Research Article Open Access

Yukun Zheng, Yiqun Liu*, Zhen Fan, Cheng Luo, Qingyao Ai, Min Zhang, Shaoping Ma

Investigating Weak Supervision in Deep Ranking

https://doi.org/10.2478/dim-2019-0010
received May 15, 2019; accepted July 18, 2019.

Abstract: A number of deep neural networks have been proposed to improve the performance of document ranking in information retrieval studies. However, the training processes of these models usually need a large scale of labeled data, leading to data shortage becoming a major hindrance to the improvement of neural ranking models' performances. Recently, several weakly supervised methods have been proposed to address this challenge with the help of heuristics or users' interaction in the Search Engine Result Pages (SERPs) to generate weak relevance labels. In this work, we adopt two kinds of weakly supervised relevance, BM25-based relevance and click model-based relevance, and make a deep investigation into their differences in the training of neural ranking models. Experimental results show that BM25-based relevance helps models capture more exact matching signals, while click model-based relevance enhances the rankings of documents that may be preferred by users. We further proposed a cascade ranking framework to combine the two weakly supervised relevance, which significantly promotes the ranking performance of neural ranking models and outperforms the best result in the last NTCIR-13 We Want Web (WWW) task. This work reveals the potential of constructing better document retrieval systems based on multiple kinds of weak relevance signals.

Keywords: document ranking, ad hoc retrieval, neural ranking model, weak supervision.

***Corresponding author: Yiqun Liu,** Department of Computer Science and Technology, Institute for Artificial Intelligence, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China, E-mail: yiqunliu@tsinghua.edu.cn
Yukun Zheng, Zhen Fan, Cheng Luo, Min Zhang, Shaoping Ma, Department of Computer Science and Technology, Institute for Artificial Intelligence, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
Qingyao Ai, College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst MA 01003, United States

1 Introduction

Document ranking is one of the core problems in information retrieval studies. Given a textual query, the goal of document ranking is to find relevant documents with respect to the query in the whole collection. Recently, researchers in the Information Retrieval(IR) community have proposed a number of neural ranking models to improve the performance of document ranking. However, the success of deep neural networks has not been widely observed in ad hoc retrieval (Pang, Lan, Guo, Xu, & Cheng, 2017a). One of the reasons lies in the shortage of labeled training data (Dehghani, Zamani, Severyn, Kamps, & Croft, 2017; MacAvaney, Hui & Yates, 2017). The large number of parameters used in neural ranking models not only lead to a better performance but also make the models extremely thirsty for data. However, collecting human-assessed relevance labels for query–document pairs costs both time and money. Therefore, recent researches turn to investigate the effectiveness of other weak but cheaper supervision signals for the training of neural retrieval models.

To generate the relevance labels of training pairs, previous studies have proposed to use several weak ranking signals such as BM25 (Dehghani et al., 2017; MacAvaney et al., 2017) and user behavior signals (Xiong, Dai, Callan, Liu, & Power, 2017; Zheng et al., 2018). In general, this kind of labels is not completely accurate and usually contains a lot of noises, so we called it weak label. Existing weak labels used in ranking models can be classified into two categories:

Heuristics-based label. Several heuristic approaches are widely applied and successful in document ranking, such as BM25 and language model. BM25 is the most common heuristic for generating weak relevance labels (Dehghani et al., 2017; MacAvaney et al., 2017). Dehghani et al. (2017) used BM25 as the heuristic to generate weak labels and reported that their fine-tuned neural models outperformed BM25. By using documents' titles as pseudo queries and BM25 scores as weak labels, MacAvaney et al. (2017) introduced a filtering method to effectively produce positive and negative query–document pairs. However, the implicit assumption that the exact matching

signals can represent relevance usually brings limitations in practical applications, because it ignores semantic matching and user factors.

Behavior-based label. User behavior, such as user clicks and mouse movement, can provide weak supervision signals for document ranking. Practical search engines usually treat user clicks as implicit relevance feedback. Compared to human-assessed relevance labels, click-through data are much easier to obtain. Meanwhile, different from heuristics like BM25, it contains abundant user preference information and implies the intent of users in search tasks. Nevertheless, there are still limitations in adopting click-through data for the training of ranking models, which cannot be ignored. User clicks are affected not only by the results' relevance to the issued queries but also by other factors, such as the documents' positions (Joachims, Granka, Pan, Hembrooke, & Gay, 2005), novelty (Zhang, Chen, Wang, & Yang, 2011), and presentation styles (Wang et al., 2013). Thus, the click-through data are strongly biased and noisy.

