Predicting Click Through Rate for Job Listings

Manish Gupta
Yahoo! HotJobs, Bangalore, India
gmanish@yahoo-inc.com

ABSTRACT

Click Through Rate (CTR) is an important metric for ad systems, job portals, recommendation systems. CTR impacts publisher's revenue, advertiser's bid amounts in "pay for performance" business models. We learn regression models using features of the job, optional click history of job, features of "related" jobs. We show that our models predict CTR much better than predicting avg. CTR for all job listings, even in absence of the click history for the job listing.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Prediction, Click Through Rate, jobs, linear regression, CTR, CPC, Treenet, GBDT, gradient boosted decision trees

1. MOTIVATION AND RELATED WORK

CTR is a common metric used to rank results in a variety of applications, especially in those with open-loop reporting systems. CTR is computed as the ratio of "clicks to get a full description of the entity" to "views of a reduced version (snippets, listings, thumbnails) of the entity". Impressions (views) and the clicks for a new entity are too low to produce a Maximum likelihood estimate (i.e. CTR) with good confidence. CTR values being too small (avg. for HotJobs [4] is about 2.29%), this estimate has a high variance. If the entity (say, a job listing) has a low shelf life, CTR wrt time does not stabilize. Attention span of users decreases rapidly as position number increases on search results page. CTR of jobs can be used to decide the rank order itself. Hence, predicting CTR fairly accurately becomes important.

Following Regelson and Fain [1], we could estimate the CTR using topic clusters (i.e. job categories). Though CTR seems to be flat over time, for every category, CTR variation within a category is high. Richardson et. al. [2] describe in detail a variety of features to be considered when predicting CTR for ads. We look at the problem in job domain.

2. REFINING PROBLEM DEFINITION

We would ideally like to predict CTR for job j per position p personalized to a user/cluster of users u and shown in some context c. This would need including properties of the user, properties of the context (like other jobs shown on the page) and their interactions with properties of jobs, in the feature vector. But this would explode the size of feature vector and cause data sparsity. Using training data across different positions, we learn CTR(job). As CTR versus position curve drops rapidly with increase in position, this predicted CTR is for a position much closer to 1. CTR for other positions can be estimated using the CTR versus position curve.

3. DATA SET USED

Job data from Aug 11, 08 to Aug 31, 08 has been taken from Yahoo! HotJobs [4]. The aim is to predict CTR of jobs on Sep 1, 08. A sample of 40K jobs (published by 7K+ companies) was randomly chosen out of the active popular jobs, maintaining the category proportions. Random set of 32K was used as train set and the remaining as test set. Each job in HotJobs has location, company name, category (like finance, healthcare), creation date, posting date, optional position wise click history, job source (feeds, newspapers, GUI), title, snippet (which contains title, location, posting date, company name) & job description (landing page). We smooth out the CTR for job listings by interpolating the missing CTR values, based on the CTR values available for the neighboring days. Missing CTR values for first or the last day of the window, are set to avg. CTR for job category.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

4. DIFFERENT MODELS

We experimented with Linear Regression and SMOReg using Weka [5]. Accuracy gain using SMOReg isn't much over simple linear regression model as against the model complexity and the time required to build the model. We also used Treenet [3] to build gradient boosted decision tree models. Treenet provides tuning of parameters like regression loss function (we used least squares), regularization shrinkage factor (we used 0.01 and 0.1), subsample fraction, nodes per tree (we used 16, 64, 256), maximum trees (we used 300, 600, 1200), atom size (minimum leaf size – we used 20, 100, 400). For feature importance, we use a. wrapper method available in Weka [5] with linear regression as the evaluator and GreedyStepwise as the search method or b. variable importance returned by GBDT of Treenet.

5. FEATURES

Features from Similar Jobs (60): CTR of jobs with same title/company/state/city+state/category and their cardinalities. To compute these features, we varied the time period of observation. Each of the these is a set of six features e.g. we have six different features based on "avg. CTR of jobs with same title posted in past 1/2 weeks or all jobs, based on the click history of past 1/2/3 weeks".

