Individual Judgments Versus Consensus: Estimating Query-URL Relevance

HENGJIE SONG, South China University of Technology, Baidu Inc.
YONGHUI XU, HUAQING MIN, and QINGYAO WU, South China University of Technology
WEI WEI, Huazhong University of Science and Technology
JIANSHU WENG, HP Labs Singapore
XIAOGANG HAN, Baidu Inc.
QIANG YANG, Hong Kong University of Science and Technology
JIALIANG SHI, Baidu Inc.
JIAQIAN GU, Guangdong Polytechnic Normal University
CHUNYAN MIAO, Nanyang Technological University
NISHIDA TOYOAKI, Kyoto University

Query-URL relevance, measuring the relevance of each retrieved URL with respect to a given query, is one of the fundamental criteria to evaluate the performance of commercial search engines. The traditional way to collect reliable and accurate query-URL relevance requires multiple annotators to provide their individual judgments based on their subjective expertise (e.g., understanding of user intents). In this case, the annotators' subjectivity reflected in each annotator individual judgment (AIJ) inevitably affects the quality of the ground truth relevance (GTR). But to the best of our knowledge, the potential impact of AIJs on estimating GTRs has not been studied and exploited quantitatively by existing work. This article first studies how multiple AIJs and GTRs are correlated. Our empirical studies find that the multiple AIJs possibly provide more cues to improve the accuracy of estimating GTRs. Inspired by this finding, we then propose a novel approach to integrating the multiple AIJs with the features characterizing query-URL pairs for estimating GTRs more accurately. Furthermore, we conduct experiments in a commercial search engine—Baidu.com—and report significant gains in terms of the normalized discounted cumulative gains.

Categories and Subject Descriptors: H.3.3 [**Information Search and Retrieval**]: Relevance Feedback

General Terms: Algorithms

Additional Key Words and Phrases: Web search, relevance feedback, performance evaluation

---

This work is supported by the JSPS Fellow Program under grant P-12350 and Baidu research grant 181315P00418.
Authors' addresses: H. Song, Y. Xu, H. Min, and Q. Wu, School of Software Engineering, South China University of Technology (P.R.C); emails: {sehjsong, x.yonghui, hqmin, qyw}@scut.edu.cn; J. Weng, HP Labs Singapore; email: jianshu@gmail.com; Q. Yang, Department of Computer Science and Engineering, Hong Kong University of Science and Technology; email: qyang@cse.ust.hk; X. Han and J. Shi, Baidu Inc.; emails: xganghan@gmail.com, shijialiang@baidu.com; W. Wei, Huazhong University of Science and Technology; email: weiw@hust.edu.cn; J. Gu, Guangdong Polytechnic Normal University; email: jiaqian314@163.com; C. Miao, Nanyang Technological University; email: ascymiao@ntu.edu.sg; N. Toyoaki, Graduate School of Informatics, Kyoto University; email: nishida@i.kyoto-u.ac.jp.

DOI: http://dx.doi.org/10.1145/2834122

ACM Reference Format:
Hengjie Song, Yonghui Xu, Huaqing Min, Qingyao Wu, Wei Wei, Jianshu Weng, Xiaogang Han, Qiang Yang, Jialiang Shi, Jiaqian Gu, Chunyan Miao, and Nishida Toyoaki. 2016. Individual judgments versus consensus: Estimating query-URL relevance. ACM Trans. Web 10, 1, Article 3 (January 2016), 21 pages.
DOI: http://dx.doi.org/10.1145/2834122

1. INTRODUCTION

Query-URL relevance, measuring the relevance of each retrieved URL with respect to a given query, is one of the fundamental criteria to evaluate the performance of commercial search engines. In practice, query-URL relevance has a wide range of applications. For instance, in the case of learning-to-rank that exploits machine learning algorithms to achieve a reasonable ranking of retrieved URLs, the training examples are typically represented by the triples Query, URL, Relevance, where the relevance associated with a query-URL pair is labeled by any value in the ordinal set {Perfect (4), Excellent (3), Good (2), Fair (1), Bad (0)} [Chapelle and Chang 2011; Xia et al. 2008]. In most cases, the performance of learning-to-rank depends not only on the number of training examples but also on the quality of the relevance associated with the training examples [Sheng et al. 2008; Yang et al. 2010]. Xu et al. [2010] studied the relationship between the error rates of query-URL relevance and the performances of the typical learning-to-rank algorithms (RankSVM, RankBoost, AdaRank, and SVM-MAP). With 30% error in the training data, the MAP score of RankBoost suffers a significant drop on three benchmark datasets (46%, 70%, and 74% on TD2004, HP2004, and NP2004 in LETOR, respectively).

Note that in practice, ranking strategy (e.g., learning-to-rank) and evaluating relevance both have some connections and remarkable differences. On the one hand, ranking strategy retrieves and ranks the URL candidates, which are the subjects in the study of relevance evaluation. On the other hand, the relevance evaluation results can be fed back into the engine as direct evidence to influence ranking [Agichtein et al. 2006; Joachims 2002], or in turn provide a basis to further improve the ranking strategy [Xu et al. 2010]. From this perspective, ranking strategy and evaluating relevance are complementary. Both of them are necessary to improve the quality of commercial search engines.

However, the traditional way to manually assess query-URL relevance is time consuming and labor intensive [He et al. 2011; Song et al. 2012] for two major reasons:

—The traditional way might be effective in small scales. Yet it becomes impractical as the manual efforts grow quickly in size and number of query-URL pairs.
—The traditional way first requires multiple annotators to provide their individual judgments based on their subjective expertise (e.g., understanding of user intents). Then for a given query-URL pair, its ground truth relevance (GTR) is derived based on the multiple annotator individual judgments (AIJs) available. In this case, the annotators' subjectivities reflected in multiple AIJs inevitably affect the quality of the GTR [Guo and Agichtein 2012].

To tackle these issues, this article proposes a novel approach to integrate the features characterizing query-URL pairs with the multiple AIJs to improve the accuracy of estimated GTRs. Unlike the prior state of the art, our study is inspired by a fact that has been largely unexplored—the inconsistency among multiple AIJs is practically considered as an indicator of whether it is necessary to recheck and have group discussion about the quality of the Web pages linked by the retrieved URLs. It means that the inconsistent AIJs possibly provide more cues to improving the accuracy of estimating GTRs, as justified and revealed in Section 3.

The contribution of our work is twofold:

(1) By quantitatively analyzing the potential correlation between multiple AIJs and GTRs, we justify the necessity and the possibility of exploiting the knowledge from multiple AIJs for estimating GTRs, which has been ignored by most existing models.
(2) A probabilistic model is proposed as the first attempt to integrate the features characterizing query-URL pairs with the expert knowledge reflected in multiple AIJs to efficiently improve the accuracy of estimating GTRs.

The rest of the article is organized as follows. Section 2 gives a brief overview about the related work and limitations. For better understanding of our research motivation, Section 3 first introduces the traditional way to collect query-URL relevance in commercial search engines. Then, we quantitatively analyze the potential correlation between the AIJs and the GTRs in the dataset prepared for this study. Inspired by this analysis result, we further present the motivation of our study and its potential benefit. Section 4 describes in detail the proposed probabilistic approach to estimating GTRs. In Section 5, we perform a set of experiments and compare the experimental results with the state-of-the-art methods used in similar tasks. A significant improvement in terms of the normalized discounted cumulative gains (NDCGs) is reported. Section 6 concludes the article with future work.

2. RELATED WORK AND LIMITATIONS

To improve the accuracy and efficiency of annotating query-URL relevance, many methods have been proposed to automatically estimate the GTRs in search engines. In general, those methods can be categorized into four types:

—*Category (a)*: *Methods that attempt to infer the user-perceived relevance for each query-URL pair by modeling the interaction (click behavior and browsing behavior) between users and retrieved URLs within the framework of click modeling.*
Typical click modeling techniques include the dynamic Bayesian network (DBN) model [Chapelle et al. 2009], the user browsing model (UBM) [Dupret and Piwowarski 2008], the task-centric click model (TCM) [Zhang et al. 2011], and Unbiased-UBM and Unbiased-DBN [Hu et al. 2011], among others. The main idea of click modeling is to infer the user's perceived relevance by constructing probabilistic models to understand user behavior (i.e., click/browse) [Agichtein et al. 2006; Srikant et al. 2010; Zhang et al. 2011]. These approaches follow the empirical observation that clicks are based primarily on the perceived relevance of a URL, which is usually derived with a user's guess according to a short summary generated by the search engine to describe the Web page linked by the URL. In this case, the perceived relevance of a retrieved URL is defined as the probability of its being clicked after examination, which is closely related to its click-through rate (CTR).

