Investigating the Reliability of Click Models

Jiaxin Mao, Zhumin Chu, Yiqun Liu*, Min Zhang, and Shaoping Ma
Department of Computer Science and Technology, Institute for Artificial Intelligence, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
Beijing, China
yiqunliu@tsinghua.edu.cn

ABSTRACT

Click models aim to extract accurate relevance feedback from the noisy and biased user clicks. Previous work focuses on reducing the systematic bias between click and relevance but few studies have examined the reliability and precision of click models' relevance estimation. So in this study, we propose to investigate the reliability of relevance estimation derived by click models. Instead of getting a point estimate of relevance, a variational Bayesian method is used to infer the posterior distribution of relevance parameters. Based on the posterior distribution, we define measures for the reliability of pointwise and pairwise relevance estimation. With experiments on both real and synthetic query logs, we show that: 1) the proposed method effectively captures the uncertainty in relevance estimation; 2) the reliability of click models' relevance estimation is affected by the size of training data, the average ranking position of documents, and the ranking strategy of search engines.

CCS CONCEPTS

• **Information systems** → **Web search engines**; *Users and interactive retrieval*; • **Theory of computation** → *Bayesian analysis*;

KEYWORDS

Click Model; Web Search; Bayesian Analysis

ACM Reference Format:
Jiaxin Mao, Zhumin Chu, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating the Reliability of Click Models. In *The 2019 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '19), October 2–5, 2019, Santa Clara, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3341981.3344242

1 INTRODUCTION

User clicks carry implicit relevance feedback that is valuable for improving the ranking performance of Web search engines. However, the click signal is *noisy* and affected by different kinds of behavioral *biases* (e.g. the position bias [8] and presentation bias [12, 13]), making it systematically different from true relevance. To extract unbiased relevance feedback from the biased click signal, a series of click models (see Chuklin et al. [4] for a survey) have been proposed to model users' click behavior on the SERP as a stochastic process. By making assumptions on how the behavioral biases affect users' clicking behavior, a click model can estimate the influence of the behavioral biases and the relevance of each query-document pair, respectively. After training the click model on query logs, we can get *less* biased relevance estimations and use them in downstream tasks. For example, we can use the click-based relevance estimation as ranking features to train a learning-to-ranking model [3] and use them as weak supervision signals to train and test data-hungry neural ranking models [9, 11].

To improve the performance of these downstream tasks, we need to ensure that the relevance estimation given by the click model is *accurate*. Generally, the *accuracy* of estimation depends on two factors: its *trueness* and *precision*, where the trueness is the estimate of the systematic error (i.e. the bias of the estimate) and the precision is the estimate of the random error (i.e. the variance of the estimate).

Previous work on click models has made a great effort in reducing the systematic bias and improving the trueness of relevance estimation by experimenting with different user behavior assumptions and building more sophisticated models [2]. However, few studies have investigated the precision and *reliability* of the relevance estimation given by click models. Intuitively, we can have a precise and reliable relevance estimation of a query-document pair if we have a large enough query log and the pair occurs many times in it. However, because the parameters of click models are often learned with *point estimators* (e.g. the maximum likelihood estimator and the expectation-maximization (EM) algorithm), it is difficult to quantitatively measure the precision of the resulted estimates. Therefore, some fundamental questions that may affect the validity of click models are left unanswered: 1) How precise is the relevance estimation of a query-document pair? 2) How many impressions are needed for obtaining a reliable relevance estimation of a single query-document pair? 3) How reliable is the estimation of the relative order of two documents in terms of their relevance?

Addressing these questions can help us figure out how to train a better click model as well as how to better utilize the relevance estimation in downstream tasks. For training the click model, we can better determine the size of the training set if we can calculate how many impressions are needed for obtaining a reliable relevance estimation. For utilizing the relevance estimation in other tasks, we can select the most credible instances to train the pairwise learning to rank model if we can measure the reliability of the pairwise relevance estimation of two documents.

