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 the participation, invited talk and referred 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 DoubleClick for $3.1B, Microsoft bought aQuantive for $6.1B and AdEcn, WPP snapped up 24/7 Real Media for about $649M and Yahoo paid $680M to get their remaining part of Right Media Inc. 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.

Hector Garcia Molina, in his keynote address at WSDM 2008 [11] mentioned internet monetization as the #3 on the list of hardest and the most important problems on the internet. Online advertising is a complicated ecosystem, which involves multiple players, including advertisers, publishers, end users and many others. 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, we have to intelligently understand the advertisements from advertisers, content from publishers and intents of the end users. Data mining techniques, which are to mine patterns and knowledge from large scale data, can provide effective solutions to the above problems in different types of online advertising, including sponsored search, contextual advertising, behavior targeting and so on [6; 12; 13; 3; 14; 9; 4].

*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

Thus there is continued interest in the need for a forum to bring together the different players in the advertisement industry, to help all the players, especially data mining researchers and practitioners from both academia and industry to share their experience in advertising, discuss and solve the cutting-edge research problems as well as practical issues.

Fain and Pedersen gave a brief survey of the history of sponsored search [5], which is to place advertisements on result pages from a web search engine. In this survey, Fain and Pedersen mention six basic elements of sponsored search, such as advertiser-provided content, bidding keywords, relevance verification, et al. Lot of research work has been conducted around these basic elements, which finally compose the whole picture of sponsored search. For example, in [6], Fuxman et al. work on bidding keyword generation. Given an advertiser who wants to launch a campaign, they propose an approach to suggest keywords related to that campaign. Their approach leverage the "wisdom of the crowd" by exploiting associations between queries and URLs, which are captured by the users' clicks. In [13], Radlinski et al. study the problem of selecting advertisements for sponsored search so that they are both relevant to the queries and profitable to the search engines. In [14], the authors study how to estimate the click-through rate of advertisement in the sponsored search, which is a key factor for advertisement ranking. Different from sponsored search, contextual advertising places advertisements within the content of a generic web page. In order to maximize revenue while improving user experience, it is crucially important to place relevant advertisements to the page content. In [3], Broder et al put forward a semantic approach to measure the relevance. They first classify the advertisements and a web page into an intermediate taxonomy. The classification results are used to calculate the "Semantic" Similarity between advertisements and web pages. Finally, they combine the "Semantic" Similarity and other syntactic features to determine the relevance score. For both sponsored search and contextual advertising, it is critical to understand the users' need, which can be inferred from the users' queries and demographics. Therefore, a group of research work has exploited this direction. Dai et al. predict users' commercial intent based their search queries [4], while Hu et al. study how to estimate users' demographic based on their browsing behaviors [9].

As we can see from the above examples, researchers have touched different aspects of the online advertising system. In 2007, we organized ADKDD 2007 to bring data mining

researchers and practitioners from both academia and industry to share their experience in advertising. To further this exchange and highlight advances in research, we organized the ADKDD 2008. ADKDD 2008 is featured with one invited talk and 7 referred paper presentations, which cover empirical analysis of ROI maximization, online effects of offline ads, user intention prediction, ad click and so on.

This year's invited speaker was Rayid Ghani from Accenture Labs, who talked about making targeted advertisement advertiser friendly. He specifically addressed two problems: (1) scalability i.e. creating, customizing and placing ads for a large number of products. In specific, he talked about systems to automatically extract product attributes and attribute values from a product description. For example, given a description of many digital cameras, is it possible to find out what the attributes of a digital camera are (megapixel, zoom, etc) and then find these attributes for a specific camera (SD-550 has 7.1 megapixel and upto 3x zoom, etc). (2) He also 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. PARTICIPATION

The workshop attracted researchers from various companies and research groups from various parts of the world - researchers from companies such as Google, HP, Microsoft and Yahoo! and countries like China, Germany, The Netherlands, Spain and USA participated in the workshop.

3. RESEARCH PAPERS

There are totally seven research papers presented at ADKDD 2008. The topics cover several hot research aspects in online advertising. Following sections give each paper a brief summary.

3.1 ROI Maximization in Sponsored Search

Understanding the empirical behavior of bidders (advertisers) in sponsored search auctions is important. Firstly, it allows search engines to develop bidding tools, user interfaces and features that help advertisers achieve their goals. Secondly, 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 maximizing 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 large fraction of advertisers may be following ROI-based strategies.

3.2 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 [10]. 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 regression analysis is put forward to estimate the effects of offline ads and a small print ad campaign illustrates the method.

3.3 Compare Performance Metrics in Organic Search with Sponsored Search

In the paper "Comparing Performance Metrics in Organic Search with Sponsored Search Advertising" [7], the authors Anindya Ghose and Sha Yang answer a question of how sponsored search advertising compares to organic listings with respect to predicting conversion rates, order values and profits from a keyword ad. They use a Hierarchical Bayesian modeling framework and estimate the model using Markov Chain Monte Carlo (MCMC) methods. Their analysis suggests that on an average while the conversion rates, order values and profits from paid search advertisements were much higher than those from natural search, most of the keyword-level characteristics have a statistically significant and stronger impact on these three performance metrics for 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.

3.4 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 revenure from the potential transactions. Traditionally, people users' individual queries to infer users' intention. In the paper, titled "An algorithm for analyzing personalized online commercial intention" from Derek Hao Hu, Qiang Yang, Ying Li, an algorithm POINT is put forward to detect users' personalized online commercial intention [8]. This algorithm is 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 similarity of different queries in a personal query log. Experiments on a real search engine query log data shows that POINT can improve the performance by 10% compared to the state-of-the-art baselines.

3.5 Consistent Phrase Relevance Measures

It is a fundamental problem to measure the relevance be-

tween a document and a phrase for online advertising, especially contextual advertising. In the paper "Consistent Phrase Relevance Measures" [15], 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 [15].

3.6 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.

3.7 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.

4. 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.

5. 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 finishing the reviews in a very short time, as well as 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 for KDD 2009 in Paris!

6. 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] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.