But when applying click modeling to infer GTRs, a significant limitation is that the perceived relevance may be inconsistent with the intrinsic relevance [Dupret and Liao 2010]. Specifically, a user may find a page irrelevant only after the page is clicked and viewed. From this point of view, most click modeling techniques usually suffer from the well-known position bias, in which a URL in a higher position is more likely to attract more user clicks even though it is not as relevant as other URLs in lower positions [Dupret and Piwowarski 2008]. To solve this problem, some approaches have been proposed to interpret noisy user feedback underlying user interactions with the search engine [Agichtein et al. 2006]. In addition, He et al. [2011] proposed a document reordering framework to efficiently collect document relevance for Web retrieval evaluation based on these click models. Specifically, the framework in He

et al. [2011] integrated four inconsistent intuitions for more efficiently collecting relevance information, and they formulated them into one reordering function. By doing so, the approach [He et al. 2011] makes it possible to reduce the number of submissions required for accurate evaluation. To minimize the costs on evaluating retrieval systems, Carterette et al. [2006] presented a novel perspective on average precision that leads to a natural algorithm for building a set of query-URL pairs. This makes it possible to evaluate a set of retrieval systems with high confidence and a minimal set of judgments. Although these approaches are promising solutions to estimate query-URL relevance, it might be debatable to simply regard CTR as the proxy of the intrinsic query-URL relevance [Chen et al. 2011].

—*Category (b)*: *Methods that attempt to infer the intrinsic relevance for each query-URL pair by analyzing the features that characterize the query-URL pairs from query level and pageview level.*

Typical examples include the methods, for example, in Dupret and Liao [2010], Song et al. [2011], and MTM [Song et al. 2012]. The main idea is to infer the intrinsic relevance of query-URL pairs by establishing the relationship between GTRs and the features characterizing query-URL pairs. These features represent the information extracted not only from the click/browse log data but also from the title, anchor text, and search log associated with the Web pages linked by the retrieved URLs. By doing so, the approaches in Dupret and Liao [2010], Song et al. [2011], and Song et al. [2012] provide more information to characterize query-URL pairs from query level and pageview level. In contrast, click modeling only considers the information about user activities in click/browse log data. As a result, the methods in category (b) reduce the negative impact of position bias on estimating GTRs.

Although these approaches [Song et al. 2011; Song et al. 2012] have been applied in many commercial search engines, such as Biadu.com, there are still some limitations. It incurs overhead in the preprocessing stage, such as the costs in acquiring features, refining features, and cleaning data. Moreover, it is unlikely for these approaches to have a unified framework for GTR estimation because extra effort is required to identify the hidden important factors that have potential impacts on estimating GTRs, such as search frequency of queries and the mouse trajectories [Jarvelin and Kekalainan 2000; Dupret and Liao 2010; Song et al. 2011; Song et al. 2012].

—*Category (c)*: *Methods that attempt to infer the intrinsic relevance for each query-URL pair by analyzing the quality assurance methods based on statistical label aggregation—that is, collecting redundant judgments from multiple workers and aggregating them via methods like majority voting or expectation maximization to produce reliable labels* [Sheng et al. 2008].

These methods are generally called *crowdsourcing*, which has recently emerged as a feasible approach to gathering relevance data in the context of information retrieval evaluation [Gao et al. 2012; Jung and Lease 2012; Jurca and Faltings 2009; Kazai et al. 2011; Le et al. 2010]. By distributing work through an open call for contributions from members of a crowd, crowdsourcing enables the gathering of relevance labels from a large population of workers at a relatively low cost. As a result, it offers a solution to the scalability problem that hinders traditional approaches based on editorial judgments. In the context of evaluating query-URL relevance, the study in Yang et al. [2010] and He et al. [2011] proposed a novel scheme to collect high-quality query-URL relevance by revealing the basic principles (i.e., whether, when, and for which query-URL pairs one should effectively produce and employ overlapping labels from multiple experts to improve Web search accuracy). In addition, Blanco et al. [2011] investigated the repeatable and reliable search system evaluation using crowdsourcing. More recently, a probabilistic matrix factorization was proposed in Jung and Lease [2012] to infer unobserved annotators' judgments and to investigate

how annotators' judgments influence the consensus labels for all examples in the TREC Relevance Feedback Track [Buckley et al. 2010].

Although best practices are gradually evolving, such as guidelines for the use of crowdsourcing in relevance assessments, the issues of attracting the "right" workers and controlling their engagement in the crowdsourcing tasks remain a challenge [Kazai et al. 2011]. In some cases, it is not immediately clear how to directly apply the crowdsourcing-based solutions in Ipeirotis et al. [2010], Jung and Lease [2012], Sheng et al. [2008], and Yang et al. [2010] to estimate the GTRs in search engines. For example, the assumption (that the annotator typically judges only a small number of examples, and hence collected judgments are typically sparse and imbalanced, with relatively few workers influencing consensus labels [Jung and Lease 2012]) widely used in crowdsourcing possibly does not always hold in the real process to collect query-URL relevance (as described in Section 3). In addition, most crowdsourcing-based solutions mainly consider a binary situation in which the relevance of retrieved URLs to a given query is either relevant or irrelevant. However, the binary situation is not sufficient to describe the multilevel query-URL relevance (i.e., Perfect, Excellent, Good, Fair, and Bad). Therefore, the existing crowdsourcing-based solutions need to be further investigated before they can be directly applied in the context of estimating query-URL relevance.

—*Category (d)*: *Methods that attempt to infer the annotation results by combining objective features and annotator judgments* [Raykar and Yu 2012; Raykar et al. 2009, 2010].

Recently, Raykar and Yu [2012] and Raykar et al. [2009, 2010] attempted to combine objective features and annotator judgments so as to iteratively eliminate the spammers and to estimate the consensus labels. Based on the work [Raykar et al. 2009] that mainly discussed the binary classification problem, Raykar et al. [2010] proposed a probabilistic framework for supervised learning with multiple annotators providing labels but no absolute gold standard. The proposed algorithm iteratively establishes a particular gold standard, measures the performance of the annotators given that gold standard, and then refines the gold standard based on the performance measures. By doing so, this work makes it possible to learn the classifier and the ground truth jointly, which is more general and can easily be extended to categorical, ordinal, and continuous data. Furthermore, Raykar et al. [2012] proposed an empirical Bayesian algorithm called *SpEM* to iteratively eliminate the spammers and estimate the consensus labels based on the good annotators. By defining a truncated Gaussian prior to the annotators' sensitivity-specificity pair with a separate precision parameter, SpEM described the annotator's confidence in terms of precision parameter. This makes it possible to apply the Bayesian approach for eliminating spammers and consolidating crowdsourced results.

3. RESEARCH MOTIVATION

To better understand our research motivation, we briefly describe the traditional way to collect reliable and accurate query-URL relevance, which consists of two primary phases, following the Cranfield paradigm [Cleverdon 1997; Harman 2010]:

—*Phase I*: For a given query-URL pair, multiple annotators provide their individual judgments on the relevance according to their expertise (e.g., understanding of the user search intents that are expressed through the search queries/context or browse behaviors). In this phase, each query-URL pair is assigned a score (e.g., Perfect (4), Excellent (3), Good (2), Fair (1) and Bad (0), scaling from highly relevant to not relevant), each of which denotes an AIJ on the URL's relevance with respect to the given query.

—*Phase II*: For a given query-URL pair, if all annotators completely agree with one another, multiple AIJs are directly taken as the GTR. However, since AIJs greatly rely on individual expertise, it is not a surprise that the inherent subjectivity in annotators' expertise leads to the inconsistent AIJs [Bailey et al. 2008]. For the query-URL pair with inconsistent AIJs, we cannot simply take the majority voting from multiple AIJs as the GTR [Raykar et al. 2010; Sheng et al. 2008; Yang et al. 2010]. Consider a scenario where a query-URL pair is with inconsistent AIJs. The majority are novices, with only one true expert. If novices give the same incorrect label to a specific pair, then the majority voting would favor the novices since they are in a majority. One could address this problem by introducing a weight capturing how good each expert is. But there is no well-recognized standard to measure the quality of one's expertise [Raykar et al. 2010]. In this case, multiple annotators have to spend extra effort in rechecking and group discussing the quality of the Web pages linked by the retrieved URLs so as to reach a GTR recognized by the majority. Although the GTR reached by this approach may not be perfect, it is the best that could be obtained practically.
—*Summary*: The real process of collecting high-quality query-URL relevance not only needs to consider the objective factors (e.g., the quality of the Web pages linked by the retrieved URLs) but also needs to take into account the subjective factors (e.g., multiple inconsistent AIJs). However, the existing approaches in categories (a) and (b) ignore the subjective factors, whereas the approaches in category (c) ignore the objective factors. In contrast, the approaches in category (d) provide solutions to iteratively eliminate the spammers and to estimate the consensus labels by combining objective features and annotator judgments. According to the promising results reported in category (d), we follow the basic idea underlying these approaches and attempt to further quantitatively analyze the correlation between individual AIJs and GTRs in the dataset prepared for this study. It justifies the necessity and the possibility to estimate GTRs by integrating AIJs and the features characterizing query-URL pairs. From this perspective, our study proposes an approach belonging to category (d).

3.1. Dataset

Since the available public datasets (Microsoft Learning to Rank, http://research.microsoft.com/en-us/projects/mslr/download.aspx; LETOR 4.0: A Benchmark Collection for Research on Learning to Rank for Information Retrieval, http://research.microsoft.com/en-us/um/beijing/projects/letor//) lack the necessary information about AIJs on the relevance of query-URL pairs that are indispensable in our proposal, we enrich the dataset used in Song et al. [2011] and Song et al. [2012] by collecting query-URL pairs from the Chinese search engine Baidu.com. The dataset prepared for this study contains 89,800 query-URL pairs, in which each query has 10 retrieved URLs in the first search result page (FSRP). For each query-URL pair in the dataset, five senior annotators and five junior annotators independently provide their AIJs on the relevance. Furthermore, the GTRs associated with the query-URL pairs are generated following the aforementioned working flow (phase I and phase II).