Zheng et al. (2018) showed that after being debiased by the click model, the click-through data can serve as a better source to help improve the performances of neural ranking models than to adopt the click-through information directly. Click models were proposed to derive feedback information of relevance from user clicks (Chuklin, Markov & Rijke, 2015). Fed with massive click-through data, click models can estimate the relevance of known query–document pairs by reducing the impacts of click sparseness, position bias, etc., which is called click model-based relevance.

In the IR community, there is no existing work to directly compare the BM25-based relevance and click model-based relevance. Existing works studying the click model-based relevance only show its effectiveness on the test data labeled with the same kind of relevance rather than on the human-assessed data. In addition, there is no existing work to leverage both BM25-based relevance and click model-based relevance in document ranking. Therefore, in this paper, we systematically investigated the difference between click model-based relevance and BM25-based relevance in training neural ranking models. We trained three proposed neural ranking models in the pairwise mode on two datasets. The first dataset consisted of billions of examples annotated by BM25 from SogouT-16 (Luo et al., 2017b) and the other one is Sogou-QCL (Zheng et al., 2018), which is annotated by five click models. We adopted partially sequential click model (PSCM), one of the best click models so far, and applied a three-step cascade ranking framework to combine the weakly supervised relevance from BM25 and click model, which achieved the state-of-the-art ranking performance on a standard test set (Luo et al., 2017a). Here, we list two main contributions of this paper:

- By conducting extensive experiments, we compared click model-based relevance with BM25-based relevance and showed their different impacts on the training of neural ranking models.
- We proposed a cascade ranking framework to effectively combine weak relevance labels generated by click model and BM25, which significantly improved neural rankers' effectiveness.

2 Related Work

2.1 Click Model

Click-through behaviors provide implicit feedback of click preferences from users (Agichtein, Brill, Dumais, & Ragno, 2006). Joachims et al. (2005) found that the click-through information is "informative yet biased". As a probabilistic model, most click models follow the *examination hypothesis* (Craswell, Zoeter, Taylor, & Ramsey, 2008): a search result at the i th position will be clicked ($C_i = 1$) only if it is relevant to the query ($R_i = 1$) and examined ($E_i = 1$), i.e.,

$$C_i = 1 \Leftrightarrow R_i = 1 \wedge E_i = 1$$

By inferring the relevance P ($R_i = 1$) and the examination probability P ($E_i = 1$), click models can estimate the click probability based on the noisy and biased search logs. Different click models are built following different assumptions on how users browse and interact with SERPs and hence have different estimations of P ($E_i = 1$). While the cascade model assumes that users are always satisfied with a single click (Craswell et al., 2008), the dynamic Bayesian network (DBN) model introduces a separate variable to model whether the user will be satisfied after a click (Chapelle & Zhang, 2009). The user browsing model (UBM) allows users to skip some of the results (Dupret & Piwowarski, 2008). Furthermore, Wang et al. (2015) looked into the revisiting behaviors of users in SERPs and incorporated non-sequential behaviors into the PSCM. Liu et al. (2017) proposed the time-aware click model (TACM), which can better capture the temporal information.

2.2 Document Ranking

A lot of learning-to-rank approaches have been proposed to address document ranking problem, such as RankNet (Burges et al., 2005), RankBoost (Freund et al., 2003), and LambdaMART (Wu, Burges, Svore, & Gao, 2010). All these learning-to-rank algorithms usually need to be trained on effective hand-crafted features in the learning process.

The IR community has applied deep learning methods to advance state-of-the-art retrieval technologies. Guo, Fan, Ai and Croft (2016) suggested that most of recent neural ranking models can be generally classified into two categories according to the network architectures. 1) *Representation-focused model* – Models in this category first learn vector representations for textual queries and candidate documents separately with deep neural networks. Then, the relevance is calculated by measuring the similarities between the two representations. This line of research includes DSSM (Huang et al., 2013), C-DSSM (Shen, He, Gao, Deng, & Mesnil, 2014), and ARC-I (Hu, Lu, Li, & Chen, 2014). 2) *Interaction-focused model* – ARC-II (Hu et al., 2014), DRMM (Guo et al., 2016), MatchPyramid (Pang et al., 2016), and K-NRM (Xiong et al., 2017) belong to this category. The term-level interactions between queries and candidate documents are calculated first in these models. Then, the neural networks learn query–document matching patterns from these interactions. Mitra, Diaz, and Craswell (2017) proposed to take advantages of both architectures in Duet. Fan et al. (2017) integrated these models into the MatchZoo, which is an open-source toolkit for text matching.

