コミュニティQ&Aサイトにおける質問検索システムの大規模オンライン評価

加藤　誠† 山本　岳洋†† 真鍋　知博††† 西田　成臣††† 藤田　澄男†††

† 京都大学国際高等教育院データ科学イノベーション教育研究センター
†† 京都大学大学院情報学研究科
††† ヤフー株式会社
E-mail: †{kato,tyamamot}@dl.kuis.kyoto-u.ac.jp, ††{tomanabe,anishida,sufujita}@yahoo-corp.jp

あらまし　本論文では，NTCIR-13 OpenLiveQ (Open Live Test for Question Retrieval) にて実施されたコミュニティQ&Aサイトにおける質問検索システムの大規模オンライン評価について述べる．本研究では特にインターリービングと呼ばれる効率的なオンライン評価手法を用いており，その実システムに適用する際の課題・解決方法やオフライン評価との差異などを報告し，より大規模な評価にあたり解決すべき問題点について述べる．

キーワード　オンライン評価，コミュニティQ&A，インターリービング

1. はじめに

コミュニティQ&Aサービスはユーザが質問を投稿し他のユーザから回答が得られるインターネットサービスである．ユーザは質問をするだけでなく，過去に投稿された質問・回答を検索することでも，自身の情報要求に適合する情報を見つけることができる．特に情報要求が特殊である場合や複雑な場合において，情報要求に類似した質問に対する回答の検索は重要な情報探索手段である．これまで多くの質問検索手法が提案されてきているが[1, 13, 14]，依然として重要な課題が残されている．

曖昧・不明瞭なクエリ　既存の研究では主に明瞭なクエリを対象としてきた．しかしながら，コミュニティQ&Aサービスで用いられる多くのクエリはWeb検索クエリのように短く，そのため，曖昧または不明瞭であり，それらのクエリを入力したユーザの情報要求は多様である可能性が高い．したがって，複数の情報要求を満たせるように質問検索結果もまた多様化されるべきである．

多様な適合性基準　伝統的な情報検索評価における適合性は**主題適合性**を指していることが多い．主題適合性とは，クエリによって表現される主題と文書に書かれている主題が合致している度合いによって計測される適合性のことである．一方，現実の質問検索者は鮮度や具体性，信憑性，簡潔さなど様々な適合性基準を持っている．したがって，従来の適合性判定では質問検索システムの実際の精度を計測することができていないのではないかと思われる．

これらの問題を解決するために，我々は *Open Live Test for Question Retrieval* (*OpenLiveQ*) [6] と呼ばれる新しい共有タスクをNTCIR-13 [5] において提案し，この共有タスクの参加者に対して，ヤフー知恵袋 (注1) におけるオープンライブテスト環境を提供した．参加者はある特定のクエリ集合に対する質問ランキングを提出し，実際のユーザからのフィードバックに基づく評価結果を得ることができる．実ユーザを評価に介入させることにより，実際の検索タスクに従事するユーザからのクエリ・フィードバックを用いることで，我々はユーザの多様な情報要求や適合性基準を考慮することが可能となった．

評価者による適合性判定を用いた従来の評価方法をオフライン評価と呼ぶのに対して，実ユーザからのフィードバックによる評価方法は**オンライン評価**と呼ばれている．評価対象のシステムを異なるユーザ群に提示するA/Bテストは代表的なオンライン評価として知られているが，近年，**インターリービング** (interleaving) というオンライン評価手法がA/Bテストよりも10〜100倍ほど効率的であることから大きな注目を集めている[2, 11]．インターリービングでは，評価対象のランキングシステムの出力を1つのランキングとして統合し，この統合された結果をユーザに提示しフィードバックを観測することによって，ランキングシステムを評価する方法である．我々の過去の実験に基づき[8]，NTCIR-13 OpenLiveQではOptimized Multileaving (OM) [12] と呼ばれるインターリービング手法を採用した．ただし，実システムへの適用に当たり，本来のOMに対していくつかの工夫が行われている．また，NTCIR-13 OpenLiveQで行われた10システムの比較において，OMの限界についても明らかとなった．

