

タイムライン形式での記事配信における課題と記事閲覧数推定による改善

Pageviews Prediction: Ranking Optimization on News Curation Service with Timeline Interface

大倉 俊平[♡] 田頭 幸浩[◇] 小野 真吾[▲]
田島 玲[▲]

Shumpei OKURA Yukihiro TAGAMI
Shingo ONO Akira TAJIMA

スマートデバイスの普及により、インターネット上で提供されるサービスのユーザインタフェースも変化しており、著名な SNS などを中心に新しい形式のインタフェースであるタイムライン形式が取り入れられている。タイムライン形式は原則として時系列順にアイテムを並べるものであるが、表示するアイテムが記事である場合には、ユーザの使い勝手を考えると、すべての表示可能な記事を時系列順に並べるわけではなく、表示可能な記事の絞り込みや、ユーザの興味を加味して順序の変更を行ったうえで記事を表示する必要がある。我々は、タイムライン形式でユーザに記事を提供するスマートフォン版 Yahoo! JAPAN トップページを対象に、上記の課題を解決するため閲覧実績や記事内容から記事閲覧数を推定する仕組みを導入し、推定値を利用したロジックによる配信を行った。その結果、より多くの記事が閲覧されるようになったことを確認した。本稿では記事を配信、表示するサービスにおける上記のタイムライン形式が持つ課題と、記事閲覧数推定による改善策の内容とその効果について報告する。

1. はじめに

タイムライン (timeline) とは、そもそも年表を意味する語である¹。転じて、SNS (Social Networking Service) を中心に用いられている時系列順にアイテムを表示するインタフェースもタイムラインと呼ばれる。本稿ではより広く、時系列を考慮してアイテムを並べ、それを次々と表示するものをタイムライン形式として扱う²。たとえば、SNS ではそれぞれのアイテムがユーザの投稿となり、それが投稿時刻に沿った形で並べられる。表示するアイテムがニュースなどの記事であっても、時系列に沿った形である場合にはタイムライン形式であるとみなす。タイムライン形式

はスマートフォンの操作性と相性が良いこともあり、多くのサービスで採用されている。

タイムライン形式では、サービスの性質に応じて、時系列順に並べたアイテムのリストに以下の変更が発生することがある。

- **割り込み** アイテムのリストに広告が差し込まれる場合がある。これは、ユーザに対しては無償でサービスを提供し、収益をオンライン広告から得るためである。
- **絞り込み** 表示可能なアイテムの量が非常に多くなる場合、その分量に応じて絞り込みが行われる場合がある。
- **並び替え** アイテムを並べる際に時刻以外の要素も考慮される場合がある。ユーザの投稿を表示するサービスにおいては、ユーザからの反響なども利用しタイムラインを表示している。

これらの要素は、サービスの特徴に加えユーザの使い勝手も考慮され、タイムライン形式に取り入れられる。たとえば、ユーザ間のテキストチャットやメッセージ交換を提供するサービスにおいては、全件表示が前提であり、並び替えも絞り込みが発生しないサービスが多い。一方でユーザの投稿を表示するようなサービスにおいては、絞り込みや並び替えの機能を用意した上でユーザ側に設定変更を可能にしたり、広告の割り込みが発生する事例が見られる。

以下では、実際にタイムライン形式で記事を配信するサービスを例として、その際に発生する課題について述べる。課題として、適切な広告の決定、記事の重複排除、記事の並びの決定の三つが挙げられるが、そのうち、記事の並びの決定のための重要な要素となる記事閲覧数の推定について焦点を当て、RNN (Recurrent Neural Networks) を用いて記事の閲覧数を推定する手法を提案する。これは、記事中の単語や、記事のこれまでの閲覧履歴を元に今後の閲覧数を予測する手法である。提案した手法について、オフライン、オンラインのそれぞれでの評価を実施した。その結果についても説明する。

2. 記事配信サービスにおける課題と取り組み

Yahoo! JAPAN のスマートフォン版トップページは、2015 年 5 月のリニューアル以降、タイムライン形式で記事の表示が行われている。図 1 にその例を示す。画面の基本的な構成要素としては、ロゴやメニューを開くためのアイコンを含むヘッダ、検索を行うためのテキスト入力欄、Yahoo! JAPAN の各サービスへのリンクとなるボタン群に続き、タイムライン形式で記事が並ぶ。上部には、編集者が人手で選択し、見出しを付与したトピックス記事が表示される。それ以降は「あなたへのおすすめ」枠として、自動的に選択、並び替えされた記事と広告が配信される。