2.3 Weakly Supervised Learning

With the development of deep neural networks, data have brought breakthroughs in a lot of machine-learning areas. With the development of deep neural networks, exponential growth of data quantity has brought breakthroughs in a lot of machine-learning areas. However, data are also the bottleneck in many cases where high-quality data are not available yet. Therefore, many works have researched into weakly supervised learning. For example, Frénay and Verleysen (2014) studied on learning from weak or noisy labels in classification tasks. Lee (2013) proposed a simple and efficient method of semi-supervised learning, training networks on labeled and unlabeled data simultaneously. In the IR community, Yin et al. (2016) took benefit of both click-through information and embedding similarities of query–document pairs for weakly supervised training in Yahoo search. Dehghani et al. (2017) chose BM25 as the heuristic to generate weak labels and reported that their fine-tuned neural models outperformed BM25. By using documents' titles as pseudo queries and BM25 scores as weak labels, MacAvaney et al. (2017) introduced a filtering method to effectively produce positive and negative query–document pairs. Compared to the previous works, we mainly study on click as an alternative weak supervision signal for document ranking.

3 Weak Supervision

3.1 BM25 Relevance

BM25 is a popular bag-of-words ranking function. By counting the term frequency (TF) and inverse document frequency (IDF) of query terms appearing in candidate documents, BM25 gives the ranking scores of these documents with respect to the query. Thus, BM25 only considers the exact matching signal from query–document pairs, regardless of the semantic relationship between the query and documents.

BM25-based relevance has the following advantages in serving as a weak supervised signal:

- In the aspect of effectiveness, BM25 is a classic ranking algorithm with proven effectiveness in document ranking.
- As for efficiency, BM25 can serve as a highly efficient approach to generate weak relevance labels in large quantities and in parallel.

3.2 Click Relevance

With a large scale of search logs collected by the search engine, various click models can be utilized to estimate the relevance of documents, i.e., $\mathrm{P}(R_i = 1)$, which is also regarded as click relevance or click model-based relevance in previous works (Dupret & Liao, 2010; Zhang et al., 2011). In this study, we propose to use it as weakly supervised relevance to train neural ranking models, hence also called the click label or click relevance. Equally, we call the BM25-based relevance the BM25 label or BM25 relevance for short in the rest of the paper.

Intuitively, without the biases of position, novelty, and attention, the more relevant a document is to the query, the more likely a user will click on it. As a weak supervision signal, click labels have the following advantages:

- Click labels contain abundant information of user preferences that heuristic methods cannot provide.

– Click labels can be easily extracted from the search logs of real search engine traffic.
– Click labels are calculated by click models based on a large number of user behaviors and do not contain any sensitive user identification information.

4 Ranking

We chose several recent neural ranking models in our experiment, i.e., ARC-I, Duet, and K-NRM, using the implementations from MacthZoo.[1]

ARC-I is a kind of representation-focused model, which extracts the vector representation of the query and document based on Convolutional Neural Network (CNN).

Duet contains two CNN-based sub-models, one of which is interaction focused for exact matching and the other one is representation focused for semantic matching.

K-NRM, an interaction-focused model, uses kernels to extract multilevel soft matching signals of query–document pairs.

Here, we will chiefly state our modifications on model implementation, including the loss function and text representation.