Note that in this study, junior annotators represent the annotators in new employee training programs. In contrast, senior annotators are the ones with minimum of 2 years of work experience in annotating the query-URL relevance. Although some qualitative company-inner standards are used as basic guidelines, there unavoidably exist differences between junior annotators and senior annotators, such as the annotators' subjective expertise in understanding search intent and sensitivity to the timeliness of search query. To some extent, we think that these AIJs could represent diverse views of Internet users on query-URL relevance. Furthermore, these senior and junior annotators follow phase I and phase II in the aforementioned working flow to achieve the

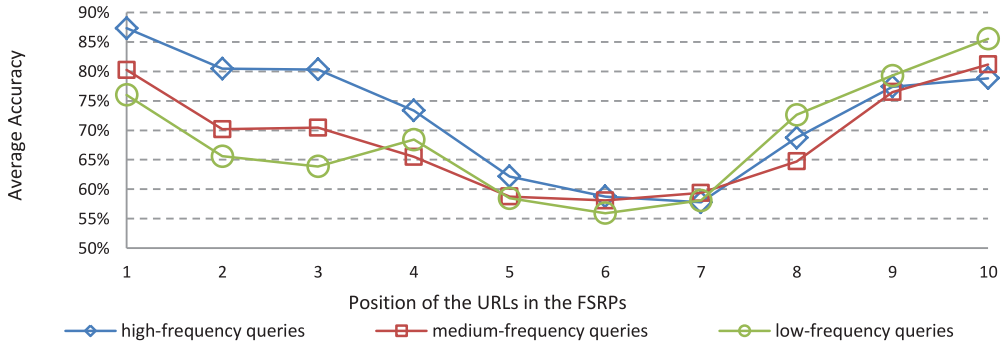

Fig. 1. Correlation among Ave_{acy}, search frequencies, and the positions of the URLs in FSRPs.

consensus on query-URL relevance (i.e., GTR). By doing so, the obtained GTR could be practically considered as the view of typical users. Although the GTR reached by this approach may not be perfect, it is the best that could be obtained practically.

3.2. Analysis

Since the distribution of the GTRs is related to the search frequency, the queries in the dataset are categorized into high-frequency ($Search_{Fre.} \geq$ 4,097/day), medium-frequency (4,097/day $>$ $Search_{Fre.} \geq$ 9 /day), and low-frequency ones (9/day $>$ $Search_{Fre.} \geq$ 1 /day), following the analysis in Song et al. [2011]. For the URLs locating on the k^{th} position in the FSRPs (URL_k), the average accuracy ($Ave_{acy,k}$) of multiple annotators is defined by Equation (1), where N_{ann} is the number of the annotators involved, $|URL_k|$ is the cardinality of URL_k, and $N_{acy}^{(i,k)}$ denotes the number of query-URL pairs (appearing at the k^{th} position) on which the i^{th} annotator has the same relevance score with GTRs.

$$Ave_{acy,k} = \frac{1}{N_{ann} \times |URL_k|} \sum_{i=1}^{N_{ann}} N_{acy}^{(i,k)} \tag{1}$$

Figure 1 gives the correlation among $Ave_{acy,k}$, search frequencies, and the positions of the URLs.

It is observed that regardless of the query frequency, $Ave_{acy,k=1,2,3,9 \text{ and } 10}$ are greater than $Ave_{acy,k=4\sim8}$. From our point of view, this observation might be explained by two reasons. First, annotation accuracy associated with the 1st, 2nd, and 3rd URLs is commonly considered as an important criterion to evaluate annotators' expertise. Therefore, annotators tend to put more effort in improving their annotation accuracy for the first three URLs. Second, under common ranking strategy, most of the GTRs for the 9th and 10th URLs are automatically categorized as Bad and Fair. In this case, an experienced annotator sometimes is able to provide $\geq$65% accuracy with respect to the 9th and 10th URLS with this simple heuristic.

To further measure the variability of the AIJs associated with the URLs at the k^{th} position, the standard deviation of the i^{th} AIJs ($Std_{i,k}$) with respect to the mean value of the relevance scores is calculated as

$$Std_{i,k} = \sqrt{\frac{1}{|URL_k|} \sum_{n_k \in URL_k} \left(RS_{i,n_k} - \overline{RS_{n_k}}\right)^2}, \tag{2}$$

where URL_k and $|URL_k|$ have the same definitions as in Equation (1). RS_{i,n_k} denotes the relevance score of the n_k query-URL pair that is assigned by the i^{th} annotator. $\overline{RS_{n_k}} = \frac{1}{L} \sum_{i=1}^{L} RS_{i,n_k}$ (L: the number of annotators who label the n_k query-URL pair)

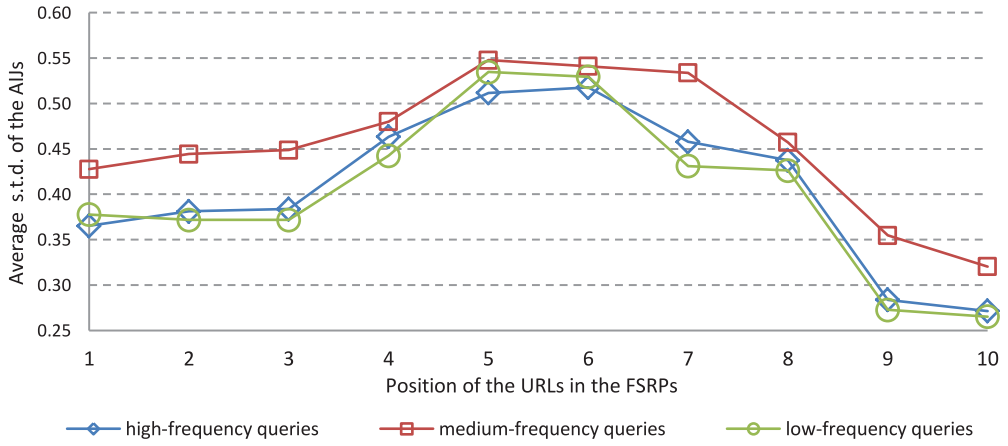

Fig. 2. Correlation among $Ave_{AIJ,k}$, search frequencies, and the positions of the URLs in FSRPs.

indicates the mean value of all individual relevance scores assigned to the n_k query-URL pair. Thus, for a given URL, the average standard deviation of the AIJs ($Ave_{Std(k)}$) is defined as

$$Ave_{Std(k)} = \frac{1}{N_{ann}} \sum_{i=1}^{N_{ann.}} Std_{i,k}. \tag{3}$$

Figure 2 depicts the correlation among ($Ave_{Std(k)}$), search frequencies, and the positions of the URLs in FSRPs.

In Figure 2, we can see that the $Ave_{Std(k)}$ of medium-frequency queries is larger than those of high-frequency queries and low-frequency queries, probably because the search intents associated with high-frequency and low-frequency queries are more explicit and specific than those associated with medium-frequency queries. It enables annotators to identify and understand search intent more easily and hence to provide AIJs with smaller $Ave_{Std(k)}$. In addition, it is also observed that for any given search frequency, $Ave_{Std(k=1,2,3,9 \text{ and } 10)}$ are much smaller than $Ave_{acy,k=4\sim8}$. The possible reason is that the retrieved URLs located at the 4th∼8th positions in FERPs are usually of more diverse qualities, which sometimes makes the individual annotator rely more on his or her subjective expertise to assess the query-URL relevance. As a result, these excessive subjectivities lead to more inconsistency between multiple AIJs.

Regarding the aforementioned analysis results, we must highlight that an alternative reason possibly also leads to the observations about $Ave_{acy,k}$ and $Ave_{Std(k)}$. It is mainly due to the use of the 5-point scale. The detailed analysis about this possible reason and its potential impacts on $Ave_{acy,k}$ and $Ave_{Std(k)}$ are beyond the scope of this work and planned for future work.

3.3. Proposal and Benefits

According to the preceding analysis, we observe that it is a challenging task to estimate the GTRs for the query-URL pairs located on the 4th∼8th positions in the FSRPs. In general, these query-URL pairs are labeled with more inconsistent AIJs (greater $Ave_{Std(k)}$). Furthermore, we believe that the disagreement among AIJs is not noise, but signal of the vagueness and ambiguity in manual labeling process, which potentially provides more cues to improve the accuracy of the automatically estimated GTRs.

Following this observation and hypothesis, we propose integrating the subjective expertise underlying multiple AIJs with the features characterizing query-URL pairs for estimating GTRs. As mentioned earlier, when assessing the query-URL relevance,

the inconsistency among different AIJs can actually be considered as an indicator of whether it is necessary to recheck and have a group discussion regarding the quality of the Web pages linked by the retrieved URLs. For example, given a query-URL pair and its relevance assigned as Perfect (4) by annotator A and Bad (0) by annotator B, it is necessary for annotator A and B to recheck the quality of the Web page linked by the given URL to achieve more accurate and reliable GTR. Actually, the similar idea has been applied in the context of information retrieval system, such as constructing minimal test collections for Cranfield paradigm evaluation [Carterette et al. 2006], and in evaluating systems accurately with minimal number of query submissions [He et al. 2011].