So in this work, we focus on investigating the reliability of the relevance estimation given by click models. We adopt an existing click model, the Bayesian browsing model (BBM) [10], in our study. This model posits the same assumptions on user behavior as the User browsing model (UBM) [6]. But instead of using the MLE or EM algorithm to obtain a point estimate of the relevance parameters,

*Corresponding author

ICTIR '19, October 2–5, 2019, Santa Clara, CA, USA

ACM ISBN 978-1-4503-6881-0/19/10…$15.00
https://doi.org/10.1145/3341981.3344242

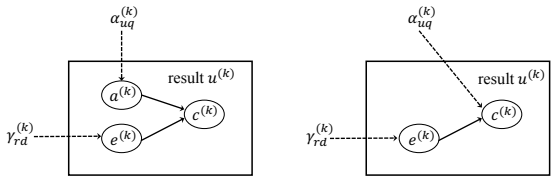


Figure 1: The graphical model representations of UBM (left) and BBM (right)

a Bayesian approach is used to infer their *posterior distribution*. With this posterior distribution, we can measure the precision of the *pointwise* relevance estimation (i.e. the relevance estimation of each query-document pair) and the reliability of *pairwise* relevance estimation (i.e. an estimation of the ordering of two documents in relevance).

To test whether the proposed model can capture the uncertainty of relevance estimation, we train the proposed model on real query logs and examine how the variance of relevance estimation changes with the number of impressions and the average ranking positions. We further conduct experiments on a synthetic dataset to analyze how the ranking performance of systems influence the reliability of click-model-based relevance estimation.

The rest of the paper will be organized as follows. In Section 2, we introduce the BBM and the variational inference method used to infer the posterior distribution of relevance parameters. Then in Section 3, we introduce the experiments on both real and synthetic query logs and report the results of these experiments. We further discuss the experiment results in Section 4 and finally conclude the paper in Section 5.

2 MODELS

In this section, we introduce the click model and the variational bayesian method we used in this study.

2.1 Bayesian Browsing Model

We adopt the Bayesian Browsing Model (BBM)[10] in this study. The BBM is inspired by a widely-used click model, the User Browsing model (UBM)[6], as they share similar assumptions on user behavior. Figure 1 shows the the graphical model representations of UBM and BBM.

The UBM assumes that the user scans the SERP (search engine result page) from the top to bottom. It follows the examination hypothesis [5] that the user will click it if and only if it is examined by the user and it is attractive. If we use three binary variables c, a, e to denote whether the user click the document, whether the document is attractive, and whether it is examined by the user, respectively, this hypothesis can be formulated as:

$$a = 1, e = 1 \iff c = 1. \tag{1}$$

The UBM further assumes a fully depends on the relevance between u and q and e depends on the position of the document r and its distance the the last click d:

$$\begin{aligned} P(a = 1|\alpha_{uq}) &= \alpha_{uq} \\ P(e = 1|\gamma_{rd}) &= \gamma_{rd} \end{aligned} \tag{2}$$

The BBM omits the latent variable a in UBM and lets the parameter α directly determine the click. Then the joint probability of observable click variable and latent examination variable $p(e, c|\alpha, \gamma)$ can be defined as the following formulae:

$$\begin{aligned} p(e = 1, c = 1|\alpha, \gamma) &= \gamma\alpha \\ p(e = 0, c = 0|\alpha, \gamma) &= 1 - \gamma \\ p(e = 1, c = 0|\alpha, \gamma) &= \gamma(1 - \alpha) \end{aligned} \tag{3}$$

2.2 Variational Inference for BBM

In the original paper that introduces the BBM, Liu et al. [10] proposed an efficient algorithm to numerically calculate $P(\alpha|Obs)$. However, in this work we choose to use a mean field variational inference method [1] to approximate the posterior distribution of α with a member of the exponential family. The advantage of using the variational inference method in this study is that we can analytically investigate the approximate posterior distribution once we learn the parameters for the exponential family distribution.

