Intent Based Relevance Estimation from Click Logs

Prakash Mandayam Comar
amazon.com, Bangalore
prakasc@amazon.com

Srinivasan H Sengamedu
amazon.cpm, Bangalore
sengamed@amazon.com

ABSTRACT

Estimating the relevance of documents based on the user feedback is an essential component of search, retrieval and ranking problems. User click modeling in search has focused primarily on factoring out the position bias. It is easy to see that the query type (generic queries vs specific queries) and user intent (purchase vs exploration) also introduce a bias in the click signal. In other words, the results not matching with the user intent will not be clicked. In this paper, we outline a technique to model the interplay of query, user intent and position bias with respect to the relevance of the retrieved search results. In particular, we define two intents namely *purchase* and *explore*, and estimate the relevance of the documents with respect to these two intents. We also relate them to the relevance estimates from considering only the position bias. We empirically demonstrate the effectiveness of the proposed approach by comparing its performance against the well-known CoEC measure and the recently proposed factor model approach for relevance estimation.

CCS CONCEPTS

• **Information systems → Learning to rank**;

KEYWORDS

Ranking, relevance, user intent

1 INTRODUCTION

Estimating the true relevance (or quality) of an item from the observed click through rates (CTR) is a challenging ML problem with applications in search and online advertising [1, 8, 9] and is central to the retrieval and ranking problems. Typically, the observed click through rate is the simplest and easiest measure of item relevance. However, it is well known that the observed CTR is corrupted by several known biases: presentation or layout bias (bias caused by layout design where not all items have equal opportunity to get noticed by user), appearance/perception bias (like attractiveness of product image, font/style of title/abstract snippet), user intent bias (navigation intent vs exploration) and click biases from the query ambiguity (not issuing right query and clicking on items). Each of these biases masks the true relevance of the results. Before we discuss relevance estimation, we provide a brief description of the different biases present in the observed CTR.

The appearance bias occurs when users click on an item that looks visually more compelling like captivating product image, attractive titles like "80% discount" or bold font title terms that match with query etc. The appearance apparently increases the perceived relevance (which could be different from the true relevance to the query) resulting in clicks on the displayed item. The study by Clarke et al. [5], supports the existence of this bias where the authors observe phenomenon called *click inversion* – items in lower position get more clicks than those in higher positions, even when lower rank items not being very relevant to query. They observe that click inversions occur when the title of lower ranked documents have more matching terms with query, thus suggesting appearance bias.

The user behavioral or intent bias occurs when the same query is used for multiple purposes by different users resulting in different observed click patterns. Broadly, there are two common behavioral intents faced by all search engines – *navigational* and *exploration* intent. In the former, a users issue the search query with intent of reaching the targeted URL or item page (with hope it shows up in top positions) and in latter the user is interested in exploring the different items presented by the retrieval system. In this paper, we deal primarily with search for online shopping and label the *navigation* intent as *purchase* intent, as users with clear purchase intent issue a specific product related query to reach the product page. In Section 3.2, discuss more on these two intents in the context of online shopping.

The query gap (or query ambiguity) bias occurs when users do not articulate or formulate their query correctly. Rather, they issue a query, peruse the retrieved items, manually filter out irrelevant results and click on items that they think are relevant to their unarticulated intent. For example, a user with intent to purchase *honey roasted peanuts* may issue a query *peanuts*. They may then skip the top few results related to peanuts and click on "honey roasted peanuts" that appear in the lower positions. In this case, relevance estimation would assume that the products or items in the top positions are less relevant for the keyword *peanuts* which is untrue. Thus, estimating the item relevance by factoring only position bias from the observed clicks, would result in poor quality relevance estimation. This suggests a requirement of model that disambiguates the different types of biases that are amalgamated in the observed clicks, in order to estimate the true item relevance.

Position bias arises out of the layout design where some items appear prominently before other items, resulting in an unequal number of views/examinations per item. For example, in a search query, when the retrieved items are arranged as top down list, typically the items presented in the top positions tend to accumulate more clicks than the ones listed in the lower positions. In practice, it has been observed that irrelevant items placed in the higher positions may incur more clicks than relevant items placed in some lower positions. As a result, the observed click through rates do not reflect the true relevance or quality or even the popularity of

CIKM'17 , November 6–10, 2017, Singapore
© 2017 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-4918-5/17/11.
https://doi.org/10.1145/3132847.3132870

an item. The prime reason for presentation bias (or position bias) is that, a typical user would not perform additional action like scrolling down the page (or moving to next page) to reach lower positions in order to examine other items.

In this paper, we propose a generative model that estimates the true item relevance, in the presence of different bias types discussed above. It should be noted that not all bias types are present in all applications as the type of bias present depends on the context of the application. For example, there is no query ambiguity bias in a statically created page which lists a set of deals for Christmas. The deals in that page can still be ranked by some algorithm and the purchases of the products will depend on the ranking of the products. In order to accommodate this aspect, the proposed model offers flexibility in specifying the different bias types as user given inputs. Our model is in discovery stage and has shown promising results on offline tests on Amazon search data.

