Journal of The Institute of Electronics and Information Engineers Vol.58, NO.12, December 2021 https://doi.org/10.5573/ieie.2021.58.12.42
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2021-58-12-6

ANN을 활용한 입찰가 변경 데이터 기반 키워드 검색광고 방문자 수 예측방법

(A Method for Predicting the Number of Visitors of Keyword Search Advertising based on Bid Change Data using ANN)

이 정 우*, 김 승 천*

(Jung-Woo Lee and Seung-Cheon Kim©)

요 약

중소규모의 쇼핑몰에 있어 온라인 광고의 중요도는 매우 높다. 통계에 따르면 특히 키워드 검색광고의 선호도와 효과가 광고주와 고객 모두에게 높은 편이다. 중소규모의 쇼핑몰은 광고비 예산이 적은 광고주들로 이들을 위해 광고 효율을 높여 적은 광고비로도 높은 광고 성과를 낼 수 있는 광고 운영을 위한 광고 성과 예측 모델이 필요하다. 검색광고 성과 예측 모델은 다양한 인공지능 모델링 기법으로 연구가 많이 되어 왔다. 그러나 대부분의 연구는 사용자 정보 데이터와 구매 이력 등의 정보를 활용하여 성과를 예측한다. 전 세계적으로 개인정보보호가 강화된 환경에서는 고객의 개인정보 데이터를 활용하여 성과를 예측하기 어려워 새로운 종류의 데이터를 사용한 광고 성과 예측을 연구가 필요하다. 본 연구에서는 온라인 검색광고의 운영을 통해 저장된 검색 키워드 입찰가 변경 로그 데이터와 성과 데이터인 방문자 수 로그 데이터를 활용하여 검색광고 성과 중 방문자 수를 예측하는 프로그램을 인공지능 모델링 기법 중 ANN을 활용하여 구현하고, 중소규모의 쇼핑몰의 실제 광고 운영 데이터를 활용하여 성능을 측정하였다.

Abstract

Online advertising is very important for small and medium sized e-commerce shopping malls. According to statistics, the preference and effectiveness of keyword search advertising are high for both advertisers and customers. Small and medium sized e-commerce shopping malls are advertisers with a small advertising budget, and for them, an advertising performance prediction model is needed for advertising operation that can achieve high advertising performance with low advertising costs by increasing advertising efficiency. The search advertising performance prediction model has been studied extensively with various artificial intelligence modeling techniques. However, most studies predict performance using user information data and purchase history. In an environment where personal information protection is strengthened worldwide, it is difficult to predict performance using customer personal information data, so research on advertising performance prediction using new types of data is needed. In this study, using the search keyword bid change log data stored through the operation of online search advertising and the number of visitors logs data as the advertising performance data, a program to predict the number of visitors among search advertisement performance was implemented using ANN among artificial intelligence modeling techniques. The performance was measured using actual advertising operation data of small and medium sized e-commerce shopping malls.

Keywords : Online advertising, Search advertising, Advertising performance prediction, RTB, ANN

Ⅰ. Introduction

The size of the Korean e-commerce market is 161 trillion won in 2020, an increase of 19.3% from the previous year, showing a high growth rate as shown in Fig 1[1]. The e-commerce market continues to

* 정회원, 한성대학교(Hansung University)
© Corresponding Author(E-mail : kimsc@hansung.ac.kr)
Received ; September 25, 2021 Revised ; October 23, 2021
Accepted ; October 27, 2021

grow rapidly, reaching an average of 20% every year thanks to the explosive growth of the mobile e-commerce market following the popularization of smartphones, the establishment of social commerce following the open market, and the expansion of internet investment by large retailers[2].

Among them, the status of the small and medium-sized e-commerce market is increasing to 12 trillion won as of 2017, and if the scope is expanded to individual shopping malls, hundreds of new ones are created every day[3]. As such, small and medium-sized e-commerce shopping malls, which are on the rise, are having difficulties in taking customer contact points. For small e-commerce companies with low customer awareness, the only contact point with customers is online advertising, although there may be differences in form, such as search advertisements, display advertisements, and advertisements through SNS such as Instagram, blogs, and Twitter. As a result, the dependence on advertisements in the e-commerce industry is high. In fact, online advertising showed a positive result in increasing the inflow to commerce shopping malls.

