Generating a Pairwise Dataset for Click-through Rate Prediction of News Articles Considering Positions and Contents

Shotaro Ishihara
Nikkei, Inc.
Tokyo, Japan
shotaro.ishihara@nex.nikkei.com

Yasufumi Nakama
Independent Researcher
Tokyo, Japan

ABSTRACT

In online news websites, the headlines and thumbnail images of articles are displayed in a list, and they are important navigation links to individual article pages. If we can predict the click-through rate (CTR) of readers to the article pages, we can assist the editors in creating article headlines and setting thumbnail images. However, the CTR that can be observed in the access log is heavily affected by the display position, and it is difficult to predict the CTR by machine learning using data on single articles. This paper proposes a method to construct a pairwise dataset based on the information such as similarity of the display positions and contents, and create a model to predict the CTR in the framework of pairwise learning-to-rank. In the experiment, we verified the usefulness of the proposed method by using the actual access log data and discuss the potential of the practical use of the CTR prediction as editing support.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

Pairwise learning-to-rank, Click-through rate prediction, Randomized controlled trial, Position bias, Multimodal

ACM Reference Format:
Shotaro Ishihara and Yasufumi Nakama. 2022. Generating a Pairwise Dataset for Click-through Rate Prediction of News Articles Considering Positions and Contents. In *Proceedings of Computation + Journalism Conference 2022*. ACM, New York, NY, USA, 5 pages.

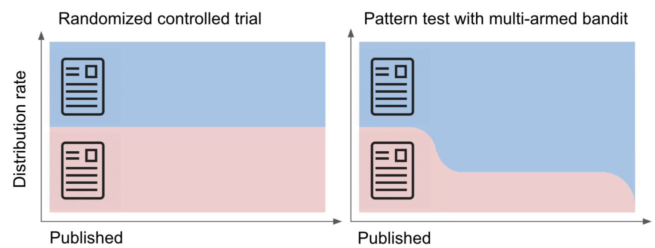

Figure 1: Methods to measure the quality of headlines and thumbnail images include randomized controlled trials and the pattern tests. Randomized controlled trials provide every option at the same rate, and the pattern tests have an advantage of adjusting the distribution rate through the experiments. Though there are practical difficulties in the delivery of multiple options, and predicting the CTR before publishing would be helpful as editing support.

1 INTRODUCTION

In online news services, article headlines and thumbnail images play an important role as leads to individual article pages. The top page of many news services displays a list of articles, and individual article pages often provide guidance on related articles. Readers also decide whether to move on to the article page based on the information displayed in the external inflow, such as social networking services and browser searches.

One method for measuring the quality of headlines and thumbnail images is to provide multiple options at the same time and compare the response. For example, the Nikkei Online Edition [1] provided by Nikkei, Inc. uses the system called "pattern test" [18]. The pattern tests provide users with multiple headlines and thumbnail images at the same time. Characteristically, the system displays more options with a higher click-through rate (CTR) in line with the response, and the comparison process is terminated when statistically significant differences are found. Multi-armed bandit [10] is used in the pattern tests and has the advantage of reducing the proportion of options with low CTR compared to a simple randomized controlled trial [3], as shown in Figure 1.

However, there are practical difficulties in the delivery of multiple options [6, 17]. Firstly, in the news media, there are situations where it is desirable to present uniform information to all readers for news of high importance. In addition, the possibility that low-quality options may negatively affect the user experience during experiments must be taken into account. The workload of editors would be increased in terms of the need to produce several candidates of sufficiently high quality to present to the readers.

Therefore, this research aims to construct a model that can predict the CTR of two alternatives and can estimate the results of the pattern tests. If the CTR of the article page can be predicted, it is possible to carry out trial-and-error before actually publishing it to the readers, and to support the editorial work such as creating headlines and setting thumbnail images. Offline evaluation also has the advantage of reducing the effect on users compared to randomized controlled trials and the pattern tests.

