Available online at www.sciencedirect.com
ScienceDirect
ICT Express 6 (2020) 175–180
www.elsevier.com/locate/icte

Identifying machine learning techniques for classification of target advertising

Jin-A Choi[a], Kiho Lim[b,*]

[a] Department of Communication, William Paterson University of New Jersey, Wayne, NJ 07470, USA
[b] Department of Computer Science, William Paterson University of New Jersey, Wayne, NJ 07470, USA

Received 1 March 2020; received in revised form 24 April 2020; accepted 27 April 2020
Available online 2 June 2020

Abstract

There have been numerous applications of artificial intelligence (AI) technologies to online advertising, especially to optimize the reach of target audiences. Previous studies show that improved computational power significantly advances granular audience targeting capabilities. This study investigates and classifies various machine learning techniques that are used to enhance targeted online advertising. Twenty-three machine learning-based online targeted advertising strategies are identified and classified largely into two categories, user-centric and content-centric approaches. The paper also identifies an underexamined area, algorithm-based detection of click frauds, to illustrate how machine learning approaches can be integrated to preserve the viability of online advertising.

Keywords: Machine learning; Artificial intelligence; Target advertising; Classification

Contents

1. Introduction

Advertising practices and strategies have been researched extensively to enhance the effectiveness of advertisements across various target audiences [1]. Artificial intelligence technologies create a competitive advantage for online advertising,

* Corresponding author.
E-mail addresses: choij21@wpunj.edu (J.A. Choi), limk2@wpunj.edu (K. Lim).
Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2020.04.012

over traditional practices, by providing improved computational power to advance the optimization of digital advertisements. Machine learning (ML) based techniques enhance the accuracy of targeting by predicting the most relevant advertisements for users based on contextual or pre-existing user data [2]. Such innovations in AI and data-driven approaches help to mitigate challenges faced by advertisers while significantly enhancing user experience. Various studies outline machine learning algorithms that optimize the delivery of targeted advertisements [3]. However, to the best of the authors' knowledge, principal machine learning techniques that are used to optimize targeting and advertising effectiveness have not been classified. Thus, this study investigates and categorizes various applications of machine learning techniques largely into two categories, user-centric and content-centric approaches to targeted advertising. Additionally, machine learning approaches for detecting click fraud are identified. The detailed classification of machine learning techniques is illustrated in Fig. 1.

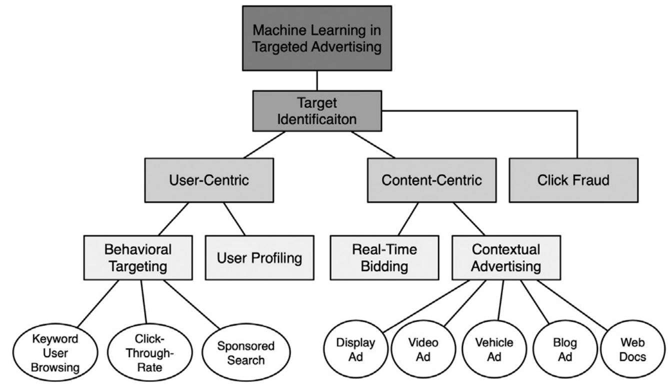

Fig. 1. Classification of target advertising strategies.

2. Target advertising classification

2.1. Target identification

The ability to accurately predict specific target audience members in an ever-cluttered digital environment is a challenge that is addressed by machine learning. Target advertising aims to present the most relevant advertising messages to consumers, [2] and machine learning-based approaches allow for the automation and optimization of processes for potential consumer identification, information extraction, and market segmentation. Applications of machine learning to user and content-based approaches hold advantages over traditional market segmentation as contents consumed and shared by individuals are more important in predicting target audiences and their purchasing behavior than demographic and geographical data alone [4]. For example, textual features of user-generated content on various social media platforms, such as Twitter, can be used to predict and classify target audiences with high accuracy [5]. Additionally, improvements in personalization and reduction in intrusiveness of advertising messages help to improve retention of customers, maximize marketing efficiencies, and improve the return on investments (ROI) [4].