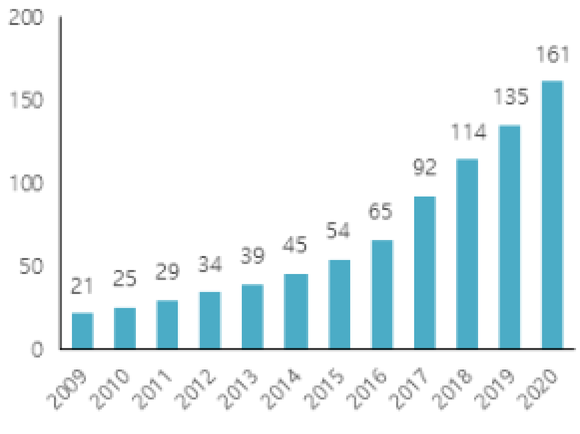

그림 1. 대한민국 e-commerce 시장 크기(조 원)
Fig. 1. Domestic e-commerce market size (KRW trillion).

According to the "Online advertising industry trend research and analysis" published by the Korea Internet & Security Agency, among various online advertising types, 'search advertisement' showed the highest preference at 62.8%. As the type of online advertising with the highest correlation to sales selected by online advertisement providers, 40.3% of advertisement providers answered 'search advertisement', and advertisers also selected 'search advertisement' as the type of online advertising with the highest connection to sales (31.5%) are doing. As such, search advertisement occupies a very important proportion among all advertising types[4]. In Korea, the amount of inflow through the N portal among internet search is significantly higher than that of other search, such as the D portal and the G portal[2]. As the ranking of information in search advertising has a great influence on the search and utilization of information, the ranking of keywords in search advertising is becoming increasingly important[5].

The biggest characteristic of search advertising compared to other advertising is that they are on-demand advertising that is exposed only when a customer searches with a will. Most other advertisings, such as display advertising, are exposed to the web page which is accessed by the customer regardless of the customer's will and are expected to receive clicks and inflows. The working principle of search advertising is as follows. When a customer enters a word to search in the search portal, advertisements from several advertisers who have registered the same text keyword as the word to search as advertising creatives are listed in the customer's search results. The advertisements listed in the search results are exposed in the order of the highest advertisement cost bid price of the corresponding text keyword set by the advertiser. When a customer clicks on one of the listed advertisements, the customer will access a web page that is set in advance by the advertiser. Search advertising is a CPC (Cost Per Click) type of advertising in which the advertiser pays the advertisement cost when a customer clicks on the advertisement. For search advertising to be effective, an advertiser registers a plurality of text keywords as advertising creatives. When selecting an advertising creative, an advertiser first selects and registers a representative product or service name, and then register keywords that customers are expected to search for as creatives and display them when they search for the keywords. The most important goal of search advertising is to make it flow to the advertiser's website through the keywords exposed when customers enter keywords at the search portal.

When a customer searches for a keyword, the same keyword of several competing advertisers is exposed as a searching result. To improve the advertising performance of keyword search advertising, keyword exposed ranking is important. This exposure rank is determined by the keyword advertising auction bid set by the advertiser. If an advertiser increases its keyword bid to increase its impression rank, it will also increase its impression rank, but if the bid is raised excessively, the advertising may not be exposed at the time it is needed due to the exhaustion of the advertising budget, or the advertising efficiency may decrease due to excessive advertising expenditure. Therefore, considering the advertising budget, it is essential to adjust the advertiser's bids to maintain an appropriate ranking.

In this study, we intend to design and implement an AI(Artificial Intelligence) program that predicts the number of visitors when bids are adjusted and to check the accuracy through experiments by using actual operation data of keyword search advertising.