Computation + Journalism Conference 2022, June 09–11, 2022, New York, NY, USA

[1] https://www.nikkei.com/

One of the approaches for achieving this is to create a machine learning model using the results of past pattern tests, but it is difficult to prepare a sufficient dataset for training due to the the practical difficulties as mentioned. Hence, this research investigates a method for constructing a training dataset that imitates the results of the pattern tests, using data on the CTR of individual articles that can be observed in daily service operations.

Since the CTR of individual articles is generally greatly influenced by the display positions of the articles [8], the training dataset need to be constructed carefully. If the raw CTR data is simply used as a training dataset, there is a concern that a prediction model would be created that focuses with more importance on the display position than on the information of the article itself including the headline and thumbnail image.

To address this problem, this research constructs a pairwise dataset using the similarity of display positions and contents, and builds a model with learning-to-rank framework: focusing more on contents information by learning to compare the two pairs of articles. The rest of this paper introduces the related works and describes the proposed method. We also verify the usefulness of the proposed method through experiments using the access log data and the actual pattern test results of the Nikkei Online Edition. Finally, we present the possibilities for using CTR prediction models as editing support.

2 RELATED WORKS

This section introduces related researches from three perspectives: (1) CTR prediction, (2) position bias, and (3) editing support. Finally, we describe the position of our research.

2.1 CTR prediction

There are many previous studies on the CTR prediction. In particular, there has been a lot of development in the area of online advertising [25]. Models that achieve high performance have been reported one after another in recent years, in line with the rapid progress of deep learning [26]. The types of datasets used are diversifying, such as text and images, and methods for handling multimodal datasets simultaneously are investigated [13].

2.2 Position bias

In machine learning research, not only improving prediction performance, but also dealing with potential bias in the observed data has attracted attention in recent years. The field of CTR prediction frequently discusses the impact of display position sometimes called position bias. Some assume that clicks occur independently of external factors such as display position [12], but a lot of models consider display position [11, 15, 21].

As one of the methods for constructing a model to predict the CTR from data including the influence of position bias, the framework of learning-to-rank has been studied [9, 23]. Three types of learning methods exist: (1) `pointwise`, where losses are calculated from individual data; (2) `pairwise`, where losses are calculated from two pairs of data; and (3) `listwise`, where losses are calculated from more than one pair.

2.3 Editing support

An example of a study on CTR prediction in the context of editing support is the study [17] on the news platform, Yahoo! News [2]. This study estimates the CTR for each of the two headline candidates and predicts the results of randomized controlled trials. Predicting with confidence using natural gradient boosting [5] enhances its usefulness as editing support. Other example of editing support is automatic heading generation [16, 24], and the use of information technology in the newsroom is in progress.

2.4 Position of our research

In line with recent trends, our research tackles the CTR prediction task using deep learning with multimodal datasets like article headlines and thumbnail images as input. It is similar to Nakamura et al. [17] in that it aims to editing support and follows their experimental settings. The major difference is that the Nikkei Online Edition requires more scrolling to browse the list of articles than Yahoo! News, and the effect of position bias is expected to be more significant. Therefore, in this research, we construct a dataset while considering the information of position, and build a model for predicting the CTR in the framework of pairwise learning-to-rank. Prospects include collaboration with automatic headline generation.

3 PROPOSED METHOD

An overview of the proposed method is shown in Figure 2. First, a training dataset is constructed using the CTR of individual articles based on the similarity of the contents while taking into account the influence of position bias. Next, a model for predicting the CTR is built in the framework of pairwise learning-to-rank, deep learning models are trained through the task of which article has a higher CTR. In this section, we explain the proposed method in two stages: (1) construction of the training dataset, and (2) pairwise learning-to-rank.

3.1 Construction of training dataset

A pairwise dataset is constructed in the following steps:

(1) Assign cluster numbers to all articles by clustering.
(2) Create candidate sets of articles based on display position and a cluster number.
(3) Extract every two pairs of articles in each set.

First, each article is vectorized to measure the similarity of the contents, and a cluster number is assigned to each article by clustering. In this research, the headlines of each article are vectorized using `TF-IDF` [19] and clustering is done by `k-means++` [1]. Here, cluster size, the number of clusters, is a hyperparameter. Since thumbnail images tend to have lower similarity than headlines in the pattern tests, each article is vectorized using the headline, even when extracting pairs related to the thumbnail images.

