Report on the Second KDD Workshop on Data Mining for Advertising *

Dou Shen
doushen@microsoft.com

Arun C. Surendran
acsuren@microsoft.com

Ying Li
yingli@microsoft.com

Microsoft adCenter Labs,
Redmond, WA 98074 USA

ABSTRACT

Following the success of our first workshop, we organized **ADKDD 2008** [1] - the second International Workshop on Data Mining and Audience Intelligence for Advertising, in conjunction with KDD 2008 at Las Vegas, Nevada, USA. This report is a summary of the workshop, including brief descriptions of the accepted papers.

1. INTRODUCTION

The past few years have seen a tremendous growth in online advertising. Especially, the last two years have seen significant changes in the advertising industry both in terms of business deals as well as new industry initiatives. In 2007 alone, Google bought ad serving company DoubleClick for $3.1B [12], Microsoft bought aQuantive for $6.1B [9] and AdECN [11], WPP snapped up 24/7 Real Media for about $649M [10] and Yahoo paid $680M to retain complete ownership of Right Media exchange [13]. Since ADKDD 2007, more deals have been announced especially in the area of targeted advertising - AOL bought Tacoda for $275M, Yahoo acquired Blue Lithium for $300M and Facebook announced their Beacon targeted advertising system. These, combined with the much publicised effort by Microsoft to buy Yahoo, made 2008 an exciting year for players in the online advertising space.

Apart from exciting business deals, online monetization has many challenging problems to solve. Hector Garcia Molina, in his keynote address at WSDM 2008 [18] listed internet monetization among the hardest and most impactful problems on the internet. Online advertising is a complex ecosystem involving multiple players, including advertisers and their agents (ad agencies), publishers and their aggregators (ad networks), ad exchanges and finally the end users. The primary channels for such online advertisement are paid search [5], content ads [3], display advertisement, ads on other media such as online video (e.g. YouTube) and advertisement on social networks. Delivering the right marketing messages to the right users in the right time is of great importance to the success of this ecosystem. To achieve this goal, the players have to successfully bring together a milieu of technologies from content understanding [3], query understanding, conducting optimal auctions [5], optimization etc. For example, the sponsored search engine has to understand a user's query and match it to an appropriate ad. The content based advertisement system has to match advertisements to page content. Both of them have to conduct optimal auctions, and build meaningful models of user behavior so that they can match advertisement to the user as opposed to just the content. The problem is tougher when the content to be matched is non-text media like video. Ad agencies need to design campaigns that optimize bids on hundreds (possibly thousands) of keywords; ad networks try to maximize the yield of publishers. Offline and online ad campaigns need learn from each other to provide synergy towards reaching advertiser's goals. Data mining and machine learning techniques, which mine patterns and learn from large scale data, provide effective solutions to the above problems. For further reading on state of the art technologies used in online advertising please refer to [6; 19; 21; 3; 22; 15; 4; 20] and last year's report on ADKDD 2007 [17]. One of the biggest challenges in online advertising is that it is diverse and siloed. There is a need for many of the practitioners, especially from both academia and industry to come together to share data and algorithms, to share their experiences, discuss and solve the cutting-edge research problems and resolve practical issues.

We addressed this need with our first workshop on data mining for advertising (ADKDD 2007) last year. Following its success, we organized the second in the series - ADKDD2008 - in conjunction with KDD 2008 to bring together data mining researchers and practitioners in online advertising.

The workshop attracted researchers from companies such as Google, Microsoft, Xerox PARC, Accenture Labs and educational instituitions such as Stanford University, NYU Stern School of Business, Hong Kong University of Science and Technology, University of Illinois at Urbana Champaign.

This year's invited speaker was Rayid Ghani from Accenture Labs, who talked about making ad system advertisement friendly. He specifically addressed two problems: First, he addressed scalability i.e. creating, customizing and placing ads for a large number of products. Specifically, he talked

*Workshop report on ADKDD 2008: the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising, held in conjunction with KDD 2008, The 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, held at Las Vegas, Nevada, USA, Aug 24-27 2008.