Ⅱ. Previous research

Studies on the introduction of AI techniques to predict the performance of search advertising have been conducted from various viewpoints. Research on predicting the optimal ranking of search keywords[6] used AI techniques to predict the bid of search keywords for domestic N-portal search advertising. They proposed a method to collect data by crawling keyword search results and to increase accuracy by applying various optimization algorithms to the collected data. In those studies, research was conducted to propose an algorithm with high prediction accuracy through comparative analysis of optimization algorithms, and through this, the ranking of search keywords was predicted. These studies help predicts the exposure ranking of search keywords. However, it does not predict how many visitors will be generated according to changes in the bid price of search keywords that are currently spending advertising expenses. The higher the exposure ranking of the search keyword, the higher the probability that the number of visitors through clicks will increase, but the bid amount to register the highest impression ranking also increases. Considering the optimization of advertising spending, it is not advisable to unconditionally raise the exposure ranking.

Previous studies to predict the number of visitors through clicks (CTR: Click Through Rate) have also been tried with various approaches. Early CTR prediction models were based on linear models such as the logistic regression model[7], naive Bayes[8], and Follow the Regularized Leader (FTRL Proximal)[9]. These models have the advantages of being simple and easy to extend, but they have limitations in effectively capturing the interactions between features and have a disadvantage in that their performance is low. Afterward, as an improved model, there is a CTR prediction model structure using the embedding technique[10, 11]. Field-Aware Factorization Machine (FFM)[12] improved the prediction performance and won two CTR prediction competitions hosted by Criteo and Avazu. However, as these methods only model second-order interactions of features, it is difficult to understand high-order feature interactions. Therefore, there is a limit that there is a decrease in performance. There are also CTR prediction studies using CNN (Convolutional Neural Network)[13~15]. Studies using CNN mainly focus on new structure proposals. CNN models are excellent at recognizing patterns of features. However, due to the characteristics of data that CTR data has no meaning in sort order, unlike image or voice data, the use of a CNN structure inevitably leads to deterioration of prediction performance[13].

In addition to this, there are also studies that suggest modeling and preprocessing techniques for predicting the winning bid amount for placing advertising in a specific ranking in keyword advertising and suggesting learning with multiple regression analysis models[16]. These studies suggested various variables for prediction but did not proceed with the prediction by reflecting the actual data.

There are also studies that predicted the minimum winning bid for keyword target ranking using LSTM and GRU[17]. Those studies compared the performance of the LSTM model and the GRU model and derived the result that the GRU model has higher accuracy than the LSTM model, but the overall performance is not high, and there is a problem with the prediction value being heavily biased.

These previous studies have a more important problem than performance degradation. They use user information (gender, occupation, etc.) and context information (product information, purchase history, etc.) as experimental data. In the case of search advertising, it is possible to target an audience that shows different search results according to users by specifying the user who searches. Recently, as the strengthening of personal information protection has become a global trend, the user's privacy policy has been strengthened. As a result, the use of personal information data in online advertising is limited. This makes accurate audience targeting difficult, and it is expected to become more difficult in the future. Therefore, in the future, search advertising should be recognized as a process in which unspecified people click on search keywords exposed through search. Studies on the performance prediction of search advertising also need new research that does not utilize user information or context information as in the past.

This study proposes a method that the number of visitors is predicted using ANN (Artificial Neural Network) among AI modeling techniques when the bid amount changes based on the past bid change history and performance history of the currently running advertising excluding user information and context information.

Ⅲ. Experimental method

The data for the experiment was collected and used by collecting search advertising operation log data and the number of visitors log which is the operation result.

In general, it is rare to save the log of the operation of search advertising, but for this study, the necessary data logs were collected while operating search advertising for mid to large sized e-commerce shopping malls that have been running search advertising on domestic N portal since 2019.

A total of 5 keywords used in the experiment were selected and collected according to the following selection criteria.

As a quantitative factor, keywords with a history of changing their bids more than once a day during 2020 were selected in the order of the most history of bid changes.

Since data was collected over a period of one year, as a qualitative factor, it was selected from keywords for products that are not sensitive to seasons or trends.

A log data of the bid change history for these keywords for one year was collected, and a log data of visitors through the keyword was collected separately. After merging the two log data in time order, the number of visitors was summed up between after the bid change and before the next change. It can figure out the difference between the number of visitors through the changed bid and the number of visitors through the previous bid. Based on the merged data, a data set was prepared with the bid change interval and bid price change rate as input values, and the sum of the number of visitors due to the bid price change as the result.