2.2. User-centric approach

2.2.1. Behavioral targeting

One way in which targeted advertisements reach desired consumers is through behavioral targeting (BT). In order to select the most relevant advertisements for consumers, BT relies on historical user behavior, such as identifying clicked links, pages visited, searches, and past purchases from the user's browsing history [2]. With the popularity of search engines, such as Google, online searches and web browsing have become two of the most common online behaviors. Web browsing behavior helps advertisers make inferences regarding users' interests and to define audience segments. Leveraging users' actual online behavior strengthens relevance and personalization of advertising messages to desired consumers [6]. Additionally, the user's search queries also help to determine which advertisements should be displayed to the user by matching them to the advertiser's keywords.

Predicting the click behavior of users is another method to enhance targeted advertising. Cost-per-click (CPC) model, where advertisers pay when users click on an advertisement, is often used as a common pricing method for online advertisements. Therefore, predicting the probability that a user will click on an advertisement, click-through rate (CTR), is unmistakably important [7]. CTR predicts the expected revenue for each displayed advertisement as well as ranking, filtering, and the placement of advertisements [8].

Sponsored search (SS) is the placement of textual advertisements based on the match between the user's search query and keywords identified by the advertiser [9]. For example, when users search for keywords purchased by the advertiser, the corresponding advertisement is then displayed. Optimizing the automated selection of advertisements helps to improve user experience [10].

2.2.2. User profiling

User profiling, a behavior-based approach, is a recommender system that discovers useful patterns in a user's behavior to determine what the user finds interesting and uninteresting [10]. Identification of user-interests is essential to suggest customized advertisements to users according to their preferences. As the users indicate an assortment of interests, explicitly (user-provided information) or implicitly (past online searches, reviews, browsing), they can be profiled in terms of attributes and predefined categories [11]. The ability to differentiate between users through behavioral targeting is important as personalization of messages is crucial in enhancing user experience.

2.3. Content-centric approach

2.3.1. Contextual advertising

Contextual advertising is another important branch in advertising which describes the placement of advertisements on a third-party web page that matches in content [12]. Delivering advertisements that are relevant to consumers is imperative as 90% of consumers find personalized advertisements to be appealing, and 95% of companies using personalization strategies saw three times ROI on their ad spend [13]. Web pages are analyzed in real-time and keywords are extracted per user visit to deliver advertisements that closely match the contents of the web page [14]. When advertisements are relevant to the contents of the web page, it not only enhances user experience but also increases probability of clicks and revenue [8]. Furthermore, machine learning approaches also address the issue of advertisers mistakenly pairing advertisements to web pages with sensitive or negative content, such as violence, war, etc., by detecting mixed-content pages [12]. As illustrated in Fig. 1, there are other subfields in targeted advertising that utilize content-centric approaches:

- **Display ads** are graphical online advertisements. Targeting effects are enhanced by the appropriate placement of display advertisements on relevant web pages [15].
- **Video ads** aim to avoid intrusiveness and irrelevance to target users by selecting advertisements that meet their interests at the right time and place in the video, and are congruent to the products shown in the video [16].
- **Blogs** are assigned to advertisements according to the interests and opinions expressed by the blogger. As blogs are likely to reflect the interests of its visitors, strong relevance between the content of the blog and advertisements creates a perception of personalization, which in turn increases the number of CTR [17].
- **Web Docs** present a wide variety of contents on a single web page. Web docs are examined for multi-topics to avoid mismatching of advertisements to sensitive contents [12].
- **Vehicle ads** utilize navigation module and context recognition module to enhance context-aware advertisement delivery. Specifically, vehicular network systems collect data generated by on-board sensors and process information to provide targeted advertising [18,19].

2.3.2. *Real-time bidding*

Real-time bidding (RTB) through machine learning allows instantaneous decisions to be made on whether to show a particular advertisement to a specific user based on insights gained from large data sets containing information regarding past impressions, clicks, searches, and purchases [20]. As RTB transactions are made in real-time, devising an optimal bidding algorithm significantly improves the advertiser's ability to strategically bid for impressions just as it is generated by a user's visit [21].