Next, the display position and cluster number are used to narrow down the candidate set of articles to be extracted as pairs. Specifically, all articles are divided into sets based on the display position and cluster number, and only sets with more than two articles and less than maximum set size are considered as candidates. The reason maximum set size is set as a hyperparameter here is that the

[2]https://news.yahoo.co.jp/

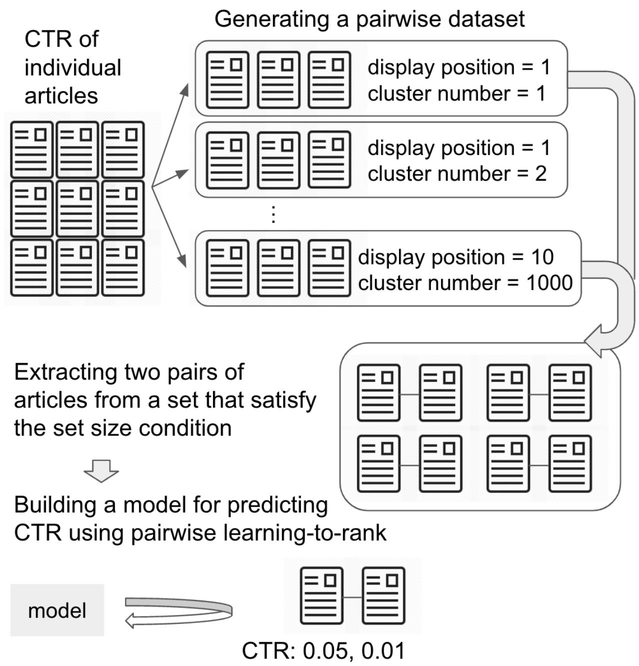

Figure 2: Overview of the proposed method. First every articles are assigned the cluster number considering the headline information. Secondly, candidate sets are created based on the display position and the cluster number. By extracting every two pairs of articles in each set, a pairwise dataset are generated. Finally, we can build a model for predicting CTR using pairwise learning-to-rank.

existence of a set with an excessively large number of articles would reduce the diversity of the training dataset as a whole. Finally, every two pairs of articles are extracted by allowing overlaps within each set and used as the training dataset.

3.2 Pairwise learning-to-rank

To tackle the task of estimating which article has a higher CTR, Margin Ranking Loss [3] is used as a loss function during model training.

$$L(x, y) = \max(0, -y(x_1 - x_2) + \text{margin})$$

Here, x contains the two inputs (x_1, x_2) and y takes 1 if $x_1 > x_2$ and -1 if $x_1 \leq x_2$. With Margin Ranking Loss, the model is trained so that one of the inputs is larger than the other by at least some `margin`, which is a hyperparameter.

4 EXPERIMENTS

In the experiment, the CTR prediction model is constructed from the access log data of the Nikkei Online Edition, and the usefulness of the proposed method is verified using the actual pattern test results.

[3] https://pytorch.org/docs/stable/generated/torch.nn.MarginRankingLoss.html

Table 1: Comparison of the information each dataset contains. ○ means always included, △ means possible missing, - means not included. The number of articles in `PairwiseCTR` depends on the hyperparameters.

	`SingleCTR`	`PatternCTR`	`PairwiseCTR`
headline	○	○	○
thumbnail image	△	△	△
published date time	○	○	○
position (1-19)	○	○	○
CTR	○	○	○
pattern test index	-	○	○
CTR (bigger or smaller)	-	○	○
the number of articles	19842	92	*

4.1 Dataset and evaluation metric

Referring to Nakamura et al. [17], we use the CTR of individual articles for training and the pattern test results for evaluation. The former, `SingleCTR`, contains 19842 articles published in the Nikkei Online Edition between 1 September 2020 and 31 August 2021. Note that the articles of the pattern tests are eliminated. The latter, `PatternCTR`, is the pattern test results conducted between September 2020 and September 2021. 25 pairs (50 articles) were targeted for headlines and 21 pairs (42 articles) for thumbnail images. `SingleCTR` is divided into two parts, one before 31 July 2021 for training and one after 1 August 2021 for validation. The dataset constructed by the proposed method from `SingleCTR` is denoted as `PairwiseCTR`.