There are two distinct and useful contributions in this work. First, we propose the `Intent Hypothesis` and using first principles derive an expression to relate the observed clicks into latent intent and relevance factor. We then propose a generative model to estimate this latent factors. The rest of the paper is organized as follows. In Section 2, we review the past work on modeling position bias and draw inspiration for our work. We also review other research on modeling user intent in search query and contrast it from our work. In Section 3, we present our *intent hypothesis* and derive a factor model to learn the true item relevance with respect to intents. We then present a generative model to learn the factors and derive EM based update formulas for model parameters. In Section 4, we present the results of relevance estimation from the *intent hypothesis* and compare the performance of our generative model against well known baselines. The proposed approach improves MRR [1] by about 5% compared to a recently proposed factor model for relevance estimation [4]. We discuss in detail about our findings and present our conclusions in Section 5. In the rest of the paper, we use the words "query" and "keywords" synonymously and the terms "document"/"item"/"products" interchangeably.

2 LITERATURE REVIEW

One of the well-studied bias in the observed click through rate is the position bias [3, 4, 6, 7]. The formal experiment on the existence of position bias was performed by Granka using eye tracking analysis of user behavior in web search [10, 12]. Subsequently, several researchers proposed methodologies to discount the position bias in order to estimate the true relevance of an item. Richardson et al. [14] proposed a model that explains the observed clicks C_{ip} and impressions I_{ip} of an item i at position p as product of item relevance (r_i) and the position bias (b_p) viz $C_{ip} = I_{ip} * r_i * b_p$. This idea was extended as *Examination hypothesis* by Craswell et al. [6]. According to this hypothesis, b_p is the probability a user examining the item placed in position p (averaged over all queries) and given that a position has been examined by a user, the click through rate of an item i is r_i, reflecting its true click through rate or relevance to the query. Thus according to this model, the relevance or true click through rate is estimated by factoring out the position bias b_p from the observed clicks C_{ip}.

Researchers have proposed several approaches to learn b_p and r_i from the observed clicks, and the most recent work by Chen et al [4] is relevant for our work. They model the observed clicks $C_{ip} \sim \texttt{Poisson}(I_{ip} * b_p * r_i)$ and estimate the position bias and item relevance by maximizing the observed data likelihood. We use it as a baseline in our experimental validation and show that the proposed approach relatively improves it by about 5%.

The assumption of constant item relevance (r_i) for a given query, was countered by Hu et al. [11] in their work on intent modeling, where they experimentally showed that for a given query, each item has distribution of relevance score (r_i) even after removing the position bias. In their experiment, they fixed a query and considered all the sessions where the same item was shown in position one (there is no position bias) and then they divided the sessions into two groups: one group consisted of all sessions with exactly one click beyond position one and other group consisted of all sessions with more than one click beyond position one. Then they showed a statistically significant difference in the true click through rate of the item placed in position one in both groups. From this experiment, they inferred the existence of additional factors or biases which had to be removed to uncover the true relevance. The authors attributed this factor to intent bias arising out of the query gap between the user issued query and the intended query. According to their model, the observed clicks C_{sip} of an item i placed in position p in a session s is factored as tri-product: $C_{sip} = \mu_s * r_i * b_p$, where $\mu_s \in [0, 1]$ is the intent bias factor of session s. In their model, the item relevance is fixed and $\mu_s = 1$ indicate absence of intent bias (no query gap bias) in the session s and $\mu_s < 1$ indicate presence of the intent bias.

Differing from the intent hypothesis proposed by Hu et al. [11], we propose an alternate and novel intent hypothesis which estimates one relevance score for each intent type. The model assumes the existence of *multiple discrete intent types, distributed across sessions, that are associated with a given query.* The aggregated effect of these intents is superimposed in the observed click through rate of items. The different types of bias in the click logs arise due to different latent intent types for a given query. Further, we hypothesize that each intent induces a distinct click pattern on the positions and has a distinct relevance score (and therefore ordering) for the retrieved items. Under this hypothesis, each item gets an explicit relevance score for each intent type.

In the next section, we state our intent hypothesis and derive an expression for relevance estimation from the observed clicks. We then propose a generative model that estimates the different user intents as distinct click patterns over positions and compute the corresponding relevance score. The performance of these relevance scores are compared against the popular baselines in Section 4.

3 PROPOSED FORMULATION

In this section, we present the mathematical notations used in this paper and state the proposed `intent hypothesis`. Using basic probability, we derive an expression for probability of click under this hypothesis. We then learn the parameters of this expression using the Poisson-Beta generative model.

[1] https://en.wikipedia.org/wiki/Mean_reciprocal_rank

3.1 Notations and Background

All through this paper, the variables I and C stand for *impressions* and *clicks* respectively and the variables E and R denote the *examination* and *relevance* respectively. Each of the random variables could be indexed by one or more of the index variables namely i and p indexing over item set and positions respectively. Let $\mathbb{I} \in \{1, 2, ..., K\}$ be the discrete variable denoting the different possible intents under consideration for a given query. As example, the term $Pr(E_p|\mathbb{I} = k)$ denote the probability of examining the position p given the intent is k and $Pr(R_p|\mathbb{I} = k)$ denote the relevance of item placed in position p given the intent under consideration is k.

Let b_{pk} denote the probability of examining position p with intent k, and r_{ik} denote the relevance of item i with respect to intent k. More formally we have,

$$Pr(R_p = 1|\mathbb{I} = k) = r_{\pi_p k} \quad \text{and} \quad Pr(E_p = 1|\mathbb{I} = k) = b_{pk}$$

where π_p denote the item placed in position p. Here we assume that the intent for a given query vary with users(session) and in the above expression, the probability is computed over all the search engine users (or sessions) with particular intent (unknown or latent). With this notation, we now formulate our *Intent Hypothesis*.

