{I,B}(\bar{Y})))
\end{aligned}$$

二行目から三行目への変形は，下記で示す定理 3 を用い，三行目から四行目は下記で示す定理 4 を用いる． □

定理 3. 条件がなりたつとき，$E_{AB,A}(Y)=E_{I,A}(Y)\wedge E_{AB,B}(Y)=E_{I,B}(Y)$ となる．

Proof. $P_{AB,A}(O=1)=w$ から $P_{AB,A}(Y=1)=P_{AB,A}(O=1)\cdot P_{AB,A}(R=1)=w\cdot P_A(R=1)$．また $P_{I,A}(O=1)=w$ から，$P_{I,A}(Y=1)=P_{I,A}(O=1)\cdot P_{I,A}(R=1)=w\cdot P_A(R=1)$．よって，$P_{AB,A}(Y=1)=P_{I,A}(Y=1)$ から $E_{AB,A}(Y)=E_{I,A}(Y)$ となる．$E_{AB,B}(Y)=E_{I,B}(Y)$ も同様． □

定理 4. 条件がなりたつとき，$V_{AB,A}(\bar{Y})+V_{AB,B}(\bar{Y})\geqq V_{I,A}(\bar{Y})+V_{I,B}(\bar{Y})$ となる．

Proof. $V_{AB,A}(\bar{Y})=V_{AB,A}(\bar{O}\cdot\bar{R})=V_{AB,A}(\bar{O})\cdot V_{AB,A}(\bar{R})+E_{AB,A}(\bar{O})^2V_{AB,A}(\bar{R})+E_{AB,A}(\bar{R})^2V_{AB,A}(\bar{O})$．ここで，$O$ はベルヌーイ分布に従うので，$V_{AB,A}(O)=E_{AB,A}(O)(1-E_{AB,A}(O))=w(1-w)$ となる．また A/B テストの総表示回数を $2n$ とすると，ランキング A は n 回表示されるので，$V_{AB,A}(\bar{O})=V_{AB,A}(O)/n$ かつ，$V_{AB,A}(\bar{R})=V_{AB,A}(R)/n$ となる．よって

$$\begin{aligned}
V_{AB,A}(\bar{Y})=w(1-w)V_{AB,A}(R)/n^2+w^2V_{AB,A}(R)/n\\
+E_{AB,A}(R)^2w(1-w)/n
\end{aligned}\tag{3}$$

となる．

一方で，インターリービングのランキング表示回数を m 回とすると，

$$\begin{aligned}
V_{I,A}(\bar{Y})=w(1-w)V_{I,A}(R)/m^2+w^2V_{I,A}(R)/m\\
+E_{I,A}(R)^2w(1-w)/m
\end{aligned}\tag{4}$$

となる．ここで，インターリービングの入力ランキングでアイテムの重複がない場合，$n=m$ となる．また入力ランキングでアイテムの重複が多く，ランキングが似ている場合では，$n<m$ であると考えられる．特に，同一のランキングを入力する場合，$m=2n$ である．以上から，$n\leqq m$ であると考えられる．

よって，$V_{AB,A}(R)=V_{I,A}(R)=V_A(R)$ から，$V_{AB,A}(\bar{Y})\geqq V_{I,A}(\bar{Y})$ となる．同様に，$V_{AB,B}(\bar{Y})\geqq V_{I,B}(\bar{Y})$ が示せるので，$V_{AB,A}(\bar{Y})+V_{AB,B}(\bar{Y})\geqq V_{I,A}(\bar{Y})+V_{I,B}(\bar{Y})$．なお，等号成立条件は，$n=m$ となる入力ランキングのアイテムに重複がないケースである． □

以上から，走査確率が静的である場合，インターリービングは A/B テストと同等以上の効率性があることがわかる．なお，インターリービングの入力ランキングのアイテムに重複がない場合に，インターリービングと A/B テストは同等の効率性を持つといえる．