Table 1 shows each keyword content and data set that is the keyword bid change and the number of visitors performance. Among the data sets, 80% was used for training data and 20% was used for test

표 1. 키워드와 데이터 셋 수
Table 1. Keyword and Number of Data Set.

Kwd No	Keyword Content	No of Data
1	women's underwear shopping mall (여성속옷쇼핑몰)	9,028
2	magic bra(뽕브라)	8,506
3	women's underwear set(여성속옷세트)	8,429
4	20's women's underwear(20대여성속옷)	8,700
5	bra panties(브라팬티)	8,618

표 2. 실험 환경
Table 2. Experiment Environment.

Category	Value
OS	Ubuntu 18.04.5
CPU	Intel(R) Xeon(R) CPU @ 2.00GHz x 2ea
GPU	None
Memory	16GB
Language	Python 3.7.11
Library	Tensorflow 2.6.0

data. The experimental environment used for the experiment is shown in Table 2.

AI modeling was performed with an ANN having two hidden layers. The node count of each layer was set to 256, and the dropout was set to 20% to reduce overfitting. The Adam algorithm was used as an optimization algorithm to minimize the difference between the actual result and the predicted result value when learning data. The activation function compared the difference using the sigmoid function and the ReLU function. The learning rate is a method that repeats the process of finding the optimal solution in which the loss function is minimized. If the learning rate is too big, overshooting that diverges without converging to the optimal value occurs. If the learning rate is too small, the convergence speed is too slow and the probability of falling into the local minimum increases. In this study, in order to find a suitable learning rate, we tried to find the value that gave the best result after applying from 0.0001 to 0.1. The loss function used mean squared error (MSE). 100 epochs were performed for each keyword, and the training was conducted in the direction of monitoring and minimizing the loss of the validation set, and it was set to stop if the performance did not increase.

Ⅳ. Experimental result

Experiments were conducted by changing the learning rate and activation function for each keyword, and the experimental results showed the accuracy shown in Table 3. And loss values were as shown in Table 4.

표 3. 정확도 비교
Table 3. Accuracy Comparison.

Kwd No	No of Data	Acti-vation Func.	Learning Rate			
			0.1	0.01	0.001	0.0001
1	9028	Sigmoid	0.8291	0.8313	0.8042	0.8286
		ReLU	0.8247	0.8308	0.8247	0.8291
2	8506	Sigmoid	0.8001	0.7995	0.7995	0.7995
		ReLU	0.7995	0.7989	0.7995	0.7989
3	8429	Sigmoid	0.7998	0.7998	0.7998	0.7998
		ReLU	0.8004	0.7992	0.8004	0.8004
4	8700	Sigmoid	0.8471	0.8471	0.8471	0.8448
		ReLU	0.8362	0.8454	0.8437	0.8448
5	8618	Sigmoid	0.4887	0.4835	0.5450	0.4870
		ReLU	0.6111	0.4046	0.5926	0.5903

표 4. 손실값 비교
Table 4. Loss Value Comparision.

Kwd No	No of Data	Acti-vation Func.	Learning Rate			
			0.1	0.01	0.001	0.0001
1	9028	Sigmoid	0.1456	0.1204	0.1423	0.3009
		ReLU	0.1140	0.0947	0.1007	0.1042
2	8506	Sigmoid	0.1184	0.1226	0.1238	0.1391
		ReLU	0.1859	0.1347	0.1346	0.1266
3	8429	Sigmoid	0.1341	0.1293	0.1306	0.2065
		ReLU	0.1921	0.1837	0.1714	0.2104
4	8700	Sigmoid	0.0852	0.0693	0.1146	0.3377
		ReLU	0.0676	0.0661	0.0650	0.0735
5	8618	Sigmoid	0.5485	4.2701	4.4080	6.8007
		ReLU	1.6179	1.5046	1.2346	1.4011

Fig. 2 is a graph showing the difference between the accuracy and loss values according to the activation function for each dataset. In the case of a

