Characterizing Expertise of Search Engine Users

Qianli Xing *, Yiqun Liu, Min Zhang, Shaoping Ma, and Kuo Zhang

State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
xingqianli@gmail.com, {yiqunliu, z-m, msp}@tsinghua.edu.cn,
zhangkuo@sogou-inc.com

Abstract. Search engine click-through data is a valuable source of implicit user feedback for relevance. However, not all user clicks are good indication of relevance. The clicks from search experts, who are more successful searching a query, tend to be more reliable in indicating document relevance than those of the non-experts. Therefore, knowing the expertise of search users is helpful to better understand their clicks. In this paper, we propose two probabilistic modelings of user expertise in the environment of web search. Inspired by the idea of evaluation metrics in classification, search users are treated as classifiers and result documents are viewed as the data samples to classify in our models. A click implies that the document is classified as relevant by the user. Therefore, the expertise of a user can be measured by how well he/she classifies the documents. We carry out experiments on a real-world click-through data of a Chinese search engine. The results show that modeling user expertise helps the click models with relevance inference, which also implies that our models are effective in identifying the user expertise.

Keywords: User expertise, Click-through data, web search

1 Introduction

Search engines collect a large amount of user interaction logs everyday when people search the Web. Among these logs, click-through data has drawn a lot of attention because of the relevance information embedded in it. Although the click data might be noisy, it can still reflect users' relevance judgments towards the documents to certain extent. Previous studies [8, 1] have presented that the relevance preferences can be extracted from the click-through data and the quality of the extracted result is even comparable to human annotations. Such implicit relevance information in click logs can be used to evaluate and improve the search engine performance. A big advantage of using user clicks for relevance is that they can be collected at low costs and the scale is far big than that the human annotation can do. Therefore, a lot of methods are proposed in an attempt to mine relevance from clicks [5–7, 2]. The core idea of these methods is to use the wisdom of the crowd in the clicks.

* This work was supported by Natural Science Foundation (60903107, 61073071) and National High Technology Research and Development (863) Program (2011AA01A205) of China.

However, not every click is equally informative to indicate relevance for various reasons. During a search, some users may used to click multiple documents at a time without careful selection; some users are bad at making relevance judgment so their clicked documents may not be as relevant as they thought. Generally, the experienced search users are more likely to accomplish a successful search while the novices may have trouble finding the relevant documents. White et al. [13] investigated the behavioral variability among search users. The results showed that the users had considerable differences in some key aspects of search, such as querying, browsing and clicking. It is also reported that the users with domain knowledge have larger percentage of successful in-domain search sessions [14]. Being aware of the diversity of search users, it is essential to take user expertise into consideration to better understand the clicks.

It is a big challenge to characterize the expertise of a search user because it is hard to define an expert in the Web search scenario. The previous studies used some simple ways to identify search experts. For example, White et al. [12] considered advanced users in Web search to be those who had issued queries with advanced syntax. As to domain search experts, the proportion of expert sites (assessed by domain experts) that a user visited was used to approximate the expertise [14]. However, these methods are neither accurate nor formal enough to be widely adopted.

In this paper, we propose two probabilistic methods to define and model the search expertise for search users. Our models assume that a click event depends on both the document relevance and the user's expertise, which most click models usually ignore. The process that a user makes relevance judgment documents is viewed as a classification task. User is the classifier and the documents are data samples to classify. The expertise of a search user can be then measured by the classifying performance. In our experiment, the parameters of expertise and relevance are estimated with a large-scale click log from a Chinese search engine. We also carry out a series of experiments to evaluate the effectiveness of the proposed models.

2 Related work

Previous studies have shown that the click-through data are useful but meanwhile noisy. Clicks from different search users may not be equally informative in relevance indication. Search users who issued queries with advanced syntax were reported to be more successful in their search sessions by [12]. In that study, the expertise of a user is viewed as the percentage of his/her queries that include advanced syntax. The experiment results showed that the average relevance of the clicked documents by the advanced users (i.e. users that issued queries with advanced syntax) are higher than that of the non-advanced users. And the more advanced syntax one uses, the more successful his/her searches are. Although this definition for user expertise was simple, it revealed the connection between user expertise and quality of the clicks.