`Intent Hypothesis`: *An item i placed in position p is clicked with intent k, if and only if the position (p) is examined with respect to the same intent and the item (i) is relevant for that intent.*

In terms of the random variables, we have

$$Pr(C_p = 1|\mathbb{I} = k) \Leftrightarrow Pr(E_p = 1, R_p = 1|\mathbb{I} = k) \tag{1}$$

According to the intent hypothesis, for a given intent k

$$Pr(C_p = 1|E_p = 1; R_p = 1, \mathbb{I} = k) = 1 \tag{2}$$

and we have the following corollaries. The probability of clicking a position p with intent k is *zero* under the following conditions.

- if the position is never examined by the user, irrespective of the relevance of item to the query. $Pr(C_p = 1|E_p = 0; R_p, \mathbb{I}) = 0$
- whenever an irrelevant item (wrt to intent k) is placed in that position, irrespective of whether the position has been examined by user. $Pr(C_p = 1|R_p = 0, \mathbb{I} = k; E_p) = 0$

Assuming independence between examination of position p and the relevance (true click through rate) of item in position p, from the above corollaries we get the following click through rate probability with respect to an intent.

$$\begin{aligned} Pr(C_p = 1|\mathbb{I} = k) &= \sum_{b,r=\{0,1\}} \big[Pr(E_p = b, R_p = r|\mathbb{I} = k) \qquad (3)\\ &\qquad \times Pr(C_p = 1|\mathbb{I} = k, E_p = b, R_p = r)\big] \\ &= \big[Pr(E_p = 1|\mathbb{I} = k) Pr(R_p = 1|\mathbb{I} = k) \\ &\qquad \times Pr(C_p = 1|\mathbb{I} = k, E_p = 1, R_p = 1)\big] \\ &= b_{pk} * r_{\pi_p k} * 1 \text{ (by equation 2)} \end{aligned}$$

In practice we do not observe C_{pk} as the intent behind a user click is never observed, rather we observe C_p, the total clicks at position p. This is obtained by marginalizing over the intents, as follows

$$\begin{aligned} Pr(C_p = 1) &= \sum_k Pr(C_p = 1|\mathbb{I} = k) * Pr(\mathbb{I} = k) \qquad (4)\\ &= \sum_k b_{pk} * r_{\pi_p k} * Pr(\mathbb{I} = k) \\ &= \sum_k \text{position bias wrt intent } k * \text{relevance wrt intent } k \\ &\qquad * \text{proportion of users/sessions with intent } k \end{aligned}$$

In this work, the term b_{pk} and r_{ik} for each query is estimated as factors and not as actual probabilities. Thus for a given query, the term denoting the proportion of sessions or $Pr(\mathbb{I} = k)$ can be absorbed into either of the factors.

Our data $\mathcal{D} = \{I_{ip}, C_{ip}\}$ consists of total impressions and clicks an item i receives at position p. Here, the observed clicks C_{ip} is modeled as superposition of clicks generated by different intents which needs to be uncovered in order to estimate the true item relevance for the corresponding intent. The generative model to perform this job is presented in the next section.

3.2 Multiple Intent Poisson-Beta Model

In this section, we propose a generative model that uncovers the true relevance of an item with respect to each intent by *estimating the position bias for each intent.* Let $P_{ipk} = b_{pk} * r_{ik}$ denote the probability of an user clicking on item i at position p with intent k. A simple and interpretable probabilistic model that can explain the observed data $\mathcal{D}$ is a Poisson model given below

$$C_{ip} \sim \sum_k \texttt{Poisson}(I_{ip} * P_{ipk}) \tag{5}$$

Here the clicks an item i receives at a position p is simple aggregation of clicks arising from the multiple intents, where each intent is a Poisson random variable with the rate P_{ipk}. This rate parameter is multiplied by the total number of impressions to get the expected number of clicks with given intent (aggregated across all sessions for a given query). Since sum of Poisson random variables is Poisson with rate parameter being simple sum of all rates, we get

$$C_{ip} \sim \texttt{Poisson}(I_{ip} * \sum_k P_{ipk}) \tag{6}$$

by letting $Y_{ip} = I_{ip} * \sum_k P_{ipk}$, we get $C_{ip} \sim \texttt{Poisson}(Y_{ip})$, where Y_{ip} denote the expected number of clicks an item i receives from position p across all intents. The unknown parameters θ in (6) are the bias and relevance terms, $\theta = r_{ik}, b_{pk}$ which can be estimated by maximizing the following likelihood function.

$$Pr(C_{ip}|I_{ip}, b_{pk}, r_{ik}) = \prod_i \prod_p \frac{Y_{ip}^{C_{ip}} \exp(-Y_{ip})}{C_{ip}!} \tag{7}$$

Given data $\mathcal{D} = \{I_{ip}, C_{ip}\}$, the log likelihood excluding the constant term is given by

$$\begin{aligned} \mathbf{L} &= \sum_i \sum_p C_{ip} \log(Y_{ip}) - Y_{ip} \qquad (8)\\ &= \sum_i \sum_p C_{ip} \log(I_{ip} * \textstyle\sum_k b_{pk} * r_{ik}) - I_{ip} * \textstyle\sum_k b_{pk} * r_{ik} \end{aligned}$$