本論文では下記の貢献について順を追って説明する：(1) クエリの背後にある異なる情報要求に対して検索結果を多様化し，多様な適合性基準に応える必要がある共有タスクを設計した（2節，3節）．(2) ヤフー知恵袋の実ユーザにより質問検索システムの評価を行った（4節）．(3) インターリービングを実システムに適用する際の課題・解決方法やオフライン評価との差異を明らかにした（5節）．

2. タスク

OpenLiveQのタスクは簡潔に「クエリと回答付きの質問集

(注1)：http://chiebukuro.yahoo.co.jp/

合が入力として与えられた時に順位付きの質問リストを出力する問題」として定義される．また，我々のタスクは下記の 3 つのフェーズから構成されていた：

（1） **オフライン訓練フェーズ** クエリ集合，各クエリに対する質問集合，クリックスルーデータから成る訓練データ（詳細は 3 節を参照）が参加者に与えられた．参加者はこの訓練データを用いて質問検索システムを構築することが期待された．

（2） **オフラインテストフェーズ** 参加者にはクエリ集合，各クエリに対する質問集合，一部のクエリに対するクリックスルーデータから成るテストデータが与えられ，ある期限までに各質問に対する順序付きの質問リストを提出することが求められた．我々は提出された質問リストを従来のオフライン評価によって評価し，どの質問検索システムをオンラインテストフェーズで評価するかを決定した．これは，現在の質問検索システムの結果よりも大きく劣っているような結果を実システム中に含めないようにするため，また，後述するようにオンラインテストフェーズで評価できるシステム数に限りがあったため，評価すべきシステムを選ぶ必要があったためである．

（3） **オンラインテストフェーズ** オフラインテストフェーズで選ばれた質問検索システムはヤフー知恵袋の実システム上で評価された．前述の通り，オンライン評価にはインターリービングが用いられた．インターリービングでは，評価対象の質問検索システム数が多くなるにつれ，評価に必要な提示回数が増加することが知られていたため [12]，我々のオンライン評価では 10 の質問検索システムを選択した．

3. データ

本節では OpenLiveQ で用いられたデータについて説明する．

3.1 クエリ

我々はヤフー知恵袋のクエリログから 2,000 クエリをランダムに抽出し，1,000 クエリを訓練データに，残りの 1,000 クエリをテストデータとして用いた．ランダム抽出の前に，我々はいくつかのフィルタリングをクエリに施し，OpenLiveQ タスクに適さないクエリを除いた．

まず，時間に大きく依存するようなクエリを除去した．参加者にはある時点での質問集合が配布されており，その質問集合を並び替えて提出することが求められていた．これらの質問集合は実ユーザの検索結果として提示されることになるため，時間に応じて適合性が変化しやすいような情報要求が予想されるクエリは望ましくない．そのため，我々は極端に時間依存なクエリを除去し，時間に依存しないようなクエリのみを OpenLiveQ タスクでは用いた．

時間依存クエリを除去する方法は以下の通りである．あるクエリ q に対する質問のうち，2017 年 7 月 16 日から 9 月 16 日の 2ヶ月間に投稿された質問の数を n^q_{recent}，2013 年 1 月 1 日から 2017 年 7 月 15 日のおよそ 4.5 年間に投稿された質問の数を n^q_{past} とする．我々は $n^q_{\text{recent}}/n^q_{\text{past}} > 1.0$ を満たすようなクエリを時間に依存するクエリとして除いた．

次に，成人向けクエリを除去した．あるクエリに対する質問のうち，10%以上の質問がヤフー知恵袋の成人向けカテゴリに分類されている場合，このクエリを成人向けクエリとして除去した．

時間依存および成人向けクエリを除去した後，倫理的に問題のあるクエリ，差別的なクエリ，個人情報に関わるようなクエリを人手により除去した．OpenLiveQ タスクの運営者 3 名が独立にクエリを目視により確認し，1 人以上が上記の問題のうちいずれかに抵触すると判断したクエリは除去された．