4.2.4 走査確率が動的な場合

走査確率が動的な場合とは，走査確率が適合度に依存する場合を指す．例えば，カスケードクリックモデルはクリックが適合度に依存し，さらにクリックが走査確率に影響する動的な場合である．ここでは，下記の条件のもと走査確率が動的な場合の誤差について考える．

条件 4. $E_A(R) > E_B(R)$*:* ランキング A の適合度の期待値がランキング B の適合度の期待値よりも大きい．

条件 5. $E_{I,A}(O) = E_{I,B}(O) \simeq f(max(E_A(R), E_B(R)))$*:* 混合ランキングの走査の期待値はランキングの適合度の期待値がより大きいランキングに依存する．ここで，f は単調減少関数である．

これらの条件が成り立つとき，インターリービングは A/B テストよりも効率がよいことを以下で示す．

定理 5. 条件がなりたつとき，$P(sign(E_A(Y), E_B(Y)) \neq sign(E_{AB,A}(\bar{Y}), E_{AB,B}(\bar{Y}))) \geqq E(sign(E_A(Y), E_B(Y)) \neq sign(E_{I,A}(\bar{Y}), E_{I,B}(\bar{Y})))$ が成り立つ．

Proof.

$$
\begin{aligned}
&P(sign(E_A(Y), E_B(Y)) \neq sign(E_{AB,A}(\bar{Y}), E_{AB,B}(\bar{Y}))) \\
&= \int_{-\infty}^{-(E_{AB,A}(\bar{Y}) - E_{AB,A}(\bar{Y}))} \mathcal{N}(x|0, V_{AB,A}(\bar{Y}) + V_{AB,A}(\bar{Y}))dx \\
&> \int_{-\infty}^{-(E_{I,A}(\bar{Y}) - E_{I,A}(\bar{Y}))} \mathcal{N}(x|0, V_{AB,A}(\bar{Y}) + V_{AB,A}(\bar{Y}))dx \\
&\geqq \int_{-\infty}^{-(E_{I,A}(\bar{Y}) - E_{I,A}(\bar{Y}))} \mathcal{N}(x|0, V_{I,A}(\bar{Y}) + V_{I,A}(\bar{Y}))dx \\
&= P(sign(E_A(Y), E_B(Y)) \neq sign(E_{I,A}(\bar{Y}), E_{I,B}(\bar{Y})))
\end{aligned}
$$

二行目から三行目は下記で示す定理 6 を用い，三行目から四行目は定理 7 を用いる．

□

定理 6. 条件がなりたつとき，$E_A(Y) > E_B(Y) \rightarrow E_{I,A}(Y) = E_{AB,A}(Y) > E_{AB,B}(Y) > E_{I,B}(Y)$ となる．言い換えると，ランキング A のクリックの期待値がランキング B よりも大きいならば，インターリービングによって得られるランキング A のクリック期待値のほうがランキング B よりも大きく評価の整合性がとれ，かつクリック期待値の差が大きくなる．

Proof. 条件より，$E_A(R) > E_B(R)$ なので $E_{I,A}(O) = f(max(E_A(R), E_B(R))) = f(E_A(R))$．A/B テストは，同じランキング同士を混ぜることと解釈すると，$E_{AB,A}(O) = f(max(E_A(R), E_A(R))) = f(E_A(R))$．よって，$E_{I,A}(O) = E_{AB,A}(O)$ から $E_{I,A}(Y) = E_{AB,A}(Y)$．

同様にして A/B テストを同じランキング同士を混ぜること解釈すると，$E_{AB,B}(O) = f(max(E_B(R), E_B(R))) = f(E_B(R))$．また，$E_{I,B}(O) = f(max(E_A(R), E_B(R))) = f(E_A(R))$．ここで，$E_A(R) > E_B(R)$ かつ f は単調減少関数なので $E_{AB,B}(O) > E_{I,B}(O)$．よって，$E_{AB,B}(Y) > E_{I,B}(Y)$ となる．

$E_A(Y) = 2E_{AB,A}(Y) \wedge E_B(Y) = 2E_{AB,B}(Y)$ なので，$E_A(Y) > E_B(Y) \rightarrow E_{AB,A}(Y) > E_{AB,B}(Y)$．

