A Mouse-Trajectory Based Model for Predicting Query-URL Relevance

Hengjie Song[1, 3], Ruoxue Liao[1], Xiangliang Zhang[2], Chunyan Miao[3], Qiang Yang[4]

[1] Baidu Inc. [2] King Abdullah University of Science and Technology
[3] Nanyang Technological University [4] Hong Kong University of Science and Technology
HJSONG@ntu.edu.sg, {songhengjie, liaoruoxue }@baidu.com, xiangliang.zhang@kaust.edu.sa,
ascymiao@ntu.edu.sg, qyang@cse.ust.hk

Abstract

For the learning to ranking algorithms used in commercial search engines, a conventional way to generate the training examples is to employ professional annotators to label the relevance of query url pairs. Since label quality depends on the expertise of annotators to a large extent, this process is time consuming and labor intensive. Automatically generating labels from click through data has been well studied to have comparable or better performance than human judges. Click through data present users' action and imply their satisfaction on search results, but exclude the interactions between users and search results beyond the page view level (e.g., eye and mouse movements). This paper proposes a novel approach to comprehensively consider the information underlying mouse trajectory and click through data so as to describe user behaviors more objectively and achieve a better understanding of the user experience. By integrating multi sources data, the proposed approach reveals that the relevance labels of query url pairs are related to positions of urls and users' behavioral features. Based on their correlations, query url pairs can be labeled more accurately and search results are more satisfactory to users. The experiments that are conducted on the most popular Chinese commercial search engine (Baidu) validated the rationality of our research motivation and proved that the proposed approach outperformed the state of the art methods.

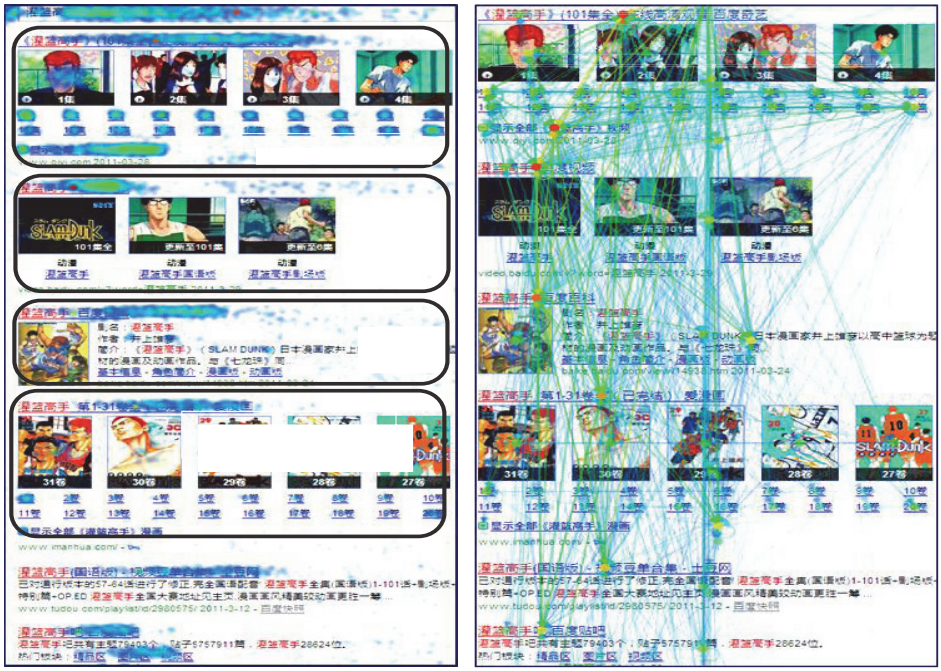
Introduction

Over the years, many learning-to-rank algorithms based on machine learning techniques have been proposed (Cao et al., 2010; Radlinski et al., 2007; Dong et al., 2010; Xia et al., 2008; Liu et al., 2008) to achieve a reasonable ranking of search results in commercial search engines. Those algorithms require a large volume of training data. A conventional way of generating training examples is to employ professional annotators to explicitly judge and label the relevance of query-url pairs. In general, each query-url pair is judged on a 5-level relevance scale from highly relevant to not relevant. The 5 levels and respective numeric values are: Perfect (4), Excellent (3), Good (2), Fair (1), and Bad (0) (Song et al., 2011; He et al., 2011; Wang et al., 2009).

However, labels manually generated have several disadvantages: 1) Due to a lack of quantitative criteria to describe the subtle differences between the label ratings, label quality greatly depends on the annotators' subjective expertise (Xun et al., 2010; Yang et al., 2010); 2) In order to achieve more accurate relevance labels, multiple annotators are always involved in the labeling process. In most cases, these individual opinions statistically differ from the consensus of all annotators involved (Yang et al., 2010; Song et al., 2011). In practice, either the variances among individual judges or the differences between an individual judgment and consensus post challenges to generating reliable relevance labels.