The main benefit of our work is that it improves the efficiency and accuracy in estimating GTRs. As mentioned earlier, in phase II, multiple annotators conventionally have to spend extra effort in rechecking and group discussing the quality of the Web pages linked by the retrieved URLs to reach more reliable GTRs. In contrast, our proposal offers a probabilistic alternative to estimating the GTRs by integrating the subjective expertise underlying multiple AIJs with the features characterizing query-URL pairs rather than rechecking and group discussing the Web pages linked by the retrieved URLs.

In addition, according to the analysis in Section 3.2, we can see that the disagreement among multiple annotator judgments always exists, especially for the medium-frequency and low-frequency queries, possibly because the linguistic expressions of these queries create a fairly wide range of possible and plausible interpretations on information needs, which leads to the significant gap between multiple annotators' views on relevance. In practice, the satisfactory results for medium-frequency and low-frequency queries can boost the head requests due to increased user satisfaction and repeat patronage [Goel et al. 2010]. Therefore, we believe that it is worth spending extra effort in improving the accuracy and efficiency of the GTRs associated with medium-frequency and low-frequency queries.

4. PROPOSED APPROACH

Following the preceding motivation, we propose a probabilistic model in which the multiple AIJs are exploited together with the features characterizing query-URL pairs to estimate the GTRs.

4.1. Notations and Problem Formulation

We start by introducing the notations used throughout the article:

—$x_i \in \mathcal{X}$ represents the feature vector that characterizes the i^{th} query-URL pair.
—$y_i^{\neg} \in \mathcal{Y}^{\neg}$ is a d-dimensional vector, representing the d-independent AIJs on the i^{th} query-URL pair.
—$y_i \in \mathcal{Y}$ represents the GTR of the i^{th} query-URL pair.

Note that y_i and each element in $y_i^{\neg}$ are described by numeric values {Perfect(4), Excellent(3), Good(2), Fair(1),Bad(0)}, scaling from highly relevant to not relevant.

Based on the preceding notations, the concerned problem can be formally defined as follows: to learn function $\mathcal{F}$ that best represents the correlation between the input $(\mathcal{X}, \mathcal{Y}^{\neg})$ and the output $\mathcal{Y}$ given training set of input-out pairs $\{x_i, y_i^{\neg}, y_i\}_{i=1,\ldots,N}^{x_i \in \mathcal{X}; y_i^{\neg} \in \mathcal{Y}^{\neg} \text{ and } y_i \in \mathcal{Y}}$, where N indicates the number of query-URL pairs in the training set. For the given set of query-URL pairs, the set of GTRs $(\mathcal{Y})$ is not only related to the feature space $\mathcal{X}$ but also is related to the set of AIJs $(\mathcal{Y}^{\neg})$.

As a result, a discriminative function $\mathcal{F}$ with generalized linear form is defined to represent the relationships among $\mathcal{Y}$, $\mathcal{X}$, and $\mathcal{Y}^{\neg}$,

$$\mathcal{F}(\mathcal{X}, \mathcal{Y}^{\neg}, U, V) = U^T \Phi(\mathcal{X}) + V^T \Psi(\mathcal{Y}^{\neg}), \tag{4}$$

where U^T and V^T are $1 \times N$ parameter vectors, respectively, and $\Phi(\mathcal{X})$ and $\Psi(\mathcal{Y}^\neg)$ are $N \times N$ matrix, representing the high-dimensional mappings induced by kernel functions $\mathcal{K}^\Phi(x_i, x_j)_{i,j \in N}$ and $\mathcal{K}^\Psi(y_i^\neg, y_j^\neg)_{i,j \in N}$, respectively. For simplicity, $\mathcal{K}^\Phi(x_i, x_j)_{i,j \in N}$ describes the pair-wise cosine similarity between the query-URL pair x_i and the query-URL pair x_j. Analogously, $\mathcal{K}^\Psi(y_i^\neg, y_j^\neg)$ also indicates the cosine similarity between the multiple AIJs of the given query-URL pair x_i and those of the query-URL pair x_j. Furthermore, each element in $\mathcal{F}_{1 \times N}$ is a real value in $[0, 4]$, denoting the corresponding GTR.

Equation (4) clearly shows the major difference between other methods and our approach belonging to category (d). From the perspective of estimating GTRs, our approach and those methods in category (d) integrate both objective factors (i.e., the features characterizing the query-URL pairs $\mathcal{X}$) and subjective factors (i.e., the multiple AIJs $\mathcal{Y}^\neg$) that have potential impacts on the quality of estimating GTRs, whereas other methods only consider one of these two factors. Therefore, the proposed approach is essentially a combination of categories (b) and (c), which represent two categories of the existing approaches in this literature (as described in Section 2).

4.2. Parameter Estimation

To find function $\mathcal{F}$ that models the correlation between input $(\mathcal{X}, \mathcal{Y}^\neg)$ and output $\mathcal{Y}$, the classical way is to apply ordinary least square (OLS) for minimizing squared loss function $\mathcal{L}(\mathcal{Y}, \mathcal{F})$, which measures the difference between the estimated GTR $f_i \in \mathcal{F}$ and the expected GTR $y_i \in \mathcal{Y}$ described by Equation (5):

$$\mathcal{L}(\mathcal{Y}, \mathcal{F}) = \frac{1}{2}(\mathcal{Y} - \mathcal{F})(\mathcal{Y} - \mathcal{F})^T = \frac{1}{2}\sum_{i=1}^{N}(y_i - f_i)^2. \tag{5}$$

Minimizing squared loss function $\mathcal{L}(\mathcal{Y}, \mathcal{F})$ needs to learn the optimal parameters U and V in Equation (4). Due to the difficulties in data collection, the number of query-URL pairs that can be used for training is limited. A small training dataset may cause an overfitting problem. In the case of applying OLS on Equation (5), the variance and magnitude of estimated parameters U and V will be unfavorably large so that estimation error is very high, although $\mathcal{L}(\mathcal{Y}, \mathcal{F})$ on training set reaches a small error. An effective way to overcome this problem is to penalize the norm of U and V as in ridge regression instead of only minimizing squared errors. Following kernel ridge regression [Saunders et al. 1998], the optimal parameters U and V can be obtained by minimizing the regularized empirical risk $\mathcal{R}_\mathcal{L}$,

$$\{U^*, V^*\} = argmin_{\{U,V\}}\mathcal{R}_\mathcal{L}, \tag{6}$$

where $\mathcal{R}_\mathcal{L} = \sum_{i=1}^{N}\{\mathcal{L}(y_i, f_i(x_i, y_i^\neg; U, V)) + \frac{\mathcal{C}_U}{2}U^2 + \frac{\mathcal{C}_V}{2}V^2\}$.

In Equation (6), $\frac{\mathcal{C}_U}{2}U^2 + \frac{\mathcal{C}_V}{2}V^2$ is the $L_2 - norm$ regularizer, and $\mathcal{C}_U$ and $\mathcal{C}_V$ are hyperparameters that control the intensity of regularization. According to the dual version for ridge regression [Saunders et al. 1998], the minimization of the regularized empirical risk $\mathcal{R}_\mathcal{L}$ in Equation (6) over the training set can be re-expressed as

$$\begin{aligned} &min_{\{U,V\}}\left\{\frac{1}{2}\sum_{i=1}^{N}\xi_i^2 + \frac{\mathcal{C}_U}{2}U^2 + \frac{\mathcal{C}_V}{2}V^2\right\} \\ &subjectto \qquad \xi_i = y_i - U^T\Phi(x_i) - V^T\Psi\left(y_i^\neg\right). \end{aligned} \tag{7}$$

By introducing Lagrangian multiplier α_i, the unstrained Lagrangian $\mathcal{H}$ of Equation (7) is described as

$$\mathcal{H} = \frac{1}{2}\sum_{i=1}^{N}\xi_i^2 + \frac{\mathcal{C}_U}{2}U^2 + \frac{\mathcal{C}_V}{2}V^2 + \sum_{i=1}^{N}\alpha_i\{y_i - U^T\Phi(x_i) - V^T\Psi(y_i^\neg) - \xi_i\}. \tag{8}$$

According to Karush-Kuhn-Tucker conditions, there exist Lagrangian multipliers α_i for which the minimum of Equation (8) equals the minimization problem of Equation (7). In other words, the minimization problem of Equation (7) can be solved in dual form, namely to find the saddle point of the Lagrangian $\mathcal{H}$. To find the optimal U, V, and ξ_i, we compute the partial derivatives of the Lagrangian $\mathcal{H}$ with respect to U and V, respectively, and let them equal 0. Then we have

$$U = \frac{1}{\mathcal{C}_U} \sum_{i=1}^{N} \alpha_i \Phi(x_i), \tag{9}$$

$$V = \frac{1}{\mathcal{C}_V} \sum_{i=1}^{N} \alpha_i \Psi\left(y_i^{\neg}\right). \tag{10}$$

Substituting Equations (9) and (10) into Equation (8), and setting the partial derivative of $\mathcal{H}$ with respect to ξ_i to be 0, we obtain

$$\xi_i = \alpha_i \qquad i = 1, \ldots, N. \tag{11}$$

Since Lagrangian multipliers α_i $(i = 1, \ldots, N)$ represent the importance of the constraint, Equation (9) indicates that U is proportional to the linear combination of features characterizing query-URL pairs (x_i), each of which is weighted with its importance α_i. Similarly, Equation (10) denotes that V is proportional to the linear combination of the d-independent AIJs $(y_i^{\neg})$, each of which is weighted with its importance α_i.

