Targetted Online Advertising Based on Historic User Trails

Udayan Khurana, Tak Yeon Lee*
Department of Computer Science
University of Maryland
College Park, MD 20742
udayan,tylee@cs.umd.edu

Abstract

In this paper we address the problem of finding the most suitable banner display advertisement option for a user given his/her current browsing session. Using the historical browsing session information for an advertisement campaign, we mine the association of different ad views, engagements and clicks. Using a probabilistic model, we find the likelihood of an ad to be clicked given a specific set of events that describe the user session. The major challenge in training the model for optimum precision is the sparsity of data ($< 0.5\%$ Click through rate) and we propose the use of ad engagement as a success event like ad click to train the model more effectively. Our results show a good click through conversion probability on test data.

1 Introduction

Online advertising has become a popular means of running marketing campaigns. The major reason for this is attributed to the ability to interactively interface with the user [3]. In a non-interactive setting like television advertising or the ones involving other media like news papers, magazine, physical banners or billboards, it is hard to determine the effectiveness of an ad campaign. However, in case of online advertising, because of the feedback that can be channeled back to the advertiser which creates an opportunity for the latter to tune the campaign to get maximum efficiency. The success of an advertisement can be measured by different levels of user interest. For example, a user click is considered to be a measure of greater success than a view. Other measures of success like engagement and conversion are discussed later.

Understanding the impact of an advertisement or a campaign on a user is not easy. Studies show [9][10] that not only clicks, but advertisement views also have an impact on overall sales for a brand. In this paper, we will not not attempt to identify or trace the impact of a specific advertisement on a user, but we will try to maximize the chances of a user clicking or engaging with an ad. Let us introduce the terminology that is going to be used in the rest of this paper:

- **AdPod** is a group of advertisements (Recipes) containing similar contents in one campaign. See Figure1.
- **Recipe** is a unit ad that can be placed any where in the screen. Ad campaign or Campaign is a set of AdPods shown during a fixed time period for a marketing campaign, e.g. Father's day campaign. See Figure1.
- **View or Impression** is considered to be the activity of an adpod being shown to a user. When we say view, it will imply that it was not an engagement or a click. Engagement is a type of event that user hovering mouse cursor over an ad, or interacting it in any way other than a click to redirect.
- **Session** is a set of actions done on one or more recipes for a given user in a fixed window of time.

*Neha Gupta worked together as her PhD project

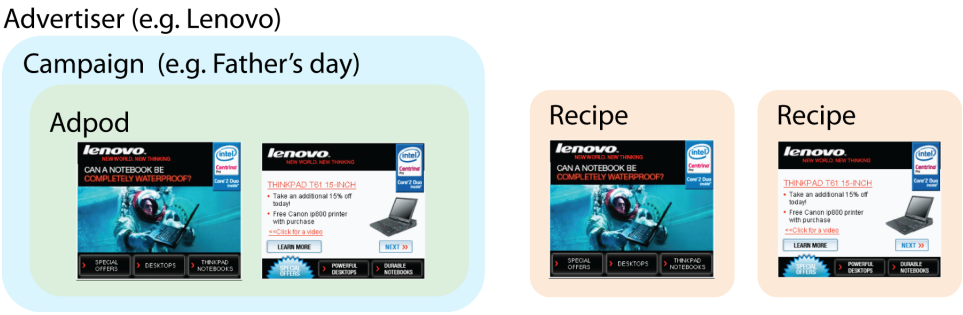

Figure 1: Hierarchy of Online Advertisement ID

- **Browsing Context** represents the state of the user with respect to ads that have been viewed/engaged or clicked by him/her during a session.

In this paper, we present a system that uses past information on a campaign's usage data to predict the choice of the most suitable ad for any user in a given context. More precisely, we aim to identify the choice of the user by the advertisements already seen, engaged or clicked by the user. And based upon that, we would suggest the ad that is most likely to be clicked if shown next. In the following section, we discuss related work in the area and our approach in the next. In section 4 and 5 we talk about our implementation and results and finally the conclusion.

