Data Mining and Audience Intelligence for Advertising *

Ying Li
yingli@microsoft.com

Arun C. Surendran
acsuren@microsoft.com

Dou Shen
doushen@microsoft.com

Microsoft adCenter Labs,
Redmond, WA 98074 USA

ABSTRACT

Growth in the global advertising industry - especially the recent rapid growth in online advertising - has generated large volumes of data, bringing along with it many challenging data mining problems. Researchers from various disciplines have brought their expertise to solve these exciting problems, leading to a plethora of novel applications and new algorithms. We strongly felt that we needed a forum where data mining researchers and practitioners, from both academia and the industry, could come together to share their experience on advertising. To this end, we organized **ADKDD 2007** [1], the First International Workshop on Data Mining and Audience Intelligence for Advertising, in conjunction with KDD 2007 at San Jose, California, USA. In this report, we will present a summary of the workshop.

1. INTRODUCTION

Global advertising is projected to exceed half-a-trillion dollars by the year 2010 [20]. Although online advertising is currently only a small part of this large enterprise, it is growing at a rapid pace [27]. The explosion in the number of participants in the online advertising marketplace has generated large volumes of data and exciting data mining problems. Earlier research on search logs, web pages, social network and blogs had focused on information organization, retrieval and understanding [2; 7; 13; 17; 16; 22; 25]. Recently there has been strong research interest in the advertisement angle to all these information sources. Researchers have tackled several challenging problems on online monetization like sponsored search [1; 12], contextual advertising for web pages [4; 14; 30], understanding user intent [6] and user demographics [9] for advertisements, mining user reviews for product pricing [3], predicting click-through rates for ads [24], just to name a few. Further, the on-line and offline advertising worlds are fast converging; for example, digital marketplaces are migrating from the online world to TV and radio [8], and audience understanding work from offline media is trickling into the online realm [9]. Since data mining researchers and practitioners in all these areas come from different communities, there is strong need for a single forum to bring together people involved in all aspects of digital advertising. We are addressing this need with the First Workshop on Data Mining and Audience Intelligence for Advertising.

The goal of this workshop is to not only increase communication between researchers working on seemingly different pieces of the advertisement pie, but to encourage data mining researchers to bring new ideas from related areas to solve the numerous challenges faced by the rapidly changing digital advertising industry. We want to bring together auction theorists, social network researchers, natural language researchers, information retrieval experts, audience understanding researchers, television advertisement analysts and many others, to promote a fruitful exchange of ideas to advance the field.

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. The topics covered varied from auction models, content understanding and matching for advertisements, understanding user attention, sentiment classification and opinion mining, to combining online and TV advertisements.

3. INVITED TALK - SOCIAL CHALLENGES IN ADVERTISING

Much of advertisement today is focused on "audience intelligence" - on what the advertisers and the data mining researchers know about about the user. We felt that it was very important for the practitioners to be aware of the converse of audience intelligence - the intelligence of the audience. More specifically we wanted to know: *what do the users know when it comes to data mining and data use?*. We invited Professor Joseph Turow, Robert Lewis Shayon Professor of Communication at the University of Pennsylvania's Annenberg School For Communication, to talk about this. Prof. Turow is a well renowed expert on public policy on "database marketing". He is also the author of nine books, among which is the well known *Niche Envy: Marketing Discrimination in the Digital Age"* (MIT Press, 2006).

Prof. Turow's talk was titled "Social Challenges of Data Mining Advertising" and addressed the challenges that pub-

*Workshop report on ADKDD 2007: the First International Workshop on Data Mining and Audience Intelligence for Advertising, held in conjunction with KDD 2007, The Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, held at San Jose, CA, Aug 12-15 2007.