For variational inference, we define a variational distribution q for each parameter and latent variable and use it to approximate the true posterior distribution of the corresponding parameter or latent variable. For the BBM, we restrict the function $q_\alpha(\alpha)$ and $q_\gamma(\gamma)$ in the independent space of Beta distribution [7], i.e. $q_{\alpha_i}(\alpha_i) \sim Be(m_{i1}, m_{i2})$, $q_{\alpha_i}(\gamma_i) \sim Be(n_{i1}, n_{i2})$, where the parameters m_{i1}, m_{i2}, n_{i1}, n_{i2} can be learned in the training process. By further using the uniform distribution (i.e. $Be(1, 1)$) as the prior of α and γ, we can ensure that the variational distribution $q_\alpha(\alpha)$ and $q_\gamma(\gamma)$ are closed in such restriction. Eq. 4-6 show the updating formulae for $q_\alpha(\alpha)$, $q_e(e)$, and $q_\gamma(\gamma)$:

$$q_{e_k}^{(t+1)}(e_k) \propto \begin{cases} e_k, & c_k = 1 \\ \exp\left[\psi(n_{k1}^{(t)}) + \psi(m_{k2}^{(t)})\right], & e_k = 1, c_k = 0 \\ \exp\left[\psi(n_{k2}^{(t)}) + \psi(m_{k1}^{(t)} + m_{k2}^{(t)})\right], & e_k = 0, c_k = 0 \end{cases} \tag{4}$$

$$q_{\alpha_{uq}}^{(t+1)}(\alpha_{uq}) \sim Be(1 + \sum_{i=1}^{M} I_i^1, 1 + \sum_{i=1}^{M} I_i^0 q_{e_i}^{(t+1)}(1)) \tag{5}$$

$$q_{\gamma_{rd}}^{(t+1)}(\gamma_{rd}) \sim Be(1 + \sum_{j=1}^{N} I_j^1 + I_j^0 q_{e_j}^{(t+1)}(1), 1 + \sum_{j=1}^{N} I_j^0 q_{e_j}^{(t+1)}(0)) \tag{6}$$

Here, $\psi(x)$ is the digamma function, and I_i^j is an indicator function that $I_i^j = 1$ if and only if $c_i = j$. To train the BBM, we iteratively update $q_\alpha(\alpha)$, $q_e(e)$, and $q_\gamma(\gamma)$ with these formulae until they finally converge.

2.3 Reliability Measures

After learning the approximate posterior distribution $q_\alpha(\alpha)$, we can derive measures the reliability of both pointwise relevance estimation (i.e. how reliable the relevance estimation of a query-document pair is) and pairwise relevance estimation (i.e. how reliable the estimation of the relative relevance order of two document is).

To measure the reliability of the pointwise relevance estimation , we compute the *variance* of the posterior distribution. Because we restrict $q_{\alpha_i}(\alpha_i) \sim Be(m_{i1}, m_{i2})$, the variance of α_i can be approximated by the variance of the Beta distribution:

$$Var[\alpha_i] = \frac{m_{i1} \cdot m_{i2}}{(m_{i1} + m_{i2})^2(m_{i1} + m_{i2} + 1)} \tag{7}$$

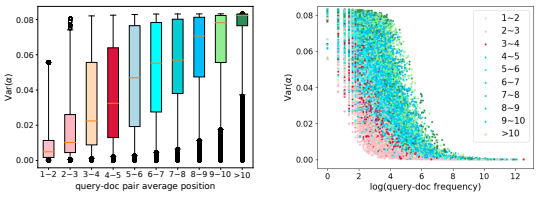
For two documents u and v in the same query q, if we know that u is more relevant to query q than v, we compute the following probability as a measure for pairwise relevance estimation:

$$P(\alpha_{uq} > \alpha_{vq}) = \int\int_{x>y} q_{\alpha_{uq}}(x) q_{\alpha_{vq}}(y) dx dy \tag{8}$$