[1] http://adlab.microsoft.com/adkdd2008

about systems that automatically extract product attributes and attribute values from a product description. For example, given descriptions of many digital cameras, he details automatic ways to find out what are the attributes of a digital camera (e.g. megapixel, zoom, etc); now given a specific camera, he finds the values of these attributes (e.g. a Canon SD-550 has 7.1 megapixel and 3x zoom, etc) [7]. Extracting this information allows advertisers to quickly automate their keyword generation and campaign optimization. Second, he talked about systems which help advertisers achieve specific business goals with advertising. He showed systems that had been used in offline, in-store systems which can be relevant to online advertisement.

2. RESEARCH PAPERS

Seven research papers were accepted to ADKDD 2008. The topics cover several interesting research aspects in online advertising. Following sections give each paper a brief summary.

2.1 ROI Maximization in Sponsored Search

Understanding the empirical behavior of bidders (advertisers) in sponsored search auctions is important. First, it allows search engines to develop bidding tools, user interfaces and features that help advertisers achieve their goals. Second, the empirical investigation can guide theoretical modeling and analysis of these auctions. The paper with the title "*An Empirical Analysis of Return on Investment Maximization in Sponsored Search Auctions*" from Jason Auerbach, Joel Galenson, Mukund Sundararajan tries to understand the bidders' behavior in terms of whether advertisers are using strategies that maximize their return on investment (ROI) across multiple keywords in sponsored search auctions [1]. Since the testing of ROI maximization relies on knowledge of advertisers' private true values per click, the authors use some necessary conditions for ROI maximizing behavior which rely only on advertisers' bids. After classifying advertisers based on the extent to which they satisfy the test conditions, they conducted a set of analysis over Version 1.0 of *Yahoo! Search Marketing advertising bidding data*, which is provided as part of the Yahoo! Research Alliance Webscope program. Their results indicate that a significant number of advertisers bid almost the same on a large percentage of keywords; many of them rarely change the bids on their keywords. The final conclusion is that a large fraction of advertisers were unable to maximize their ROI.

2.2 Online Effects of Offline Ads

Online advertising and offline ads seem to be well separated. However, their impact on users' daily life is hard to distinguish. Clearly, online advertising can affect users' offline behaviors and vice versa. Diane Lambert and Daryl Pregibon's paper "*Online Effects of Offline Ads*" proposes a methodology for assessing how ad campaigns in offline media such as print, audio and TV affect online interest in the advertisers brand [16]. As Lambert and Pregibon suggest, online interest can be measured by daily counts of the number of search queries that contain brand related keywords, by the number of visitors to the advertisers web pages, by the number of pageviews at the advertisers websites, or by the total duration of visits to the advertisers website. An increase in outcomes like these in designated market areas (DMAs) where the offline ad appeared suggests heightened interest in the advertised product, as long as there would have been no such increase if the ad had not appeared. A robust regression analysis is put forward to estimate the effects of offline ads and a small print ad campaign illustrates the method. Their method is robust enough to account for different in seasonality and pre-campaign brand awareness across DMAs, which makes the measurement very effective.

2.3 Compare Performance Metrics in Organic Search with Sponsored Search

In the paper "*Comparing Performance Metrics in Organic Search with Sponsored Search Advertising*" [8], the authors Anindya Ghose and Sha Yang answer a question of how metrics for sponsored search advertising compares to organic search listings for the same keywords. They use a Hierarchical Bayesian modeling framework and estimate the model using Markov Chain Monte Carlo (MCMC) methods to analyse the effect of various factors. Their analysis suggests that if the factors are divided into keyword based and retailer based characteristics, most of the keyword-level characteristics have a stronger impact on these the performance of organic search than paid search. This could shed light on understanding what the most "attractive" keywords are from advertisers' perspective, and how advertisers should invest in search engine advertising campaigns relative to search engine optimization.

