Translating Relevance Scores to Probabilities for Contextual Advertising

Deepak Agarwal†, Evgeniy Gabrilovich†, Robert Hall‡[1],
Vanja Josifovski†, Rajiv Khanna†

† Yahoo! Research, 2821 Mission College Blvd, Santa Clara, CA 95054, USA
‡ Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
{dagarwal | gabr | vanjaj | krajiv}@yahoo-inc.com | ‡ rjhall+@cs.cmu.edu

ABSTRACT

Information retrieval systems conventionally assess document relevance using the bag of words model. Consequently, relevance scores of documents retrieved for different queries are often difficult to compare, as they are computed on different (or even disjoint) sets of textual features. Many tasks, such as federation of search results or global thresholding of relevance scores, require that scores be globally comparable. To achieve this aim, we propose methods for non-monotonic transformation of relevance scores into probabilities for a contextual advertising selection engine that uses a vector space model. The calibration of the raw scores is based on historical click data.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*retrieval models*

General Terms

Algorithms, Experimentation, Measurement, Performance

Keywords

Relevance scores, probability of relevance, logistic regression, online advertising

1. INTRODUCTION

Conventional information retrieval (IR) systems compute document relevance scores based on the bag of words model, where *tf.idf* term weighting considers word occurrence frequencies in individual documents and in the entire corpus. Heuristic *tf.idf* weighting works well in practice when documents need to be ranked by their scores for a given query. However, in many cases, it is also necessary to consider absolute values of document scores, in addition to being able to compare their relative values. For example, in meta-search (or more generally, federated search), scores assigned by different search engines need to be reconciled. In some retrieval scenarios (notably, in online advertising) the final decision on which results are selected is based on their revenue potential, which is estimated as a function of the probability of a click and the bid amount submitted by the advertiser. In these scenarios, a principled way of comparing relevance scores is important to select the final set of results shown to the user. One way to address this problem is to *calibrate* the relevance scores using evidence from another knowledge source. Previous works focused on training regression models using human relevance judgments (cf. [8]), while ignoring the score of the initial retrieval step. As opposed to learning the probability of click directly from the query and document features, in this work we rely on the IR score to capture the information used in the first phase of the retrieval and transform that score into the probability of a click by using historical click data.

We focus on *contextual advertising*, where ads are selected for a given Web page based on its content [4]. True assessment of relevance probabilities is particularly important for ad matching as current models of ad selection often simultaneously optimize relevance and revenue. The expected revenue of an ad is computed by multiplying the estimated ad relevance by some function of the *bid*, which is the amount of money the advertiser pays if the ad gets clicked. Producing accurate probability estimates of ad relevance is therefore crucial for reliable ad ranking, as well as for determining whether or not to show ads for each particular Web page or query [3].

We translate the scores produced by a vector space model into probabilities using logistic regression over click data. We further enhance our click prediction through additional query and ad features and examine a "mixture of experts" model, where each expert is implemented using logistic regression. The resulting transformation is non-monotonic (that is, it reorders the set of retrieved ads) and gives significant cumulative gain improvement over the vector space model ordering when measured by the historical click data.

Determining the probability of relevance has been an active field of study in IR for the last few decades. Both generative and discriminative models have been proposed to estimate the probability of relevance [12, 9], however, these do not use existing relevance scores. Another approach to estimating the probability of a click is to use an aggregate of the click-through rate (CTR) over a period of time at different level of aggregation [1].

[1]Research performed during internship at Yahoo! Research.

CIKM'09, November 2–6, 2009, Hong Kong, China.

2. BACKGROUND: CONTENT MATCH

In this section we give a brief overview of the current practices in Web advertising based on [4].

Contextual advertising is an interplay of four entities: The *publisher* rents space on its Web page in return for revenue; its utility is usually measured by the revenue and user retention. The *advertiser* provides the supply of ads. The *ad network* selects the ads that are placed on publisher's Web pages. It acts as a mediator and shares the advertising revenue with the publisher. Finally, the *users* visit the Web pages of the publisher and interact with the ads. Studies has shown that users perceive as more relevant ads that relate to the content of the publisher page [5]. The vector space model has been proposed to compare the content of the web page and ads to select relevant ads [11, 4].

