Analysis of the Empirical Effects of Contextual Matching Advertising for Online News

Hyo-Jung Oh, Changki Lee, and Chung-Hee Lee

Beyond the simple keyword matching methods in contextual advertising, we propose a rich contextual matching (CM) model adopting a classification method for topic targeting and a query expansion method for semantic ad matching. This letter reports on an investigation into the empirical effects of the CM model by comparing the click-through rates (CTRs) of two practical online news advertising systems. Based on the evaluation results from over 100 million impressions, we prove that the average CTR of our proposed model outperforms that of a traditional model.

Keywords: Online advertising, contextual matching.

I. Introduction

Work in online advertising is generally focused on two main areas, contextual advertising and sponsored searches. Contextual advertising such as Google's AdSense and Yahoo's Contextual Match program is mainly concerned with the placement of ads on publisher pages. Since publisher pages are rich in content, a rich set of features can typically be extracted from the web page and used to find relevant ads [1], [2]. On the other hand, a sponsored search problem suffers from the same problem as a websearch in that the queries are short and have little context, such as Google's AdWords [3], [4].

Past studies on contextual advertising have been focused on how to extract keywords to match ad space. However, targeting mechanisms based solely on phrases found within the text of a page can lead to mismatch problems such as polysemy. Recent research has tried to classify web pages into ad taxonomy for matching with topically-relevant advertisements. In this letter, we propose a rich contextual matching (CM) model, named *AdContX*, that adopts a classification method for topic targeting and a query expansion method for semantic ad matching.

The main contribution of this letter is a detailed analysis of the empirical effects of a practical CM model for online news resulting from a series of experiments. We obtain new insight into the internal workings of the CM model and the relationship between the categories of ad campaigns and domains of news articles.

II. Practical Online Advertising Service Model

Most commercial online contextual advertising systems consist of four players [1]:

Users. Visitors at web pages published by a content provider (CP). They interact with the ads.

CP. The owner of web pages on which advertising is displayed.

Advertiser. The supplier of advertisements. As in traditional advertising, the goal of advertisers can be broadly defined as the promotion of products or services.

Ad broker (AD). A mediator between a content advertiser and publisher. This player selects the ads displayed on the web pages. The AD also shares advertisement revenue with the CP.

The AD model aligns the interests of the CP, advertisers, and the broker. In general, clicks benefit both the publisher and the AD by providing revenue and benefit the advertiser by bringing traffic to the target website. The advertisers pay for clicks on their ads, and therefore it is effective to provide ads that have a better chance of being clicked on by users.

Manuscript received Apr. 26, 2011; revised Aug. 5, 2011; accepted Aug. 22, 2011.

Hyo-Jung Oh (phone: +82 42 860 5405, ohj@etri.re.kr) and Chung-Hee Lee (forever@etri.re.kr) are with the BigData Software Research Laboratory, ETRI, Daejeon, Rep. of Korea.

Changki Lee (leeck@kangwon.ac.kr) was with the BigData Software Research Laboratory, ETRI, Daejeon, Rep. of Korea, and now is with the College of Information Technology, Department of Computer Science, Kangwon National University, Chuncheon, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.12.0211.0171

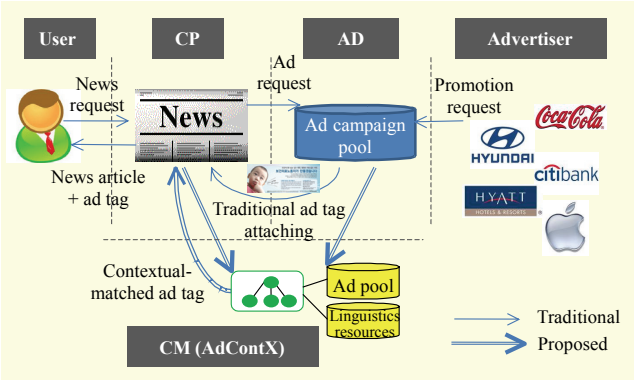

Fig. 1. Online news advertising service flow.

Unlike in a traditional model, an additional new player is added to the proposed method.

Contextual matcher (AdContX). A coordinator for matching ads based on the context of the content.

Figure 1 shows our practical online advertising service flow. Depending on how service players interact, particularly the CP, AD, and AdContX, a number of different advertising services can be generated.