Besides search expertise, it was reported that domain expertise also has impact on user clicks [14]. In their work, the affection of domain expertise to users' search behavior was studied in four specific domains (medicine, finance, law and computer science). In each domain, the users were separated into experts and non-experts based on whether they had visited one or more of the pre-defined expert sites. The results showed notable difference between domain experts and non-experts with respect to search behavior. More concretely, domain experts were found to be more successful when searching in-domain queries; the pages visited by domain experts had deeper technical depth than those visited by non-experts and so on. Building upon the analysis, a classifier was trained to predict whether a user is domain expert using his/her search interaction features. In their work, domain expertise was defined to be binary and it did not study how the expertise is related to the relevance of clicks.

There are also studies on personalized click model that treat users differently when inferring document relevance. For example, Shen et al. [10] used collaborative filtering technique to capture users' interested domains. In their personalized click model, it was assumed that the users have different domains of interests and latent factors were used to represent one's interests. The better the topics of a document match a user's interested domains, the higher the click probability. Another noise-aware click model [3] was proposed to measure the probability of a click being noisy by using both user class features and context class features. A variable N was introduced into the model, indicating whether the context is noisy. Then they made different click assumptions for different value of N. These two studies have considered the user level features when modeling clicks, but the influence of search expertise was still not taken into account.

3 Models

Our aim of introducing search expertise into click modeling is to help improve the relevance inference. Thus, the search expertise of a user should be able to reflect how well the user can make the right relevance judgment of a given document. When a user searches a query, a click indicates that he/she thinks the document is relevant and a skip (no click after examination) indicates an irrelevant judgment. If we view the this process of user making relevance judgments as a classification task, then the documents are the data samples to classify and the user is the classifier. Relevant documents correspond to positive samples and irrelevant documents correspond to negative samples. The click on a document indicates that the user classifies this document as relevant. Therefore, the expertise of a user is actually the performance of the classifier. There are many evaluation metrics for classifiers and we use *accuracy* and *confusion matrix* to measure the expertise in this paper.

3.1 Accuracy Model

Accuracy is a widely used evaluation metric in classification. It is calculated as the proportion of the correctly classified samples. In our application scenario, let a_u be the accuracy of user u making the right relevance judgment, which is a real-valued parameter ranging from 0 to 1. For the i^{th} document in the search result page, u will make the right relevance judgment with probability a_u. It can be formally denoted as:

$$\begin{aligned} a_u &= P(C_i = 1 | R_i = 1, E_i = 1, u) \\ &= P(C_i = 0 | R_i = 0, E_i = 1, u) \end{aligned} \qquad (1)$$

where C_i, R_i, and E_i are binary variables. C_i indicates whether the document is clicked; R_i indicates whether the document is relevant; and E_i indicates whether the document is examined by u. Thus, if u has examined the i^{th} document, the probability of the document being clicked can be written as:

$$\begin{aligned} & P(C_i = 1|E_i = 1, u) \\ = & \sum_{R_i \in \{0,1\}} P(R_i)P(C_i = 1|R_i, E_i = 1, u) \\ = & r_i a_u + (1 - r_i)(1 - a_u) \end{aligned} \tag{2}$$

where r_i is the probability that the i^{th} document is relevant. It means that a click happens under two situations: (1) the document is relevant and the user makes right relevance judgment; (2) the document is irrelevant and the user makes wrong relevance judgment. We call this model the accuracy model and a_u is the expertise of user u.

3.2 Confusion Matrix Model

Confusion matrix is another popular evaluation metric in classification. Unlike accuracy, it measures the classification accuracy for positive samples and negative samples separately. It is presented in the form of a matrix called the confusion matrix, as shown in Eq. 3. The element p_{ij} in the matrix is the probability that class i being classified as class j by the classifier. In our problem setting, p_{11} denotes the probability that a relevant document being classified as relevant (clicked) and p_{00} is the probability that an irrelevant document being classified as irrelevant(skipped). Therefore, in this confusion matrix model, the expertise of user u is represented by the matrix $\mathbf{M_u}$. The larger the values on the diagonal, the higher the user's expertise.

$$\mathbf{M_u} = \begin{bmatrix} p^u_{00} & p^u_{01} \\ p^u_{10} & p^u_{11} \end{bmatrix} \tag{3}$$

The values in each row of the matrix sum up to 1. Therefore, one's expertise can be represented by the following two parameters instead of the whole matrix.