Loss function. We followed Ai, Bi, Guo, and Croft (2018) and applied softmax label in cross entropy loss, which is called attention-based cross entropy loss. In the pairwise documents ranking setup, the input data instance is a series of $(q, d+, d-)$ where given the query q, the document $d+$ is ranked higher than $d-$ according to their labels. The softmax score $s'_{q,d\pm}$ is defined as

$$s'_{q,d\pm} = \frac{\exp\left(s_{q,d\pm}\right)}{\exp\left(s_{q,d+}\right) + \exp\left(s_{q,d-}\right)}$$

where $s_{q,d}$ is the relevance score of (q, d) predicted by the ranker. As the pairwise label of (d+, d-) is (1,0), the loss function is given by

$$l'_{q,d\pm} = \frac{\exp\left(l_{q,d\pm}\right)}{\exp\left(l_{q,d+}\right) + \exp\left(l_{q,d-}\right)}$$

$$\mathrm{L}\left(q, d+, d-\right) = \sum_{d \in \{d+,d-\}} l'_{q,d}\left(\log\left(l'_{q,d}\right) - \log\left(s'_{q,d}\right)\right)$$

where $l_{q,d}$ is the ground-truth label of $(q,\ d)$.

Text representation. In several ranking models (e.g., DSSM, C-DSMM and Duet, etc.), textual terms in query–document pairs are expressed by n-gram (Brown, Desouza, Mercer, Pietra, & Lai, 1992), which is not applicable for Chinese data. Instead of this, Pang et al. (2017b) suggested using word embedding in Chinese IR tasks. Thus, we changed the text representation of the distributed submodule in Duet from n-gram to word embedding, which densifies the representation matrixes in the network and reduces the calculation overhead.

5 Experiment

5.1 Dataset

Table 1
The Statistics of Cick Dataset

No. of queries	No. of documents	No. of query–document pairs	Language
537,366	5,480,860	7,736,480	Chinese

We collected three kinds of labeled data for training in our experiments: *click*, *BM25* and *human-assessed* data. The human-assessed data are labeled on a five-point scale with 2,100 distinct queries and 200,682 unique documents.

Click data. We adopted Sogou-QCL dataset (Zheng et al., 2018) in the experiment to serve as training data, which is a public dataset with multiple weak relevance labels annotated by click models, which are trained on real-world search logs sampled from a commercial search engine in China. Table 1 shows the statistics of Sogou-QCL dataset. As reported by Zheng et al. (2018), TACM and PSCM are the best two click models for predicting click probabilities of documents.

BM25 data. With the same query set as click data, we collected the first 200 documents for each query retrieved by SogouT-16 online search system (Luo et al., 2017b), a Solr[2] retrieval system using BM25 with default parameters (i.e., k1=1.2and b=0.75).

Evaluation data. We used the test data released in NTCIR-13 WWW task (Luo et al., 2017a), which is the most recent ad hoc search benchmark in NTCIR. This dataset contains 100 distinct queries and the first 1000 documents for each query, which are retrieved by BM25 in SogouT-16 corpus. We kept the first 100 documents according to their BM25 scores per query as the test set

1 https://github.com/faneshion/MatchZoo

2 http://lucene.apache.org/solr/

in our following experiment, referred to as *Test-NTCIR*. All query–document pairs in Test-NTCIR had been rated by human assessors on a four-point scale following the standard TREC criterion. We made sure that the queries for training and validation do not appear in this test set.

5.2 Baselines

We used BM25 and three types of global learning-to-rank models as our baselines: RankNet (Burges et al., 2005), RankBoost (Freund, Iyer, Schapire, & Singer, 2003), and LambdaMART (Wu et al., 2010). In this paper, we used the implementations of these models from RankLib.[3]

RankNet is a well-known ranking model using a neural network trained with pairwise losses.

RankBoost learns preferences based on the boosting approach.

LambdaMART is the state-of-the-art learning-to-rank algorithm trained with listwise losses.

5.3 Experimental Setup

Data preprocessing. We extracted the full-text of documents from HTML pages. As all the queries and documents were in Chinese, we segmented them using Jieba,[4] a popular word segmentation toolkit. Then, the data were treated as word sequences as same as English. For models with embedding vector representation, we trained the word embedding using word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) on a public Chinese web corpus, SogouT-16 (Luo et al., 2017b). For learning-to-rank baseline models, we extracted the same 46-dimensional hand-crafted features as those in LETOR 4.0 dataset (Qin & Liu, 2013) from the human-assessed data, including TF, IDF, and scores of BM25 and LMIR.