Features from Related Jobs (288): Two jobs are related if sets representing their titles have non-null intersection and cardinality of difference set is < 5. We consider avg. CTR_mn of related jobs with m=|A-B| and n=|B-A| and number of related_mn jobs as features for job with title A. Both m and n can vary from 0 to 4. Again, these features are computed for jobs posted in the past 1/2 weeks or all jobs, and based on click history of past 1/2/3 weeks.

Job Title Features (11): # words in title, # capitalized words in title. Is the job title written totally in capitals? Does it contain too much punctuation (>10% of title length)? % of long words? (words with word-size > 10). Does the title provide numbers (such as salary)? We also divided the vocabulary of words into five bins depending on the popularity of words. We then have five features: number of words in the job title that fall in each of the five bins.

Daily CTR Features for past 3 weeks (21)

Other Features (10): Job Category, age (dates of job creation, job update and job posting), location specificity, job source, and job description page features. Location speci-

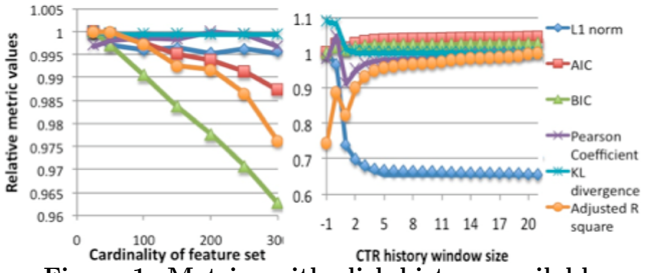

Figure 1: Metrics with click history available

ficity=# locations, the job has been posted at. User's preconceptions about the landing pages of a particular company can decide whether she would like to click on a job listing or not. So, we consider job desc page features: Presence of HTML/images/tables, # words in job description in 100s.

Other potential features: Does the title contain action/ high-marketing-pitch words such as "apply", "earn", "home", "wanted", "needed" etc.? Brand value of company. Spam feedback: Number of abuse votes against the job. Features to measure seasonal variations over the year, whether the day is a weekday/weekend, is the day special? e.g.public holiday. We can consider these features also, if available.

6. EXPERIMENTS AND RESULTS

The table presents % improvements in various metrics compared to our baseline. Predict avg. CTR (0.023) for all jobs is our baseline. Column *A* shows that predicting avg.–categorywise–CTR performs worse than our baseline. Performing linear regression using Weka [5] over 390 features (60 Similar Jobs+ 288 Related Jobs+ 11 Job Title+ 21 Daily CTR + 10 Other) produced a model that used only 142 of these features as regressors. Column *B* compares this model with our baseline. Treenet when run on this dataset produces results as shown in the column *C* using 300 regressors at 256_600_0.01_100. (Nodes per tree is 256, max trees is 600, regression shrinkage is 0.01 and atom size is 100).

An analysis of the distribution of the regressors showed that Daily CTR features are most important. Other features perform better than Related Job features. Top five Similar Jobs features include avg. CTR of jobs from the same company computed using 1/2/3 weeks click history over all jobs, avg. CTR of jobs with same title computed using 1 week click history, avg. CTR of jobs from same city+state computed using 1 week click history of jobs posted during the past 1 week. Top five Others features are creation date of the job, size of job description page, date of update, date of posting, job category. Top five Related Jobs features include avg. CTR of related_11 jobs posted in the past 1 week using their last 1 week CTR history, avg. CTR of related_11 jobs posted in the past 3 week using their last 3 week CTR history, avg. CTR of related_11 jobs posted in the past 1 week using their last 3 week CTR history, avg. CTR of related_14 jobs posted in the past 1 week using their last 1 week CTR history, avg. CTR of related_12 jobs posted in the past 1 week using their last 1 week CTR history.

Pruning the feature set shows that the accuracy does not decrease much even at featureSetSize=25 (left of figure 1). This supports our intuition that the Daily CTR history should be the best predictor of next day CTR. Column *D* of the table shows improvements using the set of 21 Daily CTR features only compared to our baseline.