A promising solution to these problems is to generate relevance labels automatically from click-through data (Agrawal et al., 2009; Cao et al., 2010; Yang et al., 2010; Song et al., 2011; He et al., 2011; Zhang et al., 2011). As users' implicit feedbacks, click-through data reflects the judgments of a large number of real world users over a variety of topics (queries). Therefore valuable statistical information could be extracted from aggregating click-through data over weeks, months, and even years to mine users' interests and preferences. By doing so, technicians could get a better understanding of user intentions and behaviors, and therefore generate more satisfactory search results to meet users' information needs (Irmak et al., 2009).

However in some cases, click-through data cannot provide vital information about user behaviors beyond the page-view level (e.g., whether a user has read the context about the url before she/he clicks it) (Granka et al., 2004). For the queries with low search frequencies, it is especially difficult to obtain sufficient click-through data to capture users' behavior pattern. Mouse trajectories can record user browsing behaviors, such as reading by moving mouse horizontally from left to right, hesitating by moving mouse horizontally back and forth. Thus mouse movements on urls ordered by a given list can reflect to what extent users are satisfied with the ranking order of searched urls and the relevance of these urls to their queries.



Against this background, this paper proposes a novel approach to describe the correlations among relevance labels, positions of urls on the list and users' behavioral features (including clicking features and browsing features) so as to generate more accurate relevance labels automatically. Comparing with most existing approaches that mainly aim to model users' clicking behaviors (Agrawal et al., 2009; Cao et al., 2010; Yang et al., 2010; He et al., 2011), our approach, as a natural extension to the method in (Song et al., 2011), has the following properties:

1) *Effectiveness:* To the best of our knowledge, the proposed approach may be the first attempt to integrate click-through data with mouse movement data for automatically generating relevance labels. Due to the scarcity of click-through data, it is hard for most previous work to generate accurate relevance labels for the queries with low search frequency. By integrating click-through data with mouse movement data, our approach makes it possible to capture and utilize user behavioral characteristics more holistically and hence provides more cues for inferring the relevance labels of query-url pairs.
2) *Robustness:* By conducting experiments on real world data that were collected through the most popular Chinese search engine Baidu.com, we are in an advantageous position to demonstrate the rationality of the research motivation. Following the evaluation metrics used in (Xu et al. 2010; Cao et al. 2010; Song et al., 2011), we prove that the proposed approach outperforms the state-of-the-art models, including SDM, FDM (Xu et al. 2010), and the model in (Song et al, 2011).

The rest of this paper is organized as follows. Section 2 uses a query 'SLAM DUNK1' as an example to analyze the differences between click-through data and mouse movement data to justify the necessity to analyze both of them in conjunction. Section 3 constructs a novel model to describe the correlations among relevance labels, the positions of retrieved urls in a browser and users' behavioral features. In Section 4, a case study about the specific query 'Romance of the Three Kingdoms2' is presented to demonstrate the effectiveness of the proposed approach. Moreover, we compare the proposed approach with other models in terms of a number of evaluation metrics. These comparisons show that the proposed approach outperforms the state-of-the-art models. Finally, the conclusion of this paper is given in Section 5.

Motivation

So far, most existing models for automatically generating relevance labels mainly focus on processing click-through data, without taking into account the other aspects of the interactions between users and search results (such as users' visual attention and mouse movements, etc).

To incorporate more user behavioral information into the analysis, eye tracking has been investigated (Granka et al., 2004; Kerry et al., 2008). By measuring users' visual attention as they navigate through the search results, eye tracking quantifies which retrieved search result is read, glanced at, skipped or ignored. However, eye tracking requires additional hardware and software setups that are not common in everyday search engine usage scenarios. These factors limit its applicability to off-line laboratory settings with a relatively small number of users. Therefore, eye tracking data is inadequate to reflect normal users' browsing behavior in a search engine.

In contrast, users' mouse movement data on search result pages can be easily collected with a high accuracy by

[1]This query is about a sports themed manga series (http://en.wikipedia.org/wiki/Slam_Dunk_(manga)). And it is submitted in Baidu with Chinese characters.

[2]This query is about a Chinese historical novel (http://en.wikipedia.org/wiki/Romance_of_the_Three_Kingdoms). And it is submitted in Baidu with Chinese characters.

remote servers on a large scale (Mueller et al., 2001). It has been found that 35% of the people moved their mouse cursor while reading a web page. This suggests that the mouse movement could be considered as an indication of the user's attention. Consequently, the strong correlation between the mouse movement data and the click-through data while browsing the search results can be used to classify user navigation behavior into several categories, such as scrolling, reading, thinking, or interacting with menus (Kerry et al., 2008). Moreover, the combination of the mouse movement data and the click-through data has been applied in the evaluation of website usability (Arroyo et al., 2006; Atterer et al., 2006) and classification of query intents (Guo et al., 2009).

To the best of our knowledge, there are few attempts to integrate click-through data with mouse movement data for automatically generating relevance labels of query-url pairs. In view of this situation, this section takes an example to intuitively show the different types of impact of click-through data and mouse movement data on inferring the relevance labels of the query-url pairs in the first search result page (FSERP). All of the data were recorded through Baidu commercial search engine.

Data preparation

To record mouse movement data, we inserted Javascript code into the Baidu search result pages. This method, as a similar approach in (Kerry et al., 2008), does not require the users to download and install additional software. By doing so, detailed data about mouse and keyboard input can be captured. Such data include the position of the mouse pointer, key pressed, browser window size, etc. The users' mouse movement data were recorded into search log files with an interval of 350ms.