III. AdContX: Contextual Matching Advertising

To select the most valuable ads for a given web news document, it is important to understand the context of the page from the viewpoint of products. For example, take a news article reporting that Samsung Electronics has been chosen as an official sponsor of the 2012 Olympics: While such news can be considered within the "sports" domain, it can be considered as a trigger of ads for "electronic products of Samsung" from an ad perspective. In this section, we describe the main components of AdContX.

1. Linguistic Analysis

Linguistic analysis is an essential process for understanding the context of web pages. When a CP publishes a web page, the page is sent to AdContX, which then converts the web page (for example, HTML or XML) into plain text while maintaining important information. To extract meaningful keywords from the content, we perform a morphological analysis, part-of-speech tagging, and named entity recognizing (NER). Since most product names are proper names, the accuracy of NER consequently influences the accuracy of CM. In this letter, 147 fine-grained named entity types are defined and organized into a hierarchical structure. This structure consists of 15 nodes at the top level, each of which consists of either 2 or 4 layers [5]. The base set of such a structure includes person, organization, location, civilization, date, time, and so on.

2. Classification Based on Ad-Taxonomy

To build a moderate taxonomy for Korean advertising, we refer to various taxonomies from online shopping malls[1] and online advertising agencies.[2] This hierarchy consists of 35 coarse (top level) categories (for example, electronics, leisure/sports, cars, furniture, and public organizations) and 435 fine classes, with each coarse class containing a non-overlapping set of fine classes. In AdContX, not only the web pages but also the ads are classified into an ad-taxonomy based on bid-phrases described by the advertiser when generating a campaign via the AD. This process allows us to semantically match web pages and ads, as well as reduce the search space. For ad-taxonomy classification, we built a hybrid classifier, structural SVM [6] and Rocchio [7], as in the following equation:

$$Tax(x) = \arg\max_{y} \left\{ (1-\alpha) * \mathbf{w}^T f(x_i, y) + \alpha * \frac{\sum_i c_y^i x^i}{\sqrt{\sum_i (c_y^i)^2} \sqrt{\sum_i (x^i)^2}} \right\},$$

where (x_i, y_i) is the training data (x is a document and y is a class label), $f(x, y)$ is a set of features for (x, y), and c_y is the centroid for class y. The terms c_y^i and x^i represent the weight of the i-th feature in the class centroid and document, respectively.

3. Ad Query Generation

The ultimate goal of CM is searching ad space to find the most appropriate ads for a given web page, and thus we need to distill ad query terms from the page content. After the main category of the page is assigned through ad classification, we can extract key bid-phrases based on the particular product class. We also extend the number of queries by referring to a linguistic resource, which consists of over 100 thousand bid-phrases connected with ad taxonomy. Extracted ad queries can be represented as a weighted vector.

4. Semantic Ad Matching

In general, we would like the match to be stronger when both the ad and page are classified into the same node and weaker when the distance between the nodes in the taxonomy increases [2]. We borrowed this idea from the relevance feedback mechanism [7].

$$Score(x, a) = \beta * TaxScore(Tax(x), Tax(a)) + (1-\beta) * KeywordScore(x, a),$$

where x is a document, a is an ad, $Tax(x)$ is the class of

1) www.enuri.com, www.bb.co.kr, www.gmarket.co.kr, www.interpark.com
2) www.overture.co.kr, www.247media.co.kr

Table 1. CTR comparisons.