Each dataset contains information on the headline, thumbnail image, published date time, display position, and the CTR, as shown in Table 1. In addition, `PatternCTR` and `PairwiseCTR` have pattern test index to link the articles for comparison and the results of the pattern test (which CTR is bigger). Considering the research objective, we set the accuracy of the pattern test results in `PatternCTR` as the evaluation metric.

4.2 Models

To verify the usefulness of the proposed method, we check the performance of the model shown in Table 2.

Baseline
First, to verify the existence of position bias, we build a model simply using headlines and thumbnail images in `SingleCTR`.

Baseline + display position + published date time
Next, by including information on display position and published date time as input, we build models that takes into account influences other than headlines and thumbnail images in `SingleCTR`.

Baseline + fixed CTR
To handle position bias, we also consider a way to correct the CTR of the training dataset dividing by the average CTR for each display position in `SingleCTR`.

Proposed method
CTR prediction model by the proposed method.

In the proposed method, the performance is evaluated with 5×6 combinations, where `margin` is fixed at 0.5, maximum set size is {10, 20, 30, 40, 50} and the cluster size is {1000, 1500, 2000, 2500,

Table 2: Performance on the evaluation dataset named PatternCTR. The proposed method showed particularly high performance for thumbnail images.

dataset	model	accuracy (headlines)	accuracy (thumbnail images)
SingleCTR	Baseline (with headline and thumbnail image)	0.560 (14/25)	0.524 (11/21)
SingleCTR	Baseline + display position	**0.640** (16/25)	0.429 (9/21)
SingleCTR	Baseline + published date time	0.560 (14/25)	0.524 (11/21)
SingleCTR	Baseline + display position + published date time	**0.640** (16/25)	0.333 (7/21)
SingleCTR	Baseline + display position + published date time + fixed CTR	0.600 (15/25)	0.524 (11/21)
PairwiseCTR	Proposed method	mean 0.590, std 0.074	mean **0.531**, std 0.100

Table 3: Effect of hyperparameters in the proposed method. The left represents accuracy for headlines and the right for thumbnail images. It is confirmed that the proposed method is sensitive to the hyperparameters.

maximum set size / cluster size	1000	1500	2000	2500	3000	3500
10	0.600 / 0.429	0.440 / 0.619	0.520 / 0.476	0.640 / 0.667	0.560 / 0.667	0.600 / 0.429
20	**0.720** / 0.476	0.480 / 0.619	0.560 / 0.524	0.640 / 0.571	0.600 / 0.571	0.520 / 0.524
30	0.600 / 0.476	0.640 / 0.476	0.560 / 0.476	**0.720** / 0.381	0.560 / **0.857**	0.480 / 0.429
40	0.600 / 0.667	0.560 / 0.479	0.640 / 0.619	0.640 / 0.429	0.680 / 0.524	0.680 / 0.476
50	0.560 / 0.476	0.520 / 0.524	0.680 / 0.524	0.680 / 0.476	0.560 / 0.476	0.480 / 0.619

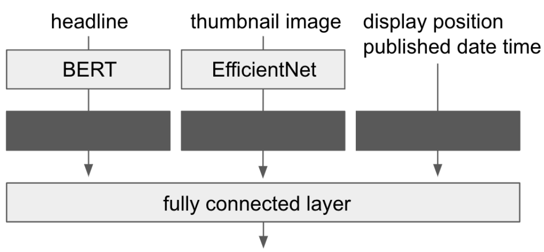

Figure 3: Neural network architecture diagram. BERT is for headlines, EfficientNet is for thumbnail images, and the outputs are concatenated with display position and published date time.