「あなたへのおすすめ」枠は、大きく分けて以下の 2 つの役割を持っている。一つは、トピックスには入りきらなかったものの、主要ユーザである日本人や日本に住む人全員にとって情報量が大きいと考えられる記事を掲出することであり、もう一つは、トピックスでは個別化ができないため、ユーザごとの興味を反映した記事を掲出することである。また、「あなたへのおすすめ」枠では、広告も配信されている。

このようにタイムライン形式で記事の配信を行う場合、以下のような課題が存在する。

- 記事の間に挿入される広告について、内容、量、位置を適切に決定 [8]
- 内容が類似した記事の削除 [12]
- ユーザにとって情報量の高い記事の推定と、それに基づく記事の並びの決定

広告に関しては、記事に対する広告の数を増やすと収益が大きくなる反面、ユーザの満足度を低下させ、離脱につながりやすくなる。この課題に関しては、広告の並び順を予測された CTR (Click-Through Rate; クリック率) だけで最適化するのではなく、広告をクリックした後の経験を重要な要素ととらえ、クリック後のページの滞在時間を予測に組み込むことで広告の CTR、滞

[♡] 非会員 ヤフー株式会社
sokura@yahoo-corp.jp

[◇] 非会員 ヤフー株式会社
yutagami@yahoo-corp.jp

[▲] 非会員 ヤフー株式会社
shiono@yahoo-corp.jp

[▲] 非会員 ヤフー株式会社
atajima@yahoo-corp.jp

¹たとえば、英英辞典を引くと A graphical representation of a period of time, on which important events are marked. と書かれている。

²文献によってはストリーム (stream) と呼ばれる場合もある。また、本論文で扱うようなニュース記事の場合は特にヘッドライン (headline) と呼ぶ場合もあるが、本稿ではタイムラインとして扱う。



図 1: スマートフォン版 Yahoo! JAPAN トップページにおけるタイムライン形式の表示例。

在時間、直帰率のいずれもが改善したという論文も報告されている [7]。また、広告自体は好印象でも、記事の並びと関係がありすぎるとサイト全体への印象が悪くなり得るという研究 [2] もあり、適切に広告を配信することは非常に難易度の高い課題である。

「あなたへのおすすめ」枠において配信される記事はさまざまな媒体から入稿されるため、同一の内容を扱った記事が複数存在することが多い。これに対処するため、記事の内容が近いものについては表示時点で削除を実施している [12]。記事のグルーピングに関する研究は盛んに実施されているが、表示可能な記事が随時更新されることを考えると、あらかじめ固定数のグループを作成するような手法は現実的ではない。また、記事を削除する際に類似度計算を行うが、単語単位での文書ベクトルを用いると意味を捉えきれなくなるため、たとえば Latent Semantic Indexing [3] のような意味を考慮した手法で記事間の類似度計算や記事の削除を行う必要がある。

上記のようにまとめ上げを行ったとしても、タイムライン上にすべての配信可能な記事を新着順で表示するというシンプルな並べ方を行うと、ユーザにとっての使い勝手が非常に悪くなる。入稿される記事の中には速報的な記事からコラムやエッセイに近いものも存在するため、非常に多くの記事が脈絡もなく並ぶことになる。また、閲覧履歴のようなユーザのフィードバックが利用されず、ユーザにとって興味のない記事が並ぶ可能性が非常に高い。これらの課題を解決し、ユーザにとって情報量が高い記事がわかりやすくなるよう配信するために、記事の入稿時刻に加え、ユーザの記事閲覧履歴、記事の閲覧数を考慮して、記事の並びを決定している。

記事の閲覧数が高いということは、多くのユーザにとって情報量が高いと考えられる。したがって、実際に記事の並びを決定する際にも閲覧数は重要な要素となり得る。しかし、閲覧数は配信してみるまでわからない上に、記事そのものは SNS 上などで話題