Substituting U, V, and ξ_i for the right-hand side of Equations (9) through (11) in Equation (8), the dual optimization of Equation (7) is described by

$$min_{\alpha} \frac{1}{2} \alpha^T \left(\frac{1}{\mathcal{C}_U} \mathcal{K}^{\Phi} + \frac{1}{\mathcal{C}_V} \mathcal{K}^{\Psi} + I \right) \alpha - \sum_{i=1}^{N} \alpha_i y_i, \tag{12}$$

where $\alpha = [\alpha_1, \ldots, \alpha_i, \ldots, \alpha_N]^T$ is a vector of Lagrangian multipliers α_i; $(\frac{1}{\mathcal{C}_U} \mathcal{K}^{\Phi} + \frac{1}{\mathcal{C}_V} \mathcal{K}^{\Psi} + I)$ is a $N \times N$ matrix; $\mathcal{K}^{\Phi}$ and $\mathcal{K}^{\Psi}$ are $N \times N$ kernel matrix with elements $k^{\Phi}(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ and $k^{\Psi}(y_i^{\neg}, y_j^{\neg}) = \Psi(y_i^{\neg})^T \Psi(y_j^{\neg})$; and I is the $N \times N$ identity matrix. Since Equation (12) is an unstrained convex quadratic program, the closed-form solution to estimate α^* is described as

$$\alpha^* = \left(\frac{1}{\mathcal{C}_U} \mathcal{K}^{\Phi} + \frac{1}{\mathcal{C}_V} \mathcal{K}^{\Psi} + I \right)^{-1} \mathcal{Y}. \tag{13}$$

4.3. Estimating the GTR of a New Query-URL Pair

For a new query-URL pair with feature vector x_{new} and multiple annotators' assessments $y_{new}^{\neg}$, we estimate the GTRs $\hat{y}_{new}$ by combining Equations (9) through (11) and Equation (13),

$$\hat{y}_{new} = argmin_{y_{new}} |y_{new} - f_{new}|. \tag{14}$$

where y_{new} represents a numeric value in the ordinal set {Perfect(4), Excellent(3), Good(2), Fair(1), Bad(0)}, and

$$\begin{aligned} f_{new} &= \left(\frac{1}{\mathcal{C}_U} \sum_{i=1}^{N} \alpha_i^* \Phi(x_i) \right)^T \Phi(x_{new}) + \left(\frac{1}{\mathcal{C}_V} \sum_{i=1}^{N} \alpha_i^* \Psi\left(y_i^{\neg}\right) \right)^T \Psi\left(y_{new}^{\neg}\right) \\ &= \frac{1}{\mathcal{C}_U} \sum_{i=1}^{N} \alpha_i^* \mathcal{K}^{\Phi}(x_{new}, x_i) + \frac{1}{\mathcal{C}_V} \sum_{i=1}^{N} \alpha_i^* \mathcal{K}^{\Psi}\left(y_{new}^{\neg}, y_i^{\neg}\right), \end{aligned} \tag{15}$$

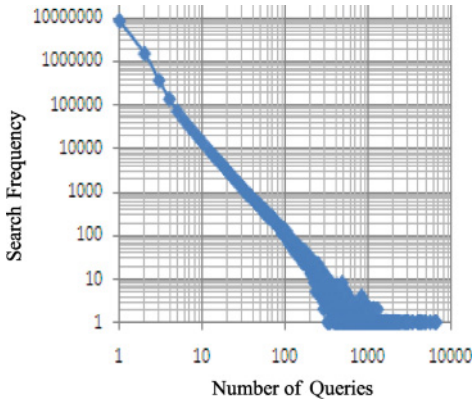

Fig. 3. Distribution of training dataset

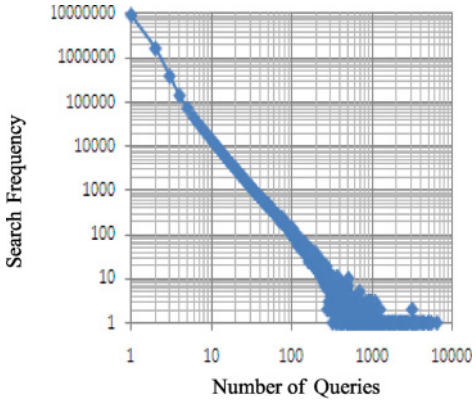

Fig. 4. Distribution of test dataset.

where $\mathcal{K}^{\Phi}(x_{new}, x_i)$, as kernel function value, denotes the pair-wise cosine similarity between the query-URL pair x_{new} and the query-URL pairs in the training dataset. Similarly, $\mathcal{K}^{\Psi}(y_{new}^{\neg}, y_i^{\neg})$ indicates the pair-wise cosine similarity between the multiple AIJs of the given x_{new} and those of the query-URL pairs in the training dataset.

5. EXPERIMENTS

In this section, we conduct extensive experiments in the search engine Baidu.com. First, we employ a specific query, "How to distinguish gold-collar, white-collar and blue-collar," as an example to illustrate the effectiveness of the proposed approach. Furthermore, we compare the experimental results against those generated by the typical baseline models that belong to categories (a) and (b) (i.e., the mouse trajectory–based model [Song et al. 2012], PCB and PCB-User models [Guo et al. 2012], and Unbiased-UBM and Unbiased-DBN [Hu et al. 2011]) and demonstrate the advantages of our approach in terms of the NDCGs (NDCG@k).

Note that this article does not compare the proposed approach to the crowdsourcing-based solutions in category (c), primarily because the crowdsourcing-based solutions have not actually been applied to estimate query-URL relevance in Baidu.com.

5.1. Experimental Setting

Training and testing datasets. As described in Section 3.1, the dataset prepared for this study consists of 89,900 query-URL pairs, in which each query has 10 retrieved URLs in the FSRP. For each query-URL pair in the dataset, five senior annotators and five junior annotators are employed to generate the GTR associated with the given query-URL pair following the working flow described in Section 3. Since machine learning methods work well on the premise that the training and testing data have similar distributions [Pan and Yang 2010], the dataset is randomly and equally divided into a training set and a testing set so that they have similar distributions regarding search frequency, as shown in Figures 3 and 4, respectively.

Note that this dataset differs significantly from the one used in Song et al. [2011] and Song et al. [2012]. Figures 3 and 4 indicate that our experiments focus more on medium-frequency queries ($4{,}097/\text{day} > Search_{Fre.} \geq 9/\text{day}$) rather than uniform distributed queries regarding search frequency in Song et al. [2011] and Song et al. [2012]. In contrast, the dataset used in this article is more consistent with reality, as the search traffic associated with medium-frequency queries is about 50% in Baidu.com (depending on the statistical time interval).

Table I. Features Used to Characterize a Query-URL Pair

	Context Level
#1	Number of query terms
#2	Number of terms in URL title
#3	Number of terms in snippet
#4	Number the terms in query covered by a URL title / #1
#5	Number the terms in query covered by a snippet / #1
#6	Ratio between #4 and max of #4 (given query and its associated different URLs)
#7	Ratio between #4 and min of #4 (given query and its associated different URLs)
#8	Ratio between #5 and max of #5 (given query and its associated different URLs)
#9	Ratio between #5 and min of #5 (given query and its associated different URLs)
#10	Mean of #4 (given query and its associated different URLs)
#11	Ratio between #4 and mean of #4 (given query and its associated different URLs)
#12	Ratio between #5 and mean of #5 (given query and its associated different URLs)
#13	Max. term frequency with respect to the Web pages linked by the retrieved URL (for a set of terms in the given query)
#14	Min. term frequency with respect to the Web pages linked by the retrieved URLs (for a set of terms in the given query)
#15	Inverse document frequency (IDF) of the Web page linked by the retrieved URL
	User Behavior Level
#16	Given query, click entropy of a URL as the first choice
#17	Ratio of skewness of a URL clicks with respect to the skewness of the query clicks
#18	Given query, click entropy of a URL as the last choice
#19	Standard score of a URL clicks with respect to query clicks
#20	For a query and its associated URL, the ratio between URL click number and the query search number
#21	For a query and its associated URL, the ratio between click entropy of a URL and click entropy of the query

5.2. Features

Except for the multiple AIJs associated with query-URL pairs, Table I lists the features used in our experiments, which are divided into context-level and user behavior–level features, respectively. All features are defined based on our intuitive judgments and empirical experiments. We describe them separately in the following.

Context-level features. These features mainly model the inherent properties about the query, the title, and the snippet that are associated with the given query-URL pair. Specifically, these features describe the readability and expression of the title and snippet associated with the retrieved URL. It is noted that although these features have direct impact on user experience [Kanungo and Orr 2009], they are not influenced by user behaviors. In contrast to click/browse data, we believe that the context-level features are possibly more suitable and objective to describe the intrinsic quality of the retrieved URLs. For example, feature #4 measures the matching degree between a query and a retrieved URL title.