以上から，$E_A(Y) > E_B(Y) \rightarrow E_{I,A}(Y) = E_{AB,A}(Y) > E_{AB,B}(Y) > E_{I,B}(Y)$ となる．

□

Table 1 クリックモデル

	$P(click = 1\|R)$			$P(stop = 1\|R)$		
R	0	1	2	0	1	2
Perfect	0.0	0.4	1.0	0.0	0.0	0.0
Navigational	0.05	0.5	0.95	0.2	0.5	0.9
Informational	0.4	0.7	0.9	0.1	0.3	0.5

定理 7. 条件がなりたつとき，$V_{AB,A}(\bar{Y}) + V_{AB,A}(\bar{Y}) \geqq V_{I,A}(\bar{Y}) + V_{I,A}(\bar{Y})$

Proof. 証明は複雑であるため，付録に示す． □

以上から，走査確率が適合度に依存し，ユーザーが適合度に応じてランキングから離脱するような場合において，インターリービングは A/B テストよりも効率が良いといえる．

5 実　　験

本節では，下記の研究疑問 (Research Question: RQ) に回答するために実験を行う．

- どのようなクリック行動のもとで，インターリービングは効率がよいか？
- 入力ランキングの適合度の差は，誤差率にどのような影響を与えるか？

5.1 データセット

本実験では，これまでのインターリービングやマルチリービングの実験に利用されてきたものと同じ複数のデータセットを使用する [9] ．各データセットは，クエリセットと各クエリに対応する文書セットから構成され，クエリの数や検索タスクが異なる．クエリは識別子のみで表現されており，文書とクエリのペアごとに適合度のラベルが提供されている．適合度のラベルは非適合 (0)，適合 (1)，大きく適合 (2) の 3 つのレベルに分かれている．すべてのデータセットは 5 つのフォールドから構成されている．

これらのデータセットの多くは，2003 年から 2008 年までの TREC Web Tracks のものである [25] [26] [27]．HP2003，HP2004，NP2003，NP2004，TD2003，TD2004 はそれぞれ 50〜150 のクエリと 1000 の文書がある．OHSUMED データセットは，オンライン医療情報データベースである MEDLINE の検索エンジンのクエリログに基づいており，106 件のクエリが含まれている．MQ2007 と MQ2008 は，TREC の Million Query Track [28] のデータセットで，それぞれ 1700 と 800 のクエリで構成されており，クエリあたりの評価文書数は他のデータセットに比べて少ない．

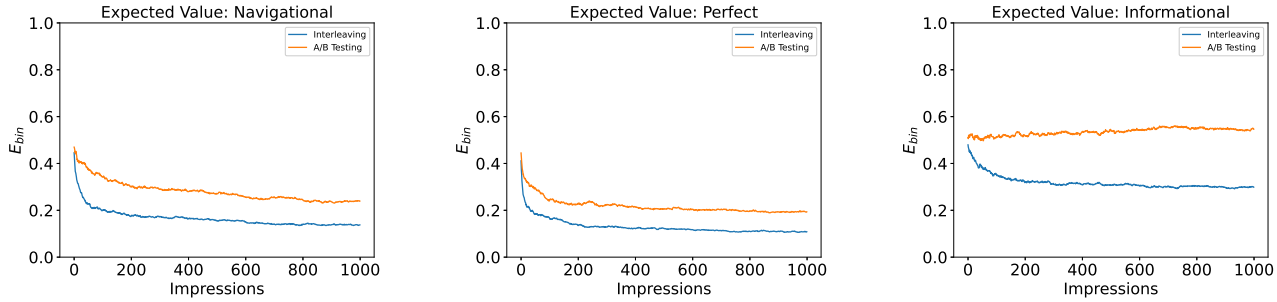

Figure 1　各クリックモデルに対する誤差率の推移

5.2　ユーザー行動

本実験では，ユーザー行動の設定についても過去のインターリービングの既存研究の手順に準拠する．実際のユーザーを使った実験が望ましいが，ほとんどの研究者は実サービス上の検索エンジンを実験に利用するのは容易ではない．そのため，インターリービングの実験では，ユーザのクリック行動のシミュレーションを用いることが多い [8] [22] [23] [29] [30]．このシミュレーションは，様々なパラメータを柔軟に変更できるほか，再現性が担保しやすいという長所もある．