For the query 'SLAM DUNK[1]', Figure 1 and Figure 2 show the click heat map and its corresponding mouse trajectories respectively. In this example, we recorded the behaviors of 4,874 users, and excluded the data from 157 of them due to abnormal behaviors (e.g., the duration between two consecutive actions from a user is longer than 10 minutes). In Figure 1, hue (from green = large to blue = small) indicates the number of mouse clicks on the corresponding positions. In Figure 2, the sequences of mouse movements and clicks are obtained by mapping the mouse/click positions to links, buttons and tags, etc.

Empirical Observations

From Figure 1 and Figure 2, we can see that the starting position of the mouse corresponds to the 'Baidu Search' textbox where users submit the query. After this point, Figure 1 clearly shows that the number of clicks on different urls decays very quickly with the decreasing ranks in the search result list.

However, it seems not easy to associate this significant change of click numbers with the relevance labels of query-url pairs directly. As Figure 1 shows, the url A.4 does not receive more clicks than A.2/A.3, even though the relevance of A.4 is much higher than that of A.2/A.3. This phenomenon is resulted from the position bias and quality bias, which explains the impact of the position and appearance of the urls (e.g., title and abstract, etc) on users' choices (Joachims et al., 2007; Dupret et al., 2008).

In contrast with Figure 1, mouse trajectory map (Figure 2) provides more information about users' behaviors beyond the page-view level. Especially for url A.4, it has i) more horizontal trajectories; ii) trajectories pointing from other URLs to it (if zooming-in Fig.2) (e.g., whether a user has read the contextual information about the url before he/she clicks it). In practice, a typical reading action could be characterized by the mouse moving horizontally from left to right. Mouse moving horizontally back and forth could represent a typical hesitating action. According to these characterizations, Figure 2 clearly shows that the url A.4 attracted more users' attentions than A.2 and A.3 did. Therefore, it seems more likely to infer that the relevance of A.4 is better than that of A.2/A.3, which is the same as the annotated relevance labels in Figure 1.

For this case, the above empirical observations show that mouse movement data and click-through data represent the different aspects of the interactions between users and search results respectively. Accordingly, we believe that integrating mouse movement data with click-through data is a promising way to reflect user behavioral characteristics more holistically and provide more cues to infer the relevance labels of query-url pairs accurately.

Proposed Approach

Based on aforementioned observations, this section attempts to construct a probabilistic model to describe the correlations underlying relevance labels, positions of urls on the ranking list and user behavioral features (including clicking features and browsing features).

Notations

First of all, we briefly introduce the notations and definitions used throughout this paper.

— $\mathbb{C} = \{\text{Bad}, \text{Fair}, \text{Good}, \text{Perfect}, \text{Excellent}\}$ represents the relevance label set of query-url pairs, as described in (Song et al., 2011; He et al., 2011; Wang et al., 2009).

— For a given k^{th} url w.r.t. the i^{th} query, X_k^i, a $1 \times \mathrm{D}$ matrix, represents the *observed* browsing and clicking features associated with url k.

— For the given k^{th} url w.r.t. the i^{th} query, Y_k^i, a $1 \times \mathrm{D}$ matrix, represents the *expected* browsing and clicking features associated with the url k.

— $S_k^{©}$ indicates the set of the urls that appear at the k^{th} position with a same relevance label $© \in \mathbb{C}$. In our approach, $S_k^{©}$ is characterized by $P(S_k^{©})$, $\overline{S_k^{©}}$ and $\Sigma_k^{©}$. $P(S_k^{©})$ is the prior probability that the relevance label of the k^{th} url is

© . Let $|S_k^{©}|$ be the number of urls in set $S_k^{©}$. $\overline{S_k^{©}}$ is the mean of feature vectors of the urls that constitute set $S_k^{©}$, calculated by $\overline{S_k^{©}} = \frac{1}{|S_k^{©}|}\sum^{N}_{\text{the label of } X_k^i \text{ is } ©} X_k^i$, where N denote the number of queries. $\Sigma_k^{©}$ denotes the corresponding feature covariance matrix. Thus, $\overline{S_k^{©}}$ is a $1 \times D$ matrix. $\Sigma_k^{©}$ is a $D \times D$ matrix.

Note that $P(S_k^{©})$, $\overline{S_k^{©}}$ and $\Sigma_k^{©}$ can be easily obtained by aggregating and calculating the features of the query-url pairs with manual annotation results.