News domain	Ad campaign	Total			Traditional			AdContX			Improv.
		Impression	Click	CTR	Impression	Click	CTR	Impression	Click	CTR	
Finance	Bank M	10,032,308	7,315	0.073%	6,025,963	4,130	0.069%	4,006,345	3,185	0.079%	16%
	Bank S	5,119,231	1,936	0.038%	2,979,323	1,017	0.034%	2,139,908	919	0.043%	26%
	Capital company H	10,004,419	3,634	0.036%	5,000,898	1,658	0.033%	5,003,521	1,976	0.039%	19%
	Automobile company H	7,010,114	9,144	0.130%	3,359,070	4,068	0.121%	3,651,044	5,076	0.139%	15%
	IT company L	10,016,607	4,819	0.048%	6,005,295	2,790	0.046%	4,011,312	2,029	0.051%	9%
Life	Education company S	10,040,972	9,170	0.091%	5,530,529	5,205	0.094%	4,510,443	3,965	0.088%	-7%
	Hospital S	5,094,385	5,856	0.115%	2,799,563	3,154	0.113%	2,294,822	2,702	0.118%	5%
	Wedding agency W	8,006,286	5,294	0.066%	3,002,048	1,713	0.057%	5,004,238	3,581	0.072%	25%
Food	Beverage D	10,006,506	6,977	0.070%	7,703,462	5,687	0.074%	2,303,044	1,290	0.056%	-24%
Leisure	Travel agency C	10,000,773	6,571	0.066%	5,000,517	2,802	0.056%	5,000,256	3,769	0.075%	35%
	Travel agency N	10,008,324	4,482	0.045%	7,003,139	3,104	0.044%	3,005,185	1,378	0.046%	3%
	Travel agency L	10,024,779	4,815	0.048%	7,356,598	3,484	0.047%	2,668,181	1,331	0.050%	5%
Total		105,364,704	70,013	0.066%	61,766,405	38,812	0.063%	43,598,299	31,201	0.072%	14%

document x, and $Tax(a)$ is the class of ad a.

In addition, $TaxScore(y, y')$ reflects a hierarchical loss that is calculated by $1–H–loss(y, y')$, and $KeywordScore(x, a)$ can be calculated as

$$KeywordScore(x,a) = \frac{\sum_i x^i a^i}{\sqrt{\sum_i (x^i)^2}\sqrt{\sum_i (a^i)^2}},$$

where the terms a^i and x^i represent the weight of the i-th feature in ad a and document x, respectively. These weights are based on the standard *tf-idf* formula. Given an article about golf, the AdContX can match the article with campaigns in "sports" node and further "golf wear" or "luxury sedan" nodes.

IV. Empirical Results

To show the efficacy of our CM model, we compared a traditional advertising model with the proposed advertising model, AdContX. For the performance comparison, we employ the click-through rate (CTR) [8]. A CTR is the ratio of the number of times an online advertisement is clicked to its number of impressions.

In a traditional advertising model, an AD considers keyword relatedness as well as the bidding amount per click when determining the exposure order. At that time, the traditional model determines whether the target content contains predefined ad keywords. On the other hand, our proposed model considers the context of the whole document and then performs semantic ad matching.

To guarantee a reliable evaluation, we built two practical online news advertising channels including fifteen commercial news service sites, including www.joins.com and www.chosun.com.[3] When a reader requests a news article, an advertising tag randomly selected between the traditional model and our AdContX model is attached via the http protocol.

Table 1 shows the results of a CTR comparison for twelve ad campaigns based on over 100 million impressions during a two-month period (November and December 2010). The ultimate CTRs which the advertisers can obtain are in the total column. Even though the number of total impressions of AdContX (43,598,299) is smaller than the traditional model (61,766,405) across most campaigns, except for the education company 'S' and the beverage product 'D', the average CTR of the AdContX model outperforms the traditional model with about a 14% improvement, 0.063 to 0.072. As shown in Fig. 2,

3) We co-worked with an online commercial advertising agency, ©interworks media (www.iwmedia.co.kr)

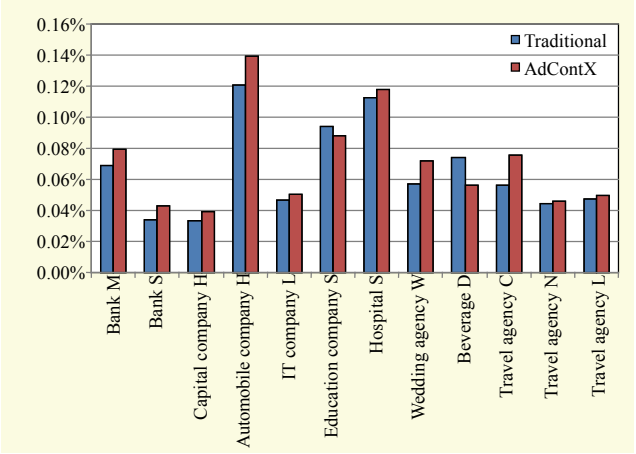

Fig. 2. CTR distribution.