$$\begin{aligned} p^u_{11} &= P(C_i = 1|R_i = 1, E_i = 1, u) \\ p^u_{00} &= P(C_i = 0|R_i = 0, E_i = 1, u) \end{aligned} \tag{4}$$

For user u, the click probability of the i^{th} document after examination becomes

$$P(\boldsymbol{C_i}|E_i = 1, u)^T = \begin{bmatrix} r_i \\ 1 - r_i \end{bmatrix}^T \mathbf{M_u} \tag{5}$$

As there are two expertise parameters, p^u_{00} and p^u_{11}, for each user. This confusion matrix model is more flexible than the accuracy model. And it can degenerate to the accuracy model if we let $p_{11} = p_{00}$.

3.3 Baseline model

To evaluate the performance of the above two proposed models, we use a baseline model for comparison. Like most of the existing click models, this baseline model does

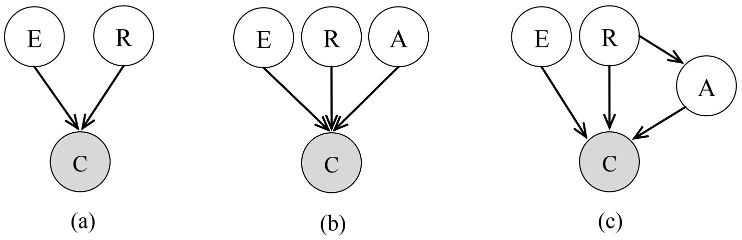

Fig. 1. The graphical representations of the models. (a) is the baseline model; (b) is the accuracy model and (c) is the confusion matrix model. R is the hidden variable for relevance; E is the hidden variable for examination; A is the hidden variable for expertise and C denotes click.

not take user expertise into consideration. Thus the click probability of a document after examination only depends on the document relevance.

$$P(C_i = 1|E_i = 1, u) = P(R_i = 1) \tag{6}$$

This assumption is widely used in click models such as [5, 7, 6]. It treats all the clicks as relevance indication.

3.4 Graphical representations

In the baseline model, R and E are the hidden variables and C can be observed from the data. The difference between our models and the baseline model is the hidden variable A which indicates whether the user made the right relevance judgment. Our models treat a click as relevance indication only when the user made the right judgment on this document. We can demonstrate the idea of the three models more clearly with the graphical representation in Figure 1. Figure 1(a) is the baseline model with the dependencies of a regular click model. (b) represents the accuracy model in which A is added as a dependent factor of C. (c) denotes the confusion matrix model in which another dependency is added from A to R, indicating that users behave differently on the relevant documents and the irrelevant documents. All these models are in probabilistic framework and can be solved in an efficient way.

4 Parameter Estimation

Having the different search expertise modelings proposed above, we now introduce how the parameters are estimated in these models. As the probability $P(C_i|E_i, u)$ has been derived for each model, we can compute the likelihood of the observed click-through data. For a search session[1], the commonly used *linear traversal hypothesis* in click models assume that users examine the documents one by one from top to bottom of a

[1] Here a session is defined as the activities of a user searching a query in a short period of time. In this paper, we set the interaction timeout to 30 minutes.

search result page. The *examination hypothesis* [4] assumes that users have to examine a document before clicking on it. These two hypotheses together implies that all the documents before a click have been examined. But for the documents after the last click in a session, we do not know whether they have been examined or not. Some click modes, such as [5, 6], use more complex assumptions to model the examination probabilities of the documents in all positions. However, considering that the focus of this paper is the influence of user expertise in modeling clicks, we decide to use the simplest examination hypothesis to reduce the influence of the other affecting factors. Therefore, we assume that in a search session, the user examined all the documents that ranked before the last clicked position. The likelihood of a search session s is then calculated as:

$$L(s) = \prod_{i=1}^{N_s} P(C_i = 1|E_i = 1, u_s)^{C_i} \times P(C_i = 0|E_i = 1, u_s)^{1-C_i} \tag{7}$$

where N_s is the last clicked position in session s and u_s is the user that conducted the session. The log-likelihood of the whole session observations is:

$$l(S) = \sum_{s \in S} \log L(s) \tag{8}$$

where S is the set of all sessions. To estimate the unknown parameters, we maximize the log-likelihood in Eq. 8 for each model.