The content match model aligns the interests of the publishers, advertisers and the network itself. In general, clicks bring benefits to the publisher and the ad network by providing revenue, and to the advertiser by bringing traffic to the target Web site. The revenue of the network, given a page p, can be estimated as:

$$R = \sum_{i=1..k} P(click|p, a_i)\ price(a_i, i)$$

where k is the number of ads displayed on page p, and $price(a_i, i)$ is the click-price of the ad a_i at position i. Note that $P(click|p, a_i)$ needs to be normalized for position, taking into account that ads in higher positions are more visible and tend to be clicked much more often. A common model of determining how much an advertiser is charged is based on the intuition that the charge should be equal to the minimum amount that the advertiser needs to bid to retain the given position. As the ads are ordered by expected revenue, this can be estimated using the bid of the next ad in the revenue-ordered list of ads:

$$price(a_i, i) = \frac{bid(a_{i+1})\ P(click|p, a_{i+1})}{P(click|p, a_i)}.$$

Note that this computation requires to estimate the ratio of the click probabilities.

3. MODELS

We propose models for estimating click probabilities $P(c|r, \boldsymbol{x})$ based on the vector space relevance score r and other features $\boldsymbol{x}$. We explore three models, all based on logistic regression. The first model fits a single global logistic regression to the entire data. Our second model partitions the data by publisher ids and fits a separate *local* logistic regression to *heavy hitters*, i.e., the publishers who make a large number of ad calls. For the tail publishers, we fall back on the global logistic regression. Our third model is based on a committee logistic regression that performs clustering and fits a local logistic regression for each member of the committee. To avoid the *cold-start* problem (new page-ad pair showing up in the test data), the cluster assignment probability of a pair is also modeled as a function of features. In the next sections, we provide a description of logistic regression model (along with features used) and committee based logistic regression. Throughout, we will assume a training set of page-ad pairs with the i^{th} pair having features $\boldsymbol{x}_i$ and relevance score r_i, each pair is accompanied by a binary variable y_i indicating whether the ad was clicked or not when shown on the page.

Logistic Regression

Logistic regression is a well known technique to estimate conditional probabilities associated with a binary outcome variable. We assume

$$\begin{aligned} y_i|\boldsymbol{x}_i, r_i, \boldsymbol{\beta} &\sim \text{Bernoulli}(p_i) \\ \log(\frac{p_i}{1-p_i}) &= \boldsymbol{x}_i^{'}\boldsymbol{\beta} + f(r_i) \end{aligned} \tag{1}$$

Here, $\boldsymbol{\beta}$ are unknown weight parameters associated with features and the function f (unknown) quantifies the relationship between relevance and CTR after adjusting for features. Empirical analysis indicates that f is well approximated by a quadratic function, however, we obtained more robust performance by approximating f through a piecewise constant function, i.e.,

$$f(r) = \sum_{k=1}^{M} \alpha_k 1(r \in B_k) \tag{2}$$

B_k is the k^{th} bin and α_k is the associated weight parameter. For our experiments, we obtained B_k by dividing $[0, 1]$ into 20-30 equi-spaced intervals. Maximum likelihood estimates of the unknown parameters ($\boldsymbol{\beta}$,$\boldsymbol{\alpha}$) is a well-studied problem [2], it can be obtained through several numerical methods like conjugate gradient, L-BFGS, Iteratively re-weighted least squares[10]. Since most of the features in our scenario are binary with a small number of them "turned on" for a given page-ad pair i, we use the L-BFGS (CG applies as well) method that exploits the sparse structure.