User behavior–level features. These features model the aggregated user behavior–associated query-URL pairs. From this perspective, behavior-level features describe the users' response to retrieved URLs. For example, similar to the definition about click entropy in Deng et al. [2009] and Duan et al. [2012], we define feature #16 associated with a given URL as follows: Feature #16 $= -P_{NFK}log(P_{NFK})$, where $P_{NFK} = (click\ num.$

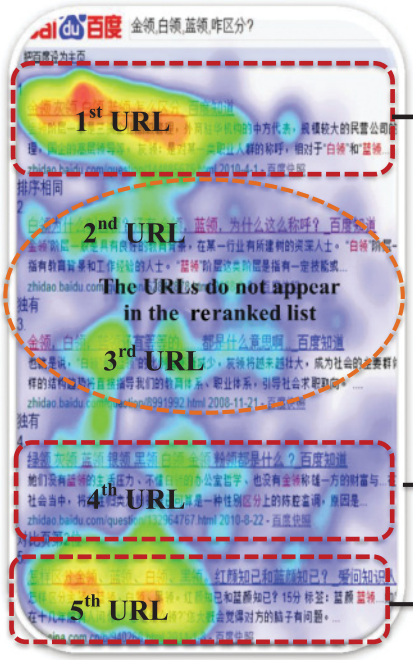

Fig. 5. Original ranked list.

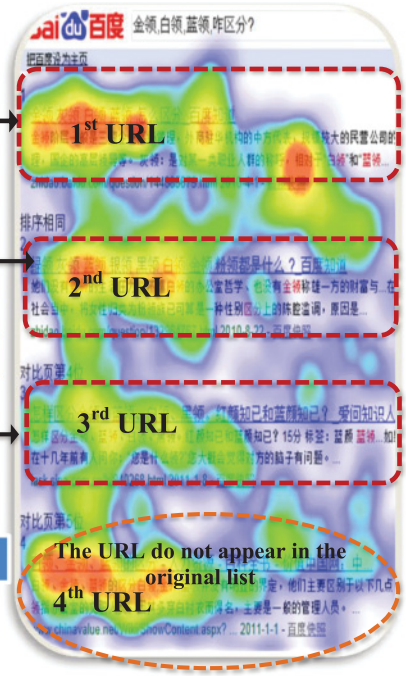

Fig. 6. Reranked list.

the URL as the first choice) / (click num. the URL). A smaller value of feature #16 indicates that users more likely click the URL as the first choice.

5.3. Case Study

To illustrate the effectiveness and performance of the proposed approach, we employ a specific query, "How to distinguish gold-collar, white-collar and blue-collar," as an example. By comparing CTRs of the original ranking list to that of the reranked list, we demonstrate the advantages of our approach in improving user experience intuitively, mainly because CTRs associated with the retrieved URLs have been practically considered as a critical and intuitive indicator of users' immediate responses to the quality of commercial search engines [Huang et al. 2012]. Note that the example query is a medium-frequency query.

First, Figure 5 depicts the click heat map of the original ranking list that is generated by the existing ranking strategy/algorithm in Baidu. Then, the proposed approach is applied to estimate the probability that the relevance of each query-URL pair (partially shown in Figure 5) is $G_{rade} = \{\text{perfect}, \text{excellent}, \text{good}, \text{fair}, \text{bad}\}$. Table II gives the calculated results in which the k^{th} column indicates the normalized probability that the k^{th} URL is labeled as G_{rade}. The probability in bold shows that the URL at the k^{th} position is most possibly labeled as G_{rade}—that is, the 1st URL associated with the given query is Excellent and the 2nd URL is Fair.

Furthermore, we rerank the URLs in the original ranking list according to the calculated results in Table II. Figure 6 depicts the click heat map of the reranked list. Note that for the given query, the 1st URL in Figure 5 is same as that in Figure 6. However, the 4th and 5th URLs in Figure 5 are moved up as the 2nd and 3rd URLs in Figure 6, respectively. In addition, the 2nd and 3rd URLs in Figure 5 do not appear in Figure 6.

Table II. Probabilities That the GTR of the URLs in the Original Ranking List Is G_{rade}

Probability of Relevance of Query-URL Pair	1st URL	2nd URL	3rd URL	4th URL	5th URL
G_{rade} is *Perfect*	0.307	0.076	0.026	0.171	0.021
G_{rade} is *Excellent*	**0.539**	0.136	0.042	0.116	0.212
G_{rade} is *Good*	0.128	0.204	0.074	**0.585**	**0.498**
G_{rade} is *Fair*	0.014	**0.423**	0.337	0.048	0.243
G_{rade} is *Bad*	0.012	0.161	**0.521**	0.079	0.026

Likewise, the 4th URL in Figure 6 does not appear in Figure 5. As a controlled experiment, Figures 5 and 6 are derived from the real users in Baidu.com. Under the same normal search circumstances, the test group (Figure 5) covers 75% of search traffic, whereas the control group (Figure 6) covers 25% of search traffic.

Note that in Figures 5 and 6, the color red marks the most heat, where the CTRs are highest, and they fade to green or nothing at all where they receive little or no attention. Comparing Figures 5 and 6, we observe that (a) the reranked list achieves higher CTRs than those of the original ranked list, and (b) the improvement in CTRs is mainly associated with the 2nd, 3rd, and 4th URLs in Figure 6.

By re-estimating the relevance associated with the URL candidates that were retrieved by the existing ranking strategy/algorithm, we further filtered the 2nd and the 3rd URLs in the original FSRP (Figure 5) out of the FSRP in the reranked list (Figure 6). In addition, a URL not present in the original FSRP (Figure 5) is moved up to the 4th one in the FSRP of the reranked list (Figure 6). The CTR associated with this URL in the reranked list (Figure 6) shows that it is also closely relevant to the given query, which intuitively proves the effectiveness of our approach on estimating query-URL relevance.

The 4th URL in Figure 5 and the 2nd URL in Figure 6 are worthy of a detailed investigation. Although these two URLs are essentially the same, the CTRs associated with them reflect significant difference. On the one hand, the much less CTR of the 4th URL in Figure 5 is mainly due to a cascade hypothesis that is extensively utilized by the cascade model [Craswell et al. 2008], stating that a URL is examined only if its upper neighbor is examined. From this perspective, the less CTR associated with the 4th URL in Figure 5 is due to the poor quality of the 3rd URL, instead of its own quality. According to the query-URL relevance estimated by our approach (Table II), the 4th URL in Figure 5 is moved up to the 2nd URL in Figure 6, which increases the CTRs of the given URL significantly. One may argue that it is mainly due to the position bias. But comparing the CTR of the 2nd URL in Figure 5 to that of the 2nd URL in Figure 6, the former is much less than the latter. Therefore, in our opinion, the position bias is not the main reason leading to the significant changes on CTRs. Regarding the 4th URL in Figure 5 that was initially retrieved and ranked by the existing ranking strategy/algorithm, our approach estimates its relevance to the given query more accurately and hence moves it up to the 2nd URL in Figure 6. By doing so, our approach improves the user experience and satisfies user search intent more efficiently.

Summary. This study case intuitively demonstrates the effectiveness of the proposed approach on improving the user experience. Regarding the URLs that have been retrieved by the existing ranking strategy/algorithm and partially shown in Figure (5), we

re-estimated their relevance to the given query using the proposed approach. Based on the estimated query-URL relevance, we adjusted these URLs' positions in the reranked list (partially shown in Figure 6). Experimental results clearly show that the proposed approach satisfies search intent more efficiently and improves user experience.

5.4. NDCG Evaluation

In this section, we employ the NDCGs (NDCG@k) as a metric to quantitatively evaluate the proposed approach. NDCG [Jarvelin and Kekalainen 2000] measures the divergence between the predicted ranking and manually labeled GTRs. It is particularly suitable for Web search applications, as it accounts for multilevel relevance and the truncation level can be compared to the manually labeled GTRs. Specifically, we followed the definition of NDCG in Burges et al. [2007]. For a set of queries Q, let $R(j,d)$ be the relevance level given to the d^{th} URL for query j. Then,

$$NDCG(Q,k) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \left(2^{R(j,m)} - 1\right) / log_2(1+m), \tag{16}$$

where $|Q|$ is the cardinality of query set Q, and where the Z_{kj} is a normalization factor so that a perfect ordering NDCG at k for query j is 1. Here, k is the ranking truncation level at which the NDCG is computed. The $NDCG(Q,k)$ is then averaged over the query set Q.

The baseline models to be compared against include the mouse trajectory–based model [Song et al. 2012], PCB and PCB-User models [Guo et al. 2012], and Unbiased-UBM and Unbiased-DBN [Hu et al. 2011], which belong to categories (a) and (b), respectively. Training of these baseline models follows the inference algorithms introduced in the respective original papers. Following that, each constructed model is employed to infer the query-URL relevance of the testing set. This article does not compare the proposed approach to the crowdsourcing-based solutions in category (c) because the crowdsourcing-based solutions have not actually been applied to estimate query-URL relevance in Baidu.com.

Note that in the NDCG evaluation, the comparisons to baseline models might not be strictly just and impartial, primarily because the available public datasets (e.g., Microsoft Learning to Rank, http://research.microsoft.com/en-us/projects/mslr/; LETOR 4.0: A Benchmark Collection for Research on Learning to Rank for Information Retrieval, http://research.microsoft.com/en-us/um/beijing/projects/letor//) lack the multiple AIJs that are necessary for the proposed approach. In this case, most of the state-of-the-art models that utilize the context-level and behavior-level features ignore the potential impact of AIJs. Even so, we still believe that the NDCG evaluation is useful to quantitatively validate the effectiveness of the proposed approach and the potentially positive impacts of the multiple AIJs on estimating GTRs.