The derivatives with respect o b_{pk} and r_{ik} is given by

$$\frac{\partial \mathbf{L}}{\partial r_{ik}} = \sum_p \frac{C_{ip}}{Y_{ip}} I_{ip} * b_{pk} - I_{ip} * b_{pk} \tag{9}$$

$$\frac{\partial \mathbf{L}}{\partial b_{pk}} = \sum_p \frac{C_{ip}}{Y_{ip}} I_{ip} * r_{ik} - I_{ip} * r_{ik}$$

the model parameters are learned iteratively by gradient ascent. Since the factors are non-negative, we derive the multiplicative updates by setting the rate parameter as given in [13]

$$b^{t+1}_{pk} = b^{t}_{pk} \frac{\sum_i \frac{C_{ip}}{Y^t_{ip}} I_{ip} r^t_{ik}}{\sum_i I_{ip} r^t_{ik}} \tag{10}$$

$$r^{t+1}_{ik} = r^{t}_{ik} \frac{\sum_p \frac{C_{ip}}{Y_{ip}} I_{ip} b^t_{pk}}{\sum_p I_{ip} b^t_{pk}}$$

The update equation for r_{ik} and b_{pk} are interdependent or cyclic, requiring multiple iterations to converge to an optimal value. At any iteration, if the optimal value for one of the parameters is reached, then the other parameter can be estimated with good accuracy. In particular, starting with good estimate of (b_{pk}) results in accurate estimate for r_{ik}. The estimator of b_{pk} given in (11) depends on I_{ip}. In order to get a good estimate of b_{pk}, we need to have sufficient number of impressions for each item in each position. In other words, we want I_{ip} to be sufficiently large for each (i, p). In practice, an item gets impressed in very few positions, making I_{ip} a sparse matrix. As a result, the b_{pk} estimates are not accurate, which in turn affects the r_{ik}. We can address this problem of noise caused due to data sparsity by adding a prior probability over the latent factors. Since we interpret these latent factors as probabilities, we want them to take values in the interval $[0 \;\; 1]$, therefore, we impose a Beta prior on the bias term. Further, the hyper-parameters of the prior distribution are independently set for each intent type, that is, $b_{pk} \sim \mathtt{Beta}(c_k, d_k)$.

Let B and T to denote the matrix whose elements are the latent factors b_{pk} and r_{ik} respectively. The posterior distribution of bias and relevance factors, given the data $\mathcal{D}$ is proportional to the likelihood,

$$Pr(B, R|\mathcal{D}) \propto Pr(C|Y) Pr(B|c_k, d_k) \tag{11}$$

The prior probability on bias factors is

$$Pr(B|c_k, d_k) = \prod_p \prod_k \frac{b_{kp}^{c_k - 1}(1 - b_{pk})^{d_k - 1}}{\mathbf{B}(c_k, d_k)} \tag{12}$$

$$\tag{13}$$

where $\mathbf{B}(c_k, d_k)$ is constant independent of b_{kp}. This changes the update equation for the bias factor as follows

$$b_{pk} = b_{pk} \frac{\sum_i \frac{C_{ip}}{Y_{ip}} r_{ik} + \frac{c_k - 1}{b_{pk}}}{\sum_i I_{ip} r_{ik} + \frac{d_k - 1}{1 - b_{pk}}} \tag{14}$$

We did not impose any priors on the relevance factor (true click through rate) as restricting both latent factors to prior may constrain the predictive power of the model. Nevertheless, we tried using the priors on both factors and the results were not better than using prior on single factor.

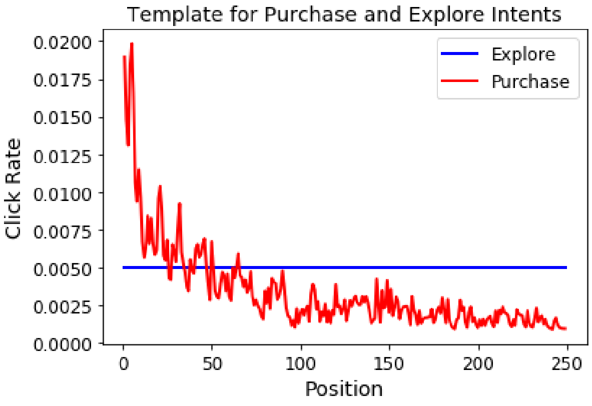

Figure 1: The plot illustrates the click through rate profile for two intent types - *purchase intent* and *explore intent*. The purchase intent falls rapidly with position where as the explore intent is a constant value.

The proposed generative model generalizes the *Examination Hypothesis* which factors the observed clicks into position bias and item relevance. The *Examination Hypothesis* can be retrieved from the proposed formulation by setting $k = 1$. The different intents or biases in the data can be identified by varying the number of intents k and plotting the latent (position bias) factor b_{pk} against p. In this work, we focus our efforts on two relevant intent types for Amazon search page, namely queries issued with intent to purchase and queries issued with intent to explore products. We evaluate the performance of our algorithm with respect to these two intents – *purchase intent* and *explore intent.*

One way to interpret the latent factors is that $\{b_{pk}\}$ terms represent the position bias profiles for different intent types and the relevance score $\{r_{ik}\}$ gives the strength of the items with respect to the intent types. Here, each intent is characterized by unique click through rate profile over the positions and the observed clicks are superposition of the click through rates of different intents. For example, users with purchase intent, typically issue a very specific query and the purchases typically happen from the early positions. Therefore, the typical profile for purchase intent is rapidly falling CTR with increase in position. On contrary, the explore intent is characterized by almost constant click through rates across positions. Figure 1 gives suggestive profile plots for the two intent types.