Committee Logistic Regression

Global logistic regression defined in Equations 1 and 2 assumes the unknown weight parameters ($\boldsymbol{\beta}$,$\boldsymbol{\alpha}$) are constant for all page-ad pairs. This maybe a limiting assumption, especially in our application where extreme heterogeneity is expected to exist due to differences in publishers, ad campaigns, user population. We relax this assumption by fitting a mixture of logistic regressions, i.e.,

$$\begin{aligned} y_i|\boldsymbol{x}_i, r_i, \{\boldsymbol{\beta}_c\}, \{\boldsymbol{\alpha}_c\} &\sim \sum_{c=1}^{K} \pi_{ic}\text{Ber}(p_{ci}) \\ \log(\frac{p_{ci}}{1-p_{ci}}) &= \boldsymbol{x}_i^{'}\boldsymbol{\beta_c} + f_c(r_i) \\ f_c(r) &= \sum_{k=1}^{M} \alpha_{ck} 1(r \in B_k) \end{aligned} \tag{3}$$

Here, $\boldsymbol{\pi}_i = (\pi_{i1}, \cdots, \pi_{iK})$ are cluster membership probabilities for the i^{th} pair, ($\boldsymbol{\beta}_c$,$\boldsymbol{\alpha}_c$) are cluster specific regression weights assigned to the features. The main advantage of such a feature based mixture allocation strategy is to avoid the *cold start* problem; we can assign a new pair to an appropriate cluster based on features alone. To complete our model specification, we provide the functional relationship between $\boldsymbol{\pi_i}$ and binary features $\boldsymbol{w}_i = (\boldsymbol{x}_i, \{1(r_i \in B_k) : k = 1, \cdots, M\})$. There are several possibilities here, we explore the simplest one that is based on a Naive Bayes assumption for each co-ordinate π_{ic}, i.e.,

$$\pi_{ic} = \prod_{j=1}^{k} \theta_{c,j}^{w_{ij}} \tag{4}$$

where $\theta_{c,j}$'s are unknown constants estimated from data.

Model Fitting

We use an EM algorithm[6] to fit the model described by Equations 3 and 4 to our training data. This is done by introducing a "latent" allocation variable z_i to each pair i which is indicative of the cluster to which i is assigned. The incomplete data log-likelihood based on $\boldsymbol{Y}$ is now optimized by working with the complete data log-likelihood $(\boldsymbol{Y}, \boldsymbol{Z})$; this splits the log of sums arising due to Equation 3 into sum of logs, usual trick employed in likelihood maximization with mixture models. The E-step computes "responsibilities" $\gamma_i(c)$ for each pair i, i.e., the weight with which pair i belongs to cluster c. Note that

$$\gamma_i(c) \propto \hat{\pi_{i,c}} \text{Ber}(\hat{p_{i,c}}),$$

where $\hat{\pi_{i,c}}$ is the estimated allocation probability based on estimates $\hat{\theta_{c,j}}$, and $\hat{p_{i,c}}$ is the estimated Bernoulli success probability based on estimates of parameters $(\hat{\boldsymbol{\beta}}_c, \hat{\boldsymbol{\alpha}}_c)$. In the M-step, the parameters $\boldsymbol{\beta}_c$'s, $\boldsymbol{\alpha}_c$'s are updated by running separate *weighted* logistic regressions in each cluster, the weight assigned to the i^{th} pair in cluster c being given by the estimated responsibility in the E-step. The allocation parameters $\{\theta_{c,j}\}$'s are updated by fitting Naive Bayes models in each cluster to the estimated responsibilities. We iterate the E and M steps until convergence. To ensure scalable model fitting to large amounts of click-log data, we exploit grid computing. In addition to the relevance score, we used the following features.

- **Taxonomy**: Each page and ad is classified into a hand-labeled taxonomy of roughly 6,000 topical classes [4].
- **Domain**: The domain of the publisher, which is a good proxy for the position of the ads on the page.
- **Ad position** within a slate of ads.
- **Words in common**:Top few words that occur both on the page and the ad selected by the ratio $i_w = \frac{\text{CTR}_{(w,ap)}}{\text{CTR}_{(w,a)}\text{CTR}_{(w,p)}} > 2$, where $\text{CTR}(w, ap)$ is the CTR across (page,ad) pairs in which w occurs on both page and ad, $\text{CTR}(w, a)$ and $\text{CTR}(w, p)$ are marginal CTR's when word w occurs on ad and page, respectively.