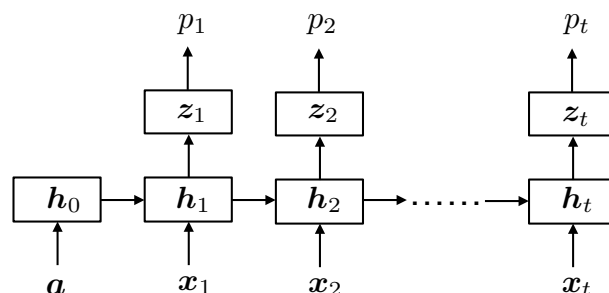


図 2: 記事の閲覧数予測に用いる RNN の構成の概要。

になったり、トピックスに取り上げられたりすることにより、大きな影響を受ける。そのため、記事の閲覧数を高精度で推定することが重要な課題となる。本稿では、この課題を解決するために、RNN を用いて記事の閲覧数を予測する手法を提案する。特徴量として、記事の単語を用いることに加え、その記事が今までどれだけ閲覧されたかを用いることで、それ以降記事がどれだけ閲覧されるかを精度良く求めることができる。

3. 提案手法

この章では、RNN (Recurrent Neural Networks) を用いた記事の閲覧数の予測手法について述べる。

以下では、 $\mathbf{x}$ のような太字の小文字でベクトルを、 $\mathbf{W}$ のような太字の大文字で行列を表すものとする。また、 $\mathbf{x}^T$ と $\mathbf{W}^T$ はそれぞれ、対応するベクトルと行列の転置を表す。

今、ある記事の時刻 t における閲覧数 y_t を、記事の静的特徴量 $\mathbf{a}$ と、その時刻までに利用可能な動的特徴量 $\{\mathbf{x}_s\}_{s=1}^t = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_t)$ を用いて予測したい。例えば、 $\mathbf{a}$ に含まれる特徴量としては、記事の文中の単語や、記事のカテゴリ、入稿元などが考えられる。同様に、 $\mathbf{x}_s$ の例として、その時点でアクセス可能な過去の y_s ($s < t$) や、時刻 t が一週間のうちの何曜日で一日 24 時間のうちの何時であるかの時刻情報などが考えられる。

この記事の閲覧数の予測問題に対して、RNN を用いたアプローチを提案する。RNN は、言語モデル [9, 10] や手書き文字認識 [4]、音声認識 [6]、機械翻訳 [1]、広告のクリック率予測 [13] などの分野でも広く用いられている、系列データを扱うための学習器である。

予測に用いる RNN は以下のように構成される。

$$\mathbf{h}_0 = f(\mathbf{W}\mathbf{a} + \mathbf{b}_a) \quad (1)$$

$$\mathbf{h}_t = f(\mathbf{U}\mathbf{h}_{t-1} + \mathbf{V}\mathbf{x}_t + \mathbf{b}_h) \quad (2)$$

$$\mathbf{z}_t = \sigma(\mathbf{F}\mathbf{h}_t + \mathbf{b}_z) \quad (3)$$

$$p_t = \mathbf{w}^T \mathbf{z}_t + b_p \quad (4)$$

ここで、 $\mathbf{W}, \mathbf{U}, \mathbf{V}, \mathbf{F}$ はパラメータ行列、 $\mathbf{w}, \mathbf{b}_a, \mathbf{b}_h, \mathbf{b}_z$ はパラメータベクトル、 b_p はパラメータである。また、 $f(\cdot)$ と $\sigma(\cdot)$ は共にベクトルに対する活性化関数であり、具体的にはベクトルの要素ごとの ReLU (Rectified Linear Unit) [11] とシグモイド関数である。

$$f(z) = \max(0, z)$$

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

この RNN の構成の概要を図 2 に示す。

記事が新規に入稿されると、式 1 を用いて、記事の静的特徴量 $\mathbf{a}$ の情報を非線形変換し、ニューラルネットの隠れ層 $\mathbf{h}_0$ にエンコードする。記事がユーザーに提示されると、時刻 t における動的特徴量 $\mathbf{x}_t$ が取得できるようになり、式 2 により、隠れ層の情報が $\mathbf{h}_{t-1}$ から $\mathbf{h}_t$ に更新される。そして、式 3 により、さらに隠れ

