Predict Commercial Promoted Contents Will Be Clicked By User

Name: Gary(Xinran) Guo Email: garyguo@stanford.edu

PREDICTING

As e-commerce, social media grows rapidly, Advertisements are everywhere in human daily activities. Clickstream data becomes incredibly large and it contains tremendous treasures that we can learn for people's interests, hobbies, and personalities. Our goal is to predict which pieces of recommended content users will be likely to click on.

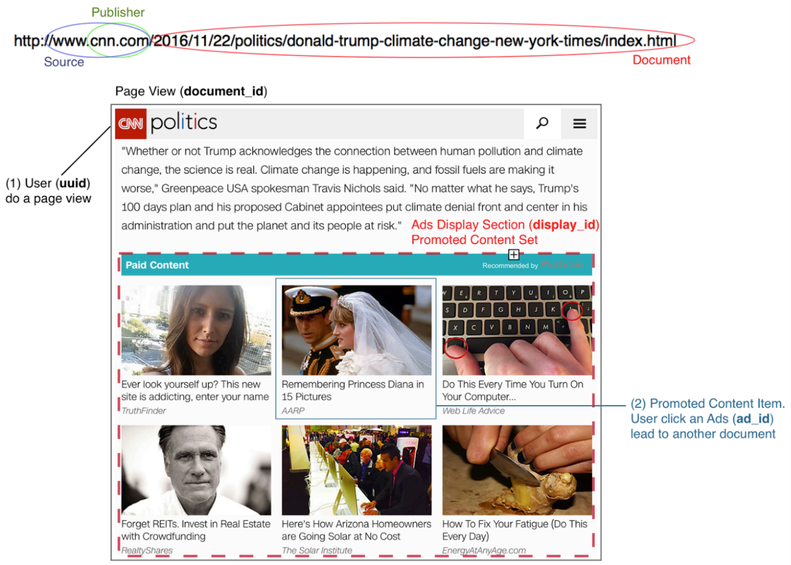


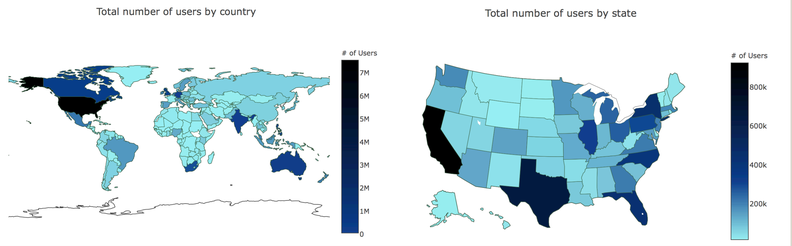
Figure 1. General process of user click ad event

Briefly, as figure 1 shown, whenever users visit a set of web pages, they will be also served by advertisements in many places (refer to Ads display section). The dataset contains numerous sets of content recommendations served to a specific user in a specific context. Each context (i.e. a set of recommendations) is given a display_id. In each such set, the user has clicked on at least one recommendation. The idea is to rank the recommendations in each group by decreasing predicted likelihood of being clicked.

FEATURES

Basic Features: To keep it simple, Select features that only related to display document and ad document. Such as, category, topics, confidence, publisher, source etc.

Advance Features: after deep exploratory data analysis, get more information about how data set distributed, select most common data features, and compare results. Such as, geographic location, platform, traffic resource, like following geo distribution.



DATA

Public data set from Ourbain.com, the web's leading content discovery platform delivers these moments or advertisements while we surf our favorite sites. **User Clickstream**: the log of users visiting documents. **Page Document**: The detail metadata of document include, sources, publisher, topics, entities, **Promoted Content**: The detail metadata of advertisement, that each ad belongs to a campaign run by an advertiser