[1]http://adlab.microsoft.com/adkdd2007

lic knowledge (or the lack of it) and perceptions raise for database advertisers and the intermediaries (such as Microsoft, Yahoo and Google). Prof. Turow outlined societies perception using several national surveys of the internet-using public and explained the evolution of these opinions from the standpoint of the ad industry's history. He tracked the changes in targeted marketing from the time it was practiced by the friendly street corner grocer, to the current practices by retailing and media regimes who use logic based on data mining. His conclusion is that as the data gets bigger and technology begins to play a bigger role, the trend in business practice is moving towards identifying and targeting niche groups.

His surveys showed that although people suspect that their data is being used (e.g. about 80% of users know that companies can track their behavior across web sites), they don't know the rules of the marketplace when it comes to using this data (e.g. 64% did not realize that supermarkets are allowed to sell other companies information on what they buy) [28]. They are also bothered by the idea of "price discrimination" i.e. they do not like that that someone else can be targeted to pay less than them for the same product. The lack of transparency makes them feel vulnerable and encourages suspicion and anger at marketers, media and even the government (only 35% trust the government to protect them from marketers who misuse their information [28]).

Prof. Turow concluded his talk by advising the industry to promote transparency (increased user control over data), limit the amount/time information stored, and solicit aspirational data i.e. find out what a person wants to be and how would they like to be treated. He referred to data miners as the new "story tellers" of the society - people who will soon have enough data to track the entire life of individuals. He encouraged data miners to take this responsibility seriously and help users reach their aspirations.

Prof. Turow's talk was well received. In fact, much of the discussion after the workshop was centered around this dilemma of social acceptance of targeted advertisements.

4. RESEARCH PAPERS

Ten research papers, split into four sessions, were presented at ADKDD 2007. The themes of the four sessions and the papers presented in them are briefly described in this section. The four sessions approximately covered four broad topics in advertising. The first covered advertising on web pages, also known as contextual advertising. The second focused on sponsored search. The third was on technologies that drive advertising on Web 2.0. The last session was focused on using techniques and principles from online advertising to content from other media such as TV.

4.1 Session 1: Understanding Content & User Attention on a Web Page

This session was focused on technologies for advertising on web pages - content filtering and targeting.

One of the problems in advertising is to decide what to show and when to show it, so that the user's attention is best captured. The first paper in this session focused on this interplay between user attention and information on a web page. The paper titled "*The Economics of Attention: Maximizing User Value in Information-Rich Environments*" was presented by Bernardo Huberman and Fang Wu from HP Labs [10]. The basic hypothesis of the paper is that in an information-rich environment, there is competition for a user's attention. Treating the user's attention as a limited resource, the paper models the problem as a dual-speed restless bandit problem and presents an automatic mechanism that generates the most relevant information to be presented to users. The proposed solution in this paper guarantees to maximize the users' total expected utility from the information they receive.

The core problem of content-targeted advertisement is to accurately match the content of a page to the sparse content of advertisements. The second paper in this session addressed this issue. It was presented by Vanessa Murdock, Massimiliano Ciaramita and Vassilis Plachouras from Yahoo! Research Barcelona, and was titled "*A Noisy Channel Approach to Contextual Advertising*" [18]. The system assumes that each web page has been provided with several ad candidates of varying quality. Their goal is to re-rank the candidates so that the best ads are at the top of the list. This paper present a language independent machine learning system that re-ranks ad candidates based on a noisy-channel model using features derived from machine translation technologies. The system is validated through the experiments on a large number of advertisements appearing on real web pages.

The last paper in this session addresses an important problem in online advertising - one of *filtering out* unwanted content. Specifically, how to detect whether a publisher web page contains content that is inappropriate for showing advertisement(s) on it. The paper is titled "*Sensitive Webpage Classification for Content Advertising*" and was presented by Xin Jin, Ying Li, Teresa Mah and Jie Tong, from Microsoft adCenter Labs [11]. The paper presented a classification based approach to this problem. First, it presented a hierarchical sensitive content taxonomy that was customized for this task. Next, it outlined an iterative training procedure based on active learning to learn from a large collection of labeled and unlabeled data. Finally, the paper presented experimental results to compare the performance of classifiers like SVMs and logistic regression on this task.