3000, 3500}. For all models, we use a neural network illustrated in Figure 3, including BERT [4] for headlines and EfficientNet [22] for thumbnail images. BERT model was pretrained with about six years of articles of the Nikkei Online Edition from 21 September 2013 to 20 September 2019, and EfficientNet model is `tf_efficientnet_b0_ns` from PyTorch Image Models [4].

4.3 Results

The performance on the evaluation dataset named PatternCTR is shown in Table 2.

Baseline
Model that simply used headlines and thumbnail images in SingleCTR had accuracy of around 0.5 and relatively low performance for PatternCTR, suggesting the existence of position bias.

Baseline + display position + published date time
Models that included information on display position and published date time as input showed improvement for headlines, while no clear performance improvement could be confirmed for thumbnail images.

Baseline + fixed CTR
Fixing the CTR by display position did not contribute to the performance.

Proposed method
The proposed method showed particularly high performance for thumbnail images. There was also a certain improvement for headlines compared to the baseline, in some cases obtaining results as good as 0.720.

The results for different values of maximum set size and cluster size are shown in Table 3. It can be seen that changing the hyperparameters results in a large variation. It is confirmed that the proposed method is sensitive to the hyperparameters, although this is partly due to the fact that the size of PatternCTR is not large.

5 USE CASE OF HEADLINE GENERATION

The CTR prediction model is also considered for application to automatic headline generation. One way to utilize automatic headline generation as editing support is to enumerate headline candidates and the human check and edit them if necessary [7, 16]. Within this framework, the editor's decision-making can be assisted if a prediction of the CTR is included along with the candidates as shown in Figure 4. The metric called ROUGE [14] is often used for evaluating generated headlines, but the predicted CTR should also be available as for one perspective. We can also present a visualization of the weights of BERT attention mechanism in the case of headlines as supplemental information like Figure 5. The same can be done by Grad-CAM [20] in the case of thumbnail images.

On the other hand, when using the CTR prediction models in news media, it is necessary to be aware of the clickbait issues [2]. Even if the CTR is high, headlines and thumbnail images that do not match the body text would damage the user experience.

[4] https://github.com/rwightman/pytorch-image-models

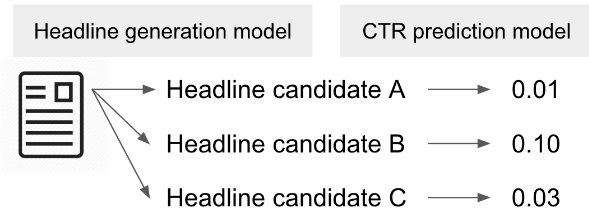

Figure 4: Workflow in automatic headline generation including headline generation model and CTR prediction model.

Figure 5: Visualization of the weights of BERT attention mechanism. Tokens with higher values are painted in denser colors.

6 CONCLUSIONS AND FUTURE WORK

This research proposed a method to generate a pairwise dataset for creating the CTR prediction model in the framework of pairwise learning-to-rank considering position bias. The experiment reported the better performance potential of the proposed method, and the practical use as editing support was explained. The future work is to expand the evaluation dataset. By accumulating the results of the daily pattern tests, larger-scale performance evaluation and the search for appropriate hyperparameters can be achieved.

ACKNOWLEDGMENTS

We would like to thank many people involved in Nikkei, Inc.

REFERENCES

[1] David Arthur and Sergei Vassilvitskii. 2006. k-means++: the advantages of careful seeding. In *Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms*. Stanford, USA, 1027–1035.

[2] Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. 2017. From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. Association for Computational Linguistics, Copenhagen, Denmark, 84–89.

[3] T C Chalmers, H Smith, Jr, B Blackburn, B Silverman, B Schroeder, D Reitman, and A Ambroz. 1981. A method for assessing the quality of a randomized control trial. *Control. Clin. Trials* 2, 1 (May 1981), 31–49.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[5] Tony Duan, Avati Anand, Daisy Yi Ding, Khanh K Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. 2020. NGBoost: Natural Gradient Boosting for Probabilistic Prediction. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119. PMLR, Virtual Event, 2690–2700.

[6] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline A/B Testing for Recommender Systems. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, New York, NY, USA, 198–206.