2.4. *Click fraud*

Fraudulent online activities, such as click fraud, not only deplete ad budgets dishonestly, but it damages the integrity of the online advertising industry. Click fraud occurs when illegitimate clicks, in which the user does not have a genuine interest in the advertisement, are generated to gain dishonest commission [22]. False clicks can be simulated on advertisements using manual or automated techniques, such as automated clicking tools or bots that imitate a legitimate user [23]. Click fraud can damage an advertiser's return on investment significantly. As publishers are often paid by PPC (pay-per-click), the number of clicks an advertisement receives, inflation of clicks or impressions may be falsely generated. It was found that 30% of ad revenue is wasted on click frauds [24]. Therefore, it is crucial to implement click fraud detection and prevention systems to preserve trust between advertisers and publishers, and to ensure the legitimacy of each click made [25]. Therefore, machine learning-based click fraud detection algorithms are essential tools to identify and protect advertisers from fraudulent behaviors [23].

3. Machine learning techniques for advertising

In this section, various machine learning techniques that enhance targeted online advertising are investigated and classified into two categories, user-centric and content-centric approaches. Additionally, several machine learning-based approaches that detect and/or prevent click fraud are explored.

3.1. *User-centric advertising*

Targeted online advertising can be enhanced by keyword extraction. An approach for extracting keywords from online broadcasting contents [6] finds meaningful language patterns from a content corpus, and the patterns mined are used to extract keywords from online broadcasting contents. In order to produce a set of candidate language patterns, a sequential pattern mining (SPM) algorithm is applied in the language pattern mining step. The discovered meaningful language patterns are used to help keywords extraction in real time.

To deal with CTR prediction, Jiang et al. [14] proposed a deep architecture model, called DBNLR, which combines deep belief network (DBN) and logistical regression (LR). In this model, DBN is used to learn the features and relation of user information and click logs, and the probability value of CTR prediction is calculated using a regression model. Two-stages of learning to rank approach based CTR prediction algorithm for contextual advertising were also proposed [8]. This algorithm utilizes clicked requests to build a ranking model where ads are ranked as lists, and to build a regression model where the predicted value of the ranking model is converted into CTR by a sigmoid function. Another approach based on deep neural networks via attention mechanism was studied by Wang et al. [9]. The proposed approach uses a reduced dimension method to cluster similar users, queries and advertisements, and build a tensor model after dimension reduction. A hybrid model, called ASAE was proposed for advertising CTR estimation that trains a deep component and attentional factorization machine component together. An online learning algorithm [7], called Follow-The-Regularized-Factorized-Leader (FTRFL) was also proposed for CTR prediction.

This scheme integrates the Follow-The-Regularized-Leader-Proximal (FTRL-Proximal) with per-coordinate learning rate into factorization machines, and shows a faster rate of convergence compared to the factorization machines with stochastic gradient descent.

In the study [26], the attributes of target audiences that are not annotated are identified. To predict the age, gender, and five personality attributes, the following machine learning algorithms are used: Linear Regression, Naive Bayes, and Support Vector Machine. In order to predict client-side profiles for personalized advertising, Bilenko et al. [11] proposed a practical solution for keyword advertising platform, which also addresses the issues of privacy and control of the user data. In this proposed solution, a parameterized function is trained on user data including ad clicks and impressions to optimize utility estimation. After simulating the profile construction process, the true labels are collected from the presence of the keyword candidate. And then, the learning algorithm tries to identify the predictor parameters that minimize training set error. The proposed approach was evaluated with three different learning algorithms, including max-margin averaged perceptron, logistic regression trained using L-BFGS algorithm with L1 and L2 regularization, and boosted decision trees. Another approach [10] was proposed to develop agent-based referral systems to identify user interests by user profiling. The system is behavior-based with a binary class model (0: user not interested and 1: user interested). In respect to the binary model, machine learning is used to find potential user interests. Depending on the issues, including the nature of items, the specific type of the service, and type and amount of available information (e.g., documents, multimedia, or with additional textual information), different or hybrid machine learning algorithms are used for a recommender system in the proposed approach.