上記のフィルタリング作業の後，我々は 2,000 クエリをランダムに選定した．

3.2 質問

2016 年 12 月 1 日から 9 日にかけて，我々は現行のヤフー知恵袋検索システムに各々のクエリを入力し，得られた上位 1,000 件の質問をランキング対象の質問集合とした．この時点での質問に関する情報は記録され，OpenLiveQ タスクの参加者に配布された．以下にその詳細を示す．

- クエリ ID（その質問を検索するのに用いられたクエリを指す）
- ヤフー知恵袋の検索結果中の順位
- 質問 ID
- 質問のタイトル
- 検索結果に表示される質問の要約
- 質問の状態 (回答受付中，投票受付中，解決済),
- 質問の最終更新時間
- 質問への回答数
- 質問のページビュー数
- 質問のカテゴリ
- 質問の本文
- 質問へのベストアンサーの本文

質問の総数は 1,967,274 件であった．前述の通り，参加者は各テストクエリに対して，これらの質問を順位付けし提出することを求められた．

3.3 クリックスルーデータ

いくつかの質問に対してクリックスルーデータが利用可能であった．このデータを用いることで，質問のクリック確率やどのようなユーザがある質問をクリックしたのかを知ることができる．このクリックスルーデータは 2016 年 8 月 24 日から 11 月 23 日の 3ヶ月間に収集されたものである．

クリックスルーデータは以下の情報を含んでいる．

- クエリ ID（その質問を検索するのに用いられたクエリを指す）
- 質問 ID
- ヤフー知恵袋の検索結果中の最頻順位
- クリックスルー率
- 質問をクリックしたユーザにおける男性の割合
- 質問をクリックしたユーザにおける女性の割合
- 質問をクリックしたユーザにおける 10 歳以下の割合
- 質問をクリックしたユーザにおける 10 代の割合
- 質問をクリックしたユーザにおける 20 代の割合
- 質問をクリックしたユーザにおける 30 代の割合
- 質問をクリックしたユーザにおける 40 代の割合

図 1　適合性判定システムのスクリーンショット

- 質問をクリックしたユーザにおける 50 代の割合
- 質問をクリックしたユーザにおける 60 歳以上の割合

このクリックスルーデータには，クエリ ID によって特定されるクエリが入力された時の，質問 ID によって特定される質問に対するクリックの統計情報が含まれている．同じクエリが入力された場合であっても，ある質問の検索結果中の順位は変わりうる．そのため，3 番目の情報はヤフー知恵袋の検索結果中の「最頻」順位となっている．クリックスルーデータ中のクエリと質問の組数は 440,163 であった．

4. 評価方法

本節では適合性判定結果を用いてシステムを評価するオフライン評価と，インターリービングによって実ユーザにより評価を行うオンライン評価について述べる．

4.1 オフライン評価

オフライン評価は，後述のオンライン評価の前に行われ，ここでの評価によって参加者のどのシステムがオンライン評価に含まれるかが決定された．評価は通常のアドホック検索タスクと同様に，適合性判定結果と nDCG (normalized discounted cumulative gain), ERR (expected reciprocal rank), Q-measure といった評価指標を用いて行われた．参加者はオフラインテストフェーズ中，OpenLiveQ の Web サイト (注2) からシステム出力を 1 日 1 回提出することが可能であり，その場で評価結果を受け取ることができた．

オンライン評価と同じような評価結果を得るために，オフライン評価では以下のような指示により，各質問の適合性が得られた：「もしあなたが <query> を入力したと想定した場合に，以下の質問のうち，あなたがクリックしたいと思う質問を全て選んでください」．評価者には質問の本文を提示せず，ヤフー知恵袋の検索結果ページに似たページ上で各質問の適合性を判断させた．このような適合性判定の結果は，従来の適合性判定とは異なり，よりオンライン評価での評価結果に近くなることが期待された．各クエリについて，5 名の評価者を割り当て，各質問の適合度は「その質問を選択した評価者数」とした．例えば，5 名中 2 名の評価者がある質問を選んだ場合，その質問の適合度を「2」とした．

適合性判定にはクラウドソーシングサービスである Lancers(注3) が用いられた．図 1 に適合性判定システムのスクリーンショットを示す．評価者はタイトルや青い四角の部分をクリックすることによって質問を選択することができる．この適合性判定は，コストの問題から，1,000 件のテストクエリのうち，100 件のテストクエリにのみ行われた．