We leverage the above observation in the initialization of the terms b_{pk} and r_{ik}. Naive initializations of $\{b_{pk}, r_{ik}\}$ result in low-quality solutions [2]. In order to avoid it, for $k = 2$, we initialize the first intent factor $\{b_{p1}\}$ with observed CTR over positions and second intent factor $\{b_{p2}\}$ with constant values of $1e - 3$. The relevance terms need not be initialized and can be recursively estimated from the update equation (11).

4 PERFORMANCE EVALUATION

In this section, we present the performance of our intent based relevance estimation algorithm on the Amazon search dataset. The data consists of queries sampled from Amazon product search queries over a period of 30 days belonging to *Sports* category. We compare the proposed algorithm against the following baselines, where the prefix **SI** and **MI** respectively refer to algorithms with single and multiple intents.

- **SI-CoEC**: This is the well known click over expected clicks given by

$$\beta_p = \frac{\sum_i C_{ip}}{\sum_i I_{ip}}, \qquad CoEC_i \quad = \quad \frac{\sum_p C_{ip}}{\sum_i I_{ip} * \beta_p} \tag{15}$$

- **SI-Gamma**: refer to the factor model [4]. The hyper parameters on gamma prior reported by author did not give good results, and we set to `Gamma(1,1.01)`. This gave optimal results across different query subsets.
- **SI-Beta**: refer to the single intent Poisson model with beta prior on the bias factor. The Beta prior was set to `Beta(2,50)`
- **MI-Purchase** and **MI-Explore**: are the two intents from the proposed Multi Intent Poisson-Beta model run with $K = 2$. The factor $\{b_{p1}\}$ is initialized to observed CTR over positions and the factor $\{b_{p2}\}$ is set to constant value of $1e - 3$. The hyper parameters c_k and d_k are set as follows. $c_1 = 2, c_2 = 1/2$ and $d_1 = d_2 = 50$.

In order to train the generative model, we have collected the click logs data for period of 30 days, by aggregating the total number of impressions and clicks for each (item, position) pair across all sessions for a given query/keyword. For each query/keyword, the retrieved items could vary from 5 items to few thousands of items, impressed over multiple pages. An item could be placed in different positions across sessions. In this work, we clip the maximum position to 1500 as items placed beyond this position receive very little attention. For a given query, we discard items that received less than 5 aggregated impressions (across all sessions) from a position as these are spurious placements. For each query, the click/impression of an item in given position is smoothed across days as follows

$$\begin{aligned} \text{Weighted Click/Impr(d)} \quad &= \quad \alpha * \text{Weighted Click/Imp(d-1)} \\ &+ \quad (1 - \alpha) * \ \text{Click/Imp(d)} \end{aligned} \tag{16}$$

where parameter α is used as tradeoff between using the recent data and past data. The relevance score of each item for a given query is computed using the five different techniques listed above.

In order to evaluate the model performance, we consider the session level log data of all the train queries on the 31st day (one day following the last day in the train set) as test set. Here, for a given (query, session) the item that got clicked has ground truth relevance of 1 and all other items have ground truth relevance of zero. We use the MRR (Mean Reciprocal Rank) as metric to evaluate the performance which is computed as follows: For each (query, session), the items are ranked in the decreasing order of relevance scores, i.e. the item with highest relevance scores get a rank of 1 and second highest score gets rank 2 and so on. We take the reciprocal rank of the item that received a click as score for that session. If more than one item received a click, then max of reciprocal rank is taken as score for the session. The MRR is computed as mean of reciprocal rank across all sessions for a given query. If more than one item gets same rank, then the reciprocal rank is further weighted by the inverse of number of items with same rank. For example, if n items in a session receives same relevance rank k, and if any one of the n items receives a click then the reciprocal rank is $1/nk$. Table 1 gives the MRR that is averaged across all the queries/keywords.

It is clear from Table 1 that the **SI-CoEC** performs poor compared to all other algorithms across different query categories. This is because of the data sparsity where the occasional and spurious clicks on the lower positions artificially inflates the item relevance or click through rates. The Poisson model with `Gamma` and `Beta` priors smooths these spurious bursts occurring due to data sparsity and estimates the item relevance. Both these priors can model different shapes of densities and give similar performance on our query corpus. However, the single intent Poisson model with beta prior performs slightly better than the Gamma prior especially on queries with more words.

The proposed multi-intent model with two intents gives the best result on this data set. In particular, the **MI-Purchase** corresponding to *purchase intent* gives the highest MRR across all the query categories. It outperforms all other relevance measures when the query length increases. Recall that the purchase intent does not reflect actual purchases, rather it corresponds to click pattern or template that is characterized by rapidly falling CTR with more clicks in the early positions. In general, query with more words tend to be very specific targeting few relevant products with higher intent to purchase. For such queries, most user actions (clicks, purchases, add to carts) happens on items displayed in early positions. In next section, we highlight the difference between the templates for *purchase intent* and well known position bias.