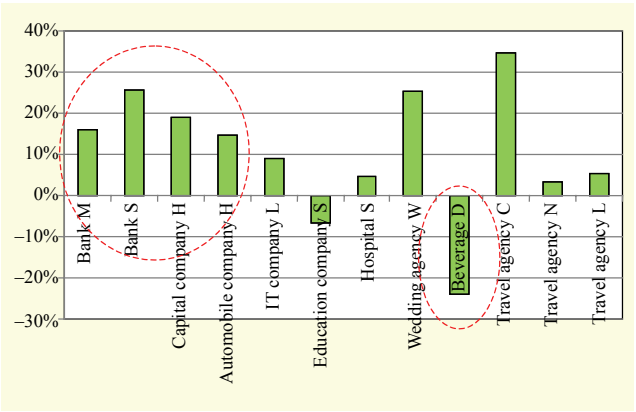

Fig. 3. Improvement distribution.

the automobile ad campaign achieved the highest CTR of both the traditional model (0.121%) and our AdContX (0.139%), whereas the capital company promotion ad 'H' obtained the lowest CTR (0.033 and 0.039, respectively). That is, the automobile ad is a more effective promotion than the capital company campaign.

The amount of improvement for the finance domain was particularly greater, as shown in Fig. 3. This indicates that the readers of finance news have a high tendency to click on ad campaigns when they are related with the context of the article. We argue that advertisements through AdContX might achieve the greatest promotional effect in the area of finance news.

On the contrary, the CTR of the beverage campaign 'D' was reduced by our proposed model because news articles related with that beverage are very rare; accordingly, the chance to display such an advertisement is also very limited. Such results have been similar with ad campaigns corresponding to the leisure domain. When the impression times of AdContX and the traditional model are very similar (the case of travel agency 'C'), the CTR of AdContX is much higher (a 35% improvement). However, for travel agencies 'N' and 'L', the chances for AdContX to display their ads are less, and thus the gains achieved in terms of the CTR decrease (3% and 5% improvements).

Another interesting point is that the CTR has a lesser relationship when an advertisement is produced by a popular celebrity such as "Rain," who is a very famous singer and actor in South Korea. In our experiments, a famous teen star promoted the 'S' education company, and thus the readers clicked the ad regardless of the news stories or context.

V. Conclusion

This letter reported on an investigation into the empirical effects of a CM ad model by comparing the CTR of two practical online news advertising systems. Based on the evaluation results of over 100 million impressions, the average CTR of the proposed AdContX model outperforms the traditional model by about 14%, 0.063 to 0.072. We also observed wide variations in improvements of CTR depending on the news domains. The gains are greater in the finance domain. In addition, we distilled a correlation between the number of impressions and the CTR in the AdContX model. Further studies will be focused on the sensitivity of AdContX components to internal key elements.

References

[1] W.T. Yih, J. Goodman, and V.R. Carvalho, "Finding Advertising Keywords on Web Pages," *Proc. 15th ACM WWW*, 2006, pp. 213-222.

[2] A. Broder et al., "A Semantic Approach to Contextual Advertising," *Proc. 30th ACM SIGIR*, 2007, pp. 559-566.

[3] D. Fain and J. Pedersen, "Sponsored Search: A Brief History," *Proc. 2nd Workshop Sponsored Search Auctions*, 2006.

[4] H. Raghavan and R. Iyer, "Evaluating Vector-Space and Probablistic Models for Query to Ad Matching," *Proc. 1st Workshop Info. Retrieval Advertising ACM SIGIR*, 2008, pp. 7-14.

[5] H.J. Oh, S.H. Myaeng, and M.G. Jang, "Enhancing Performance with a Learnable Strategy for Multiple Question Answering Modules," *ETRI J.*, vol. 31, no. 4, 2009, pp. 419-428.

[6] C.K. Lee and M.G. Jang, "A Modified Fixed-Threshold SMO for 1-Slack Structural SVMs," *ETRI J.*, vol. 32, no. 1, 2010, pp.120-128.

[7] C. Buckley, G. Salton, and J. Allan, "The Effect of Adding Relevance Information in a Relevance Feedback Environment," *Proc. 17th ACM SIGIR*, 1994, pp. 293-300.

[8] M. Hollis, "Ten Years of Learning on How Online Advertising Builds Brand," *J. Advertising Research*, vol. 45, no. 2, 2005, pp. 255-268.