2 Related Work

The problem of ad optimization is closely linked with the problem of online user behavior modeling. Various Markov models have been proposed to capture the behavior of user sessions [6][8]. Different models differ in the way they model links between events or whether they work towards the notion of a successful session or not. In online ad revenue maximization, different approaches have been tried, for example [1], is based on predicting the optimum time for an particular ad to be shown to a specific user. [7] works on creating decision rules to optimize the click through rate via optimum ad selection in a given situation and models like [4], or Google Ads use content information on the host page, like search query in order to give a more meaningful suggestion. More session based approaches that use propagation algorithms on graphs for user behavior modeling have been proposed in the recent past [5][2]. [3] specifically addresses the problem of optimum ad selection by traversing through aggregated user trails. [11] suggests useful data mining techniques for frequent episodes in event sequences.

3 Approach

We sessionize the raw user action records to form a chronological chain of events for each user. A typical session can be pictorially represented as:

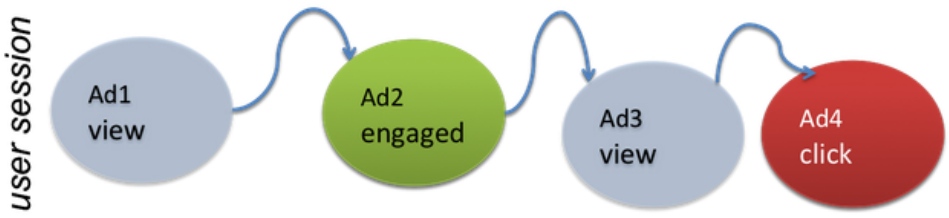

Figure 2: Typical User Session

The session above describes that a particular user viewed ad1, then engaged with another ad2, then viewed ad3 before finally clicking ad4. Based upon this information we say that given that the ad4 was clicked, ad1 was viewed in the same session. If C is the count, then,

$$C_{user_x}(ad1_{view}|ad4_{click}) = 1$$

This count is conditioned on a clicked ad, because we will be counting such instances only for the sessions where ad4 was clicked. The same can be said for Ad2 being engaged with and ad3 being viewed. By counting all such occurrences over the training data set, we can say that,

$$\begin{aligned} C(ad1_{view} \mid ad4_{click}) &= \text{Number of sessions where ad1 is viewed and ad4 is clicked} \\ C(ad4_{click}) &= \text{Number of sessionswhere ad4 is clicked} \\ P(ad1_{view} \mid ad4_{click}) &= \frac{C(ad1_{view} \mid ad4_{click})}{C(ad4_{clicked})} \end{aligned}$$

Now, our prediction engine needs to be able to suggest which ad is the one most likely to be clicked, given other set of ads that have been shown to the user. So, we want to determine $P(ad4_{click}|ad1_{view})$ for each such adpod ad4. Using Naive Bayes,

$$P(ad4_{click} \mid ad1_{view}) = \frac{P(ad1_{view} \mid ad4_{click}) \times P(ad4_{click})}{P(ad1_{view})}$$

Now, given ad1, the quantity above on the right hand side is independent of the denominator for different choices of ad4, i.e. $P(ad1_{view})$ is same for all instance of ad4. Thus, we reduce our problem to the following expression to find ad4 such that,

$$\arg\max_{ad4} P(ad1_{view} \mid ad4_{click}) P(ad4_{click})$$

To summarize, we are trying to find the best suited ad for a user given the prior knowledge that he/she has seen ad1. We define the context here to be the ad1. If we are given a context of three ads as in the figure above, we can write the optimization function as:

$$\arg\max_{ad4} P(ad1_{view} + ad2_{engaged} + ad3_{view} \mid ad4_{click}) P(ad4, click)$$

In theory, if the context is greater, a model should become more discriminative. But, in reality, a larger context means requirement of more training examples to get a good classifier. In order to get the best accuracy of prediction with such a model, there is a trade-off on the size of context that should be used given the size of a training set. If we were to make a simplifying assumption of the mutual independence of occurrence of events here, i.e. ad1 view is independent of ad2 engagement, which is independent of ad3 view, and so on, we can rewrite the expression for finding the best recipe or adpod, ad4 as:

$$\begin{aligned} &\arg\max_{ad4} P(ad1_{view} + ad2, engaged + ad3_{view} \mid ad4_{click}) \times P(ad4_{click}) = \\ &\quad \arg\max_{ad4} P(ad1_{view} \mid ad4_{click}) \times P(ad2, engaged \mid ad4_{click}) \times P(ad3_{view} \mid ad4_{click}) \times P(ad4_{click}) \end{aligned}$$