層を重ねて非線形変換を行い、 z_t を作成する。最終的に式4を用いて、 z_t から閲覧数の予測値 p_t が得られる。

上記のRNNのパラメータを、データから学習する。 i 番目の記事に対応するデータを、上付き文字を用いて、 $(\mathbf{a}^{(i)}, \{\mathbf{x}_t^{(i)}\}_{t=1}^{T_i}, \{y_t^{(i)}\}_{t=1}^{T_i})$ と表す。

これらのデータから、以下の目的関数を最適化するように、パラメータ $\theta = (W, U, V, F, w, b_a, b_h, b_c, b_p)$ を学習する。

$$\hat{\theta} = \arg \min_{\theta} \sum_i^N \sum_{t=1}^{T_i} L(y_t^{(i)}, p_t^{(i)}) \quad (5)$$

ただし、 N は記事数、 $L(y, p) = (y - p)^2$ は二乗誤差の損失関数である。

式5をミニバッチのSGD (Stochastic Gradient Descent) で最適化する。

4. 実験

この章では、提案手法の評価実験について述べる。

2015年5月から6月の期間に、Yahoo! JAPANのトップページに掲載された記事データを28日分用意し、前の21日分を学習データとし、後ろの7日分をテストデータとした。それぞれのデータは同じ規則にしたがい、フィルタリングとサンプリングを行った。その結果、学習データに含まれる記事数が約13万、テストデータに含まれる記事数が約5万となった。

各記事の掲載開始から24時間目までの、1時間ごとの閲覧数を $y_t^{(i)}$ とした。なお、実際の閲覧数は範囲が大きいため、自然対数を適用して用いた。つまり、全ての記事において $T_i = 24$ となり、 i 番目の記事に対するデータは $(\mathbf{a}^{(i)}, \{\mathbf{x}_t^{(i)}\}_{t=1}^{24}, \{y_t^{(i)}\}_{t=1}^{24})$ と表される。

記事の静的特徴量 $\mathbf{a}$ として、記事のタイトルおよび本文の形態素と、入稿時に付与されるメタデータを用いた。入稿時に付与されるメタデータとしては、記事種別、カテゴリ、入稿元、著者、画像の有無などが含まれる。また、動的特徴量 $\mathbf{x}_t$ として、過去の実績情報 $\{y_s\}_{s=1}^{t-2}$ と、時刻 t が表す時間の曜日と24時間の中の何時かを表す時刻情報、その記事が入稿されてから経過した時間を用いた。ここで、実システムでは閲覧数の集計結果を得るまでに時間がかかるという制約を加味して、アクセス可能な実績閲覧数を2時間前の y_{t-2} までに制限している。この制約のため、時刻 $t \leq 2$ の予測では実績閲覧数を用いることができず、静的特徴量 $\mathbf{a}$ と時刻 t の時刻情報のみから閲覧数の予測を行うことになる。

評価指標として、ピアソンの相関係数と平均二乗誤差 (MSE; Mean Square Error) を用いた。

$$r = \frac{\sum_i^N \sum_{t=1}^{24} (y_t^{(i)} - \bar{y})(p_t^{(i)} - \bar{p})}{\sqrt{\sum_i^N \sum_{t=1}^{24} (y_t^{(i)} - \bar{y})^2} \sqrt{\sum_i^N \sum_{t=1}^{24} (p_t^{(i)} - \bar{p})^2}}$$

$$\text{MSE} = \frac{1}{24N} \sum_i^N \sum_{t=1}^{24} (y_t^{(i)} - p_t^{(i)})^2$$

ただし、 $\bar{p} = \sum_i^N \sum_{t=1}^{24} p_t^{(i)} / 24N$ 、 $\bar{y} = \sum_i^N \sum_{t=1}^{24} y_t^{(i)} / 24N$ はそれぞれの平均値である。ピアソンの相関係数は値が大きい方が、平均二乗誤差は値が小さい方が、精度良く予測ができているといえる。