Baseline model: The baseline model has no expertise parameters. The only parameters to estimate are the relevance parameter $\{r\}$. By maximizing Eq. 8, the relevance of document d can be estimated as:

$$r_d = \frac{\#\text{Click on } d}{\#\text{Impression of } d \text{ before position } l} \tag{9}$$

where l is the position of the last click of each session that includes document d. The denominator calculates how many times d has appeared before a clicked document in all related sessions. r_d can be computed very efficiently by scanning the click-through data only once. In fact, this estimated relevance is exactly the same as that in the dependent click model proposed by Guo et al.[7], which was reported to be a very effective and efficient model in estimating relevance.

Accuracy model: In the accuracy model, C is dependent on two hidden variables A and R. The click probability $P(C_i = 1|E_i = 1, u)$ becomes a sum of several parts and the MLE method can no longer lead to closed form solution in parameter estimation. Therefore, the expectation-maximization algorithm (EM) is used here. In order to have a better control on the value of the estimated expertise parameters, beta distribution is used as conjugate prior for the expertise parameters. In an EM iteration, the parameters are updated in the following way:

$$
\begin{aligned}
r_d^{new} &= \frac{1}{|S_d|}\sum_{s\in S_d} I_{C=1}\frac{r_d a_u}{1-r_d-a_u+2r_d a_u} + I_{C=0}\frac{r_d(1-a_u)}{a_u-r_d a_u+r_d-r_d a_u} \\
a_u^{new} &= \frac{1}{\sum\limits_{s\in S_u} N_s(\alpha+\beta-1)}\sum_{s\in S_u}\sum_{i=1}^{N_s} I_{C_i=1}\frac{\alpha r_d a_u+(\beta-1)(1-r_d)(1-a_u)}{1-r_i-a_u+2r_i a_u} + \\
&\qquad I_{C_i=0}\frac{\alpha(1-r_d)a_u+(\beta-1)r_d(1-a_u)}{a_u-r_i a_u+r_i-r_i a_u}
\end{aligned}
\tag{10}
$$

where S_d is the set of sessions that include document d; S_u is the set of sessions of user u and N_s is the number of the examined documents in session s; I is the indicator function; α and β are the parameters of the beta prior. When $\alpha = 1$, $\beta = 1$, the pdf of beta distribution becomes a constant value (i.e. equal to using no prior).

Confusion matrix model: For the confusion matrix model, the capacity of a user making the right relevance judgment is represented by two parameters p_{11} and p_{00}, while in the accuracy model we only use one parameter. p_{11} and p_{00} are independent from each other. The estimation of the parameters using EM algorithm is similar to that in the accuracy model so the details are not listed here due to the space limitation. For simplicity, we let p_{11} and p_{00} share the same beta prior.

During the training of accuracy model and confusion matrix model, we run the EM algorithm for a fixed number of 20 iterations. The EM algorithm has a fast convergency speed so that after 20 iterations, the change ratio of the objective function is smaller than 0.1%, which is regarded as a signal of convergence.

5 Experiments

5.1 Experimental settings

A sampled click log of a commercial Chinese search engine in November 2011 is used for our experiments. The click log records the interaction information of users with the search engine, such as user's cookie ID, session ID, query string, presented documents, clicked documents and timestamps. For the protection of users' privacy, all sensitive attributes are processed into numbers. A user is identified by cookie ID. As some users might have cleaned the cookie during the period that the data was collected, we removed the users who have fewer than 10 distinct queries to avoid noise when estimating the parameters of user expertise. After that, we finally obtain 23,534 unique users, 253,045 unique queries, 1,034,598 query sessions, 1,173,426 clicks and 476,737 skips. For each user, all his/her sessions are sorted by timestamp and we split them into two parts at the ratio of 4:1 for training and testing respectively.

For the prior, α and β together control the shape of beta distribution. In the training phase, we try different combinations of α and β and the best performance is obtained when $\alpha = 2, \beta = 2$ with respect to the estimated relevance. Therefore, we only use the results of $\alpha = 2, \beta = 2$ for demonstration in this section.