A higher $P(\alpha_{uq} > \alpha_{vq})$ indicates that the trained click model reliably captures this ordered relationship in relevance and a $P(\alpha_{uq} > \alpha_{vq})$ that is near 0.5 may suggest that the pairwise relevance estimation is very

uncertain. However, if we do not know which document is more relevant to the corresponding query q, we can use $max\{P(\alpha_{uq} > \alpha_{vq}), P(\alpha_{uq} < \alpha_{vq})\}$ as a measure for the pairwise relevance estimation.

3 EXPERIMENTS



(a) different average ranking position **(b) different query-doc pair frequency**

Figure 2: The variance of relevance estimations $Var[\alpha]$ under different (a) average ranking positions and (b) query-doc pair frequencies.

In this section, we introduce the experiment settings and results on both real and synthetic query logs.

3.1 Experiment On Real Dataset

We first conduct experiment on real query logs from a commercial search engine. Table 1 shows the statistics of the query logs.

Table 1: The statistics of the real search logs

# unique queries	1,909
# unique query-document pairs	75,204
# query sessions	1,453,647
# document impressions	15,440,560
# clicks	1,879,532

We train the BBM on the query logs using the variational inference method described in Section 2.2. Figure 2 shows the variance of relevance estimations for the query-document pairs with different average ranking positions and different frequency. In Figure 2a, we can see the average ranking position affects relevance estimation's reliability. The document ranked in the lower position tends to have a relevance estimation with higher variance. Figure 2b further shows that the variance of relevance estimation also depends on the frequency of the query-document pair (i.e. # impressions). From these two figures, we can see that the BBM trained with variational inference method can capture the uncertainty in relevance estimation. Theses results are consistent with our intuition that the query-document pair that are more likely to be examined by users has a more precise relevance estimation.

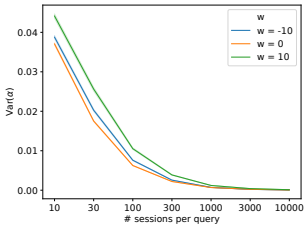


Figure 3: The variance of relevance estimations $Var[\alpha]$ with different ranking performance.

3.2 Experiments on Synthetic Dataset

To further analyze the reliability of click-model-based relevance estimation, we also conduct experiment on synthetic datasets so that we can compare the relevance estimations with ground-truth relevance labels under different settings.

3.2.1 Data generation. We adopt the Position-Based Model (PBM) introduced in [5] as our generative model. The Position-Based Model assumes that the click probability on a document depends on its the attractiveness α_{uq} and ranking position r:

$$p(c = 1) = \alpha_{uq} \cdot \gamma_r \tag{9}$$

We trained the parameter for position-bias γ_r on real search logs used in Section 3.1. For the relevance parameters α_{uq}, we first sample the parameters $\beta_q^{(1)}$, $\beta_q^{(2)}$ for query q uniformly from the interval $[2, 4]$ and then randomly sample the relevance parameter for each document u in query q from the Beta distribution: $\alpha_{uq} \sim Be(\beta_q^{(1)}, \beta_q^{(2)})$. In total, we generate the ground-truth relevance parameters for 500 queries and 5,000 unique query-document pairs

We want to investigate how the ranking performance influence the reliability of relevance estimation, so after sampling the ground-truth relevance parameters α_{uq}, we use the following method to generate ranking list with different ranking performance. Considering a particular query q, we rank the candidate documents from top to bottom. Let S_r be the set of document already ranked among the top-r positions ($S_0 = \Phi$), we select a document u and rank it at $(r + 1)$-th position with the following probability:

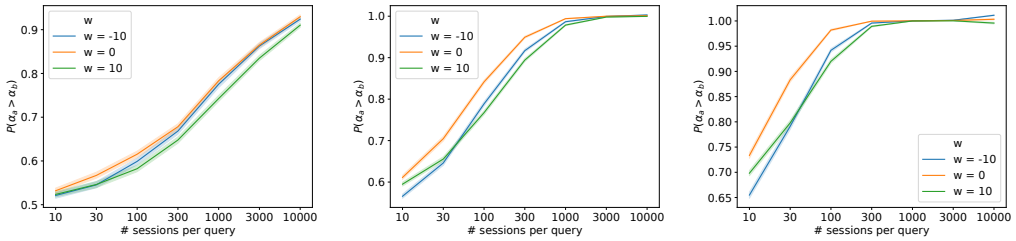
$$P(R_u = r + 1 | S_r, \alpha) = \frac{I(u \notin S_r) e^{w\alpha_{uq}}}{\sum_{u' \notin S_r} e^{w\alpha_{u'q}}} \tag{10}$$

The hyperparameter w determines the ranking performance. A larger w is associated with a better ranking performance. When $w = 0$, the generation process downgrades to the random permutation of documents. When $w < 0$, the documents will be ranked in an inverse order, i.e. less relevance documents will have a larger probability to be ranked at higher positions. In this study, we set $w = -10, 0, 10$ to simulate different levels of ranking performance.

For each query session, we repeat this process to generate a ranking list and simulate user clicks with Eq. 9. We also vary the number of sessions per query from 10 to 10,000 to analyze how the reliability of relevance estimation will change if we have more training data.

3.2.2 Pointwise analysis. We first investigate the reliability of pointwise relevance estimation. Figure 3 shows the mean and standard derivation (in shaded error bands) of the variance $Var[\alpha]$ under different settings of w and # sessions per query. We can see that when the session number increases, the mean and standard derivation of the variance $Var[\alpha]$ both decrease. When the number of sessions per query is over 3,000, the variance of α is nearly zero, suggesting that the confidence interval of relevance estimation becomes extremely tight. We also find that the ranking performance does influence the variance of pointwise relevance estimation. The variance under the random ranking orders (i.e. $w = 0$) is lower than other settings and a better ranking performance seems to harm the reliability of relevance estimation as we observe the highest variance when we set $w = 10$.

3.2.3 Pairwise analysis. We also investigate the reliability of pairwise relevance estimation by investigating whether the click model can consistently capture a small ($0 < \alpha_a - \alpha_b \leq 0.1$), medium ($0.1 < \alpha_a - \alpha_b \leq 0.3$), and large ($\alpha_a - \alpha_b > 0.3$) difference in relevance between two documents a and b in the same query. Figure 4 shows the probability $P(\alpha_a > \alpha_b)$ for two documents with the small, medium, and large different, respectively. From the results, we can see that when the number of sessions per query increase, the reliability of pairwise relevance estimation measured by $P(\alpha_a > \alpha_b)$ increases accordingly. For the document pair with a large(medium) difference



(a) small difference ($0 < \alpha_a - \alpha_b \leq 0.1$)

(b) medium difference ($0.1 < \alpha_a - \alpha_b \leq 0.3$)

(c) large difference ($\alpha_a - \alpha_b > 0.3$)

Figure 4: The reliability of pairwise relevance estimation measured by $P(\alpha_a > \alpha_b)$ for document pairs with (a) small, (b) medium, and (c) large difference in relevance.

in relevance, it needs around 100(300) sessions per query to obtain a reliable pairwise relevance estimation ($P(\alpha_a > \alpha_b) > 0.9$). However, it is difficult to reliably detect the small difference in relevance as we may need over 10,000 sessions to achieve the same criteria. We also find that the random ranking orders results in a more reliable pairwise relevance estimation, which is consistent with the findings in the analysis on pointwise relevance estimation.