One curious observation in the Table 1 is that the MRR from *explore intent* matches closely with *purchase intent* and is higher than all other single intent models. This is because the MRR is averaged across all the keywords and our finding is that not both intent types are equally relevant for each keyword. We need a mechanism to classify each keyword into major intent types so that the right relevance scores are used for evaluating the performance. In the Table 1, the **MI-Oracle** is computed as maximum of MRR between the two intent types, reflecting the MRR by an oracle that assigns right intent types to the keywords.

The left plot in the Figure 2 shows the change in MRR with respect to the query length. As query length increases, it becomes more specific and fewer results are retrieved, therefore the maximum reciprocal rank in general for increases for all the algorithms. For example, when 100 items are retrieved, the minimum possible MRR is 0.01(1/100) whereas when only 10 items are retrieved, the minimum possible MRR is 0.1. The right plot in Figure 2 illustrates this fact where the MRR decreases with increase in number of retrieved items.

It should be noted that each query can be issued with either of the intent types and we need to identify the dominating intent type for each query. The `Intent Hypothesis` as expressed in equation (5) has provision to capture the proportion of sessions expressing a given intent type. However, the proposed generative model treat all

Table 1: MRR on different slices of the search data with 50K queries from sports category

Algorithm	All queries	Single word queries	Queries with 1-3 words	Queries with 4+ words	Queries with 7+ words
MI- Oracle	0.669	0.595	0.657	0.714	0.907
MI- Purchase	0.562	0.491	0.546	0.621	0.87
MI- Explore	0.549	0.476	0.534	0.606	0.83
SI- Poisson-Beta	0.542	0.458	0.526	0.603	0.84
SI- Poisson-Gamma	0.532	0.451	0.515	0.593	0.82
SI- CoEC	0.440	0.362	0.418	0.517	0.79

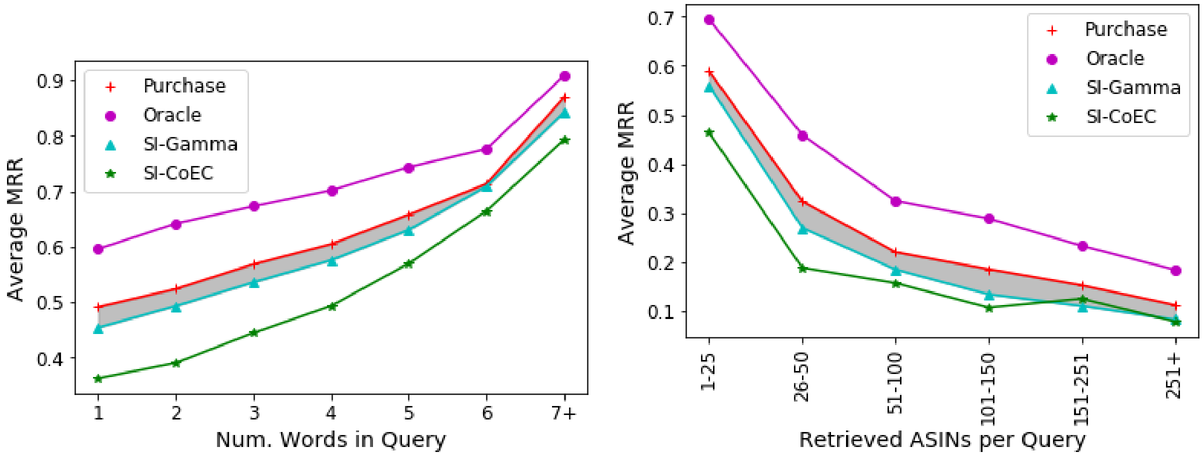

Figure 2: [Best viewed in color.] Plot comparing the MRR scores between *purchase intent*, Poisson-Gamma model [4] and CoEC. The shaded region is the lift between the MRR of *purchase intent* and SI-Gamma model.

the probabilities as factors and this proportion is unevenly absorbed into the bias and relevance scores.

4.1 Analysis on Intent Templates

In this section, we present the intent templates estimated by the proposed model. As mentioned earlier, the purchase and explore intent templates were initialized to observed CTR and constant values. The templates are recursively updated using equation (11) and (14). In Figure 3 we plot the final intent profiles estimated by the algorithm at convergence. The plot is shown for queries *Fishing*, *Fishing poles*, and *Fishing float tubes* As can be seen, the queries are in the order of increasing specificity. In these plots, the purchase and explore intent are shown in red and blue color. The actual CTR is shown in green color and the position bias estimated by Poisson-Beta model is shown in cyan color.

The query *Fishing* is a generic word which retrieves lot of items from several different categories namely – fishing gadgets or equipment, fishing related apparels, fishing related books etc. Clearly, this query has a strong exploratory intent and in practice, we have noticed that the users have clicked items placed at several different positions across multiple pages. The right plot in Figure 3 shows the intents and position bias for query *fishing*, where the *purchase intent* (red line) rapidly decays to zero compared to the position bias (cyan color) from **SI-Beta** model. As the purchase intent decreases, the *explore intent* (blue line) increases and reaches a constant level. This suggests that in early positions the purchase intent dominates over the exploration intent. Here, the magnitude of the intents does not convey the strength of intent and it should be seen with the magnitude of the corresponding relevance factors, however the trend conveys the intent type.