本研究では，3 つのステップでユーザの行動をシミュレートする．まず，ユーザーはクエリを発行した後にランキングが表示される．次に，ユーザーはランキングのアイテムをクリックするか否かを決定する．ユーザーがアイテムをクリックした場合，ユーザーがそのアイテムに満足すれば，ランキングから離脱してセッションを終了する．ユーザー行動の詳細は以下のとおりである．

5.2.1　ランキングの表示

まず，ユーザはデータセットからクエリを一様にサンプリングして擬似的なクエリを発行する．その後，インターリービングにより混合ランキングを生成し，ユーザに混合ランキングを表示する．本実験では TDI をインターリービング手法として用いた．各ランキング表示において，最大で 5 個のアイテムがユーザに表示される．このランキング表示のあと，ユーザーはクリックモデルを用いてクリックをシミュレートする．

5.2.2　クリックモデル

クリックのシミュレーションには，カスケードクリックモデルとランダムクリックモデルを使用する．このカスケードクリックモデルでは，ユーザは表示された順番にアイテムを走査することを想定している．各文書について，ユーザはそれがクリックに値するかどうかを決定する．これは条件付き確率 $P(click = 1|R)$ としてモデル化され，R はデータセットが提供する適合度のラベルである．カスケードクリックモデルは，適合度ラベルの度合いに応じてクリックの確率を増加させる．

Table 1 は本稿で使用するカスケードクリックモデルの 3 つのインスタンスを表している．Perfect モデルは，表示されたアイテムを全て考慮し，全ての適合アイテムをクリックする．Navigational モデルは，1 つの適合度の高いアイテムを探しているユーザーをモデル化している．Navigational モデルは，適合度の高いアイテムをクリックした後は，ランキングから離脱するモデルである．Informational モデルは，複数のドキュメントをクリックする情報収集型のユーザーを表している．

5.3　結　　果

5.3.1　RQ1: どのようなクリック行動のもとで，インターリービングは効率がよいか？

Figure 1 は，各クリックモデルに対するクリックの期待値を用いた誤差率の推移を表している．横軸は，ランキングの表示回数で，縦軸が評価指標として期待値を用いた際の誤差率を表している．結果は，各データセットに対して 10 回集計を行い，全てのデータセットの結果を平均化した値を表している．

この結果から，Navigational クリックモデルは，A/B テストに比べて評価の効率がよくかつ評価の誤差が小さいことがわかる．Navigational クリックモデルは，ユーザーは適合アイテムを 1 つクリックしたあとで離脱するというモデルであった．そのためこの Figure 1 の結果は，インターリービングのモデルの節で説明した，ランキングからの離脱によって，インターリービングの効率性がもたらされるという仮説とも一致する．

以上から，RQ1 の答えは，インターリービングは適合度に応じてユーザーがランキングから離脱するような場合において，特に効率が良いといえる．

5.3.2　RQ2: 入力ランキングの適合度の差は，誤差率にどのような影響を与えるか？

Figure 2 は，nDCG の差に対する誤差率を表している．横軸は，ランカーペアに対する nDCG の差を表し，縦軸は誤差率を表している．結果は，各クリックモデルごとに表示している．

全てのクリックモデルにおいて，nDCG の差が 0.3 より大きい場合にインターリービングの誤差率は A/B テストの誤差率を下回っている．特に，Navigational クリックモデルにおいて，インターリービングの誤差率は A/B テストよりも誤差率の差が大きくなっていることがわかる．一方で，nDCG が 0.2 のときに Navigational クリックモデルと Perfect クリックモデルでは，インターリービングの誤差率は A/B テストよりも大きくなっている．これらの結果は,，適合度の小さなキングが適合度の大きいランキングにアイテムの表示機会を奪われることで，誤差確率が小さくなるというインターリービングのモデルでの考察と一致する．