Model settings. We implemented all the ranking models using TensorFlow (Abadi et al., 2016). The parameters in models were optimized by using Adam optimizer (Kingma & Ba, 2014) to compute gradients in the backpropagation. We splitted click data and BM25 data into 200 queries for validation and others for training. The human-assessed data were split into 100 queries for validation and 2,000 documents for training. We tuned all hyperparameters of models on their validation sets. We kept the first 10 and 1,000 terms in queries and documents, respectively, in all models. The embedding size was 50, and the first one million most frequent terms in our corpus were kept, while others were replaced by an identical word, UNK, which was common in lots of Natural Language Processing(NLP) and IR tasks (Guo et al., 2016; Xiong et al., 2017). The initial learning rate was set to 0.001. To prevent overfitting, we used dropout in all models with 0.5 as the dropout rate.

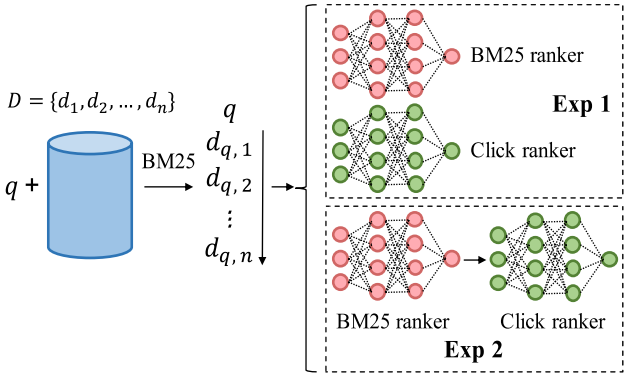

Figure 1. The experimental framework. The BM25/click ranker represents the neural ranking model trained on the data using BM25/click labels.

Evaluation. We evaluated neural rankers by nDCG, Q-measure, and nERR. We found that the results of all three metrics present similar findings. Owing to page limit, we only report the results of nDCG in this paper. The student's t-test was used to examine the differences in model performance.

5.4 Results and Analysis

In this section, we seek to answer the following research questions:

RQ1: What is the difference between the click label and the BM25 label as a weak supervision signal?

RQ2: Is click label effective in training neural ranking models when evaluated on human-assessed test data?

RQ3: How can we combine click label and BM25 label to contribute jointly to model promotion?

We designed our experimental framework as shown in Figure 1. For a query q in the test set, the candidate documents in the D set will be ranked by BM25 at the first step. Then, the top-ranked documents will be judged by neural rankers. Specifically, to answer RQ1 and RQ2, we conducted a case study and trained all the ranking models on click labels and BM25 labels, respectively, which is *Exp 1* in Figure 1. In *Exp 2*, we employed the ranker trained on click labels (click ranker) and after that trained on BM25

3 https://sourceforge.net/p/lemur/wiki/RankLib/
4 https://github.com/fxsjy/jieba

Table 2

Examples from Training Data Annotated with BM25 Labels and Click Labels. (The Red Terms Represent the Exact Matching Parts in the Document with Respect to the Query Terms. We Normalize the Values of BM25 Labels and Click Labels into the Range of [0, 1] within the Query.)

Case	A	B	C
Query	What is the medicine for esophagitis?	Nikon D800 camera setup tutorial	The difference between battleship and cruiser
Document	Hello, I suggest you take a drug for treatment. Omeprazole 1 capsule twice a day (taken on an empty stomach); Lizhu Dele 2 packets per day 2 times a day (taken on an empty stomach); Amoxicillin 2 times a day 2 times a day; Clarithromycin 1 capsule twice a day. Also pay attention to several aspects: 1, life is regular, optimistic, quit smoking and avoid alcohol, do not overeating or hunger and unequal; 2, eat less meals, avoid foods that are difficult to digest and irritate, such as coffee, spicy things; 3, have stomach swell, pantothenic acid, suffocation, should use morphine or metoclopramide, take half an hour before meals; 4, people with stomach pain, can use pain relievers or other stomach drugs with analgesic effect; 5, a very small number of patients with chronic atrophic gastritis have malignant gastric cancer, so a gastroscopy review is required every year. I wish you a speedy recovery.	Nikon D800 D810 Photography Tutorial | Getting Started Tutorial | Usage Tutorial Related tutorials recommended: Nikon D800 detailed setup tutorial (a total of 21 lessons) (VIP) 2016.9.10 update Nikon D800 D800E Photography Tutorial Using the tutorial (VIP) 2015.7.25 update Nikon D810 detailed setup tutorial | Usage tutorial (9 lessons) (VIP) 2015.3.25 update Nikon D810 digital SLR operation tutorial (20 lessons in total) (free video sharing) 2016.2.14 update Nikon D810 detailed setup tutorial (24 lessons in total) (VIP) 2016.2.17 update	Comparing the armor and artillery caliber, the battleship is clearly dominant. The speed of the cruiser is slightly higher than that of the battleship. The size of the artillery is the smallest. The armor is roughly the same as the main gun, so it is also the thinnest, but the range may exceed the battleship (for example, the Hipper Class in Germany hit the British battle Hood on 29km). In battles where no battleships and aircraft carriers fought, cruisers often replaced battleships for artillery and other missions, and their number of artillery was sometimes comparable to that of a battleship. Most of the battleship's tonnage is much higher than the cruiser, but there are still some "perverts" in the cruiser, reaching 20,000 tons of German-class German-class armored cruisers. The speed of the cruiser is never much higher than that of the battleship, because most of the time they need to follow the battleship.
BM 25 label	0.21	0.94	0.98
Click label	0.95	0.47	0.01