When the query *Fishing* is refined to *Fishing poles*, the purchase intent drops more rapidly compared to the purchase intent in the *Fishing* query as shown in the middle plot of Figure 3. Empirically we have observed that as we move from generic to specific query the purchase intent rapidly drops and the exploration intent starts to dominate from early positions. This is because for very specific query, there are few relevant items and most of them are placed in early positions. Users going past these early positions are assumed to explore the limited choice of items presented to them.

Finally, the query *Fishing float tubes* has very sparse clicks resulting in spiked observed CTR in two different positions. Both the bias correction algorithm and multi intent algorithm smooths this erratic CTR by assigning an almost constant click through rate (there is some variability in red/blue lines, but not seen because of magnitude of green curve) to all positions. Here, all the observed clicks are absorbed into the relevance factors (r_{ik}). From intent perspective, since there is lack of information (clicks) the model assigns single intent as seen from the similar trend for both the intent factors. For this query the relevance terms corresponding to the two intents have different magnitude but they have same ordering of items, resulting in large positive correlation between the relevance scores of both intent types.

For all the queries, the final intent templates computed by the EM model is quite different from the initialized values. The *purchase intent* templates are initialized to green curve and converges to the red curve. The *explore intent* parameters are initialized to constant value and converges to the blue curve. In both the cases, the interpretation of the model identified intent templates remain the same. The iterative updates uses the observed clicks data to move the template from initial values to the final state. For queries with click sparsity, both the intent templates end up being looking similar as seen for the query *Fishing float tubes*.

The EM based iterative updates does not always converge to distinct intent signature for each query as evident from the final templates for the query *Fishing float tubes*. Figure 4 shows the model estimated intent template for three queries *Tennis shoes, Men's tennis shoes and nike men's tennis shoes*. Here again, we observe two distinct position bias profiles for first two queries and not for the third query. For this query, both the estimated intent template curves are identical to *explore intent* but of different magnitude. In this case, the estimated relevance scores are highly correlated resulting in same MRR value across intent types. This shows that the model does not force fit relevance for the initialized intent template, rather the data plays an important role in identifying the underlying intent template.

4.2 Analysis on Relevance Scores

In this section, we present our analysis on the relevance score $\{r_{ik}\}$ corresponding to the intent factors. Figure 5 shows the change in MRR as function of correlation between the intent factors. A large positive correlation between the two intent factors implies presence of single intent. A large negative correlation implies strong presence of multiple intents where the two intent factors show an opposing trend (increasing/decreasing) in the early positions. When two distinct intents are present, the *purchase* intent outperforms the MRR from *explore* intent.

5 CONCLUSIONS

Estimating the true item relevance for a given query is important problem in search relevance ranking. Most of the past work has focused on removing the position bias. In this work, we define intent types as click signatures over positions and propose an approach to estimate relevance with respect intent type. In particular we define purchase and explore intents and empirically found that the relevance scores from purchase intent outperform the relevance scores from position bias correction algorithms. In addition, we have shown that the ranking performance (MRR) can be boosted by more than 20% by knowing the cardinal intent behind each keyword. The proposed work is in discovery stage and has shown promising results on amazon internal data sets.

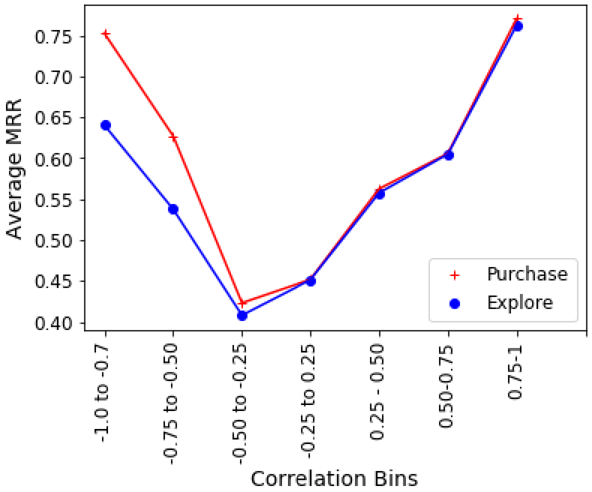

Figure 5: [Best viewed in color.] Plot shows the MRR as function of correlation between the purchase and explore relevance factors.

REFERENCES

[1] S. Athey and D. Nekipelov. 2010. A structural model of sponsored search advertising auctions. In *Proceedings of the 6th Ad Auctions Workshop*.

[2] Jean-Patrick Baudry and Gilles Celeux. 2015. EM for mixtures. *Stat. Comput* (2015), 713–726.

[3] Olivier Chapelle and Ya Zhang. 2009. A Dynamic Bayesian Network Click Model for Web Search Ranking. In *Proceedings of the 18th International Conference on World Wide Web*. 1–10.

[4] Ye Chen and Tak W. Yan. 2012. Position-normalized Click Prediction in Search Advertising. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. 795–803.

[5] Charles L. A. Clarke, Eugene Agichtein, Susan Dumais, and Ryen W. White. 2007. The Influence of Caption Features on Clickthrough Patterns in Web Search. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 135–142.

[6] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An Experimental Comparison of Click Position-bias Models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM '08)*. ACM, New York, NY, USA, 87–94. DOI:http://dx.doi.org/10.1145/1341531.1341545

[7] G. E. Dupret and B. Piwowarski. 2008. A user browsing model to predict search engine click data from past observations.. In *SIGIR*. 331 –âĂŞ 338.