Note that this is a strong independence assumption. Whether this should hold true or not depends on order in which the ads were shown in the first place and whether it had an impact on the user decision to click or not. Oblivious of the relevant information needed to reason this independence and the complex psychological impact on the user, we will verify this assumption by verifying if the model is consistent outside its training examples.
Another option here is not to make a complete independence assumption, saying that successive events are not independent of each other but all events except those immediately before or after are independent of an particular event. So, in this case, we can say that ad1 view is independent of ad3 view but not ad2 engagement. The objective function of this model can be approximated as:

$$\arg\max_{ad4} P(ad1_{view} + ad2, engaged + ad3_{view} \mid ad4_{click}) \times P(ad4, click) =$$
$$\arg\max_{ad4} P(ad1_{view} + ad2, engaged \mid ad4_{click}) \times P(ad2, engaged + ad3_{view} | ad4_{click}) \times P(ad4, click)$$

The feature space in this model is lesser than that required in the model without any independence assumption, but more than that with a complete Independence assumption. In our experiments, we will evaluate the models with complete independence assumption and the one with only one degree of dependence.
In spite of the independence assumptions, data sparsity in online advertising is still an issue, as the click through rate is extremely low. To use a richer training set, we consider the case of an engagement being a successful event like a click, with lesser confidence though. Based on the earlier discussion [B] [C], where we say that a click is not the only measure of influence on the user, we experiment with the assumption of an engagement being partially effective as a click. As we will see in the results section, this increases greatly improves our training data size.

4 Implementation

In the following subsections, we provides details of data processing, details of the experiment and the results obtained.

4.1 Data Processing

Raw data from Tumri is a tab delimited log file that contains following headers: (* items are used)

```
*USERID    ADVERTISERID ADVERTISERNAME TIMESTAMP    CAMPAIGNNAME CAMPAIGNID
ADPODNAME    *ADPODID ADTYPE RECIPENAME *RECIPEID LOCATIONNAME LOCATIONID
EXTERNALADID EXTERNALBUYID EXTERNALPAGEID EXTERNALSITEIEXTERNALCREATIVEID
COUNTRY    STATE *EVENTLOGTYPE
```

First, all the ads under a single CAMPAINNAME (e.g. ThinkPad ChristmasSale) are selected and sessionized by USERID so that each data point represents historical trail of a unique user. As the data has been recorded in a single day, we treated one users activities as a single session.
Among various types of ID, ADPODID and RECIPEID were chosen because they both consists main hierarchy of ads in a campaign. Each ADPOD has a set of RECIPEIDs, and they are mutually exclusive across different ADPODs.

EVENTLOGTYPE represents how user reacted to the ad with three values:

- 0 CLICKS : user clicked the ad
- 1 ENGAGEMENTS : mouse pointer moved around the ad
- 2 IMPRESSIONS (VIEW) : ad just shown on screen

Basic stats of the data set is shown below. As mentioned earlier, a problematic issue is the rarity of positive records (CLICK events $< 0.1\%$ of entire records). A good new is the large number of Engagements over 10% of entire records. In Figure 3 shows most ads are clicked less than 100 times and reasonably well-distributed.

```
# of records : 2,652,799
# of RECIPEID : 105
# of ENGAGEMENTS : 260433
# of CLICKS : 2517
```

4.2 Experimental Setup

The goal of the experiment is to find the best configuration on the test dataset and generalize how each parameter affects on performance. We chose four parameters of extracting features as described below:

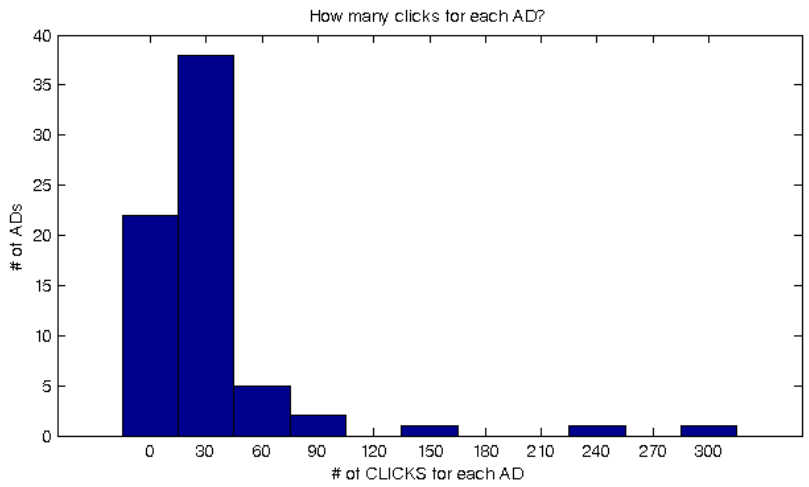

Figure 3: Distribution of Clicks For Each Advertisement

- **ID of Advertisement** (RecipeID | AdPodID)
  Both IDs are available for each ad, however, the total number of RecipeID is three times larger than AdPodID. It may deteriorate prediction accuracy.
- **Feature space** (1-ad | 2-ads | 1&2-ad)
  From a sequence of ads, we can use either a single ad or multiple consecutive ads as a feature. Using 2-ads is more expressive than 1-ad, however, it can make training set sparser. To compensate it, Fall-back method, using 1-ad feature when 2-ads feature is not available, is applied.
- **Notion of Success** (Click event only | Click + Engage)
  When Click event is too rare to generate an accurate model, Engage (users mouse activity on the ad.) can be interpreted as a measure of partical success. To be fair, different weights (Click:5, Engage:1) are assigned.
- **Action Code** (Use Action Code | Do not use)
  Each ad in the feature set has either View or Engage action code. Using action codes may provide more expressive model, however, it can make training set sparser.

24 (2*3*2*2) training sets were generated. Below is two examples of feature extraction.

- **ORIGINAL SESSION DATA**
  (AdPod1,Recipe1)view (AdPod1,Recipe2)view (AdPod2,Recipe3)engage (AdPod2,Recipe4)view (AdPod2,Recipe3)click
- **CONFIGURATION A**
  ID Type of Ads:*RecipeID*, Feature space:*1-ad*, Notion of success:*Click + Engage*, Action code:*Use*

$$P(Recipe1_{view} \mid Recipe3_{success}) = 1 + 5 = 6$$
$$P(Recipe2_{view} \mid Recipe3_{success}) = 1 + 5 = 6$$
$$P(Recipe3_{engage} \mid Recipe3_{success}) = 5$$
$$P(Recipe4_{view} \mid Recipe3_{success}) = 5$$

- **CONFIGURATION B**
  ID Type of Ads:*AdPodID*, Feature space:*2-ad*, Notion of success:*Click only*, Action code:*Do Not Use*

$$P(AdPod1 + AdPod1 \mid Recipe3_{success}) = 5$$
$$P(AdPod1 + AdPod2 \mid Recipe3_{success}) = 5 + 5 + 5 + 5 = 20$$
$$P(AdPod2 + AdPod2 \mid Recipe3_{success}) = 5$$

When extracting features from session data, 4-folds technique is used in order to spare 25% of data points as testing set. For unseen events whose probabilities are zero, a small smoothing constant ($\lambda = 0.01$) for regularization.