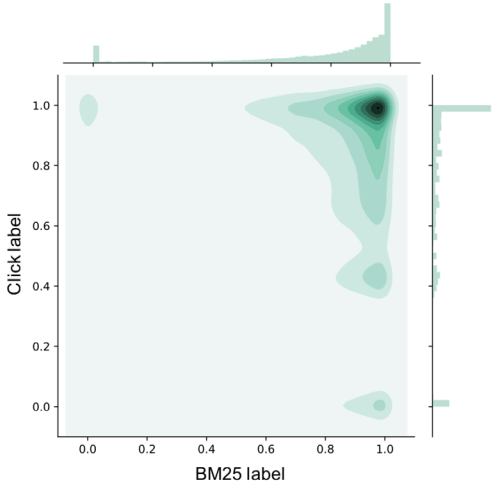

Figure 2. The kernel density distribution of BM25 labels and click labels. We normalized BM25 labels and click labels into the range of [0, 1] within the query. The darker the color, the more the documents with corresponding values of BM25 relevance and click relevance.

labels (BM25 ranker) in a cascade ranking framework to investigate RQ3.

What is the difference between click label and BM25 label as a weak supervision signal? To answer this research question, we ranked all documents of Sogou-QCL according to their click labels and kept the first five documents per query. Then, we looked into the relationship of distributions between BM25 label and click label, which is shown in Figure 2. BM25 label and click label are positively correlated in general. The closer the labels are to 1, the more the documents with corresponding BM25 labels and click labels. When comparing the histograms beside the kernel density map, we found that the distribution of documents' click labels is more concentrated than that of their BM25 labels. From the kernel density distribution, we can see that among the documents with higher click (or BM25) labels, the range of their BM25 (or click) labels is more dispersed. Meanwhile, there exist a number of documents with the highest click (or BM25) labels and the lowest BM25 (or click) labels. All of these findings indicate the big difference between click model and BM25 in judging highly relevant documents.

For a clearer understanding of the difference between click label and BM25 label, we conducted a case study as given in Table 2. Case A shows an example with a low BM25 label but a high click label. The document has no exact matching with important terms in the query, such as "esophagitis" and "medicine", so its BM25 score is rather low. It, however, provides the right answer to the query, which leads to a high probability to be clicked by search engine users. In case B, the document is annotated with a high BM25 label and a low click label. Although the document has a lot of exact matching signals with respect to the query terms, it cannot satisfy the information need of users. Thus, the click relevance of this document is rather low. These two cases show the limitation of the BM25 label that sometimes exact matching signals cannot represent relevance in neural model training. Based on a large scale

of practical click logs, the click model can estimate more accurate relevance labels for query–document pairs with more user clicks without knowing the content of queries and documents. However, there are also disadvantages of click label. When clicks of a document are rare, its relevance estimated by click models may be rather inaccurate, such as case C shown in Figure 2. Although the document in case C is highly relevant in both exact and semantic matching, its click relevance is almost 0 because of its low-ranking position in the result list and rare clicks.