4 DISCUSSION

Before discussing the experiment results and implications, we acknowledge some limitations of this study. First, in this study, we only adopt one click model, the BBM, as an example click model. In future work, we can investigate the reliability of other, presumably more sophisticated, click models. By comparing the reliability of different models, we may investigate the tradeoffs between bias and variance in estimating relevance. Second, besides the ranking position, numbers of impressions, and ranking performance investigated in this study, many other factors may affect the reliability of click models. For example, a high level of noise in user clicks may harm the reliability of relevance estimation derived by the click models. We can extend the experiments on synthetic datasets to analyze the influence of the noise in click signals in future work.

From the experiments on both real and synthetic datasets, we find that: 1) there exists a considerable level of uncertainty in the relevance estimation of click models and it can be captured by the proposed Bayesian approach (Section 3.1). 2) we can get more reliable relevance estimation for the query-document pairs that are ranked in higher positions and presented more frequently (Section 3.1); 3) for click models, it is difficult to reliably detect small difference in relevance (Section 3.2.3); 4) The ranking performance may influence the reliability of click models and randomly shuffling the rankings can help to reduce the variance of relevance estimation (Section 3.2.3).

These findings suggests that we should consider the inherent uncertainty when using the relevance estimation of click models in other tasks. For example, when training a pairwise learning to rank model with the click relevance, we can use the pairwise reliability measures (e.g. $|2P(\alpha_a > \alpha_b) - 1|$) to filter out some unreliable pairs of documents. Besides, the results also suggest that we can also proactively improve the reliability of click models by deliberately putting new documents at top positions or incorporating random explorations in ranking (e.g. randomly shuffling the ranking list).

5 CONCLUSIONS

In this study, we investigate the reliability of the click model using a Bayesian approach. We infer the posterior distribution of the relevance estimation given by click models with the variational inference method and propose two measures for the reliability of pointwise and pairwise relevance estimations. We further conduct experiments on both real and synthetic query logs to show how the size of training data, the ranking position, and the ranking performance of systems affect the precision of pointwise and pairwise relevance estimation. Experiment results emphasize the importance of considering the reliability of click models and provide some useful implications.

ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China (2018YFC0831700) and Natural Science Foundation of China (Grant No. 61622208, 61732008, 61532011) . This research is partly supported by the Tsinghua-Sogou Tiangong Institute for Intelligent Computing.

REFERENCES

[1] Matthew James Beal et al. 2003. *Variational algorithms for approximate Bayesian inference*. university of London London.
[2] Alexey Borisov, Ilya Markov, Maarten de Rijke, and Pavel Serdyukov. 2016. A neural click model for web search. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 531–541.
[3] Olivier Chapelle and Ya Zhang. 2009. A dynamic bayesian network click model for web search ranking. In *WWW '09*. ACM, 1–10.
[4] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. 2015. Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 7, 3 (2015), 1–115.
[5] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *WSDM'08*. ACM, 87–94.
[6] Georges E Dupret and Benjamin Piwowarski. 2008. A user browsing model to predict search engine click data from past observations.. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 331–338.
[7] Arjun K Gupta and Saralees Nadarajah. 2004. *Handbook of beta distribution and its applications*. CRC press.
[8] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR'05*. Acm, 154–161.
[9] Xiangsheng Li, Jiaxin Mao, Chao Wang, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Teach Machine How to Read: Reading Behavior Inspired Relevance Estimation. In *SIGIR'19*.
[10] Chao Liu, Fan Guo, and Christos Faloutsos. 2009. Bbm: bayesian browsing model from petabyte-scale data. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 537–546.
[11] Cheng Luo, Yukun Zheng, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2017. Training deep ranking model with weak relevance labels. In *Australasian Database Conference*. Springer, 205–216.
[12] Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Constructing Click Models for Mobile Search. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 775–784.
[13] Chao Wang, Yiqun Liu, Min Zhang, Shaoping Ma, Meihong Zheng, Jing Qian, and Kuo Zhang. 2013. Incorporating vertical results into search click models. In *SIGIR'13*. ACM, 503–512.