以上から，RQ2 の答えは，適合度の差が大きいランキングのペアほどインターリービングの誤差率は小さくなるといえる．

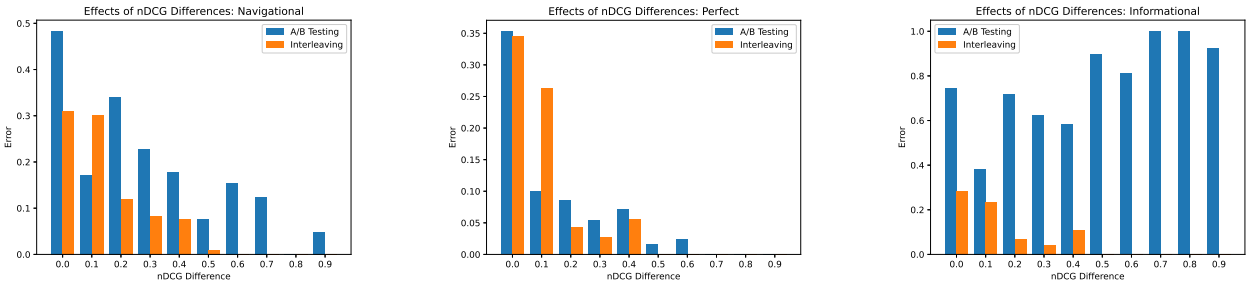

Figure 2　nDCG の差に対する誤差率

6　まとめと今後の課題

本論文では，オンライン評価手法の１つであるインターリービングの正確性と効率性について理論的な考察を与えた．まずインターリービングの動作を説明するモデルを構築し，インターリービングのランキング生成方針の特徴を利用してモデルの解析を行った．我々は，このモデルの解析を通して，ユーザーがアイテムの適合度に依存してランキングから離脱するケースにおいて，インターリービングが A/B テストよりも誤差が小さくなり，正確性と効率性の面で優れていることを示した．数値実験では，複数のクリックモデルに対しインターリービングの正確性と効率性を検証した．結果として，理論的な考察と同様に，ユーザーがアイテムの適合度に依存してランキングから離脱しやすい Navigational クリックモデルにおいて，インターリービングが A/B テストよりも正確性と効率性がよいことを確認した．

References

[1] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, p. 123–132, 2013.

[2] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. Embedding-based news recommendation for millions of users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, p. 1933–1942, New York, NY, USA, 2017. Association for Computing Machinery.

[3] Mihajlo Grbovic and Haibin Cheng. Real-time personalization using embeddings for search ranking at airbnb. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '18, p. 311–320, New York, NY, USA, 2018. Association for Computing Machinery.

[4] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In *Proceedings of the 17th ACM conference on Information and knowledge management*, pp. 43–52, 2008.

[5] Olivier Chapelle, Thorsten Joachims, Filip Radlinski, and Yisong Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems (TOIS)*, Vol. 30, No. 1, pp. 1–41, 2012.

[6] Kojiro Iizuka, Yoshifumi Seki, and Makoto P. Kato. *Decomposition and interleaving for variance reduction of post-click metrics*, p. 221–230. 2021.

[7] Masahiro Sato. *Online evaluation methods for the causal effect of recommendations*, p. 96–101. 2021.

[8] Anne Schuth, Floor Sietsma, Shimon Whiteson, Damien Lefortier, and Maarten de Rijke. Multileaved comparisons for fast online evaluation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 71–80, 2014.

[9] Harrie Oosterhuis and Maarten de Rijke. Sensitive and scalable online evaluation with theoretical guarantees. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, p. 77–86, New York, NY, USA, 2017. Association for Computing Machinery.

[10] Fan Guo, Chao Liu, and Yi Min Wang. Efficient multiple-click models in web search. In *Proceedings of the second acm international conference on web search and data mining*, pp. 124–131, 2009.