As for the data we used in the experiment, since click data are sampled from a commercial search engine, the documents in it are more likely to be highly relevant to the queries and preferred by users than those in the BM25 data, which are retrieved by BM25 from a web collection. However, the average number of documents for each query (the depth of the document pool) in click data is much smaller because the search logs mostly record the results in the first SERPs. Thus, we assume that click label has more potential to be effective when ranking on highly relevant documents, while BM25 label can improve the ranking quality especially by capturing the exact matching signal to the queries.

Is click label effective in training neural ranking models when evaluated on human-assessed test data? Existing works (Xiong et al., 2017; Zheng et al., 2018) that leverage click-through information in document ranking focus on the model performance on test data assessed with click-through rates or click model-based relevance, instead of human-assessed data. Therefore, we would like to investigate this research question in our experiment. Table 3 shows the performances of ranking models on Test-NTCIR in *Exp 1*. We report the performances of ranking models trained on BM25 labels and click labels, as well as several baseline methods, including four learning-to-rank models and BM25. BM25 achieves the best performance among all models. For the BM25 rankers, although their objective is to rank like BM25, they are neural approaches using semantic matching and representation learning without the external knowledge, such as TF and IDF, used in BM25, which we consider as the reason for their worse performance compared to BM25. For the click rankers, their training data have different data distribution from the test data, such as fewer documents per query and higher quality documents. For those documents with low quality in the test set, which usually do not attract any user click, it is difficult for click rankers to predict their relevance effectively. For the baseline models trained on human-assessed data, we attribute their worse performance compared to BM25 to the limited size of human-assessed training data.

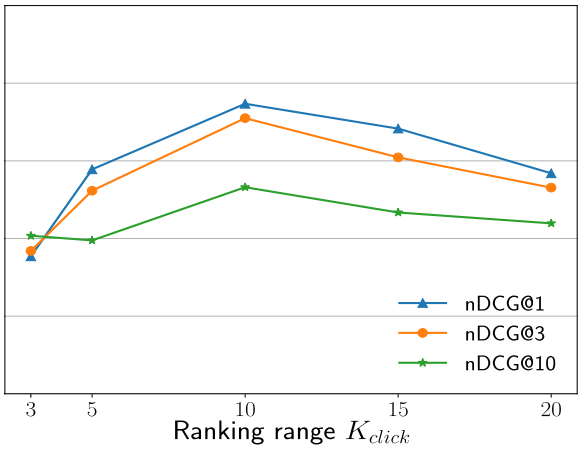

Figure 3. The performance curves of K-NRM*click* in the cascade ranking framework with different ranking ranges K_{click} .

Table 3
*The Performances of Ranking Models on Test-NTCIR. (△ Indicates Statistical Significance over BM25 with p ≤ 0.05. * Indicates Tested on the Click Rankers over the same BM25 Rankers (e.g., ARC-I$_{click}$ vs. ARC-I$_{BM25}$) with p ≤ 0.05.)*

Data	Model	nDCG@1	nDCG@3	nDCG@10
BM25	ARC-I	0.5836	0.5791	0.5778$^{\triangle}$
	Duet	0.5918	0.5913	0.5980$^{\triangle}$
	K-NRM	0.5459	0.5630$^{\triangle}$	0.5957$^{\triangle}$
Click	ARC-I	0.5160$^{\triangle *}$	0.5300$^{\triangle *}$	0.5390$^{\triangle *}$
	Duet	0.5954	0.5893	0.5892$^{\triangle}$
	K-NRM	0.5797	0.5872	0.5900$^{\triangle}$
Human	RankNet	0.5993	0.5970	0.6104
	RankBoost	0.5950	0.6117	0.6193
	LambdaMart	0.5870	0.6165	0.6213
	ARC-I	0.5248$^{\triangle}$	0.5317$^{\triangle}$	0.5376$^{\triangle}$
	Duet	0.5254$^{\triangle}$	0.5214$^{\triangle}$	0.5467$^{\triangle}$
	K-NRM	0.5135$^{\triangle}$	0.5325$^{\triangle}$	0.5323$^{\triangle}$
	BM25	**0.6109**	**0.6196**	**0.6386**

With the poor performance of click rankers on the whole Test-NTCIR, we further employed the models in only re-ranking on Test-NTCIR-TopK$_{BM25}$, a subset of Test-NTCIR, where only the first K results ranked by BM25 for each query were kept. We selected the re-ranking range K from $\{10,20,30,50,70,90\}$ and got the best performances of all click rankers when $K=10$, while all BM25 rankers performed best when K=20. Table 4 shows the model performances on Test-NTCIR-TopK$_{BM25}$.