[8] B. Edelman, M. Ostrovsky, and M. Schwarz. 2005. Internetadvertising and the generalized second price auction:selling billions of dollars worth of keywords. In *American Economic Review*. 242 –âĂŞ 259.

[9] T. Graepel, J. Q. Candela, T. Borchert, , and R. Herbrich. 2010. Web-scale Bayesian click-through rate prediction for sponsored search advertising in MicrosoftâĂŹs Bing search engine.. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*. 13 –âĂŞ20.

[10] Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking Analysis of User Behavior in WWW Search. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*. 478–479.

[11] Botao Hu, Yuchen Zhang, Weizhu Chen, Gang Wang, and Qiang Yang. 2011. Characterizing Search Intent Diversity into Click Models. In *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*. ACM, New York, NY, USA, 17–26. DOI:http://dx.doi.org/10.1145/1963405.1963412

[12] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately Interpreting Clickthrough Data As Implicit Feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*. 154–161.

[13] Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by nonnegative matrix factorization. *Nature* 401 (1999), 788–791.

[14] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting Clicks: Estimating the Click-through Rate for New Ads. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. 521–530.

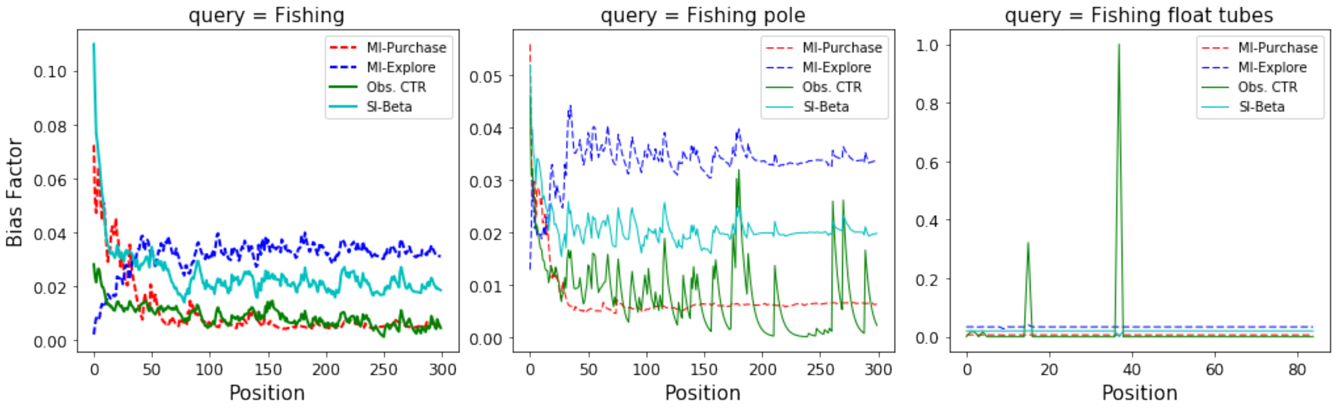

Figure 3: [Best viewed in color.] Plot showing the click through rate /intent templates at each position for three different queries. The CTR for *Fishing* and *Fishing pole* queries are smoothed for better presentation. The green curve corresponds to observed CTR over position averaged across sessions. The cyan curve is the position bias obtained from SI-Beta model. The red and blue lines corresponds to the *purchase* and *explore* intent templates obtained from the proposed generative model. The purchase intent factors were initialized to green curve and converges to the red curve. The explore intent factor was initialized to constant and converges to blue curve.

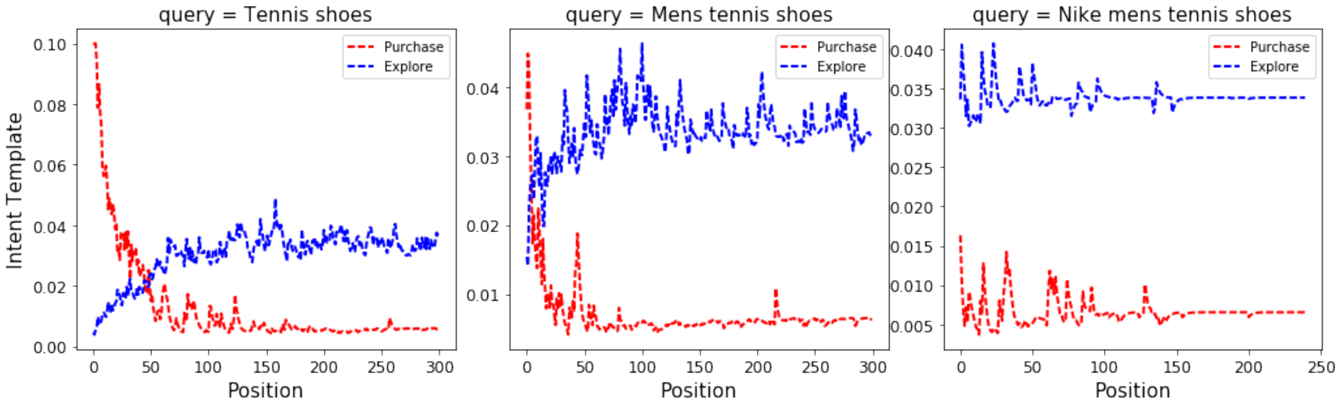

Figure 4: [Best viewed in color.] Plot showing the intent templates estimated by the model for three queries *Tennis shoes, Men's tennis shoes and nike men's tennis shoes.*