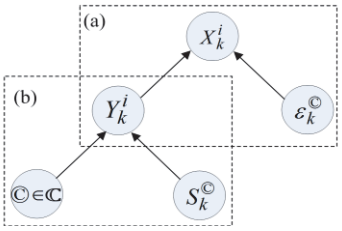


Figure 3. Graphical model to generate relevance labels of query-url pairs. (a) The observed browsing and clicking features X_k^i is modeled by the sum of expected browsing and clicking features Y_k^i and a Gaussian deviation $\varepsilon_k^{©}$ with covariance $\Theta_k{}^{©}$. (b) The expected browsing and clicking features Y_k^i depends on the class label © ∈ ℂ and the set of the urls appearing at the k^{th} position with class label ©

Formulation

Based on the above notations, Figure 3 depicts the graphical model of our approach. For the k^{th} url associated with query i, if its relevance label is © ∈ ℂ, Figure 3. (a) represents the relationship between its expected features Y_k^i and the observed features X_k^i. This relationship can be described by,

$$X_k^i = Y_k^i + \varepsilon_k^{©} \tag{1}$$

where $\varepsilon_k^{©}$ is assumed to be Gaussian deviation with zero mean and a diagonal covariance matrix $\Theta_k^{©}$:
$P(\varepsilon_k^{©}) = \mathcal{N}\left(\varepsilon_k^{©}\middle|0, \Theta_k^{©^{-1}}\right)$; $\Theta_k^{©} = \text{diag}(\theta_{k,1}^{©}, \dots, \theta_{k,D}^{©})$ is the inverse of covariance. In our opinion, $\Theta_k^{©}$ potentially measures the impacts of position bias[3] and quality bias[4] on the k^{th} url with relevance label © ∈ ℂ.

Therefore, Eq. 1 essentially describes the conditional probability of X_k^i given Y_k^i and $\Theta_k^{©}$, which is described as,

$$P(X_k^i|Y_k^i, \theta_k^{©}) = \prod_{d=1}^{D} \mathcal{N}\left(X_{k,d}^i\middle|Y_{k,d}^i,\ \theta_{k,d}^{©\ -1}\right) \tag{2}$$

Since the score distribution for relevant documents can be well approximated by a Gamma distribution (Kanoulas et al., 2010), each $\Theta_k^{©}$ in our approach is assumed to be an independent Gamma distribution, thus we derive,

$$P(\Theta_k^{©}) = \prod_{d=1}^{D} \Gamma(\theta_{k,d}^{©}|a_{k,d}^{©}, b_{k,d}^{©}) \tag{3}$$

where $\Gamma(\theta_{k,d}^{©}|a_{k,d}^{©}, b_{k,d}^{©}) \propto {b_{k,d}^{©}}^{a_{k,d}^{©}} {\theta_{k,d}^{©}}^{a_{k,d}^{©}-1} e^{-b_{k,d}^{©}\theta_{k,d}^{©}}$ denotes a Gamma distribution with hyper-parameters $a_{k,d}^{©}$ and $b_{k,d}^{©}$.

In contrast, Figure 3. (b) describes the distribution of the expected features associated with the url appearing at the k^{th} position. It is considered as a Gaussian mixture model governed by class label © ∈ ℂ and the relevant parameters of url set $S_k^{©}$ ($P(S_k^{©})$, $\overline{S_k^{©}}$ and $\Sigma_k^{©}$). Hence for the given k^{th} url with relevance label ©,

$$P(©, S_k^{©}, Y_k^i) = P(S_k^{©})\mathcal{N}(Y_k^i|\overline{S_k^{©}}, \Sigma_k^{©}) \tag{4}$$

As mentioned in the last Section, $P(S_k^{©})$, $\overline{S_k^{©}}$ and $\Sigma_k^{©}$ in Eq. 4 characterize the urls appearing at the k^{th} position with class label © ∈ ℂ which constitute set $S_k^{©}$. By combining (a) and (b) in Figure 3, the problem of generating relevance labels for query-url pairs is formulated as,

$$P(©|X_k^i) = \sum_{S_k^{©}} \int\int P(©, S_k^{©}, Y_k^i, \Theta_k^{©}|X_k^i) dY_k^i d\Theta_k^{©} \tag{5}$$

Calculation

Since Eq. 5 has no analytic solution, we apply variational methods (Hua et al., 2005; Fergus et al., 2006) to approximate the integrand $P(©, S_k^{©}, Y_k^i, \Theta_k^{©}|X_k^i)$ in Eq. 5 so that we can derive the tractable integration. Variational methods have been widely applied in visual tracking (Hua et al., 2005), factor analysis for modeling correlations in multidimensional data (Ghahramani et al., 2000) and image deblurring (Fergus et al., 2006). By using standard variational methods, the integrand $P(©, S_k^{©}, Y_k^i, \Theta_k^{©}|X_k^i)$ in Eq. 5 can be approximated by a factorized distribution $Q(©, S_k^{©}, Y_k^i, \Theta_k^{©})$,

$$\underbrace{P(©, S_k^{©}, Y_k^i, \Theta_k^{©}|X_k^i)}_{\mathcal{P}} \leftarrow \underbrace{Q(©, S_k^{©}, Y_k^i, \Theta_k^{©})}_{Q}$$
$$= Q(©, S_k^{©}, Y_k^i)Q(\Theta_k^{©}) \tag{6}$$

Accordingly, this approximation objective could be considered to minimize the Kullback-Leibler divergence $KL(\mathcal{P} \parallel Q)$. According to Eq. 6, we can see that the minimization of $KL(\mathcal{P} \parallel Q)$ involves the iterations for estimating $Q(©, S_k^{©}, Y_k^i)$ and $Q(\Theta_k^{©})$ respectively.