5 Result

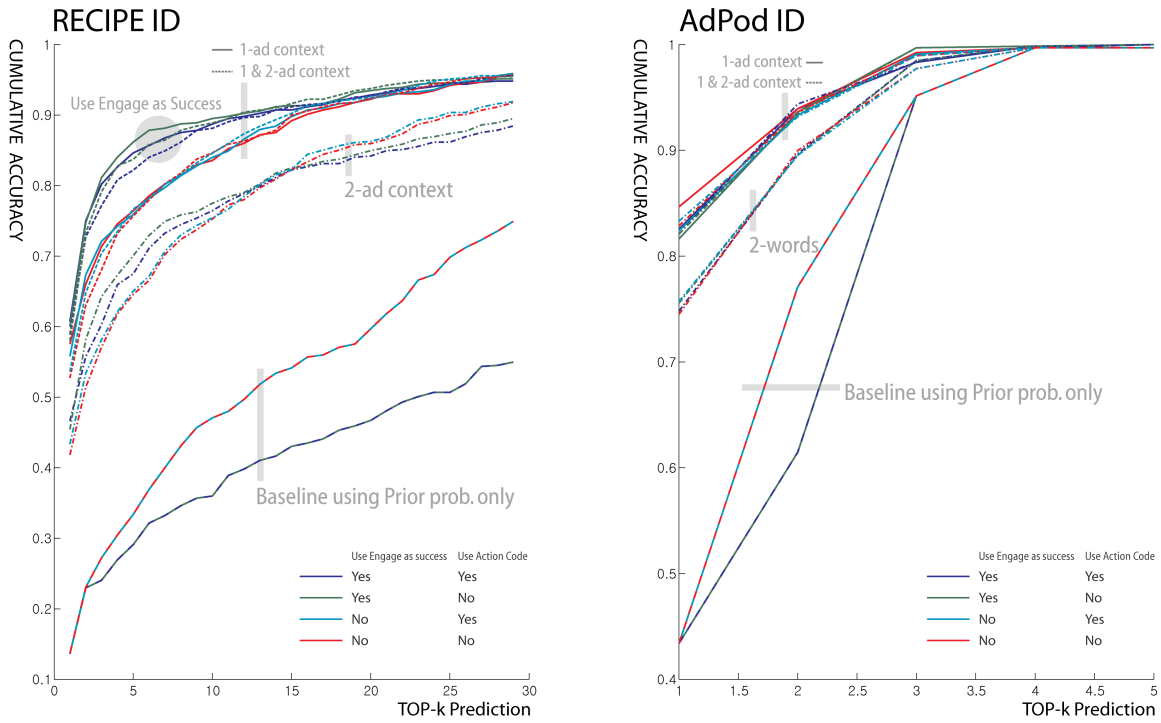

Figure 4: Cumulative Accuracy over Top-k predictions. left: RecipeID, right: AdPodID

Our prediction model ranks all the ads by predicted probabilities of being clicked under the given context of prior events in the session. To evaluate the accuracy of the ranked results, we used the position of the correct ad (which was actually clicked) in the list. We thus expressed the performance of each prediction model as a graph of cumulative accuracy over top-k predictions. For example, if k is set to 5, the performance of the model is a total occurrences of correct ads that appear in top 5 predictions. Figure 5 illustrates k=5 is the sweet spot over various prediction thresholds. In both figures of RecipeID and AdPodID, it is clear that using only the prior probability is not a good move. Although using Click+Engage as notion of success improves the performance a little (accuracy=0.3338, k=5), Baseline model still under-performs all the other tested models.

While 1-ad and 1 & 2-ads configurations show almost the same performance(0.7$\sim$0.9, k=5), 2-ads is not as good as them (0.6 $\sim$ 0.7, k=5). This result suggests that using more expressive feature for a sparse training dataset gives deteriorated performance when compared to a simpler model. Using both Click and Engage as successful case is a clear winner(0.8$\sim$0.9, k=5).

Another result of using AdPod ID (figure2) shows much better performance overall, because the total number of AdPod ID is much smaller than Recipe ID. Although it does not have as much detail as Recipe ID graph shows, general tendency is similar. In Table 1, characteristic results suggest that shadowed parameters (1-ad, 1&2-ads, Click+Engage, k=5) give out the best accuracy in the test.