複数の評価指標が OpenLiveQ では用いられていたが，OpenLiveQ の Web サイトでは各システム出力に対する nDCG@10 の値のみが表示されていた．最終的に，各参加者のシステム出力のうち，最も nDCG@10 が高いシステム出力がオンライン評価で用いられることとなった．

4.2 オンライン評価

オンライン評価においてシステム出力は，マルチリービング（3 つ以上のシステム出力に対するインターリービング手法の総称）の一種である，Optimized Multileaving (OM) [12] によって評価された．順位付けされた質問は 1 つの検索結果として統合され，オンラインテストフェーズの間に実ユーザへ提示され，観測されたユーザのクリックに基づいて評価された．マルチリービングの中でも OM を採用したのは，我々の予備実験 [8] の結果，ヤフー知恵袋の環境において，他のマルチリービングよりも優れた性能を発揮したためである．オフラインテストフェーズ中に提出された質問ランキングはそのままオンライン評価で用いられた．ただし，オンライン評価前またはオンライン評価中に何らかの理由で質問が削除された場合にはそれらの質問はオンライン評価で用いられなかった．

一般的なマルチリービング手法はランキング集合 $\mathcal{I} = \{I_1, I_2, \ldots, I_n\}$ を入力として受け取り，混合ランキング集合 $\mathcal{O} = \{O_1, O_2, \ldots, O_m\}$ を出力する．ただし，各々の混合ランキング O_k は l 個の要素から構成されている．入力ランキング I_j，および，混合ランキング O_k の i 番目の要素はそれぞれ $I_{j,i}$，$O_{k,i}$ と表記することにする．あるユーザがあるクエリを入力したとき，1 つの混合ランキング O_k を確率 p_k で提示し，O_k に対するユーザのクリックを観測する．もし $O_{k,i}$ がユーザによってクリックされた場合，各入力ランキング I_j に対してクレジット $\delta(O_{k,i}, I_j)$ を与える．基本的には高いクレジットを得た入力ランキングがより優れていると評価される．各マルチリービング手法は，入力ランキング集合 $\mathcal{I}$ から混合ランキング集合 $\mathcal{O}$ を構成する方法，および，各混合ランキング O_k の提示確率 p_k，クレジット関数 δ によって構成され，これらが異なる様々な方法が提案されてきている．

OM [12] はアルゴリズム 1 によって混合ランキングを出力し，**Bias** をなくしつつ，混合ランキングの **Sensitivity** を最大化するような提示確率 p_k を用いるマルチリービング手法である．混合ランキングの Sensitivity とは，それがユーザに提示された際にどの入力ランキングが優れているかを識別できる力を

(注2)：http://www.openliveq.net/

(注3)：http://www.lancers.jp/

Algorithm 1: Optimized Multileaving (OM)

Require: 入力ランキング $\mathcal{I}$，混合ランキング数 m，各混合ランキングの長さ l

```
O ← {};
for k = 1, ..., m do
    for i = 1, ..., l do
        Select j randomly;
        r = 1;
        while I_{j,r} ∈ O_k do
            r ← r + 1
        end
        if r ≦ |I_j| then
            O_{k,i} = I_{j,r};
        end
    end
    O ← O ∪ {O_k};
end
return O;
```

指す．直感的には，混合ランキングに対してランダムなクリックが与えられた際に，同じ量のクレジットを各入力ランキングに与えるような混合ランキングの Sensitivity は高いと言える．高い Sensitivity は，評価結果の早期収束のために望ましい性質である．混合ランキング集合の Bias とは，ランダムなクリックに対して得られる各入力ランキングのクレジットの差を指す．もし Bias がある場合，ランダムなクリックが実際に与えられた場合であっても，特定の入力ランキングに対して不当に高いクレジットが与えられてしまう．そのため，マルチリービング手法はこの Bias をできる限り小さくする必要がある．