Using the standard variational minimization (Ghahramani et al., 2001), $Q(©, S_k^{©}, Y_k^i)$ and $Q(\Theta_k^{©})$ that minimize the $KL(\mathcal{P} \parallel Q)$ have the following forms respectively:

$$Q(©, S_k^{©}, Y_k^i) \propto e^{\int logP\left(©, S_k^{©}, Y_k^i, X_k^i, \Theta_k^{©}\right) Q(\Theta_k^{©}) d\Theta_k^{©}} \tag{7}$$

$$Q(\Theta_k^{©}) \propto e^{\sum_{S_k^{©}} \int logP\left(©, S_k^{©}, Y_k^i, X_k^i, \Theta_k^{©}\right) Q\left(©, S_k^{©}, Y_k^i\right) dY_k^i} \tag{8}$$

In Eq. 7 and Eq. 8, $logP(©, S_k^{©}, Y_k^i, X_k^i, \Theta_k^{©})$ is the log likelihood of the joint probability of all variables that is described as,

$$logP(©, S_k^{©}, Y_k^i, X_k^i, \Theta_k^{©})$$
$$= logP(X_k^i|Y_k^i, \Theta_k^{©}) + logP(Y_k^i|S^{©}) + logP(S_k^{©}) + logP(\Theta_k^{©})$$
$$= -\frac{1}{2}\sum_{d=1}^{D}\left(X_{k,d}^i - Y_{k,d}^i\right)^2 \theta_{k,d}^{©} + \frac{1}{2}\sum_{d=1}^{D}\theta_{k,d}^{©}$$
$$-\frac{1}{2}\left(Y_k^i - \overline{S_k^{©}}\right)\Sigma_k^{©^{-1}}\left(Y_k^i - \overline{S_k^{©}}\right)^T - \frac{1}{2}|\Sigma_k^{©}|$$
$$+ logP(S_k^{©}) + \sum_{d=1}^{D} log\left(\theta_{k,d}^{©}\right)^{a_{k,d}^{©}-1} - \sum_{d=1}^{D} b_{k,d}^{©}\theta_{k,d}^{©} \tag{9}$$

By substituting Eq. 9 into Eq. 7 and Eq. 8 respectively, we can estimate $Q(©, S_k^{©}, Y_k^i)$ and $Q(\Theta_k^{©})$ using the following iterations:

Step 1: Update $Q(©, S_k^{©}, Y_k^i)$ with fixed $Q(\Theta_k^{©})$

In this step, $Q(\text{C}, S_k^{©}, Y_k^i)$ is described as a Gaussian mixture model,

$$Q(©, S_k^{©}, Y_k^i) = \widetilde{P(S_k^{©})}\mathcal{N}\left(Y_k^i\middle|\widetilde{\overline{S_k^{©}}}, \widetilde{\Sigma_k^{©}}\right) \tag{10}$$

In Eq. 10, Y_k^i is calculated by,

$$Y_k^i \leftarrow \frac{1}{\text{T}}\sum^{\text{T}}_{\text{the label of } X_k^i \text{ is } ©} X_k^i \tag{11}$$

Figure 4. Click heat map regarding query 'The Romance of the Three Kingdoms' before experiment

Figure 5. Click heat map regarding query 'The Romance of the Three Kingdoms' after the experiment

where T is the iteration numbers. In addition, $\widetilde{P(S_k^{©})}$, $\widetilde{\overline{S_k^{©}}}$ and $\widetilde{\sum_k^{©}}$ are estimations of the corresponding variables and respectively described by,

$$\widetilde{P(S_k^{©})} \propto P(S_k^{©}) \frac{exp\left(-\frac{1}{2}\left(X_k^i-\widetilde{\overline{S_k^{©}}}\right)\left(\widetilde{\sum_k^{©}}+\langle\Theta_k^{©}\rangle^{-1}\right)^{-1}\left(X_k^i-\widetilde{\overline{S_k^{©}}}\right)^T\right)}{\left|\widetilde{\sum_k^{©}}+\langle\Theta_k^{©}\rangle^{-1}\right|^{1/2}} \quad (12)$$

$$\widetilde{\overline{S_k^{©}}} \leftarrow \left(X_k^i-\widetilde{\overline{S_k^{©}}}\right)\left(\widetilde{\sum_k^{©}}+\langle\Theta_k^{©}\rangle\right)^{-1}+\widetilde{\overline{S_k^{©}}} \quad (13)$$

$$\widetilde{\sum_k^{©}} \leftarrow \langle\Theta_k^{©}\rangle^{-1}\left(I_D-\left(\widetilde{\sum_k^{©}}+\langle\Theta_k^{©}\rangle\right)^{-1}\langle\Theta_k^{©}\rangle^{-1}\right) \quad (14)$$

In our approach $\widetilde{\overline{S_k^{©}}}$ and $\widetilde{\sum_k^{©}}$ are initialized by $\overline{S_k^{©}}$ and $\sum_k^{©}$ respectively.