Results in right of figure 1 show the effect of reducing click history window size. Here, x axis represents click history window size. x=-1 represents metrics for the avg.–CTR–for–all strategy while x=0 represents category wise avg. CTR strategy. x=1 implies that only today's CTR is used to predict next day's CTR. We go on from x=1 to x=21 which means that we use past 3 weeks' CTR to predict next day's CTR. In order to represent all the metrics in a single figure, we have normalized the original values by dividing them by the max in that group in case of positive values and the minimum in that group in case of negative values.

Metric	A	B	C	D	E	F	G
L1 norm	-3.53	32.7	35.9	32.2	34.9	13.8	22.1
AIC	-2.30	2.17	5.54	1.97	6.49	-5.21	-2.02
BIC	-1.48	-0.15	0.29	1.11	3.96	-5.40	-4.39
Pearson Coeff	-6.18	-2.83	4.97	-4.57	4.24	-26.4	-12.9
KL divergence	-0.69	7.59	8.04	7.58	7.89	5.60	6.69
Adjusted R-Sq	-15.9	15.5	35.4	12.6	36.8	-35.6	-9.18

Applying Weka's [5] wrapper based feature selection using linear regression and GreedyStepwise search strategy, we identify a minimal set of Daily CTR features. Results using this feature set are shown in column *E* of the table. This model has as low as 11 features which include all of the "past week CTR" features, which is quite intuitive.

In absence of click history of the job, we cannot evaluate the Daily CTR features. So, we trained another model using all features except Daily CTR. Using this 369-sized feature vector, Linear Regression (uses 187 regressors) and Treenet (uses 282 regressors at 256_600_0.01_20) to provide results as shown in columns *F* and *G* of the table. Similar job features, Others, Related features perform better than title features.

As shown in left of figure 2 pruning the feature set to 50 does not cause any substantial change in accuracy.

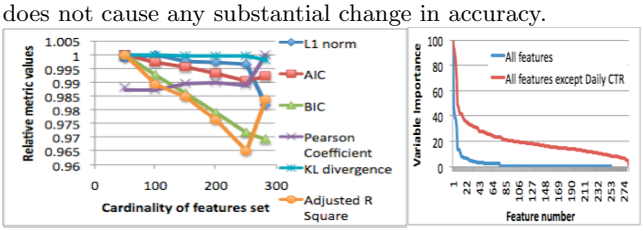

Figure 2: Metrics with click history unavailable

The right of figure 2 shows that the variable importance takes a steep drop in case when CTR history is available. But, when Daily CTR feature values are missing, drop in variable importance curve is gentle and so the tradeoff between number of regressors and accuracy starts becoming conspicuous when we reduce the #regressors below 50.

Experiments further show that none of the sets of features alone is capable of predicting CTR with reasonable accuracy. Having a mix of features from all the sets is the right approach when CTR history for the job is not available.

7. CONCLUSION AND FUTURE WORK

Models for predicting CTR can be further improved by recognizing and incorporating more features as described in the section 5. We can use dyadic models to predict user-personalized CTR where the dyads could be of the form (job feature vector, user feature vector). We can put in auto model update mechanisms to handle the model drift.

We built a machine learning system to predict CTR for a job listing and presented our results using a variety of regression metrics. Results from our models can be combined with the empirical CTR values to achieve better predictions.

8. REFERENCES

[1] M. Regelson, D. Fain. Predicting click-through rate using keyword clusters. *Proceedings of Second Workshop on Sponsored Search Auctions*, Jan. 2006.

[2] Matthew Richardson and Ewa Dominowska and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. *WWW '07*, 521–530, 2007.

[3] Treenet: http://www.salford-systems.com/Treenet.php

[4] Yahoo! HotJobs home page http://hotjobs.yahoo.com

[5] Ian H. Witten and Eibe Frank Data Mining: Practical machine learning tools and techniques, 2nd Edition, 2005