[7] Shotaro Ishihara, Yuta Matsuda, and Norihiko Sawa. 2021. Editors-in-the-loop News Article Summarization Framework with Sentence Selection and Compression. In *Proceedings of the fifth IEEE Workshop on Human-in-the-Loop Methods and Future of Work in BigData*. IEEE Computer Society, Los Alamitos, CA, USA, 3522–3524.

[8] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. Association for Computing Machinery, New York, NY, USA, 154–161.

[9] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased Learning-to-Rank with Biased Feedback. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, New York, NY, USA, 781–789.

[10] Michael N Katehakis and Arthur F Veinott. 1987. The Multi-Armed Bandit Problem: Decomposition and Computation. *Mathematics of OR* 12, 2 (May 1987), 262–268.

[11] Haruka Kiyohara, Yuta Saito, Tatsuya Matsuhiro, Yusuke Narita, Nobuyuki Shimizu, and Yasuo Yamamoto. 2022. Doubly Robust Off-Policy Evaluation for Ranking Policies under the Cascade Behavior Model. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, New York, NY, USA, 487–497.

[12] Shuai Li, Yasin Abbasi-Yadkori, Branislav Kveton, S Muthukrishnan, Vishwa Vinay, and Zheng Wen. 2018. Offline Evaluation of Ranking Policies with Click Models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, New York, NY, USA, 1685–1694.

[13] Xiang Li, Chao Wang, Jiwei Tan, Xiaoyi Zeng, Dan Ou, Dan Ou, and Bo Zheng. 2020. Adversarial Multimodal Representation Learning for Click-Through Rate Prediction. In *Proceedings of the Web Conference 2020*. Association for Computing Machinery, New York, NY, USA, 827–836.

[14] Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. Association for Computational Linguistics, USA, 71–78.

[15] James McInerney, Brian Brost, Praveen Chandar, and Benjamin Carterette. 2020. Counterfactual Evaluation of Slate Recommendations with Sequential Reward Interactions. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 1779–1788.

[16] Kazuma Murao, Ken Kobayashi, Hayato Kobayashi, Taichi Yatsuka, Takeshi Masuyama, Tatsuru Higurashi, and Yoshimune Tabuchi. 2019. A Case Study on Neural Headline Generation for Editing Support. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 73–82.

[17] Yuki Nakamura, Tomohide Shibata, Hayato Kobayashi, Nobuyuki Shimizu, and Hiroaki Taguchi. 2020. Click-Through Rate Prediction with Confidence for News Headlines Using Natural Gradient Boosting. In *Proceedings of the Annual Conference of JSAI*, Vol. JSAI2020. 4M2GS503–4M2GS503.

[18] Sawa Norihiko. 2020. Test headlines on News Media by Multi-Armed Bandit. In *Computation + Journalism Symposium 2020*.

[19] Juan Enrique Ramos. 2003. Using TF-IDF to Determine Word Relevance in Document Queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. 29–48.

[20] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int. J. Comput. Vis.* 128, 2 (Feb. 2020), 336–359.

[21] Alexander L Strehl, John Langford, Lihong Li, and Sham M Kakade. 2010. Learning from logged implicit exploration data. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*. Curran Associates Inc., Red Hook, NY, USA, 2217–2225.

[22] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*. proceedings.mlr.press, 6105–6114.

[23] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. 2018. Position Bias Estimation for Unbiased Learning to Rank in Personal Search. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, New York, NY, USA, 610–618.

[24] Kosuke Yamada, Yuta Hitomi, Hideaki Tamori, Ryohei Sasano, Naoaki Okazaki, Kentaro Inui, and Koichi Takeda. 2021. Transformer-based Lexically Constrained Headline Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 4085–4090.

[25] Yanwu Yang and Panyu Zhai. 2022. Click-through rate prediction in online advertising: A literature review. *Inf. Process. Manag.* 59, 2 (March 2022), 102853.

[26] Weinan Zhang, Jiarui Qin, Wei Guo, Ruiming Tang, and Xiuqiang He. 2021. Deep learning for click-through rate estimation. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, California, 4695–4703.