4.2 Session 2: Sponsored Search

The second session of the morning addressed two diverse problems in the practice of placing advertisement next to the search results (also known as "sponsored search").

The first paper explored an alternative to the "pay-per-click" model that is popular today. The paper was titled "*Pay-per-Action Model for Online Advertising*", and was co-authored by Mohammad Mahdian and Kerem Tomak from Yahoo! Research, Santa Clara [15]. They discussed the challenges involved in designing a "pay-per-action" business model, where payment is made based on user *conversion* or purchase. Although such a model is a natural progression from the two dominant business models: the pay-per-impression model and the pay-per-click model, and has been discussed often in the advertising industry, it is not widely used yet. This paper discussed the challenges faced in implementing such a system - for example the challenge of encouraging advertisers to invest in tracking and disclosing conversions truthfully to the auctioneer - and presented mechanisms that can be implemented to overcome these problems. For example, when the conversion rate becomes part of the mechanism to rank ads, advertisers with higher conversion

rates tend to be ranked higher than others with the same bid, and this encourages truthful reporting.

The next paper was on maximizing advertiser return on investment (ROI), specifically about how to measure the effectiveness of methods implemented to increase ROI. For example, online search systems that display ads continually offer new features that advertisers can use to fine-tune and enhance their ad campaigns. However, an important question is whether a new feature actually helps advertisers or not. This is further complicated by biased sampling (using whitelisted advertisers). The paper in this session entitled "*More Bang for Their Bucks: Assessing New Features for Online Advertisers*" from Daryl Pregibon and Diane Lambert from Google address this problem [19]. The authors introduce metrics and ways to correct for biases due to the selection process, which allow the system to make valid inferences from whitelist trials about the effects of a new features on advertiser happiness.

4.3 Session 3: Sentiment Classification, Opinion Mining and Analyzing Information Flow in Blogs

Social aspects of advertising are becoming increasingly important. Web 2.0 sites like blogs and review sites are important areas for targeted advertisement [3].

For example, understanding the sentiment of a blog or a review is important in determining whether to target an ad to its content. For example, a digital camera manufacturer may not want to advertise on a site that has a negative review of its product. The first paper in this session authored by Stephan Raaijmakers from TNO, Netherlands - "*Sentiment Classification with Interpolated Information Diffusion Kernels*" - addresses exactly this problem [23]. In this paper, Raaijmakers presents a novel approach to global sentiment classification using information diffusion kernels. Diffusion kernels are similarity metrics in non-Euclidean information spaces, which have been found to produce state of the art results for document classification. Through extensive experiments on a well-known movie review data set, Raaijmakers concludes that interpolation of unigram and bigram information is beneficiary for sentiment classification.

The second paper "*Extracting Opinion Topics for Chinese Opinions using Dependence Grammar*" authored by Guang Qiu, Kangmiao Liu, Jiajun Bu, Chun Chen and Zhiming Kang from Zhejian University, China, addresses the twin problem of sentiment mining - to find the object of the sentiment or opinion [21]. For example, a review blog on a digital camera may praise one feature, while panning another. How is an advertiser to know automatically, which features are liked by the reviewers and which are not? In this paper, the authors propose a rule-based approach to extracting topics from opinion sentences by assuming the sentences are identified from texts in advance. They build a sentiment dictionary and define several rules based on the syntactic roles of words using the Dependence Grammar which is more suitable for Chinese natural language parsing. The experiments in the paper validate the effectiveness of their proposed solution.

The last paper in this session is "*Discovering Information Diffusion Paths from Blogosphere for Online Advertising*" by Avare Stewart, Ling Chen, and Raluca Paiu Wolfgang Nejdl [26] from L3S, Germany. The authors propose to discover information diffusion paths from the blogosphere to track how information flows from blog to blog. This is useful for determining the most effective places to advertise in the blog world. Their method first analyzes the content of blogs to detect trackable topics. Then they model a blog community as a blog sequence database, and formalize and solve the problem of discovering diffusion paths as one of frequent pattern mining. Experiments conducted on real life dataset show that their algorithm can discover the information diffusion paths efficiently.