2.4 Personalized Online Commercial Intention

Understanding users' intention, especially their online commercial intention through their search queries is very important to online advertising. It can help search engines provide proper search results and advertisements; help Web users obtain the right information they desire; and help the advertisers make revenue from the potential transactions. Traditionally, systems use individual queries to infer a user's intention. In the paper, titled "*An algorithm for analyzing personalized online commercial intention*" from Derek Hao Hu, Qiang Yang, Ying Li, present an algorithm to detect users' personalized online commercial intention (POINT) [14] based on a skip-chain conditional random field model, which can comprehensively consider the evidences from the target query, the profile of the user issuing the query, as well as the semantic similarity of different queries in a personal query log. Experiments on a real search engine query log data shows that the new algorithm can improve the performance by 10% compared to the state-of-the-art baselines.

2.5 Consistent Phrase Relevance Measures

It is a fundamental problem to measure the relevance between a document and a phrase for online advertising, especially contextual advertising. In the paper "*Consistent Phrase Relevance Measures*" [23], Wen-tau Yih and Christopher Meek solve this problem by exploiting two approaches to provide consistent relevance scores for both in and out-of document phrases. The first approach is a similarity-based method which represents both the document and phrase as term vectors to derive a real-valued relevance score. The second approach takes as input the relevance estimates of some in-document phrases and uses Gaussian Process Regression to predict the score of a target out-of-document phrase. More details about these two approaches can be found in [23].

2.6 Variable Selection for Ad Prediction

Knowing the probability of a click for an advertisement can greatly improve user experience and advertiser revenue in online advertising. However, the probability of a click is usually a function of a large number of variables. Suma Bhat and Kenneth Church investigate a forward selection method to select a subset of variables to better predict the click probability in their paper "*Variable Selection for Ad Prediction*" [2]. Their forward selection method proceeds sequentially in a way that rewards a set of variables by how much information it provides regarding the outcome, but penalizes the set based on the number of variables in it.By using this method in the context of a logistic regression model, they can provide an estimate of the click-through-rate. Experimental results demonstrate the efficacy of their approach, even when compared to a brute force exhaustive search for variable subset selection.

2.7 Sponsored Ad-Based Similarity

The paper "*Sponsored Ad-Based Similarity: An Approach to Mining Collective Advertiser Intelligence*" is authored by Jessica Staddon. This paper presents a method for mining the intelligence of advertisers to detect product similarities and generate accurate recommendations. The basic assumption is that if object A and object B each lead to the display of sponsored ad C, then this is an indication of similarity between A and B. With this assumption, Staddon proposes a general framework for leveraging linked advertisements to detect object similarity. Experimental results show that the proposed approach yields useful product recommendations.

3. CONCLUSION

ADKDD 2008 - The Second International Workshop on Data Mining and Audience Intelligence for Advertising was conducted in conjunction with KDD 2008 in Las Vegas, Nevada, USA. Papers presented at this workshop addressed various challenging data mining and machine learning problems in advertising, including analysis of empirical bidding behaviors, study of the online effects of offline ads, comparison between organic search and sponsored search, personalized user commercial intention detection, relevance measurement between phrase and documents, advertisement click through rate prediction and so on. Participants in this workshop were from top industry and research labs around the world. ADKDD 2008, as we have expected, provided an excellent forum for researchers and industry practitioners in advertising to come together to exchange ideas on this fast growing business.

4. ACKNOWLEDGEMENTS

We thank everyone who submitted papers to ADKDD 2008. The high quality of the submissions enabled us to put together a strong technical program. We would like to express our sincere gratitude to all the program committee members for helping us put together a strong research program, and for their feedbacks and valuable suggestions. The program committee members include: Eugene Agichtein, Rayid Ghani, Tao Hong, Kartik Hosanagar, Rong Jin, Vanja Josifovski, Ramakrishnan Srikant, Ankur Teredesai, Michael Wellman, Qiang Yang, Yi Zhang. We thank all the participants of this workshop for making this a resounding success. We look forward to doing this again with KDD 2009 in Paris!