Step 2: Update $Q(\Theta_k^{©})$ with fixed $Q(\mathbb{C}, S_k^{©}, Y_k^i)$

Given a fixed $Q(©, S_k^{©}, Y_k^i)$, $Q(\Theta_k^{©})$ is described by a Gaussian distribution,

$$Q(\Theta_k^{©}) = \prod_{d=1}^{D} Q(\Theta_{k,d}^{©}) = \prod_{d=1}^{D} \Gamma\left(\theta_{k,d}^{©}\middle|\widetilde{a_{k,d}^{©}}, \widetilde{b_{k,d}^{©}}\right) \quad (15)$$

For simplicity, we let the shape parameter $\widetilde{a_{k,d}^{©}}$ be a constant and update the scale parameter $\widetilde{b_{k,d}^{©}}$ as,

$$\widetilde{a_{k,d}^{©}} = a_{k,d}^{©} \quad (16)$$

$$\widetilde{b_{k,d}^{©}} \leftarrow \frac{1}{2}\langle\left(X_{k,d}^i - Y_{k,d}^i\right)^2\rangle + \widetilde{b_{k,d}^{©}} \quad (17)$$

Experimental Results

In this section, we begin with a specific query 'Romance of the Three Kingdoms[2]' performed in Baidu search engine as an example to validate the rationality of the research motivation and demonstrate the effectiveness of our approach intuitively. Furthermore, we compares our method with the baseline models, including SDM, FDM (Xu et al. 2010) and the model in (Song et al, 2011), which only use click-through data to infer relevance labels of query-url pairs.

Case Study

As mentioned in Section 1, an important application of automatically generating relevance labels is to re-rank the retrieved search result list so as to improve the user experiences. For a given query, the number of mouse clicks on the different urls could be considered as a critical and intuitive indicator to measure users' immediate responses. Therefore, this case study recorded and compared the significant changes between the number of mouse clicks on the original ranking list and that of the re-ranked list. Data used for calculation is the same training data set used in (Song et al., 2011) that will be introduced in next section.

A specific query 'Romance of the Three Kingdoms[2]' is studied as an example because its main search intent (i.e., video search) is similar with that of 'SLAM DUNK[1]', which has been used to explain our motivation. In order to validate the rationality of our research motivation, the first step is to use the proposed approach to calculate the estimated probability $\widetilde{P(S_k^{©})}$. As shown in Table 1, the k[th] row gives the normalized probability that the set of urls appear at the k[th] position having different relevance labels © ∈ ℂ. The probability in bold shows that the url at k[th] position is most possibly labeled by the corresponding ©, i.e., the 1[st] url associated with query 'Romance of the Three Kingdoms' is *Bad*, the 2[nd] url is *Fair*, and the 3[rd] url is *Perfect*.

Note that we only estimate the relevance labels of the top-three urls whose re-ranking will significantly change the search result list. The effectiveness of our approach can be demonstrated by having positive impacts on user experiences through re-ranking the top-three urls. Then in the second step, we adjust the original rank list (Figure 4) according to these estimated probabilities (Table 1).

Figure 4 and Figure 5 show the click heat maps of the original rank list and the re-ranked list respectively. Comparing the 1[st] url in the Figure 4 with the 1[st] url in Figure 5, we can clearly see that the latter gets much more clicking amounts. That is to say, the re-ranked 1[st] url can more accurately satisfy users' query. In addition, the re-ranked 1[st] url in Figure 5 does not affect other urls too

much. This means that the re-ranked list still keeps the capability to satisfy diversified needs on query 'Romance of the Three Kingdoms' (for example, the introduction or discussion about 'Romance of the Three Kingdoms' is still present in the re-ranked list).

Table 1: The probabilities of the urls appearing at the 1st, 2nd and 3rd position with different relevance labels © ∈ ℂ

	Relevance label © ∈ ℂ ℂ = {Bad, Fair, Good, Perfect, Excellent}				
	© is Bad	© is Fair	© is Good	© is Perfect	© is Excellent
$P(\widehat{S_1^{©}})$	**0.3619**	0.2627	0.3411	0.0128	0.0213
$P(\widehat{S_2^{©}})$	0.2433	**0.4434**	0.2590	0.0351	0.0188
$P(\widehat{S_3^{©}})$	0.1015	0.1026	0.2417	**0.3146**	0.2403

The above evaluation results show that the proposed approach accurately estimates the relevance labels of the top-three urls associated with query 'Romance of the Three Kingdoms'. Based on these estimated results, the re-ranked list improves user experiences significantly.

Comparisons of Generated Relevance Labels

In this section, we compare the proposed approach with the state-of-the-art models, including SDM, FDM (Xu et al. 2010) and the model in (Song et al, 2011), which only use click-through data to infer the relevance labels of query-url pairs. In order to achieve the fairness of experimental comparisons, we adopt the same data set used in (Song et al., 2011), and follow the same experimental procedures and evaluation metrics (*accuracy*, *consistency* and *correlation* that were used by Xu et al. 2010, Cao et al. 2010 and Song et al, 2011).

The data set has been collected through Baidu search engine, which consists of 4,723 unique queries and corresponding urls[3]. The relevance labels of these query-url pairs are annotated by 3 well-trained editors. This data set is divided into a training set and a test set that have similar distributions regarding search frequency, as shown in Figure 6 and Figure 7 (the x-axis and the y-axis indicate the number of queries and the search frequency respectively).