3.2. Content-centric advertising

To integrate relevance with click feedback, a class of models for a contextual advertising system [27] was proposed. Specifically, the ad-page scoring function is augmented with extra parameters from a logistic regression model on data in the advertisements and web pages. The regression integrates click feedback and semantic information available from both advertisements and web pages to determine relevancy. The proposed scheme consists of three steps: feature extraction, feature selection, and coefficient estimation for feature through a logistic regression.

A blog-centric contextual advertising framework (BCCA) [17] was proposed to integrate contextual advertising with text mining approaches for selecting advertisements based on personal interests shown in a blog and ranking them based on their relevance. The proposed scheme consists of four modules to find the best matching advertisements: (a) intention recognition, (b) sentiment detection, (c) term expansion, and (d) target-ad matching. For page-ad matching, advertisements are ranked by $P(a|q)$, where the probability of an advertisement a is interpreted as the likelihood that it is relevant to the query q.

Zhang et al. [12] proposed a scheme to learn sub-document classification for contextual advertisement applications when only page level labels are available. The proposed scheme utilizes MIL (Multiple Instant Learning)-Boost to address different problems in contextual advertising. One of the problems addressed was sensitive content detection so undesired contents can be avoided even if they occur in a small part of the web page. Another problem addressed was opinion mining from review sites so that negative opinions about products can be detected and avoided. This system helps to save time and cost significantly by providing good quality block level labels.

A simple and scalable response prediction framework [28] was proposed for display advertising to provide simplicity, scalability, and efficiency. The proposed framework uses Maximum Entropy (also known as logistic regression) with a two-phase feature selection algorithm (feature selection and value selection) for increasing automation and reducing the need for domain expertise. The proposed machine learning framework can address display advertising issues effectively with a limited memory and can scale to a large number of samples and parameters.

Huang et al. [18] introduced an architecture for context-aware advertisement and delivery in vehicular networks. The architecture functions as an application on network devices and consists of various modules including a navigation module for map data, a context recognition module for sensors within the vehicle, a data reception module for advertising, a configuration module for presentation slot availability, and an advertisement management module for authentication of advertisement content and value.

A video representation scheme was proposed in the study [29] to capture the latent semantics of a video advertisement in unsupervised learning. The proposed scheme integrates poster probability information objects and logos into Latent Dirichlet Allocation (LDA), which is named ppLDA (posterior probability involved in LDA). Real-world video advertisement data sets were used with four typical classifiers, random forest, kNN, SVM, and AdaBoost for performance evaluation. The proposed scheme demonstrated that better classification performance was achieved and ad categorization can be effectively supported. DeepLink [30] introduced a deep learning-based framework for video advertising. In the proposed scheme, deep convolutional neural networks (CNNs) are used to link sitcom-stars and online shops with clothing retrieval. For composing multiple sub-modules including human pose selection, human-body detection, face verification, clothing detection and retrieval from ad-images, DeePLink adopted several deep CNN models. The deep CNN models were transferred to the data domain and then corresponding models were trained based on the constructed large-scale clothes dataset, so the clothing retrieval can be used for video advertising system efficiently.

As RTB is a favored way of delivering online advertising for ad exchanges and search providers, Andrey et al. [20] proposed an adaptive targeting for RTB to maximize

Table 1
Machine learning approaches for target advertising.

Machine learning approaches for target advertising			
User-centric approach		Content-centric approach	
Behavioral targeting	[2]	Contextual advertising	[12,27]
Keyword browsing	[6]	Display	[28]
Click-through-rate	[7–9,14]	Video	[29,30]
Sponsored search	[2,7]	Vehicle	[18]
User profiling	[10,11,26]	Blog	[17]
		Web documents	[12]
Click fraud	[22–25]	Real-time bidding	[20,21]

the CTR. The results of evaluation show that the advertisement campaign with the proposed adaptive strategy noticeably increases the CTR. A model-based reinforcement learning model [21] was introduced to learn the bidding strategy in RTB display advertising. In this proposed scheme, the bid decision process is formulated as a reinforcement learning problem. By modeling the state transition via auction competition, the Markov Decision Process framework was established to optimize advertising performance for real-time bidding.