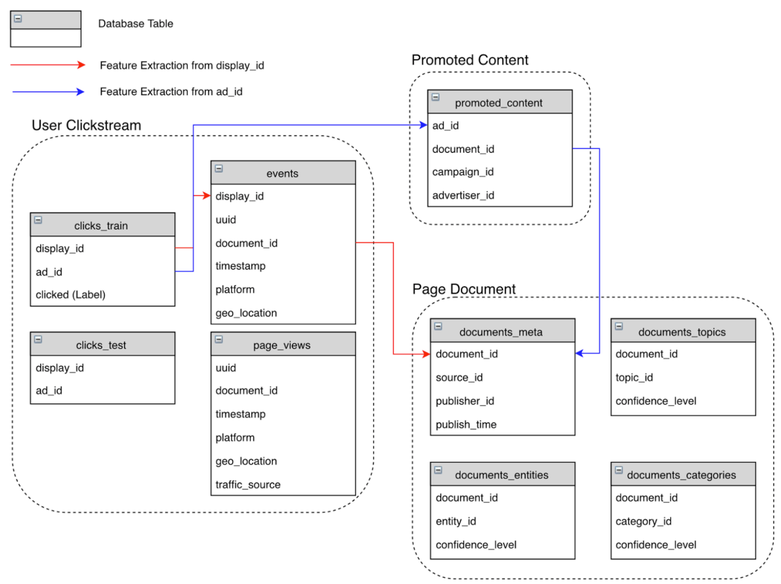


Figure 2: Dataset Structure

MODELS

Regularized Click Probability: only focus on ad itself, suppose not to consider any other related features. Calculate the click probability of each single ad via full training set. Test data click probability will ranked based on table.

REG = 10
Defined Probability table, Ad_id as key, Probability as value.
During Training given (display_id, ad_id, clicked):
For each ad (ad_id):
    Probability (ad_id) = train[train.clicked==1].ad_id.value_counts / (train.ad_id.value_counts + **REG**)

During Testing given (display_id, ad_id):
Set_of_ad = Testing set group by "display_id"
For each ad in Set_of_ad:
    If ad_id not in Probability table: Probability (ad_id) = 0
    Sort by Probability(ad_id) from high to low

logistic-regression: Applied main algorithm as Follow-The-Regularized-Leader - proximal. An adaptive-learning-rate sparse logistic-regression with efficient L1-L2-regularization.

Input: Parameter α, β, λ_1, λ_2
$(\forall i \in \{1,...,d\})$, initialize and
For t = 1 to T do
    Receive feature vector and let
    for $i \in I$ compute
        If $|z_i| \le \lambda_1$, then $\omega_{t,i} = 0$
        else, $\omega_{t,i} = -(\frac{\beta+\sqrt{n_i}}{\alpha} + \lambda_2)^{-1}(z_i - sgn(z_i)\lambda_1)$
    Predict $p_t = \sigma(x_t * w)$ using the $\omega_{t,i}$ computed above observe label $y_i \in \{0,1\}$
    For all $i \in I$ do
        $g_i = (p_t - y_t)x_i$ # gradient of loss
        $\sigma_i = \frac{1}{\alpha}(\sqrt{n_i + g_i^2} - \sqrt{n_i})$
        $z_i = z_i + g_i - \sigma_i \omega_{t,i}$
        $n_i = n_i + g_i^2$
    End for
End for

RESULTS

Sample Output and Feature and Model prediction comparison

```
Sample Output
display_id  ad_id   clicked (probability)
16874594    66758   0.167213874715
16874594    150083  0.110528833553
16874594    162754  0.247692716539
16874594    170392  0.301802838584
...

display_id     ad_id (Rank probability)
16874594       170392 172888 162754 66758 150083
16874595       8846 143982 30609
16874596       289915 11430 289122 132820 57197
16874597       305790 285834 143981 182039 155945
...
```

Prediction Score	Regularized Click Probability	Logistic-Regression
Basic Features	0.63854	0.66897
Advance Features	N/A	0.63269

DISCUSSION

- This project is application based. So the major task is to applying different machine learning models to a practical problem and make prediction. Features selection and model comparison will be primary focus of project.
- Doing exploratory data analysis (EDA) is one of the most important tasks for this project. The cumulative volume of data sets exceed 20GB. Knowing feature distributions helps significantly on features selection.
- After EDA, finding that the average page view of single user is just 2.835, it indicates that user-based recommendation profile is not suitable for click-prediction, instead, document categories and topics have uniform distribution among data.
- In Regularized Click Probability model, adding a regularized term REG that penalizes ads with small amounts of data, therefore making it prefer an ad with large amounts of training data and a reliable probability. It improves score by 2%
- More advance features is not guaranteed to make prediction better. Comparing the result of basic feature, and advance features, The prediction score is actually decreasing when apply more unrelated features.

FUTURE

- Do more exploratory data analysis to discover more features distributions in huge amount of data set, such as calculate mean, sum, standard deviation.
- Join various tables, and select different features to do prediction and compare result, test error.
- Apply Keras, deep Learning library for Theano and TensorFlow, high-level neural networks library. To deeply dig more into user click stream (page view).
- Changing the parameters of the model in logistic-regression algorithm and compare results.

REFERENCES

[1] H. Brendan McMahan, Gary Holt, D. Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, Jeremy Kubica. Ad Click Prediction: a View from the Trenches. Pages 2.