[11] Aleksandr Chuklin, Ilya Markov, and Maarten De Rijke. *Click models for web search*. Morgan & Claypool Publishers, 2015.

[12] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 1st ACM International Conference on Web Search and Data Mining*, pp. 87–94, 2008.

[13] Georges Dupret, Vanessa Murdock, and Benjamin Piwowarski. Web search engine evaluation using clickthrough data and a user model. In *WWW2007 workshop Query Log Analysis: Social and Technological Challenges*, Vol. 2. Citeseer, 2007.

[14] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, Vol. 27, No. 1, pp. 1–27, 2008.

[15] Georges E Dupret and Benjamin Piwowarski. A user browsing model to predict search engine click data from past observations. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 331–338, 2008.

[16] Olivier Chapelle and Ya Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th international conference on World wide web*, pp. 1–10, 2009.

[17] Jeff Huang, Ryen W White, Georg Buscher, and Kuansan Wang. Improving searcher models using mouse cursor activity. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 195–204, 2012.

[18] Qianli Xing, Yiqun Liu, Jian-Yun Nie, Min Zhang, Shaoping Ma, and Kuo Zhang. Incorporating user preferences into click models. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pp. 1301–1310, 2013.

[19] Alexey Poyarkov, Alexey Drutsa, Andrey Khalyavin, Gleb Gusev, and Pavel Serdyukov. Boosted decision tree regres-

sion adjustment for variance reduction in online controlled experiments. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 235–244, 2016.

[20] Brian Karrer, Liang Shi, Monica Bhole, Matt Goldman, Tyrone Palmer, Charlie Gelman, Mikael Konutgan, and Feng Sun. Network experimentation at scale. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3106–3116, 2021.

[21] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 133–142, 2002.

[22] Katja Hofmann, Shimon Whiteson, and Maarten De Rijke. A probabilistic method for inferring preferences from clicks. In *Proceedings of the 20th ACM International on Conference on Information and Knowledge Management*, pp. 249–258, 2011.

[23] Brian Brost, Ingemar J Cox, Yevgeny Seldin, and Christina Lioma. An improved multileaving algorithm for online ranker evaluation. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 745–748, 2016.

[24] Filip Radlinski and Nick Craswell. Optimized interleaving for online retrieval evaluation. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, pp. 245–254, 2013.

[25] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. Letor: A benchmark collection for research on learning to rank for information retrieval. August 2016.

[26] Charles L Clarke, Nick Craswell, and Ian Soboroff. Overview of the trec 2009 web track. Technical report, WATERLOO UNIV (ONTARIO), 2009.

[27] Ellen M Voorhees and Donna Harman. Overview of trec 2003. In *Trec*, pp. 1–13, 2003.

[28] James Allan, Ben Carterette, Javed A Aslam, Virgil Pavlu, Blagovest Dachev, and Evangelos Kanoulas. Million query track 2007 overview. Technical report, MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE, 2007.

[29] Katja Hofmann, Shimon Whiteson, and Maarten De Rijke. Fidelity, soundness, and efficiency of interleaved comparison methods. *ACM Transactions on Information Systems (TOIS)*, Vol. 31, No. 4, pp. 1–43, 2013.

[30] Anne Schuth, Robert-Jan Bruintjes, Fritjof Buüttner, Joost van Doorn, Carla Groenland, Harrie Oosterhuis, Cong-Nguyen Tran, Bas Veeling, Jos van der Velde, Roger Wechsler, et al. Probabilistic multileave for online retrieval evaluation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 955–958, 2015.