5.2 Perplexity

Perplexity is an evaluation metric often used by click models [4]. It measures how well the predicted probabilities fit the real data. Smaller perplexity means better performance and the ideal value for perplexity is 1. We calculate the perplexity for all models on both training set and test set. Table 1 shows the results. The accuracy model (AM) without prior and the confusion matrix model (CMM) without prior have close perplexity together with the baseline model on test set. We notice that the models with prior have worse perplexity. The reason is that perplexity is very similar to likelihood, which is the optimization objective of the models without prior. When the posterior becomes the optimization objective for the models with prior, the perplexity is no longer optimized. Therefore, it is not surprising that CMM without prior obtains the best perplexity given that CMM has greater flexibility than AM. Although the models without prior perform better in perplexity, it does not mean they are better in inferring relevance. In fact, perplexity can not directly reflect the quality of the estimated relevance. Wang et al. have pointed out that perplexity might not be a trustable metric for click models because it is defined based on the absolute value of the predicted probabilities and thus is sensitive to scaling [11]. Therefore, we use perplexity as a reference but it is not the main evaluation metric in this paper. For the evaluation, we will focus on the quality of the estimated relevance, which is the aim of modeling the clicks.

Table 1. Perplexity of different models.

	Training set			Test set		
	all	click	skip	all	click	skip
	1,339,912	950,934	388,978	222,275	165,758	56,517
Baseline	1.203	1.934	1.380	1.302	2.779	1.579
AM(no prior)	1.206	1.911	1.379	1.304	2.795	1.583
AM($\alpha = 2, \beta = 2$)	1.760	2.044	1.838	1.752	2.085	1.831
CMM(no prior)	1.201	1.900	1.372	1.299	2.795	1.576
CMM($\alpha = 2, \beta = 2$)	1.710	2.141	1.825	1.703	2.208	1.820

5.3 Effectiveness of the estimated relevance

To evaluate the effectiveness of the estimated relevance, we use the manually labeled relevance as ground truth. To create the relevance labels, we first divide all queries into seven groups according to log-frequency of the query and randomly select 30 queries from each group. For the 210 selected queries, the related documents (clicked or skipped in a session) are then extracted from the click log, which gives us 1,133 unique query-document pairs. We manually label all the query-document pairs with three relevance scales: *2=very relevant, 1=relevant, 0=irrelevant*.

As the estimated relevance is supposed to help improve the search engine ranking performance, the relative order of relevance of document pair can be used to evaluate the effectiveness of the estimated relevance [1]. The idea is that for a document pair

(d_i, d_j) under a query, if d_i is more relevant than d_j, the estimated relevance of d_i should be higher than the estimated relevance of d_j as well. With the relevance labels, we investigate the agreement between the estimated relevance preference pairs and the labeled relevance preference pairs. Let r_i be the estimated relevance of d_i and l_i be the labeled relevance of d_i. A concordant pair means that $r_i > r_j, l_i > l_j$ or $r_i < r_j, l_i < l_j$. If $r_i > r_j, l_i < l_j$ or $r_i < r_j, l_i > l_j$, it is a discordant pair. Otherwise, it is neither a concordant nor a discordant pair. The more concordant pairs a model obtains, the better the model is in estimating relevance.

Table 2. Relevance preference pairs

	#concordant pair	#discordant pair	precision	%improve over baseline
baseline	780	449	63.5%	-
AM(no prior)	778	451	63.3%	-0.3%
AM($\alpha = 2, \beta = 2$)	**862**	**367**	**70.1**%	**10.5**%
CMM(no prior)	767	462	62.5%	-1.6%
CMM($\alpha = 2, \beta = 2$)	**817**	**412**	**66.5**%	**4.74**%

Table 2 shows the number of concordant and discordant relevance preference pairs obtained by each model. Precision is defined as the proportion of concordant pairs. We observe that without prior, AM and CMM are even worse than the baseline model which does not consider user expertise at all. It indicates the necessity of introducing prior. With a proper beta prior $\alpha = 2, \beta = 2$, both AM and CMM gain good precision. The precision of the accuracy model even reaches 70.1%, which improves the baseline by 10%. The confusion matrix model also improves the baseline by 4.7%. This fact verifies the effectiveness of our models in estimating relevance. We notice that the confusion matrix model, which has more modeling flexibility, fails to outperform the accuracy model. We will analyze the reason in next section by investigating the estimated user expertise parameters.