Sensitivity の最大化は，下記のように定義される，混合ランキング O_k 上の順位依存のランダムクリックによって与えられるクレジットの分散，すなわち， Insensitivity の最小化によって達成できる [12]：

$$\sigma_k^2 = \sum_{j=1}^{n} \left(\left(\sum_{i=1}^{l} f(i)\delta(O_{k,i}, I_j) \right) - \mu_k \right)^2, \tag{1}$$

ただし，$f(i)$ はユーザが i 番目の要素をクリックする確率であり，元論文に従い，$f(i) = 1/i$，および，$\delta(O_{k,i}, I_j) = 1/i$ $(O_{k,i} \in I_j)$，$\delta(O_{k,i}, I_j) = 1/(|I_j|+1)$ $(O_{k,i} \notin I_j)$ とする．混合ランキング O_k のクレジットの平均 μ_k は下記のように計算できる：

$$\mu_k = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{l} f(i)\delta(O_{k,i}, I_j). \tag{2}$$

各混合ランキング O_k は提示確率 p_k でユーザに提示されるため，OM は下記で与えられる Insensitivity の期待値を最小化する：

$$\mathbb{E}[\sigma_k^2] = \sum_{k=1}^{m} p_k \sigma_k^2. \tag{3}$$

混合ランキング集合 $\mathcal{O}$ の Bias は，入力ランキングへ与えられるクレジットの期待値の差によって定義される．上位 1 件，2 件，$\cdots$，r 件の要素に対する順位非依存のランダムクリックによって，ある入力ランキング I_j が得られるクレジットの期待値は下記のように定義される：

$$\mathbb{E}[g(I_j, r)] = \sum_{k=1}^{m} p_k \sum_{i=1}^{r} \delta(O_{k,i}, I_j). \tag{4}$$

もし各入力ランキングのクレジットの期待値が異なれば，マルチリービングによって得られる評価結果には偏りが生じる．したがって，元々の提案手法ではすべての入力ランキングへ与えられるクレジットの期待値が同じでなくてはならない，という下記の制約を与えていた：

$$(\forall r, \exists c_r, \forall j)\ \mathbb{E}[g(I_j, r)] = c_r. \tag{5}$$

Schuth らによる論文 [12] と彼らが公開している実装 (注4) によると，元々の OM はまず式 (5) の制約を満たすことで Bias を 0 にし，制約が満たされた上で Sensitivity を最大化するような解を求める．しかしながら，我々の予備実験において，90%以上の場合において，Bias に関する制約を満たせないことがわかっている．すなわち，多くの場合，式 (5) の制約を満たすような解を見つけることができなかったのである．

そこで我々は，下記で定義される Bias と Insensitivity の線形和を最小化するような，より実用的な OM の実装を提案する (注5)：

$$\min_{p_k} \alpha \sum_{r=1}^{l} \lambda_r + \sum_{k=1}^{m} p_k \sigma_k^2 \tag{6}$$

$$\text{subject to } \sum_{k=1}^{m} p_k = 1,\ 0 \leqq p_k \leqq 1\ (k = 1, \ldots, m),$$

$$\forall r, \forall j, j', -\lambda_r \leqq \mathbb{E}[g(I_j, r)] - \mathbb{E}[g(I_{j'}, r)] \leqq \lambda_r\ ,$$

ただし，α は Bias と Insensitivity の優先度を決める超パラメータであり，λ_r は任意の入力ランキングペアにおける期待クレジットの差異の最大値である．もし，λ_r が 0 に近ければ，入力ランキングの期待クレジットは近くなり，そのため， Bias も小さくなる．

4.3 評価設定

ヤフー知恵袋におけるオンライン評価は 2017 年 5 月 9 日から 2017 年 8 月 8 日にかけて実施された．オンライン評価には合計 410,812 件のインプレッションが利用された．

評価対象となった質問検索システムは NTCIR-13 OpenLiveQ タスクに投稿された 85 システム（7 参加チーム）である．前述の通り，これらのシステムに対してオフライン評価が行われ，このうち 10 システムに対するオンライン評価も行われた．NTCIR-13 OpenLiveQ タスクに投稿されたシステムの詳細は，それぞれの参加チームの論文にて説明されている：