Table III reports the arithmetic mean of NDCG@k ($k = 1, \ldots, 10$) that are generated by respective models for queries with different frequencies (high, medium, and low). For the baseline models, NDCG scores slightly differ from those reported in the respective original papers, possibly because of the intrinsic difference between the Chinese search engine and English search engine (e.g., on understanding user behavior and search intents) [Xiao et al. 2008]. The impact of the difference between the Chinese search engine and English search engine is planned for future study. In addition, Table III reports the statistical significance (p-values) between the baseline models and our approach. Since the NDCG@k is the averaged over the query set Q with high frequency, medium frequency, and low frequency, respectively, p-values are computed for z-tests.

For queries with different frequencies, NDCG@1 and 2 achieved by the proposed approach are similar to those of Unbiased-UBM and the mouse trajectory–based model.

Table III. NDCG Comparisons to Baseline Models

High Frequency	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10	*p*-Value
PCB	0.655	0.643	0.652	0.683	0.688	0.694	0.707	0.719	0.727	0.733	0.00011
PCB-User	0.606	0.652	0.672	0.687	0.693	0.688	0.715	0.731	0.746	0.751	0.00382
Unbiased-UBM	**0.691**	**0.713**	0.715	0.707	0.721	0.721	0.723	0.727	0.714	0.722	0.00417
Unbiased-DBN	0.681	0.677	0.670	0.664	0.683	0.691	0.721	0.727	0.735	0.732	0.00037
Mouse trajectory model	0.687	0.706	0.692	0.712	0.717	0.724	0.731	0.733	0.736	0.717	0.00997
Our approach	0.686	0.711	**0.721**	**0.745**	**0.747**	**0.752**	**0.746**	**0.747**	**0.753**	**0.755**	—

Medium Frequency	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10	*p*-Value
PCB	0.624	0.527	0.534	0.515	0.549	0.554	0.611	0.683	0.687	**0.724**	2.3e-05
PCB-User	0.592	0.560	0.542	0.593	0.604	0.631	0.684	0.697	0.712	0.707	0.00042
Unbiased-UBM	0.656	0.623	0.625	0.637	0.641	0.623	0.637	0.607	0.614	0.605	2.7e-13
Unbiased-DBN	0.632	0.627	0.637	0.646	0.626	0.631	0.637	0.634	0.608	0.617	6.2e-14
Mouse trajectory model	**0.663**	0.633	0.644	0.637	0.644	0.625	0.638	0.659	0.686	0.692	3.4e-06
Our approach	0.657	**0.644**	**0.698**	**0.724**	**0.732**	**0.728**	**0.721**	**0.718**	**0.724**	0.722	—

Low Frequency	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10	*p*-Value
PCB	0.572	0.474	0.472	0.491	0.524	0.525	0.545	0.539	0.569	0.565	1.52e-08
PCB-User	0.517	0.453	0.469	0.529	0.578	0.601	0.647	0.678	**0.673**	**0.676**	0.04929
Unbiased-UBM	**0.633**	**0.550**	0.544	0.549	0.534	0.557	0.574	0.592	0.583	0.572	0.00011
Unbiased-DBN	0.552	0.531	0.532	0.543	0.551	0.571	0.549	0.538	0.539	0.522	2.8e-09
Mouse trajectory model	0.621	0.544	0.533	0.556	0.548	0.569	0.612	0.654	0.594	0.613	0.00605
Our approach	0.631	0.547	**0.568**	**0.613**	**0.627**	**0.646**	**0.681**	**0.688**	0.658	0.677	—

This observation is reasonable, as the proposed approach is insensitive to the consistent AIJs associated with the 1st and 2nd URLs (as discussed in Section 3.2). For the same reason, the performance of the proposed approach on NDCG@ 8, 9, and 10 are comparable to those of the baseline models, especially for the high-frequency queries. Note that the improvement associated with high-frequency queries is definitely nontrivial when taking into account the hundreds of millions of search requests served by Baidu.com every day and the approximate 30.4% of search traffic associated with high-frequency queries. Since the proposed approach is initially designed to improve the GTR estimation for URLs at the 3rd through 7th positions, it is acceptable that the proposed approach's performance is comparable to baseline methods over URLs at other positions. In contrast, the gains achieved by the proposed approach on NDCG@3∼7 are more significant than those on NDCG@1, 2, 8, 9, and 10, especially for the medium-frequency and low-frequency queries, as shown in Table III.

In Figures 7 and 8, the relative NDCG improvement of the proposed approach $e_{pro.}$ over respective baseline model e_{other} is further measured as $(e_{pro.} - e_{other})/e_{other} \times 100\%$. For both the medium-frequency and low-frequency queries, we can see that the proposed approach outperforms the baseline models with different degree of improvements (mostly more than 10%) in terms of NDCG for URLs at the 3rd through 7th positions.

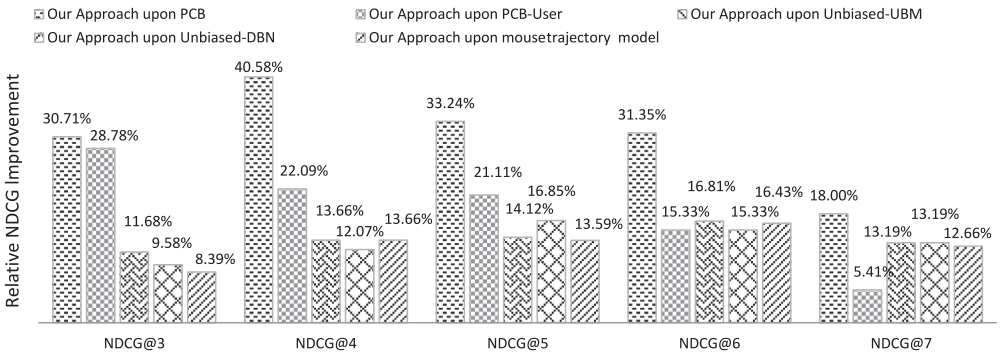

Fig. 7. Relative NDCG improvements on medium-frequency queries.

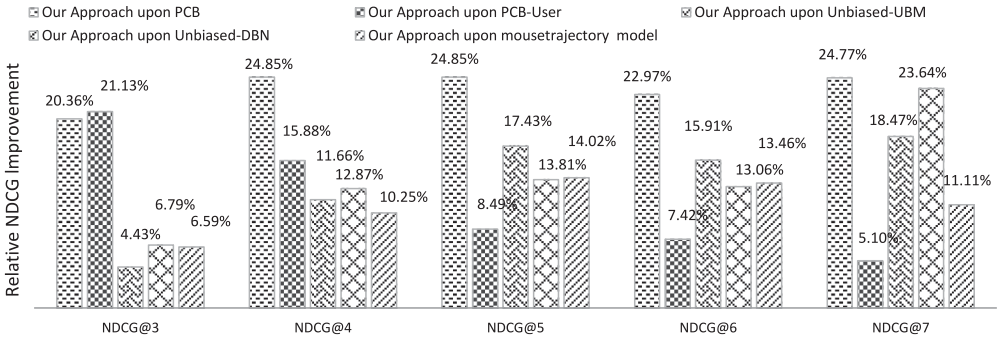

Fig. 8. Relative NDCG improvements on low-frequency queries.

We also observe that the relative NDCG@3∼6 improvements of medium-frequency queries are more significant than those of low-frequency queries. Considering the higher $Ave_{Std(3\sim6)}$ associated with medium-frequency queries than those associated with low-frequency queries (Section 3.2), it justifies that our research motivation is reasonable—that is, the inconsistency among multiple AIJs does provide more cues to improve the accuracy of estimating GTRs. It is also noted that the relative NDCG@7 improvements of low-frequency queries are more significant than those of medium-frequency queries, primarily because the click/browse behaviors associated with the 7^{th} URLs of low-frequency queries are much sparser than those of medium-frequency queries. Therefore, this phenomenon possibly proves that the features extracted from the title, anchor text, and search log associated with the Web pages linked by the retrieved URLs play an important role in estimating the GTRs, especially for the URLs with lower ranking of the low-frequency queries.

6. CONCLUSION

Different from the state-of-the-art models on estimating GTRs, we investigate the necessity of employing the knowledge from multiple AIJs and propose a probabilistic model in which the multiple AIJs are integrated with the features characterizing query-URL pairs to estimate the GTRs. By conducting experiments with a dataset collected from Baidu.com, we demonstrate that the proposed approach consistently and significantly outperforms related works, which verifies the effectiveness of the proposed approach.