The inferred relevance preference pairs are useful in improving the search engine ranking performance. They can either be used as training samples in the pairwise learning algorithms, or they can be used directly to re-rank the search results. When used as training samples, the automatically generated preference pairs are of particular advantage because they are faster and easier to get compared to manual relevance labeling.

5.4 Effectiveness of the estimated user expertise

Besides document relevance, our models estimate the expertise of users as well. In this section, we evaluate how close the estimated expertise parameters are to the ground truth. Before the evaluation, we first need to obtain the ground truth of a user's expertise. With the labeled relevance in the previous section, we can calculate the ground truth of the expertise parameters using their definitions. Let L be the set of labeled query-document pairs, for a user u, the ground truth of a_u in the accuracy model is calculated as the proportion of correct click/skip decisions made by u on the all documents in

L; the ground truth of p_{11} and p_{00} are also calculated in L according to their own definitions. To avoid noise, we do not evaluate the users with fewer than ten query-document pairs in L. We note that the calculated ground truth will be unavoidably biased to certain extent because of the limited size of L . However, this has been the best ground truth we can obtain with the data we have.

For each user u, we now have the estimated a_u, p_{11}^u, p_{00}^u and the calculated ground truth. We use multiple metrics to measure the gap between the estimated value and the ground truth, such as correlation, Kendall's τ, mean absolute error (MAE) and rooted mean square error (RMSE). Table 3 shows the results.

Table 3. Comparing the estimated expertise parameters with the ground truth

	Kendall's τ	correlation	MAE	RMSE
AM-$a_u(\alpha = 2, \beta = 2)$	0.425	0.609	0.277	0.298
AM-a_u(no prior)	0.413	0.407	0.129	0.201
CMM-$p_{11}(\alpha = 2, \beta = 2)$	0.395	0.591	0.280	0.299
CMM-p_{11}(no prior)	0.419	0.512	0.113	0.182
CMM-$p_{00}(\alpha = 2, \beta = 2)$	0.272	0.540	0.162	0.201
CMM-p_{00}(no prior)	0.321	-0.077	0.516	0.556

Kendall's τ [9] measures the similarity of the orderings of the data ranked by two quantities. We find that all the models obtain a positive τ. The correlation coefficient measures the dependence between two variables. Except for p_{00} in CMM without prior, all the estimated parameters have a relatively high correlation coefficient with the ground truth, especially for the models with prior. For the accuracy model with prior $\alpha = 2, \beta = 2$, which is the best performing model in relevance estimation, the estimated parameter a_u obtains the largest τ and correlation coefficient with the ground truth among all the models. MAE and RMSE reflect the distance of the absolute values. It is not surprising to find that for a_u and p_{11}, the models with prior have larger MAE and RMSE than the models without prior, which is not consistent with the result of τ and correlation. One explanation is that adding prior helps the models do better with the relative order of the estimate parameters rather than the absolute value, which we care more in the evaluation. We notice that the result of p_{00} in Table 3 is quite noisy. After investigation, we find that the number of users who have ground truth calculated for p_{00} is much smaller than that for a_u and p_{11} (i.e. the relevant query-document pairs in L are more accessed than the irrelevant pairs by users).

Since the calculated ground truth of expertise for individual users can be noisy due to the lack of data. We now evaluate the expertise of user groups. In this evaluation, we first divide users into ten user groups according to estimated expertise in descending order such that each user group has the same number of users. Then we treat each user group as a single unit and compute its ground truth of expertise on set L. As a user group is supposed to have sufficient amount of data, the calculated ground truth is more reliable than that calculated for individual users. Figure 2 shows the performance of different user groups.

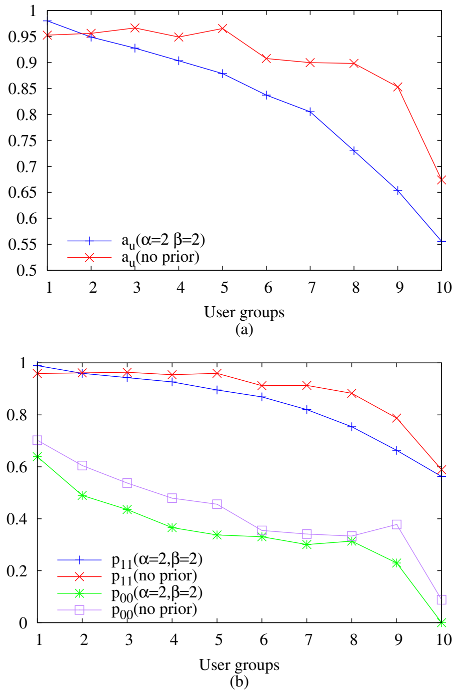

Fig. 2. The ground truth of expertise for different user groups. (a) shows the ground truth of a_u in the accuracy model. (b) shows the ground truth of p_{11} and p_{00} in the confusion matrix model. Smaller group number indicates higher estimated expertise.