5. REFERENCES

[1] J. Auerbach, J. Galenson, and M. Sundararajan. An empirical analysis of return on investment maximization in sponsored search auctions. In *ADKDD'08: Proceedings of the Second International Workshop on Data Mining and Audience Intelligence for Advertising*, Las Vegas, Nevada, USA, 2008.

[2] S. Bhat and K. Church. Variable selection for ad prediction. In *ADKDD'08: Proceedings of the Second International Workshop on Data Mining and Audience Intelligence for Advertising*, Las Vegas, Nevada, USA, 2008.

[3] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 559–566, New York, NY, USA, 2007. ACM.

[4] H. K. Dai, L. Zhao, Z. Nie, J.-R. Wen, L. Wang, and Y. Li. Detecting online commercial intention (oci). In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 829–837, New York, NY, USA, 2006. ACM.

[5] D. C. Fain and J. O. Pedersen. Sponsored search: A brief history. In *Proceedings of the Second Workshop on Sponsored Search Auctions*, 2006.

[6] A. Fuxman, P. Tsaparas, K. Achan, and R. Agrawal. Using the wisdom of the crowds for keyword generation. In *WWW '08: Proceedings of the World Wide Web Conference 2008*, 2008.

[7] R. Ghani, K. Probst, Y. Liu, M. Krema, and A. Fano. Text mining for product attribute extraction. *SIGKDD Explor. Newsl.*, 8(1):41–48, 2006.

[8] A. Ghose and S. Yang. Comparing performance metrics in organic search with sponsored search advertising. In *ADKDD'08: Proceedings of the Second International Workshop on Data Mining and Audience Intelligence for Advertising*, Las Vegas, Nevada, USA, 2008.

[9] http://advertising.microsoft.com/news-events/microsoft-aquantive announcement.

[10] http://blog.searchenginewatch.com/blog/070517120022.

[11] http://www.microsoft.com/Presspass/press/2007/jul07/07-26AdECNPR.mspx.

[12] http://www.nytimes.com/2007/04/14/technology/14DoubleClick.html.

[13] http://www.variety.com/article/VR1117964019.html.

[14] D. H. Hu, Q. Yang, and Y. Li. An algorithm for analyzing personalized online commercial intention. In *ADKDD'08: Proceedings of the Second International Workshop on Data Mining and Audience Intelligence for Advertising*, Las Vegas, Nevada, USA, 2008.

[15] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 151–160, New York, NY, USA, 2007. ACM.

[16] D. Lambert and D. Pregibon. Online effects of offline ads. In *ADKDD'08: Proceedings of the Second International Workshop on Data Mining and Audience Intelligence for Advertising*, Las Vegas, Nevada, USA, 2008.

[17] Y. Li, A. C. Surendran, and D. Shen. Data mining and audience intelligence for advertising. *SIGKDD Explor. Newsl.*, 9(2):96–99, 2007.

[18] H. G. Molina. Web information management: Past, present and future. In *WSDM 2008*, 2008.

[19] H. Nazerzadeh, A. Saberi, and R. Vohra. Dynamic cost-per-action mechanisms and applications to online advertising. In *WWW '08: Proceedings of the World Wide Web Conference 2008*, 2008.

[20] D. Pregibon and D. Lambert. More bang for their bucks: Assessing new features for online advertisers. In *ADKDD'07: Proceedings of the First International Workshop on Data Mining and Audience Intelligence for Advertising*, San Jose, California, USA, 2007.

[21] F. Radlinski, A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, and L. Riedel. Optimizing relevance and revenue in ad search: a query substitution approach. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 403–410, New York, NY, USA, 2008. ACM.

[22] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: estimating the click-through rate for new ads. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 521–530, New York, NY, USA, 2007. ACM.

[23] W. tau Yih and C. Meek. Consistent phrase relevance measures. In *ADKDD'08: Proceedings of the Second International Workshop on Data Mining and Audience Intelligence for Advertising*, Las Vegas, Nevada, USA, 2008.