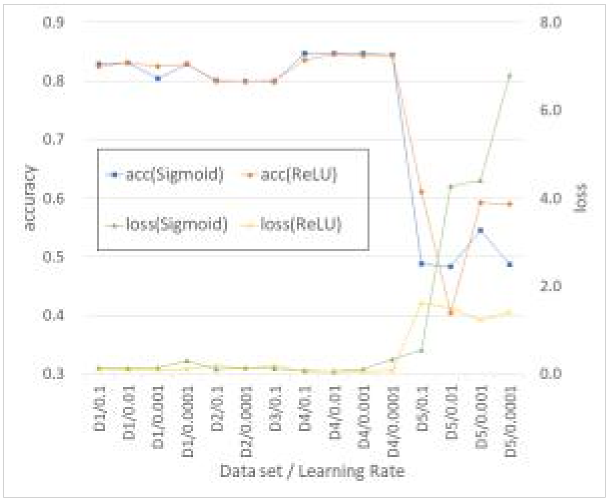

그림 2. 활성화 함수 비교
Fig. 2. Activation Function Comparison.

dataset with high accuracy and low loss value, it means high reliability, there is no significant difference depending on the change of the activation function. In the case of a dataset with low accuracy and large loss value, it was seen that when the activation function is set to ReLU, it has higher accuracy and lower loss value than when set to sigmoid. Through, the experiments conducted in this study, it has been shown that using the ReLU function as the activation function can secure higher reliability.

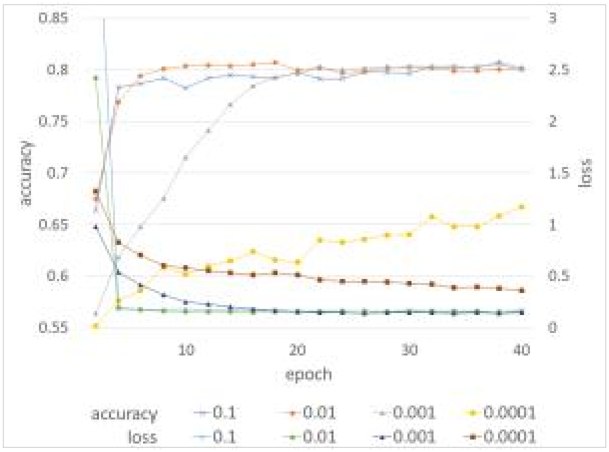

그림 3. 학습률 비교
Fig. 3. Learning Rate Comparison.

Fig. 3 shows the change of the accuracy and loss values up to 40 epochs for each learning rate with the first keyword data set. In the case of the learning rate, it was seen that the accuracy and loss function are sufficiently reliable when set to a certain value or less, except for the case of 0.0001 value. When the learning rate is 0.001, it takes longer to reach high accuracy compared to other values. As can be seen from Table 2 and Table 3, it was seen that the highest reliability is shown when the learning rate is 0.001 when the last keyword with low accuracy is considered.

V. Conclusion

In this study, in accordance with the strengthened privacy policy, a program was designed to predict the number of visitors through clicks among search advertising performance using information other than user information, namely, bid change information, and conducted an experiment using actual operational data. As a result, some keywords showed a relatively low prediction accuracy of about 60%, but most keywords showed a prediction accuracy of about 80%, confirming that significant results were obtained. This study showed lower performance than previous studies using several advanced artificial intelligence modeling techniques, which showed more than 90% prediction accuracy, but it is a meaningful study in that predictions were attempted only from the operation logs without using personal information or context information. Started from this study, it is expected that research showing improved performance through the application of different datasets and various artificial intelligence modeling will continue. When this is applied to the actual industry, it is possible to help the strategy of changing the advertising cost by providing predictive data on whether it is better to keep, increase, or lower the current advertising cost for each keyword in order to increase the number of visitors.

In the future, if the accuracy can be increased by diversifying experimental data, improving AI modeling, and changing parameters through additional research, it is expected that it can be used in the real industry. Through this, it will be possible to efficiently execute advertising costs without wasting the advertising costs of advertisers.

Acknowledgement

This research was financially supported by Hansung University.