4.4 Session 4: TV and Other Broadcast Content, and their Relation to Online Advertisement

The final session focused on two diverse topics. The first paper was about targeting online transcripts of TV news. Traditional content advertising has been targeted for web pages, and has relied on keyword extraction [30]. However, the existing technologies cannot be easily applied to find keywords from online broadcasting content, which usually contain more specific phrases and wordings in certain communities than in general Web-page content. "*Finding Keyword from Online Broadcasting Content for Targeted Advertising*" by Hua Li, Duo Zhang, Jian Hu, Hua-Jun Zeng and Zheng Chen from Microsoft Research Asia, presented a sequential pattern mining-based method to discover language patterns from online broadcasting content. Starting with selected keyword seeds, and by iteratively applying the language pattern mining and keyword extraction steps, the proposed technique avoids any tedious labeling work for this task. Experiments on some real-world data show that the proposed keyword extraction algorithm can significantly outperforms some baseline methods.

The second and last paper in this session is "*From TV to Online Advertising: Recent Experience from the Spanish Media*" authored by Jorge Sueiras, Fausto Morales and Juan-Carlos Ibanez from Neo Metrics in Spain. The advance of the Internet as a competitor with traditional media (radio, TV, newspapers and magazines) has attracted many advertisements. However, the traditional analytical tools for media planning may not be directly applicable on online advertising. In this paper, the authors describe their experience in the Spanish TV domain and its evolution into the Internet arena.

5. CONCLUSION

ADKDD - The First International Workshop on Data Mining and Audience Intelligence for Advertising was conducted in conjunction with KDD 2007 in San Jose, CA. Papers presented at this workshop addressed various challenging data mining and machine learning problems in advertising. Papers covered a wide variety of topics from pay-per-action business models for sponsored search, to effective targeting for content-based advertisement; from modeling user attention to classifying user sentiment, and from online advertisement to TV advertisement. Our invited speaker addressed the crucial issue of social challenges created by targeted advertisements. Participants in this workshop were from top industry and research labs around the world. Overall, as the first workshop of its kind in the data mining community, ADKDD 2007 was an excellent forum for researchers and industry practitioners in advertising to come together to exchange ideas on this fast growing business.

6. ACKNOWLEDGEMENTS

We thank everyone who submitted papers to ADKDD 2007. 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: Zheng Chen, Xiaoming Jin, Vanja Josifovski, Rajan M. Lukose, Mohammed Mahdian, Nitin Sharma, Guirong Xue, Yunhong Zhou. We thank all the participants of this workshop for making this a resounding success. Our sincerest thanks go to the organizers of KDD 2007, especially Prof. Qiang Yang, the Workshop Chair. We look forward to doing this again for KDD 2008 in Las Vegas!

7. REFERENCES

[1] G. Agarwal, K. Hosnagar, D. Pennock, M. Schwarz, and R. Vohra, *Third Workshop on Sponsored Search Auctions*, Banff, Canada, 2007.

[2] M. Banko, M. J Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open Information Extraction from theWeb," in *IJCAI '07*, Hyderabad, India, 2007.

[3] N. Archak, A. Ghose, and P. Ipeirotis, "Show me the money: Deriving the Pricing Power of Product Features by Mining Consumer Reviews", *KDD '07*, San Jose, CA, 2007.

[4] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In *SIGIR '07* pp. 559–566, New York, NY, USA, 2007.

[5] J. J. Carrasco, D. Fain, K. Lang, and L. Zhukov. Clustering of bipartite advertiser-keyword graphs. Workshop on Large Scale Clustering at IEEE International Conference on Data Mining, 2003.