実験では、RNNを用いた提案手法と、いくつかの手法を比較した。単純に、時刻 t においてアクセス可能な直近の実績値 y_{t-2} を、予測値 p_t とする手法を *Pre-2* と表す。時系列を考慮せずに単純な回帰の問題として捉え、線形SVR (Support Vector Regression) [5] で予測するアプローチをベースラインとした。*SVR1* は、実績閲覧数を用いずに、静的特徴量と時刻情報、経過時間のみを用いて予測を行う手法である。*SVR2* は、*SVR1* で用いた特徴量に、直近の実績値 y_{t-2} を加えて予測を行うモデルである。*SVR3* は、*SVR1* で用いた特徴量に、アクセス可能な過去の全ての実績閲覧数 $\{y_s\}_{s=1}^{t-2}$ と、どの時間の閲覧数が利用可能かを表

現する欠損フラグを加えて、予測を行う。提案手法を *RNN* と表す。なお、式2の $\mathbf{h}_{t-1}$ を全て $\mathbf{h}_0$ で置き換え、時系列の依存性をなくしたものを比較手法として用い、*NN* と表す。

実験結果を表1に示す。実績閲覧数を用いることができない場合 ($t \leq 2$) と、できる場合 ($t \geq 3$) のそれぞれの結果についても示す。ただし、実績値を予測時に必ず用いる *Pre-2* と *SVR2* は、実績値ありの場合の結果のみ示している。各列ごとに、一番優れている手法に対応する値を太字で記した。

結果を見ると、提案手法である *RNN* がそれぞれの場合において、各指標で優れている結果となった。各手法で、実績値なしと実績値ありの二つの場合を比較すると、実績値ありの方が相関係数と *MSE* の両方において良い傾向が見られた。しかし、*SVR1* は実績閲覧数にアクセス可能である場合 ($t \geq 3$) においても、この情報を予測のための特徴量として用いない手法であるため、奮わない結果となった。このことから、実績閲覧数は強い特徴量だということが分かる。実績値ありの場合においても、提案手法の *RNN* は、線形SVRを用いた手法である *SVR2* よりも良い性能を示した。また、同じく実績値ありの場合において、以前の時刻の隠れ層からの情報の伝播をなくした *NN* も *SVR2* より優れた結果を示し、*RNN* に迫る性能であった。このことから、ニューラルネットの非線形性が記事閲覧数の予測に寄与していると考えられる。一方、実績値なしの場合で比較すると、*NN* は *RNN* に引き離される結果となった。提案手法は、*RNN* を用いることで誤差を時系列上で逆向きに伝播させることにより、実績値がない場合においても記事の静的な情報から閲覧数を予測できるように学習を行うことが可能になったのだと考えられる。

上記の実験結果を受けて、提案手法を実システムに導入し、予測された記事の閲覧数を元に記事の並びを変更した。図3に、閲覧数の推定対象になった記事に対する、オンライン環境におけるセッションCTR (= クリックされた記事数 / セッション数) の推移を示す。なお、ビジネス上の理由からセッションCTRは実数値ではなく、サービスリリース当初の値を基準とした比率で示す。提案手法により、サービスリリース当初と比較してセッションCTRが2倍以上になったことが図からわかる。

5. おわりに

本稿では、スマートフォン版Yahoo! JAPANトップページを具体的な事例として、タイムライン形式でユーザに記事を配信する際の課題について述べた。

課題は複数存在するが、そのうちの記事の並びの決定について重要な要素である記事閲覧数についてRNNを用いて、閲覧実績や記事内容から推定する手法を提案した。提案した手法について、オフライン、オンラインのそれぞれでの評価を実施し、性能の向上を確認した。

本論文では閲覧数の推定に焦点を当てたが、先に述べたとおり、適切な広告の選択や記事の重複削除、また、推定された記事閲覧数に加えさまざまな要素を考慮した記事の並び替えの方法の検討など、多くの課題が存在している。これらについても引き続き研究を行っていく予定である。

【謝辞】

本稿の作成にあたり、日々Yahoo! JAPANの各サービスを利用し、フィードバックをくださっている、全てのユーザの皆様に感謝いたします。

【文献】

- [1] Michael Auli, Michel Galley, Chris Quirk, and Geoffrey Zweig. Joint language and translation modeling with recurrent neural networks. In *EMNLP*, 2013.
- [2] Henriette Cramer. Effects of ad quality & content-relevance on perceived content quality. In *Proceedings*

表 1: オフラインでの実験結果.