REFERENCES

[1] Y. S. Jeon, “E-commerce market trend that converges with software,” Software Oriented Society, pp. 14-23, Jun. 2021.
[2] Jeoung-Kuk Kim, Sang-Won Lee, and Dong-Oun Choi, “Relevance Analysis of Online Advertisement and e-Commerce Sales,” Journal of the Korea Entertainment Industry Association, vol. 10, no. 2, pp. 27-35, Dec. 2016.
[3] Mi Young Son, “Case Study on the Operation of Small Capital Online Fashion Stores - Focusing on 4P Mix,” International Journal of Coustume, vol. 70, no. 2, pp. 117-134, Dec. 2020.
[4] Korea Internet & Security Agency, “Online advertising industry trend research and analysis,” 2017.
[5] Jeeyun Oh and Sung Gwan Park, “The Effects of Search Engine Credibility and Information Ranking on Search Behavior,” Korean Journal of Journalism and Communication Studies, vol. 53, no. 6, pp. 26-49, Dec. 2009.
[6] M. Ji and H. Park, “Prediction Algorithm for Optimal Ranking of Search Keyword by TensorFlow,” The Journal of Korean Institute of Communications and Information Sciences, vol. 43, no. 8, pp. 1347-356, Jan. 2018.
[7] M. Richardson, E. Dominowska, and R. Ragno, “Predicting clicks: Estimating the Click-Through Rate for New Ads,” in Proc. of the 16th international conference on World Wide Web - WWW ’07, pp. 521-29, 2007.
[8] D. J. Hand and K. Yu, “Idiot’s Bayes: Not So Stupid after All?,” International Statistical Review, vol. 69, no. 3, pp. 385-98, Dec. 2001.
[9] H. B. Mcmahan et al., “Ad click prediction: A view from the trenches,” in Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1222-230, Aug. 2013.
[10] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, “A Neural Probabilistic Language Model.,” J. Mach. Learn. Res., vol. 3, no. 6, pp. 1137-155, Aug. 2003.
[11] S. Rendle, “Factorization machines,” in 2010 IEEE International Conference on Data Mining, pp. 995-000, 2010.
[12] Y. Juan, D. Lefortier, and O. Chapelle, “Field-aware factorization machines in a real-world online advertising system,” in 26th International World Wide Web Conference 2017, WWW 2017 Companion, pp. 680-88, Jan. 2017.
[13] Tae-Suk Kim, “Visualization Analysis on CNN Based CTR Prediction,” Journal of The Korean Data Analysis Society, vol. 22, no. 6, pp. 2603-2614, Dec. 2020.
[14] Q. Liu, F. Yu, S. Wu, and L. Wang, “A convolutional click prediction model,” in International Conference on Information and Knowledge Management, Proceedings, pp. 1743-1746, Oct. 2015.
[15] B. Liu, J. Yu, R. Tang, H. Guo, Y. Zhang, and Y. Chen, “Feature generation by convolutional neural network for click-through rate prediction,” in The Web Conference 2019 - Proc. of the World Wide Web Conference, WWW 2019, pp. 1119-1129, May 2019.
[16] D. H. Kim, T. S. Kim, and D. Park, “Research regarding Prediction Model of Bidding Price in SearchAD using Machine Learning,” in Proc. of the Korean Information Science Society Conference, vol. 2017, no. 06, pp. 1910-1913, Jan. 2017.
[17] G. R. Chang, “Prediction for Suggesting Minimum Bidding Price of Keyword Using Deep Learning.” Gyeongsang National Univ. Graduate School, Jan. 16, 2020.

저 자 소 개

이 정 우(정회원)
1995년 한남대학교 전자공학과 학사 졸업.
1997년 한남대학교 전자공학과 석사 졸업.
2011년 Aalto Univ. 경영학 석사 졸업.
<주관심분야 : 온라인 광고, 인공지능>

김 승 천(정회원)
1994년 연세대학교 전자공학과 학사 졸업.
1996년 연세대학교 전자공학과 석사 졸업.
1999년 연세대학교 전기컴퓨터 공학과 박사 졸업.
<주관심분야 : 네트워크 보안, 블록체인 서비스, 사물인터넷 보안, 5G 이동통신망 서비스>