4. EXPERIMENTAL EVALUATION

We now present an extensive evaluation of our methods on a sample of data obtained from an actual content match system of a major US search engine. To demonstrate the effectiveness of converting relevance scores to probabilities, we use historical data to compare retrieval results using conventional *tf.idf* scores and using our probability estimates. Our training data consisted of about 2 million contextual advertising ad slates spanning 15 days. All models were fitted using the training data, and results are reported by computing metrics on the test data, which consists of approximately 1 million slates spanning 7 days. Overall, our data consisted of approximately 400 publishers with the top-20 accounting for 70% of the total number of clicks.

We refer to the different models as follows. We use **VSM** to denote the baseline system that uses a vector space model with *tf.idf* weights (similar to the approach proposed by Broder et al. [4]). Global logistic regression without word-in-common features will be called **Global**, while **GlobalW** is a global logistic regression that includes the word-in-common features. The variants of committee logistic regressions with and without word-in-common would be denoted by **EMW** and **EM**, respectively. Finally, **PART** denotes the model that runs *local* logistic regression for the top 20 publishers that account for approximately 70% of clicks, and falls back to **Global** for the tail publishers that obtain the remaining 30% of clicks.

We compare the ad ranking of the VSM model to the ranking based on $P(click|p, a_i)$ for the various models using the Discounted Cumulative Gain (DCG) metrics. For a given slate with l positions ordered by priority (position 1 being the best), the DCG is defined by

$$\sum_{i=1}^{l} w_i(2^{r_i} - 1),$$

where r_i is the relevance of ad at position i and w_i is a position-specific weight, assumed to be $1/log_2(i+1)$ in the standard literature. The relevance r_i is binary and takes the value 1 if the ad at slate position i is clicked, and 0 otherwise. Normalized DCG (NDCG) is defined as DCG/IDCG, where IDCG is the ideal DCG attained using the best relevance ranking.

Logarithmic decay by positions is a reasonable assumption when ads are presented in a list format (e.g., in Web Search). This is not the case in our application where the presentation of ad slates depends critically on the page layout, which can vary greatly. An ideal estimate of w_i should be based on CTR differential, that is, the drop in CTR when a typical ad is moved from position 1 to i after randomizing over all other factors. This is intractable to be calculated at a granularity of a page, hence we make a simplifying assumption that decay is constant across all pages of one publisher and use the per-publisher decay curve through historic data. In our data we observed clear heterogeneity in CTR in different publishers due to the positional effects. We note that the bidding mechanism automatically induces a certain degree of randomness into the system, hence we believe such global decay estimates of positional effects are reasonable for the purpose of evaluation.

More specifically, we let $w_i = \text{CTR}(i)/\text{CTR}(1)$, where $\text{CTR}(i)$ denotes the global CTR at position i. We refer to the modified NDCG formula that uses these weights as Emp-NDCG, which stands for NDCG with *empirically* estimated decay weights. We refer to the standard formulation of NDCG as Log-NDCG. Figure 1 shows the weight curve obtained through logarithmic weighting and through global positional CTR estimates. The weights estimated empirically from the actual data are further smoothed through an isotonic regression to ensure monotonicity. The CTR-based decay tapers off much faster than the standard logarithmic weighting resulting in higher relative reward for Emp-NDCG at the premium positions (i.e., positions 1 and 2).

Table 1 provides the overall NDCG numbers for two different choices of decay (Logarithmic and Empirical) for all methods. We note that **Global**, **GlobalW**, **EM**, **EMW** all have similar performance. **PART** has the best performance and has a relative improvement of 13% in Log-NDCG, 2% in Emp-NDCG over the VSM baseline. The improvement of **PART** over **Global** is marginal on Log-NDCG scale (1%)

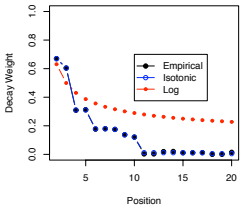

Figure 1: Decay weights for Log-NDCG and Emp-NDCG. *Empirical* refers to Emp-NDCG with weights estimated from the data; these are smoothed to be monotonic through *Isotonic* regression.