3.3. Click fraud detection in advertising

Click fraud is a challenging issue in advertising because it can negatively impact ad budget and harm the integrity of the online advertising market. In order to detect click fraud, an ensemble learning-based approach [23] was proposed for click fraud detection in mobile advertising. A set of new features is derived to detect click fraud from existing properties. For the evaluation, a final ensemble model based on six different learning algorithms was analyzed in terms of three different performance indicators to show that the proposed model can detect fraudulent partners with a high rate. In a more recent work, Taneja et al. [22] proposed a mobile advertising framework to detect fraudulent partners based on click data associated with web surfing information on mobile phones. The proposed framework consists of Recursive Feature Elimination (RFE) as the feature selection technique and Hellinger Distance Decision Tree (HDDT) as the classifier to identify deceitful publishers.

Paulo et al. [25] also proposed a system that detects and prevents click fraud practices and implemented it on an ad network with three main agents. The main idea of the proposed system is to test every received click with every rule to determine if the click is suspicious. The system demonstrated good performance against different types of attacks in a testing environment, and it suggested that the weights of the rules for accurately classifying attacks should be appropriately selected by optimization in a real world scenario.

4. Discussion and conclusion

In this paper, applications of various machine learning techniques to online advertising strategies were investigated to optimize online target advertising. Target advertising strategies and corresponding machine learning techniques are summarized in Table 1. The increasing interest in and demand for the integration of artificial intelligence and data-driven approaches to digital advertising gave rise to the need for the current investigation. Targeted online advertising strategies and respective machine learning-based techniques are identified and classified into two broad categories, user-centric and content-centric approaches. The paper also identifies the need for machine learning-based click-fraud detection systems to protect advertisers from illegitimate clicks. This classification serves as groundwork for future research examining machine learning and artificial intelligence techniques to further optimize targeting and effectiveness of online advertising strategies, and to address security and privacy issues in advertising, such as click-fraud detection.

Acknowledgments

This work was supported in part by the A.R.T. Program and the College of the Arts and Communication Center for Creative Activity & Research Summer Grant, William Paterson University of New Jersey.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J.A. Choi, R. Lewis, Culture and the star-power strategy: Comparing American and Korean response to Celebrity-Endorsed advertising, J. Glob. Mark. 30 (1) (2017) 3–11.
[2] Y. Chen, M. Kapralov, J. Canny, D.Y. Pavlov, Factor modeling for advertisement targeting, in: Advances in Neural Information Processing Systems, 2009, pp. 324–332.
[3] A. Capatina, M. Kachour, J. Lichy, A. Micu, A. Micu, F. Codignola, Matching the future capabilities of an artificial intelligence-based software for social media marketing with potential users' expectations, Technol. Forecast. Soc. Change 151 (2020) 119794.
[4] S.L. Lo, D. Cornforth, R. Chiong, Effects of training datasets on both the extreme learning machine and support vector machine for target audience identification on twitter, in: Proceedings of ELM-2014 Volume 1, Springer, Cham, 2015, pp. 417–434.
[5] S.L. Lo, R. Chiong, D. Cornforth, Using support vector machine ensembles for target audience classification on Twitter, PLoS One 10 (4) (2015).
[6] H. Li, D. Zhang, J. Hu, H. Zeng, Z. Chen, Finding keyword from online broadcasting content for targeted advertising, in: Proceedings of the 1st International Workshop on Data Mining and Audience Intelligence for Advertising, 2007, pp. 55–62.
[7] A. Ta, Factorization machines with follow-the-regularized-leader for CTR prediction in display advertising, in: Proceedings of 2015 IEEE International Conference on Big Data (Big Data), IEEE, 2015, pp. 2889–2891.
[8] Y. Tagami, S. Ono, K. Yamamoto, K. Tsukamoto, A. Tajima, Ctr prediction for contextual advertising: Learning-to-rank approach, in: Proceedings of the Seventh International Workshop on Data Mining for Online Advertising, 2013, pp. 1–8.
[9] Q. Wang, F. Liu, S. Xing, X. Zhao, A new approach for advertising CTR prediction based on deep neural network via attention mechanism, Comput. Math. Methods Med. (2018).

[10] A. Addis, G. Armano, E. Vargiu, Profiling users to perform contextual advertising, in: Proceedings of the 10th Workshop dagli Oggetti agli Agenti (WOA 2009), 2009.