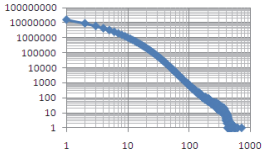


Figure 6. Distribution of training set regarding to search frequency of queries

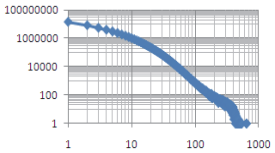


Figure 7. Distribution of test set regarding to search frequency of queries

For the queries with different search frequencies, Table 2 compares the consistencies and the accuracies which are generated by three annotators, the model in (Song et al., 2011) and the proposed approach respectively[4]. Table 2 shows that the proposed approach outperforms the model in (Song et al., 2001) over all queries with diverse frequencies. The consistency and accuracy have been improved by 1%-5%. Comparing with the human judges, the proposed approach has significant improvement on consistency (8%-26%) except when queries are extremely rare (search freq.=1) or extremely frequent (search freq. >32769), as well as on accuracy (1%-20%) except on frequent queries on the last row of Table 2.

Table 2: Comparisons of Consistency and Accuracy

Model† represent the model in (Song et al., 2011); #P represents the proposed approach

Search Frequency	Consistency to the consensus			Accuracy		
	Manual labels	Model†	#P	Manual labels	Model†	#P
1	72.5%	61.2%	***64.3%***	77.2%	65.6%	***67.4%***
2~8	63.2%	68.3%	***71.5%***	71.6%	68.3%	***71.7%***
9~64	45.7%	52.7%	***58.6%***	57.2%	61.4%	***66.1%***
65~512	49.8%	57.4%	***62.1%***	63.3%	66.2%	***71.4%***
513~4096	40.4%	61.3%	***66.2%***	58.7%	73.7%	***78.8%***
4097~32768	53.6%	63.8%	***67.3%***	58.3%	64.5%	***66.7%***
≥32769	77.2%	75.1%	***75.5%***	78.4%	80.3%	***81.1%***

Especially, the performance improvements on the queries with medium and low search frequency (9 times/day ≤ search freq. ≤ 4,096 times/day) are more significant than the improvements on the top queries (search freq. ≥ 4,097 times/day) and tail queries (1 times/day ≤ search freq. ≤ 8 times/day). The frequency-related performance is mainly due to the fact that navigational queries (the intent of the search query is to find a particular website or webpage) are in the majority of the top queries. In this situation, most users will directly click the targeted urls. For the tail queries, it is hard to measure users' immediate responses effectively with a small amount of click-through data and mouse movement data. In general, this experimental result proves that the mouse movement data can be considered as an important complement to click-through data so as to describe the diverse users' behaviors more holistically, and infer the relevance labels more accurately.

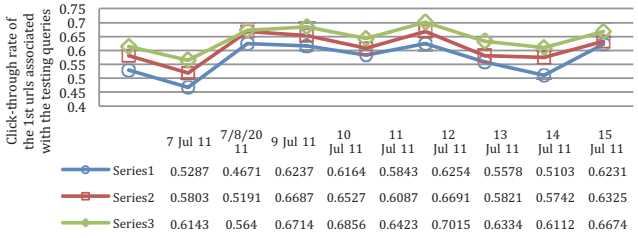


Figure 8. Comparisons of the mean CTRS of the 1st urls

Table 3: Comparisons of Correlation

SDM and FDM are proposed in (Xu et al. 2010);
Model† is proposed in (Song et al. 2011); #P represents the proposed approac

	SDM	FDM	Model†	#P
Correlation	0.69	0.75	0.78±0.023	0.83±0.017

Moreover, according to the estimated relevance labels, we re-rank the search result lists with respect to the testing

[3] For the detailed information about the data set, experimental procedures and evaluation metrics please refer to (Song et al., 2011).

[4] Note that the consistencies and accuracies generated by annotators (manual labels) and Model† in Table 2 slightly differ from the results in (Song et al., 2011). It is mainly due to the fact that this experiment only considers 10 urls per query (in the FSERP) while the previous experiments (Song et al., 2011) studied 21.12 urls per query on average.

queries. Figure 8 compares the mean click-through rates (CTRs) of the 1st urls of the original search result lists (Series 1), the lists re-ranked by the Model† (Series 2), and the lists re-ranked by the proposed approach (Series 3) respectively. Following the evaluation metric in (Song et al., 2011; Xu et al. 2010), Table 3 compares SDM, FDM, Model† and the proposed approach in terms of correlation. Statistically, the closer the correlation is to 1, the stronger the relationship between the estimated labels and the consensus among the annotators. These comparisons show that the proposed approach improves CTRs of the 1st urls and the correlation associated with the testing queries. In another word, the proposed approach outperforms the baseline model SDM, FDM and Model†.

Conclusion

To more accurately generate relevance labels of query-url pairs, this paper proposed a novel approach to reveal the correlations underlying manual annotation results, positions of urls and user behavioral features by integrating click-through data with mouse movement data. This approach is inspired by the intuitive observations that the combination of mouse movement data and click-through data is able to reflect user behavioral characteristics more holistically and potentially provide more cues to infer the relevance labels. The experiments on real world data have shown that the proposed approach outperforms the state-of-the-art models in terms of accuracy, consistency and correlation of the relevance labels, especially for the queries with medium or low search frequencies.