- YJRS [7]
- Erler [3]

(注4)：https://github.com/djoerd/mirex

(注5)：https://github.com/mpkato/interleaving にて公開されている．

- SLOLQ [4]
- OKSAT [10]

5. 評価結果

図 2(a)，2(b)，2(c) に nDCG@10，ERR@10，Q-measure によるオフライン評価の結果を示す．これらの図にて，ベースラインの結果は赤色で示されている．1 つ目のベースラインである ORG 4 は，実サービス中で用いられているランキングと全く同じランキングを用いる手法である．他のベースライン手法は，座標降下法によって最適化された線形モデルである [9]．ベースライン手法が複数存在しているのは，同じモデルで異なるパラメータ設定を用いているためである．どの評価指標においても，線形モデルによるベースライン手法，ORG 7 がベースライン手法の中で最も高い性能を示している．いくつかのチームは ORG 7 を超える性能を示しており，nDCG@10 および ERR では OKSAT，YJRS，cdlab が，Q-measure では YJRS と Erler が ORG 7 を上回る性能を示している．

図 3 にオンライン評価で得られたクレジットの累積値を示す．オンライン評価では，オフライン評価に基づき以下の質問検索システムが採用された：KUIDL 18，TUA1 19，Erler 22，SLOLQ 54，cdlab 83，YJRS 86，OKSAT 88．これに加え，オフライン評価で最も高い性能を示した ORG 7，現行のランキング結果である ORG 4，単純に回答数の降順を行うベースライン手法（オフライン評価には含まれていない）に対してオンライン評価が行われた．実験結果から，YJRS と Erler が最も高性能なベースライン手法である ORG を上回っていることが読み取れる．ただし，YJRS と Erler には統計的有意差は認められなかった．検定方法については後述する．

オンライン評価の結果はオフライン評価の結果と異なる傾向を示している．特に，OKSAT はオンライン評価では高い評価を得ているにも関わらず，オンライン評価では比較的低い性能を示している．これに関して，オフライン評価では 100 件のクエリに対する適合性判定を行っているため，検索システムがこの 100 件の結果に過剰適合してしまったことや，オフライン評価とオンライン評価の評価方法の違いに起因するものであると考えられる．

図 4 に t 検定（対応あり，両側，有意水準 5%）によって統計的有意差が認められなかった検索システムペアの数を示す．この図において，x 軸はオンライン評価を実施してから経過した日数を表している．この検定は多重比較となっているため，p 値は Bonferroni 法によって補正されている．オンライン評価ではほとんどの検索システム間において統計的有意差を 10 日以内に発見できていることがわかる（$37/45 = 82.2\%$）．20 日が経過した時点では，$41/45 = 91.1\%$ の検索システムペアに対して，また，64 日が経過した時点では，3 システムペア以外，KUIDL-TUA1，KUIDL-OKSAT，YJRS-Erler 以外の統計的有意差が認められている．

検定の結果から高い識別性能が示唆されているものの，実験規模に比べ統計的有意差を得るために多くの日数が必要となっていることがわかる．このため，より多くの検索システムを同時に比較するためには，更に効率の良い手段が期待される．オンライン評価の評価対象となる検索システムの中には全く違うランキングを行うシステムや非常に似たランキングを行うシステムが存在し，前者は少ないインプレッションでも統計的有意差を得やすいが，後者は多くのインプレッションを必要とする．したがって，重点サンプリングなどの方法を用いて，似た検索システムペアからの結果を含む混合ランキングに，多くのインプレッションを割り当てるような手法が考えられる．