[11] M. Bilenko, M. Richardson, Predictive client-side profiles for personalized advertising, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011.

[12] Y. Zhang, A.C. Surendran, J.C. Platt, M. Narasimhan, Learning from multi-topic web documents for contextual advertisement, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 1051–1059.

[13] 50 Stats Showing the Power of Personalization, 2020, https://www.forbes.com/sites/blakemorgan/2020/02/18/50-stats-showing-the-power-of-personalization/#3fce20942a94.

[14] Z. Jiang, S. Gao, W. Dai, Research on CTR prediction for contextual advertising based on deep architecture model, J. Control Eng. Appl. Inform. 18 (1) (2016) 11–19.

[15] O. Chapelle, E. Manavoglu, R. Rosales, Simple and scalable response prediction for display advertising, ACM Trans. Intell. Syst. Technol. (TIST) 5 (4) (2014) 1–34.

[16] H. Zhang, X. Cao, J.K.L. Ho, T.W.S. Chow, Object-level video advertising: an optimization framework, IEEE Trans. Ind. Inf. 13 (2) (2016) 520–531.

[17] T. Fan, C. Chang, Blogger-centric contextual advertising, Expert Syst. Appl. 38 (3) (2011) 1777–1788.

[18] Q. Huang, D. Greene, J. Liu, H. Calabria, Vehicle network advertising system, 2007, U.S. Patent Application 11/213, 129.

[19] K. Lim, D. Manivannan, An efficient protocol for authenticated and secure message delivery in vehicular ad hoc networks, Veh. Commun. 4 (2016) 30–37.

[20] A. Pepelyshev, Y. Staroselskiy, A. Zhigljavsky, Adaptive targeting for online advertisement, in: Proceedings of International Workshop on Machine Learning, Optimization and Big Data, Springer, Cham, 2015, pp. 240–251.

[21] H. Cai, K. Ren, W. Zhang, K. Malialis, J. Wang, Y. Yu, D. Guo, Real-time bidding by reinforcement learning in display advertising, in: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, 2017, pp. 661–670.

[22] M. Taneja, K. Garg, A. Purwar, S. Sharma, Prediction of click frauds in mobile advertising, in: Proceedings of 2015 Eighth International Conference on Contemporary Computing (IC3), IEEE, 2015, pp. 162–166.

[23] K.S. Perera, B. Neupane, M.A. Faisal, Z. Aung, W.L. Woon, A novel ensemble learning-based approach for click fraud detection in mobile advertising, in: Mining Intelligence and Knowledge Exploration, Springer, Cham, 2013, pp. 370–382.

[24] C.M.R. Haider, A. Iqbal, A.H. Rahman, M.S. Rahman, An ensemble learning based approach for impression fraud detection in mobile advertising, J. Netw. Comput. Appl. 112 (2018) 126–141.

[25] P.S. Almeida, J.J. Gondim, Click Fraud detection and prevention system for Ad Networks, J. Inf. Secur. Cryptogr. 5 (1) (2018) 27–39.

[26] R.B. Tareaf, P. Berger, P. Hennig, J. Jung, C. Meinel, Identifying audience attributes: predicting age, gender and personality for enhanced article writing, in: Proceedings of the 2017 International Conference on Cloud and Big Data Computing, 2017, pp. 79–88.

[27] D. Chakrabarti, D. Agarwal, V. Josifovski, Contextual advertising by combining relevance with click feedback, in: Proceedings of the 17th international conference on World Wide Web, 2008, pp. 417–426.

[28] C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman, F. Provost, Machine learning for targeted display advertising: Transfer learning in action, Mach. Learn. 95 (1) (2014) 103–127.

[29] S. Hou, S. Zhou, W. Liu, Y. Zheng, Classifying advertising video by topicalizing high-level semantic concepts, Multimedia Tools Appl. 77 (19) (2018) 25475–25511.

[30] H. Zhang, Y. Ji, W. Huang, L. Liu, Sitcom-star-based clothing retrieval for video advertising: a deep learning framework, Neural Comput. Appl. 31 (11) (2019) 7361–7380.