Model	VSM	PART	Global	GlobalW	EM	EMW
Log-NDCG	.692	.781	.773	.773	.771	.773
Emp-NDCG	.549	.561	.556	.555	.555	.555

Table 1: Overall NDCG for different models compared to the vector space model (VSM) baseline

but relatively better on Emp-NDCG scale (0.9%). This is largely because **PART** provides separate estimates of positional effects for top publishers, and is more effective in re-ranking ads that get clicked at the bottom to the top of the list. In fact, if we just confine the evaluation to the top 20 publishers where **PART** differs from **Global**, the Emp-NDCG scores for **PART** and **Global** are 0.564 and 0.556, respectively, a 1.4% relative improvement.

To test statistical significance, we conducted a bootstrap procedure [7], i.e., we computed average NDCG by taking a random sample of size n (we used $n = 100,000$) from the clicked slates with replacement. We took B such bootstrap replications (we used $B = 50$) and computed empirical distributions of relative improvements between the NDCG's. The improvements reported above between (**PART**,**VSM**) and (**PART**,**Global**) are statistically significant. In Figure 2, we look at the distribution of average NDCG per publisher for **PART** and **VSM**. Here again, the NDCG's are significantly better for the best probabilistic model across publishers.

The experiments show that modeling using the relevance score and the four features described above produces a significant improvement of the ad ordering. One of the questions posed by this result is what contributes to the improvement - the prediction modeling or the extra information provided by the features used in the model. Of the four sets of features

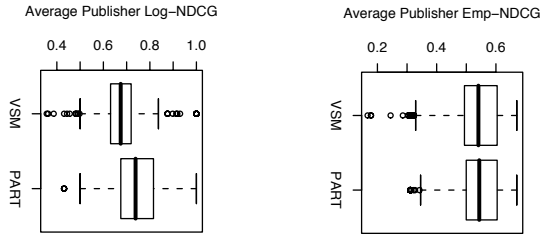

Figure 2: Model performance across publishers

used in the modeling, two (words in common and taxonomy features) are already used in the vector space model and do not contribute new information to the re-ranking process. The other two (publisher domain and ad position) pertain only the page and its layout and are not used in ad selection. While it is feasible to assume that ad selection could be influenced by the ad position on the page (.e.g. ads in all capital letters do better at the bottom of a page), this is an unlikely cause of the improvement of ad ranking. Therefore we can conclude that the NDCG improvements are brought by the prediction modeling having better differentiation than the cosine similarity used in the VSM.

5. CONCLUSIONS

We described a method for converting vector space relevance scores into probabilities for contextual advertising. The transformation function uses the original (VSM-based) score as well as other features of the page to achieve non-monotonic transformation that significantly improves the NDCG over the click data. In fact, we obtain a 13% relative gain in NDCG over the vector space model using our best probabilistic model.

6. REFERENCES

[1] D. Agarwal, A. Z. Broder, D. Chakrabarti, D. Diklic, V. Josifovski, and M. Sayyadian. Estimating rates of rare events at multiple resolutions. In *KDD*, 2007.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, August 2006.

[3] A. Broder, M. Ciaramita, M. Fontoura, E. Gabrilovich, V. Josifovski, D. Metzler, V. Murdock, and V. Plachouras. To swing or not to swing: Learning when (not) to advertise. In *CIKM*, 2008.

[4] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In *SIGIR'07*, pages 559–566. ACM Press, 2007.

[5] P. Chatterjee, D. L. Hoffman, and T. P. Novak. Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Science*, 22(4), 2003.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[7] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1994.

[8] F. C. Gey. Inferring probability of relevance using the method of logistic regression. In *SIGIR'94*, 1994.

[9] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. 1998.

[10] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge Univ. Press, 2007.

[11] B. Ribeiro-Neto, M. Cristo, P. B. Golgher, and E. S. de Moura. Impedance coupling in content-targeted advertising. In *SIGIR'05*, 2005.

[12] S. Robertson and K. Sparck Jones. Relevance weighting of search terms. *JASIS*, 27:129–146, 1976.