RecipeID, ActionCode		k=1	k=5	k=10	k=20
1-ad	Click	0.5753	0.7641	0.8356	0.9224
	Click+Engage	0.6073	0.8463	0.8950	0.9285
1&2-ad	Click	0.5274	0.7568	0.8465	0.9240
	Click+Engage	0.5890	0.8219	0.8813	0.9269
2-ad	Click	0.4183	0.6458	0.7527	0.8595
	Click+Engage	0.4666	0.6748	0.7644	0.8419

Table 1: Characteristic Results

6 Conclusion

We proposed an online advertisement recommendation system that delivers ads customized based on users past activities. Naive Bayes model is a very simple yet effective method to predict a list of ads ranked by probability of being clicked, however, the rarity of click events deteriorate accuracy. In order to overcome the issue, we applied four parameters (ID type of ads, Feature space, Notion of Success, and Action code) of feature extraction from data set.

In the result all the tested models outperform the baseline model to considerable degrees. It means historical trails of users session are helpful to improve CTR(Click Through Rate). More specifically, using engagements as partial success is proven to be a valid technique for training Naive Bayes model with insufficient training data. On the other hand, complex (more expressive) models are better to be applied in comparison with simpler models, because complex models require more training data before gaining accuracy.

With more data set and time we could extend this study to various aspects. Applying the model on broader data set, using demographic features, graphical models, and online testing are all intriguing topics to explore.

Those insights gained through the project are not only beneficial to an online advertisement recommendation system but also suggestive to other domains dealing with sequential events of rare positive cases such as medical history inference and emergency prediction.

7 Future Work

While we gained some interesting insights about nature of the problem of ad selection, we believe further understanding of the nature of online user behavior in sessions has a potential to improve the efficiency of such a system.

First of all, we can extend our test to multiple campaigns over longer time span such as one month. This will either support our findings or provide alternative interpretation.

Using demographic data such as users sex, age and location might help collaborative filtering. Combining demographic features with the event information is an open challenge though.

Graphical models have gained popularity in recent years. Works done by [3] and [5] show a deeper understanding of the impact of user actions on predicting successful events while using a graphical flow model. While current Bayesian inference model considers only the existence of a certain event in the session, graphical models utilize the impact of order. However, since graphical models are more expressive than Naive Bayes model, more data might be need to construct a rich instance of such a model.

Finally, external online testing with real users can show whether our model would actually improve CTR or not. Due to the limited schedule we could not conduct it as a part of this project, but hope the plans of Tumri adopting it on their recommendation system to be realized soon.

8 Acknowledgements

We are very thankful to Prof Lise Getoor for valuable insights during the course of this project. We are also grateful to Tumri Inc., for sharing their user logs for the purpose of this research.

References

[1] N. Abe and A. Nakamura. Learning to optimally schedule internet banner advertisements. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 1221, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[2] A. Anagnostopoulos, L. Becchetti, C. Castillo, and A. Gionis. An optimization framework for query recommendation. In *Proceedings of the third ACM international conference on Web search and data mining*, page 161170, 2010.

[3] N. Archak, V. S. Mirrokni, and S. Muthukrishnan. Mining advertiser-specific user behavior using adfactors. In *Proceedings of the 19th international conference on World wide web*, page 3140, 2010.

[4] A. Ashkan, C. L. Clarke, E. Agichtein, and Q. Guo. Estimating ad clickthrough rate through query intent analysis. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, page 222229, 2009.

[5] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. In *Proceeding of the 17th ACM conference on Information and knowledge management*, page 609618, 2008.

[6] H. Cao, D. Jiang, J. Pei, E. Chen, and H. Li. Towards context-aware search by learning a very large variable length hidden markov model from search logs. In *Proceedings of the 18th international conference on World wide web*, page 191200, 2009.

[7] K. Dembczynski, W. Kotlowski, and D. Weiss. Predicting ads click-through rate with decision rules. In *Proceedings of the 17th International Conference on World Wide Web (WWW 2008), Beijing, China*, 2008.

[8] A. Hassan, R. Jones, and K. L. Klinkner. Beyond DCG: user behavior as a predictor of a successful search. In *Proceedings of the third ACM international conference on Web search and data mining*, page 221230, 2010.

[9] K. L. Keller. Brand equity and integrated communication. *Integrated communication: Synergy of persuasive voices*, page 103132, 1996.

[10] R. Lewis and D. Reiley. Retail advertising works! 2008.

[11] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3):259289, 1997.