手法	全て		実績値なし ($t \leq 2$)		実績値あり ($t \geq 3$)	
	相関	MSE	相関	MSE	相関	MSE
<i>Pre-2</i>	-	-	-	-	0.9468	0.7518
<i>SVR1</i>	0.7870	2.3879	0.7882	3.6448	0.7756	2.2736
<i>SVR2</i>	-	-	-	-	0.9574	0.4653
<i>SVR3</i>	0.8952	1.2389	0.7988	4.8902	0.9152	0.9069
<i>NN</i>	0.9523	0.5875	0.8426	2.4642	0.9630	0.4169
<i>RNN</i>	0.9549	0.5564	0.8644	2.1235	0.9631	0.4139

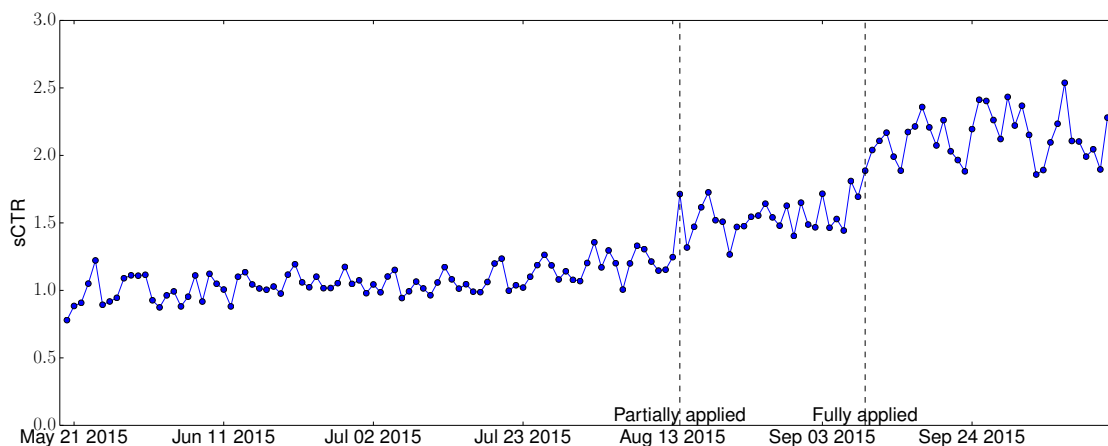


図 3: オンライン環境におけるセッション CTR の推移. ビジネス上の理由からセッション CTR は実数値ではなく比率で示している.

- of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015.
- [3] Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *JAsIs*, 1990.
- [4] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2009.
- [5] Chia-Hua Ho and Chih-Jen Lin. Large-scale linear support vector regression. *The Journal of Machine Learning Research*, 2012.
- [6] Stefan Kombrink, Tomas Mikolov, Martin Karafiát, and Lukáš Burget. Recurrent neural network based language modeling in meeting recognition. In *INTER-SPEECH*, 2011.
- [7] Mounia Lalmas, Janette Lehmann, Guy Shaked, Fabrizio Silvestri, and Gabriele Tolomei. Promoting positive post-click experience for in-stream yahoo gemini users. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.
- [8] Cheng Li, Yue Lu, Qiaozhu Mei, Dong Wang, and Sandeep Pandey. Click-through prediction for advertising in twitter timeline. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.
- [9] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTER-SPEECH*, 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010, 2010.
- [10] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Honza Černocký, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011.
- [11] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010.
- [12] Shumpei Okura, Yukihiro Tagami, and Akira Tajima. Article de-duplication using distributed representations. In *Proceedings of the 25th International Conference on World Wide Web, WWW '16 Companion*, 2016.
- [13] Yuyu Zhang, Hanjun Dai, Chang Xu, Jun Feng, Taifeng Wang, Jiang Bian, Bin Wang, and Tie-Yan Liu. Sequential click prediction for sponsored search with recurrent neural networks. *arXiv preprint arXiv:1404.5772*, 2014.

大倉 俊平 Shumpei OKURA

ヤフー株式会社 データ&サイエンスソリューション統括本部

田頭 幸浩 Yukihiro TAGAMI

ヤフー株式会社 データ&サイエンスソリューション統括本部

小野 真吾 Singo ONO

ヤフー株式会社 データ&サイエンスソリューション統括本部

田島 玲 Akira TAJIMA

ヤフー株式会社 データ&サイエンスソリューション統括本部