By utilizing multiple AIJs to estimate the relevance of query-URL pairs, this work is complementary to our previous work in Song et al. [2011] and Song et al. [2012]. With regard to the medium-frequency queries and low-frequency queries, the proposed approach provides a promising solution to estimate GTRs more accurately. It makes it possible to address the problem that multiple annotators have to spend extra effort on rechecking and group discussing the quality of the Web pages linked by the retrieved URLs to achieve more reliable GTRs. Despite its promising performance, the proposed approach still has space for improvement. For example, the overhead in the preprocessing stage, including the costs of acquiring features, refining features, and cleaning data and collecting AIJs with low cost, among others, can be further reduced. We plan to address this in the future. In addition, our study is conducted without missing assessors' judgments, although it is common in crowdsourcing settings [Raykar and Yu 2012]. This is mainly because all annotators in our study are required to provide their individual relevance judgments on all query-URL pairs (actually, these annotators are professional employees who are trained according to some company inner standards, so their judges have a certain degree of bias). From this perspective, concerns expressed in this article are different from cases in crowdsourcing settings. For cases with missing assessors' judgments, future work is planned based on the Baidu crowdsourcing platform (http://test.baidu.com/crowdtest/).

Note that the potential impact on ranking performance imposed by the "objective" features and "subjective" features, respectively, has not been analyzed explicitly, which is a weakness inherent in our work. Due to experimental limits, we plan to address this in future work. We also plan to apply the proposed approach to other datasets and quantitatively validate its benefits for learning-to-rank algorithms. We also plan to integrate it with our previous work in Song et al. [2011] and Song et al. [2012] to define a more general and unified framework to estimate query-URL relevance in search engines. In addition, it is important to investigate the distinguishable impact of different features on the accuracy of estimated GTRs by running ablation tests, which is also planned for future work.

ACKNOWLEDGMENTS

The authors wish to thank Xinfeng Ou, Dan Chen, and Ming Li of Baidu Inc. for their support and insights drawn from platform data. We also thank the anonymous reviewers for their valuable help and comments. H. Song, H. Min, W. Wei, and J. Gu serve as the corresponding authors of this article.

REFERENCES

E. Agichtein, E. Brill, and S. Dumais. 2006. Improving Web search ranking by incorporating user behavior information. In *Proceedings of the Annual International ACM SIGIR Conference (SIGIR'06)*. 19–26.

P. Bailey, N. Craswell, I. Soboroff, P. Thomas, and E. Yilmaz. 2008. Relevance assessment: Are judges exchangeable and does it matter. In *Proceedings of the Annual International ACM SIGIR Conference (SIGIR'08)*. 667–674.

R. Blanco, H. Halpin, D. Herzig, P. Mika, J. Pound, and H. S. Thompson. 2011. Repeatable and reliable search system evaluation using crowdsourcing. In *Proceedings of the Annual International ACM SIGIR Conference (SIGIR'11)*. 923–932.

C. Buckley, M. Lease, and M. D. Smucker. 2010. Overview of the TREC 2010 Relevance Feedback Track (Notebook). Retrieved December 2, 2015, from https://www.ischool.utexas.edu/~ml/papers/trec-notebook-2010.pdf.

C. J. Burges, Q. V. Le, and R. Ragno. 2007. Learning to rank with nonsmooth cost functions. In *Proceedings of the Neural Information Processing Systems Conference (NIPS'07)*. 193–200.

B. Carterette, J. Allan, and R. Sitaraman. 2006. Minimal test collections for retrieval evaluation. In *Proceedings of the Annual International ACM SIGIR Conference (SIGIR'06)*. 268–275.

O. Chapelle and Y. Chang. 2011. Yahoo! learning to rank challenge overview. In *Proceedings of the JMLR Workshop (JMLR'11)*. 14:1–14:24.

O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09)*. 620–631.

W. Chen, Z. Ji, S. Shen, and Q. Yang. 2011. A whole page click model to better interpret search engine click data. In *Proceedings of the 25th Conference on Artificial Intelligence (AAAI'11)*.

C. Cleverdon. 1997. The Cranfield tests on index language devices. In *Readings in Information Retrieval*. Morgan Kaufman, San Francisco, CA, 47–59.

N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the International Conference on Web Search and Web Data Mining (WSDM'08)*. 87–94.

H. Deng, I. King, and M. R. Lyu. 2009. Entropy-biased models for query representation on the click graph. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*. 339–346.

H. Duan, K. Emre, and C. Zhai. 2012. Click patterns: An empirical representation of complex query intents. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM'12)*. 1035–1044.

G. Dupret and C. A. Liao. 2010. Model to estimate intrinsic document relevance from the click-through logs of a Web search engine. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10)*. 181–190.

G. Dupret and B. Piwowarski. 2008. A user browsing model to predict search engine click data from past observations. In *Proceedings of the Annual International ACM SIGIR Conference (SIGIR'08)*. 331–338.

Q. Guo and E. Agichtein. 2012. Beyond dwell time: Estimating document relevance from cursor movements and other post-click searcher behavior. In *Proceedings of the World Wide Web Conference (WWW'12)*. 569–578.

A. Gao, Y. Bachrach, P. Key, and T. Graepel. 2012. Quality expectation-variance tradeoffs in crowdsourcing contests. In *Proceedings of the 26th Conference on Artificial Intelligence (AAAI'12)*.

S. Goel, A. Broder, E. Gabrilovich, and B. Pang. 2010. Anatomy of the long tail: Ordinary people with extraordinary tastes. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10)*. 201–210.

D. Harman. 2010. Is the Cranfield paradigm outdated? In *Proceedings of the Annual International ACM SIGIR Conference (SIGIR'10)*. 1.

J. He, W. X. Zhao, B. Shu, X. M. Li, and H. F. Yan. 2011. Efficiently collecting relevance information from clickthroughs for Web retrieval system evaluation. In *Proceedings of the Annual International ACM SIGIR Conference (SIGIR'11)*. 275–284.

B. T. Hu, Y. C. Zhang, W. Z. Chen, G. Wang, and Q. Yang. 2011. Characterize search intent diversity into click models. In *Proceedings of the World Wide Web Conference (WWW'11)*. 17–26.

J. Huang, R. W. White, G. Buscher, and K. Wang. 2012. Improving searcher models using mouse cursor activity. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'12)*. 195–204.

P. Ipeirotis, F. Provost, and J. Wang. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOM'10)*. 64–67.

K. Jarvelin and J. Kekalainen. 2000. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the Annual International ACM SIGIR Conference (SIGIR'00)*. 41–48.

T. Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*. 133–142.

H. J. Jung and M. Lease. 2012. Inferring missing relevance judgments from crowd workers via probabilistic matrix factorization. In *Proceedings of the Annual International ACM SIGIR Conference (SIGIR'12)*. 1095–1096.

R. Jurca and B. Faltings. 2009. Mechanisms for making crowds truthful. *Journal of Artificial Intelligence Research* 34, 209–253.

T. Kanungo and D. Orr. 2009. Predicting the readability of short Web summaries. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM'09)*. 202–211.

G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. 2011. Crowdsourcing for book search evaluation: Impact of HIT design on comparative system ranking. In *Proceedings of the Annual International ACM SIGIR Conference (SIGIR'11)*. 205–214.

J. Le, A. Edmonds, V. Hester, and L. Biewald. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *Proceedings of the Workshop on Crowdsourcing for Search Evaluation (SIGIR'10)*. 21–26.

S. J. Pan and Q. Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10, 1345–1359.

V. C. Raykar and S. Yu. 2012. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research* 13, 491–518.

V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy. 2009. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *Proceedings of the 26th International Conference on Machine Learning (ICML'09)*. 889–896.

V. C. Raykar, S. Yu, L. H. Zhao, and G. H. Valadez. 2010. Learning from crowds. *Journal of Machine Learning Research* 11, 1297–1322.

C. Saunders, A. Gammerman, and V. Vovk. 1998. Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning (ICML'98)*. 515–521.

V. S. Sheng, F. Provost, and P. G. Lpeirotis. 2008. Get another label? Improving data quality and data mining using multiple noisy labelers. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD'08)*. 614–622.

H. J. Song, R. X. Liao, X. L. Zhang, C. Y. Miao, and Q. Yang. 2012. A mouse-trajectory based model for predicting query-URL relevance. In *Proceedings of the 26th Conference on Artificial Intelligence (AAAI'12)*. 143–149.

H. J. Song, C. Y. Miao, and Z. Q. Shen. 2011. Generating true relevance labels in Chinese search engine using clickthrough data. In *Proceedings of the 25th Conference on Artificial Intelligence (AAAI'11)*. 1230–1236.

R. Srikant, S. Basu, N. Wang, and D. Pregibon. 2010. User browsing models: Relevance versus examination. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD'10)*. 223–232.

F. Xia, T. Y. Liu, J. Wang, W. Zhang, and H. Li. 2008. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning (ICML'08)*. 1192–1199.

L. Xiao, G. R. Xue, W. Y. Dai, Y. Jiang, Q. Yang, and Y. Yu. 2008. Can Chinese Web pages be classified with English data source? In *Proceedings of the World Wide Web Conference (WWW'08)*. 969–978.

J. Xu, C. Chen, G. Xu, H. Li, and E. Abib. 2010. Improving quality of training data for learning to rank using click-through data. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10)*. 171–180.

H. Yang, A. Mityagin, and K. M. Svore. 2010. Collecting high quality overlapping labels at low cost. In *Proceedings of the Annual International ACM SIGIR Conference (SIGIR'10)*. 459–466.

Y. Zhang, W. Chen, D. Wang, and Q. Yang. 2011. User-click modeling for understanding and predicting search-behavior. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD'11)*. 1388–1396.

Received May 2013; revised September 2015; accepted October 2015