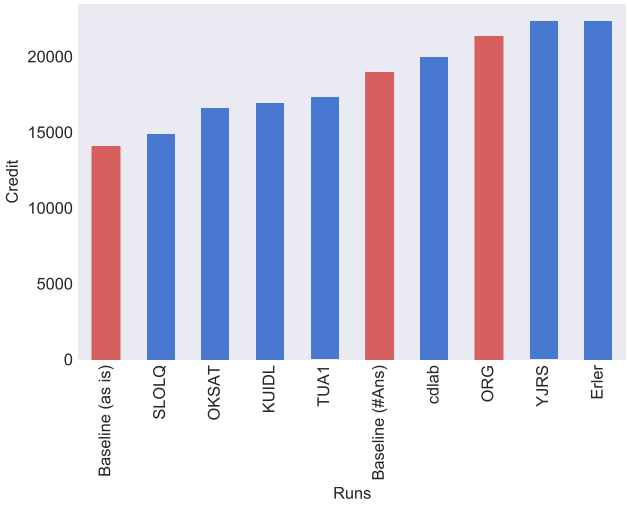


図 3　オンライン評価の累積クレジット

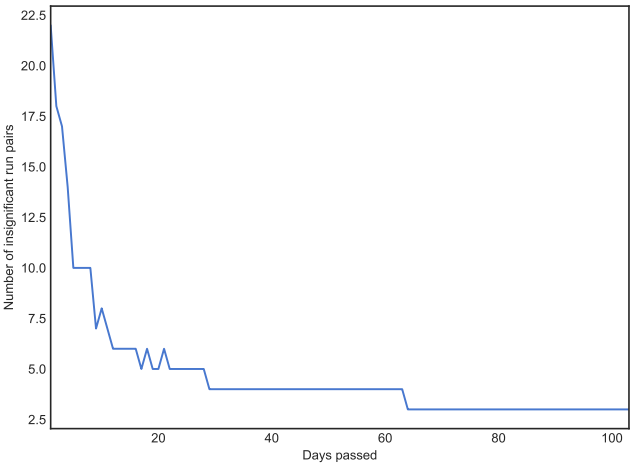
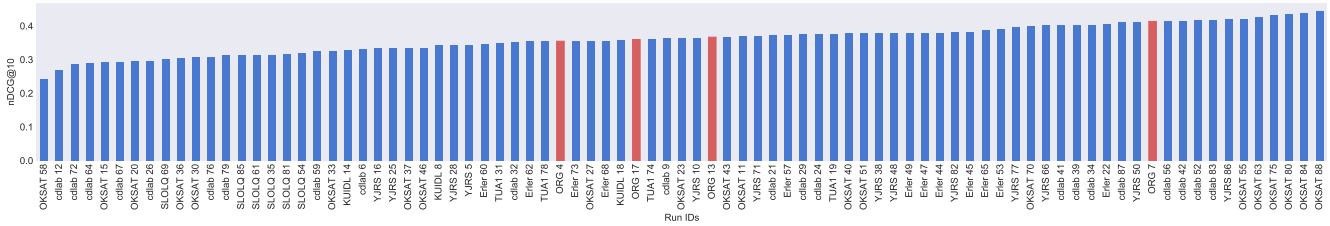


図 4　実験に要した日数と統計的有意差のない検索システムペア数の関係

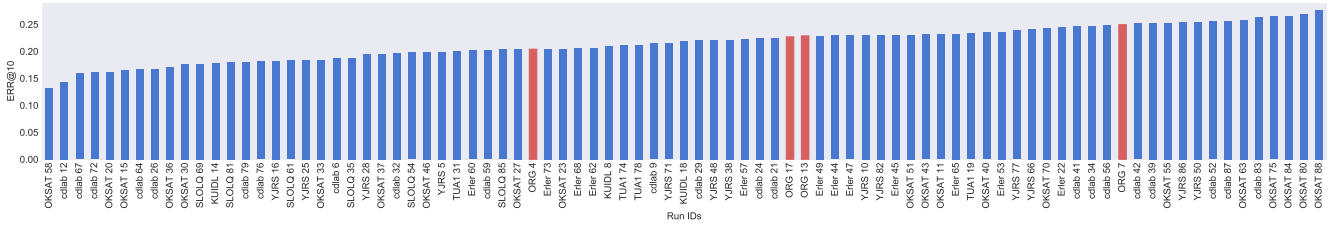
6. まとめ

本論文では，NTCIR-13 OpenLiveQ (Open Live Test for Question Retrieval) にて実施されたコミュニティQ&A サイトにおける質問検索システムの大規模オンライン評価について述べた．本研究では特にインターリービングと呼ばれる効率的なオンライン評価手法を用いており，その実システムに適用する際の課題・解決方法やオフライン評価との差異などを報告し，より大規模な評価にあたり解決すべき問題点について述べた．