From Figure 2(a) we see that for the accuracy model with prior, the ground truth of a_u strictly decreases as the group number increases. It indicates that the estimated a_u is very effective in ordering the users such that the group expertise reflects the ground truth very well. If we let the estimated expertise of user group i be $(1 - i/10)$, the group level correlation coefficient between the estimated expertise and the ground truth reaches as high as 0.949, which is much higher than the values reported in Table 3 for individual users. And the Kendall's τ even reaches the optimal value 1, which means perfect ranking for the user groups. This result validates the effectiveness of estimated expertise in the accuracy model. For the confusion matrix model with prior, we also observe the similar trend for p_{11} in Figure 2(b); the trend for p_{00} is basically consistent but not as clear as that of p_{11}. It indicates that the estimated p_{00} in the confusion matrix model does not reflect the ground truth well as p_{11}. And this may be the reason that the confusion matrix model failed to outperform the accuracy model in relevance estimation. In Figure 2, we also plot the result for the models without prior, which shows more inconsistency and weaker correlation with the ground truth. This fact again verifies the advantage of using prior in our models.

To conclude, we find that the accuracy model with beta prior achieves the best performance in inferring relevance and user expertise. It implies that the assumption of the accuracy model is more suitable to the real situation. We also find that the estimated expertise can better reflect the ground truth when used in level of user groups.

6 Conclusion and Future Work

Clicks from different search users in Web search are not equally informative in indicating relevance. Search experts are supposed to be more likely to find relevant documents than the others so their clicks are more reliable in inferring relevance. In this paper, we propose two probabilistic modelings for users' search expertise which are inspired by the evaluation metrics of classification. The experimental results on a real-world click-through data show that our models are effective in estimating both the relevance and the user expertise. Our best performing model improves the baseline by 10% in inferring relevance preference pairs. And the estimated expertise is highly consistent with the ground truth, especially when used in group level. The user expertise information can be useful in helping the search engine improve personal search experience, which is the direction of our future work.

References

1. E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of SIGIR'06*, pages 19–26, 2006.
2. O. Chapelle and Y. Zhang. A dynamic bayesian network click model for web search ranking. In *Proceedings of WWW'09*, pages 1–10, 2009.
3. W. Chen, D. Wang, Y. Zhang, Z. Chen, A. Singla, and Q. Yang. A noise-aware click model for web search. In *Proceedings of WSDM'12*, WSDM '12, pages 313–322, 2012.
4. N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of WSDM'08*, 2008.
5. G. Dupret and B. Piwowarski. A user browsing model to predict search engine click data from past observations. In *Proceedings of SIGIR'08*, 2008.
6. F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y. Wang, and C. Faloutsos. Click chain model in web search. In *Proceedings of WWW'09*, pages 11–20, 2009.
7. F. Guo, C. Liu, and Y. Wang. Efficient multiple-click models in web search. In *Proceedings of WSDM'09*, pages 124–131, 2009.
8. T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting click-through data as implicit feedback. In *Proceedings of SIGIR'05*, 2005.
9. M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.
10. S. Shen, B. Hu, W. Chen, and Q. Yang. Personalized click model through collaborative filtering. In *Proceedings of WSDM'12*, 2012.
11. H. Wang, C. Zhai, A. Dong, and Y. Chang. Content-aware click modeling. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 1365–1376, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
12. R. White and D. Morris. Investigating the querying and browsing behavior of advanced search engine users. In *Proceedings of SIGIR'07*, pages 255–262, 2007.
13. R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *Proceedings of WWW'07*, WWW '07, pages 21–30, 2007.
14. R. W. White, S. T. Dumais, and J. Teevan. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of WSDM'09*, WSDM '09, pages 132–141, 2009.