[6] H. Dai, L. Zhao, Z. Nie, J.-R. Wen, L. Wang and Y. Li, Detecting Online Commercial Intention, *WWW '06*, Edinburgh, Scotland, 2006.

[7] K. Fujimura, T. Inoue, and M. Sugisaki. The eigenrumor algorithm for ranking blogs. *WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem*, 2005.

[8] "Google Announces TV Ads Trial", http://www.google.com/intl/en/press/annc/tv_ads_trial.html

[9] J. Hu, H-J., Zeng, H. Li, C. Niu, and Z. Chen, Demographic Prediction based on User's Browsing Behavior, *WWW '07*, Banff, Canada.

[10] B. Huberman and F. Wu. The economics of attention: Maximizing user value in information-rich environments. *ADKDD '07* San Jose, CA, USA, 2007.

[11] X. Jin, Y. Li, T. Mah, and J. Tong. Sensitive webpage classification for content advertising. *ADKDD '07* San Jose, CA, USA, 2007.

[12] A. Joshi and R. Motwani. Keyword generation for search engine advertising. *ICDMW '06: Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops*, pp. 490–496, Washington, DC, USA, 2006.

[13] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *WWW '03*, pages 568–576, New York, NY, USA, 2003.

[14] A. Lacerda, M. Cristo, M. A. Gonçalves, W. Fan, N. Ziviani, and B. Ribeiro-Neto. Learning to advertise. *SIGIR '06* pp. 549–556, New York, NY, USA, 2006.

[15] M. Mahdian and K. Tomak. Pay-per-action model for online advertising. *ADKDD'07* San Jose, CA, USA, 2007.

[16] C. D. Manning, P. Raghavan, and H. SchÜZe, *Introduction To Information Retrieval*. Cambridge University Press, 2007.

[17] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. *WWW '06*, Edinburgh, Scotland, 2006.

[18] V. Murdock, M. Ciaramita, and V. Plachouras. A noisy channel approach to contextual advertising. *ADKDD'07* San Jose, CA, USA, 2007.

[19] D. Pregibon and D. Lambert. More bang for their bucks: Assessing new features for online advertisers. *ADKDD '07* San Jose, CA, USA, 2007.

[20] "Global Entertainment and Media Outlook: 2007-2011", PriceWaterhouseCoopers Report, January 2007

[21] G. Qiu, K. Liu, C. C. Jiajun Bu, and Z. Kang. Extracting opinion topics for chinese opinions using dependence grammar. *ADKDD '07* San Jose, CA, USA, 2007.

[22] Y. Qiu, H.-P. Frei, Concept based query expansion, *SIGIR '93* pp. 160–169, Pittsburgh, PA, 1993.

[23] S. Raaijmakers. Sentiment classification with interpolated information diffusion kernels. *ADKDD '07* San Jose, CA, USA, 2007.

[24] M. Richardson, E. Dominowska, and R. Ragno Predicting Clicks: Estimating Click-through Rate for New Ads *WWW '07*, Banff, Canada, 2007.

[25] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen, Building bridges for web query classification, *SIGIR '06* pp. 131–138, Seattle, WA, 2006.

[26] A. Stewart, L. Chen, R. Paiu, and W. Nejdl. Discovering information diffusion paths from blogosphere for online advertising. *ADKDD '07* San Jose, CA, USA, 2007.

[27] TSN Media Intelligence Report, 2007, http://www.tns-mi.com/news/01082007.htm

[28] "Americans and Online Privacy: The System is Broken", A Report from the Annenberg Public Policy Center of the University of Pennylvania, by Joseph Turow, 2003.

[29] "Open to Exploitation: American Shoppers Online and Offline," Report of the Annenberg Public Policy Center, by Joseph Turow, June 2005.

[30] W. Yih, J. Goodman and V. R. Carvalho, Finding Advertising Keywords on Web Pages, *WWW '06*, Edinburgh, Scotland, 2006.