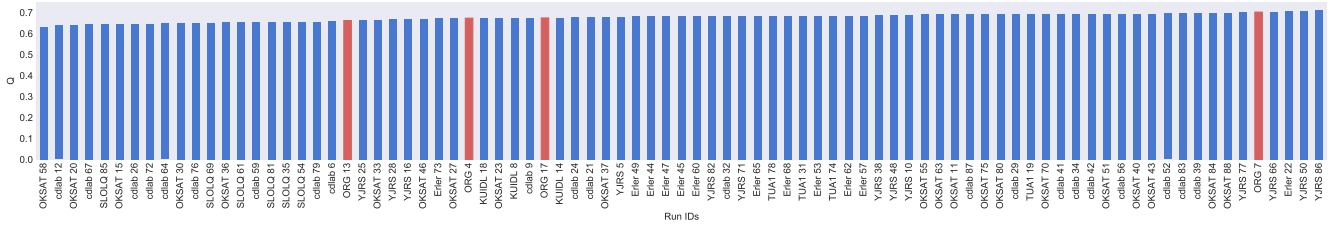
謝辞　本研究は JSPS 科研費 JP26700009, JP16H01756 の助成を受けたものです．ここに記して謝意を表します．また，



(a) nDCG@10



(b) ERR@10



(c) Q-measure

図 2　オフライン評価の結果

NTCIR-13 運営者の皆様および NTCIR-13 OpenLiveQ タスクに参加の皆様にも心から感謝申し上げます.

文　　献

[1] X. Cao, G. Cong, B. Cui, and C. S. Jensen. A generalized framework of exploring category information for question retrieval in community question answer archives. In *WWW*, pages 201–210, 2010.

[2] O. Chapelle, T. Joachims, F. Radlinski, and Y. Yue. Large-scale validation and analysis of interleaved search evaluation. *ACM TOIS*, 30(1):6, 2012.

[3] M. Chen, L. Li, Y. Sun, and J. Zhang. Erler at the NTCIR-13 OpenLiveQ Task. In *NTCIR-13 Conference*, 2017.

[4] R. Kashimura and T. Sakai. SLOLQ at the NTCIR-13 OpenLiveQ Task. In *NTCIR-13 Conference*, 2017.

[5] M. P. Kato and Y. Liu. Overview of ntcir-13. In *NTCIR-13 Conference*, 2017.

[6] M. P. Kato, T. Yamamoto, T. Manabe, A. Nishida, and S. Fujita. Overview of the ntcir-13 openliveq task. In *NTCIR-13 Conference*, 2017.

[7] T. Manabe, A. Nishida, and S. Fujita. YJRS at the NTCIR-13 OpenLiveQ Task. In *NTCIR-13 Conference*, 2017.

[8] T. Manabe, A. Nishida, M. P. Kato, T. Yamamoto, and S. Fujita. A comparative live evaluation of multileaving methods on a commercial cqa search. In *SIGIR*, pages 949–952, 2017.

[9] D. Metzler and W. B. Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.

[10] T. Sato. OKSAT at NTCIR-13 OpenLiveQ Task. In *NTCIR-13 Conference*, 2017.

[11] A. Schuth, K. Hofmann, and F. Radlinski. Predicting search satisfaction metrics with interleaved comparisons. In *SIGIR*, pages 463–472, 2015.

[12] A. Schuth, F. Sietsma, S. Whiteson, D. Lefortier, and M. de Rijke. Multileaved comparisons for fast online evaluation. In *CIKM*, pages 71–80, 2014.

[13] K. Wang, Z. Ming, and T.-S. Chua. A syntactic tree matching approach to finding similar questions in community-based qa services. In *SIGIR*, pages 187–194, 2009.

[14] G. Zhou, Y. Liu, F. Liu, D. Zeng, and J. Zhao. Improving question retrieval in community question answering using world knowledge. In *IJCAI*, pages 2239–2245, 2013.