付　　録

定理 7 の証明を与える．

Proof. まず，$V_{AB,A}(\bar{Y}) \geqq V_{I,A}(\bar{Y})$ を示す．$P_{AB,A}(O) = w_A, P_{AB,B}(O) = w_B$ とする．$V_{AB,A}(\bar{Y}) = V_{AB,A}(\bar{O} \cdot \bar{R}) = V_{AB,A}(\bar{O}) \cdot V_{AB,A}(\bar{R}) + E_{AB,A}(\bar{O})^2 V_{AB,A}(\bar{R}) + E_{AB,A}(\bar{R})^2 V_{AB,A}(\bar{O})$．ここで，$O$ はベルヌーイ分布に従うので，$V_{AB,A}(O) = E_{AB,A}(O)(1 - E_{AB,A}(O)) = w_A(1 - w_A)$ となる．

A/B テストの総表示回数を $2n$ とすると，ランキング A は n 回表示されるので，$V_{AB,A}(\bar{O}) = V_{AB,A}(O)/n$ かつ，$V_{AB,A}(\bar{R}) = V_{AB,A}(R)/n$ となる．以上をまとめると，

$$V_{AB,A}(\bar{Y}) = w_A(1 - w_A)V_{AB,A}(R)/n^2 + w_A^2 V_{AB,A}(R)/n + E_{AB,A}(R)^2 w_A(1 - w_A)/n$$

同様に，インターリービングのランキングを m 回とすると

$$V_{I,A}(\bar{Y}) = w_A(1 - w_A)V_{I,A}(R)/m^2 + w_A^2 V_{I,A}(R)/m + E_{I,A}(R)^2 w_A(1 - w_A)/m$$

となる．ここで，インターリービングの入力ランキングでアイテムの重複がない場合，$n = m$ となる．また入力ランキングでアイテムの重複が多く，ランキングが似ている場合では，$n < m$ であると考えられる．特に，同一のランキングを入力する場合，$m = 2n$ である．以上から，$n \leqq m$ であると考えられる．$V_{AB,A}(R) = V_{I,A}(R) = V_A(R)$ であるから，$V_{AB,A}(\bar{Y}) \geqq V_{I,A}(\bar{Y})$．

つぎに，$V_{AB,B}(\bar{Y}) > V_{I,B}(\bar{Y})$ を示す．上記と同様に，A/B テストの総表示回数を $2n$ とすると，

$$V_{AB,B}(\bar{Y}) = w_B(1 - w_B)V_{AB,B}(R)/n^2 + w_B^2 V_{AB,B}(R)/n + E_{AB,B}(R)^2 w_B(1 - w_B)/n$$

$P_{AB,A}(O) = P_{I,A}(O) = P_{I,B}(O) = w_A$ なので，インターリービングのランキングの表示回数を m 回とすると，

$$V_{I,B}(\bar{Y}) = w_A(1 - w_A)V_{I,B}(R)/m^2 + w_A^2 V_{I,B}(R)/m + E_{I,B}(R)^2 w_A(1 - w_A)/m$$

w_B は A/B テストにおけるランキング B の表示確率であるため $w_B = P_{AB}(S = B)P_B(O) \leqq 0.5$．このとき，$w_A < w_B \leqq 0.5$，$n \leqq m$，$V_{AB,B}(R) = V_{I,B}(R)$ から $w_B(1 - w_B)V_{AB,A}(R)/n^2 \geqq w_A(1 - w_A)V_{I,B}(R)/m^2$ かつ $w_B^2 V_{AB,B}(R)/n \geqq w_A^2 V_{I,B}(R)/m$ となる．また $w_A < w_B \leqq 0.5$，$n \leqq m$，$E_{AB,B} = E_{I,B}(R)$ から $E_{AB,B}(R)^2 w_B(1 - w_B)/n \geqq E_{I,B}(R)^2 w_A(1 - w_A)/m$．よって，$V_{AB,B}(\bar{Y}) \geqq V_{I,B}(\bar{Y})$ となる．

以上から，$V_{AB,A}(\bar{Y}) + V_{AB,B}(\bar{Y}) \geqq V_{I,A}(\bar{Y}) + V_{I,B}(\bar{Y})$ となる． □

謝　　辞

この研究は，株式会社 Gunosy の皆様による協力のもと行われました．また，筑波大学の加藤研究室の皆様には，温かいご指導ご鞭撻を賜りました．この場を借りて深く御礼申し上げます．