References

Agrawal, R., Kenthapadi, K., Mishra, H., and Tsaparas, P. 2009. Generating labels from clicks. In *Proc. of WSDM'09,* 172 181. Barcelona, Spain: ACM.

Arroyo, E., Selker, T. and Wei, W. 2006. Usability tool for analysis of web designs using mouse tracks. In *Proc. of Ext. Abstracts CHI'06*, 484 489, 2006.

Atterer, R., Wnuk, M., and Schmidt, A. 2006. Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction. In *Proc. of WWW'06*, 203 212, 2006.

Cao, B., Shen, D., Wang, K. S, and Yang. Q. 2010. Clickthrough log analysis by collaborative ranking. *In Proc. of AAAI'10*, 224 229. Atlanta, Georgia: AAAI Press.

Dong, A., Chang, Y., Zheng, Z., Mishne, G., Bai, J., Zhang, R., Buchner, K., Liao, C. and Diaz, F. 2010. Towards recency ranking in web search. In *Proc. of WSDM '10*, 11 20.

Dupret, G. and Piwowarski, B. 2008. A user browsing model to predict search engine click data from past observations. In *Proc. of SIGIR'08*, 331 338, 2008.

Fergus, R., Singh, B., Hertzmann, A. 2006. Roweis, S. T. and Freeman, W. T. 2006. Removing camera shake from a single photograph. In *Proc. SIGGRAPH'06*, 787 794.

Guo, Q. and Agichtei, E. 2008. Exploring mouse movements for inferring query intent. In *Proc. of SIGIR'08*, 707 708, 2008.

Granka, L., Joachims, T. and Gay, G. 2004. Eye Tracking analysis of user behavior in WWW Search. In *Proc. SIGIR'04*, 478, 2004.

Ghahramani, Z. and Beal, M. J. 2000. Variational inference for Bayesian mixtures of factors analysis. In *Proc. NIPS'00*, 449 455.

Ghahramani, Z. and Beal, M. J. 2001. Propagation Algorithms for Variational Bayesian Learning. In *Proc. NIPS'01*, 507 513.

He, J., Zhao, W. X., Shu, B., Li, X. M., Yan, H. F. 2011. Efficiently Collecting Relevance Information from Clickthroughs for Web Retrieval System Evaluation. In *Proc. of SIGIR'11*.

Hua, G. and Wu, Y. 2005. Variational maximum a posteriori by annealed mean field analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(11):1 15.

Irmak, U., Brzeski, V. V. and Kraft, R. 2009. Contextual ranking of keywords using click data. In *Proc. of ICDE'09*, 457 468.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F. and Gay, G. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2), April 2007.

Kerry, R., Fu, X., Anne, A. and Ian, S. 2008. Eye Mouse coordination patterns on web search results pages. In *Proc. of CHI'08*, 2997 3002, 2008.

Kanoulas, E., Dai, K., Pavlu, V., Aslam, J. A., 2010. Score Distribution Models: Assumptions, Intuition, and Robustness to Score Manipulation. In *Proc. of SIGIR'10*, 242 249, 2010.

Liu, N. N., Yang, Q., 2008. EigenRank: a ranking oriented approach to collaborative filtering. In *Proc. of WWW'08,* 83 90, 2008.

Mueller, F. and Lockerd, A. 2001. Cheese: tracking mouse movement activity on websites, a tool for user modeling. In *Proc. of CHI'01*, 2001.

Radlinski, F., and Joachims, T. 2007. Active exploration for learning rankings from clickthrough data. In *Proc. of KDD '07*, 570 579.

Song, H. J., Miao, C. Y. and Shen, Z. Q. 2011. Generating True Relevance Labels in Chinese Search Engine Using Clickthrough Data. In *Proc. of AAAI'11*. San Francisco, California: AAAI Press.

Wang, K., Walker, T., and Zheng, Z. 2009. Pskip: estimating relevance ranking quality from web search clickthrough data. In *Proc. of KDD '09*, 1355 1364.

Xu, J., Chen, C., Xu, G., Li, H., Abib, E. 2010. Improving quality of training data for learning to rank using click through data. In *Proc. of WSDM'10*. 171 180, 2010.

Xia, F., Liu, T.Y., Wang, J., Zhang, W. and Li, H. 2008. Listwise approach to learning to rank: theory and algorithm. In *Proc. of the ICML'08*, 1192 1199.

Xun, J. F., Chen, C. L., Xu, G., Li, H., and Elbio, A. 2010. Improving quality of training data for learning to rank using click through data. *In Proc. of WSDM'10*, 171 180. New York, US: ACM.

Yang, H., Mityagin, A., Svore, K. M. Collecting high quality overlapping labels at low cost. In *Proc. of SIGIR'10*, 459 466. Geneva, Switzerland: ACM.

Zhang, Y. C., Chen, W. Z., Wang, D., Yang, Q., 2011. User Click Modeling for Understanding anf Predicting Search Behavior. In *Proc. of KDD'11*, 1388 1396, 2011.