Table 4
The Performances of Ranking Models on Test-NTCIR-TopK_{BM25} (i.e., the Set of the First K Documents Per Query in Test-NTCIR Ranked by BM25). (and ** Indicate Statistical Significance over the same Models Evaluated on Test-NTCIR with p ≤ 0.05 and p ≤ 0.01, Respectively.)*

Data	Model	*K*	nDCG@1	nDCG@3	nDCG@10
BM25	ARC-I		0.6151*	0.6192**	0.6229**
	Duet	20	0.6337**	0.6321**	0.6396**
	K-NRM		0.6052*	0.6192**	0.6322**
Click	ARC-I		0.6111**	0.6241**	0.6405**
	Duet	10	0.6423**	0.6428**	0.6448**
	K-NRM		**0.6448***	**0.6637****	**0.6520****
	BM25	–	0.6109	0.6196	0.6386

Almost all models with weak supervision have significant improvements on all evaluation metrics compared to themselves tested on the original Test-NTCIR. The reasons for this may be two-fold: (1) the top documents ranked by BM25, which contain lots of exact matching signals, are more likely to be relevant to the queries and (2) with the increase in the size of test set, the ranking task for the models also gets more difficult because more documents dissimilar from the training data will be involved in. K-NRM is the best performed model trained on click labels. Our results reveal the different behaviors of click rankers and BM25 rankers in re-ranking. Although the improvements of BM25 rankers are statistically significant compared to themselves tested on the whole Test-NTCIR, they are still smaller to those of click rankers. In this experiment, click label shows its effectiveness in training neural ranking models when re-ranking the top retrieved documents.

How can we combine click label and BM25 label to contribute jointly to model promotion? Owing to different impacts of click label and BM25 label on model training, we propose to combine two kinds of weak labels by employing click rankers after BM25 rankers in a cascade ranking framework, i.e., *Exp 2* in Figure 1. At the first stage, a large scale of documents will be ranked by BM25 and the top-ranked documents will be kept (i.e. the generation process of Test-NTCIR). Then, these kept documents will be sorted by BM25 rankers at the second stage. At the third stage, click rankers will be adopted to predict the final ranking lists of documents on the set of top documents from the second stage.

We selected the most effective model K-NRM in the previous experiment. We used $Duet_{BM25}$ at the second stage in our cascade ranking framework, because it is the best model trained on BM25 label and slightly outperforms the strongest baseline BM25. At the second stage, we fixed the ranking range at 20, while at the third stage, we chose ranking ranges K_{click} from $\{3,5,10,15,20\}$ and got the best performance of all click models when $K_{click}=10$. Table 5 shows the best model performances and statistical significance of our cascade ranking framework. The statistical significance is tested over the performances of themselves on the whole Test-NTCIR and the BM25 ranker at the second stage, respectively. In the cascade ranking framework, all click rankers have significant improvement on all nDCG metrics compared to their previous performances in Table 3. Meanwhile, the performance of $K\text{-}NRM_{click}$ is also significantly improved on all nDCG metrics compared to $Duet_{BM25}$. Our results show the effectiveness of cascade ranking framework in enhancing the performances of weakly supervised ranking models and prove our hypothesis that click label is more likely to be applied to re-rank highly relevant documents. Our cascade ranking framework outperforms the best result in NRCIR-13 WWW task (**Luo** et al., 2017a) and RUCIR-C-NU-Base-1, with 3.31% improvement on nDCG@10. Figure 3 shows the performance curves of $K\text{-}NRM_{click}$ in our cascade ranking framework with different K_{click}. With the increase in ranking range K_{click} from 3 to 10, $K\text{-}NRM_{click}$ performs better in all evaluation metrics and achieves the best performance when $K_{click}=10$ .When K_{click} is larger than 10, there will be more nonrelevant or somewhat relevant documents in the candidate list, causing worse performance of click rankers.

5.5 Preference Test

Since search engine is an interactive system with users, we conducted a preference test among cascade ranking frameworks $K\text{-}NRM_{click}$, $Duet_{BM25}$, and BM25 to investigate whether our cascade ranking framework and click neural ranker can win the preference of search engine users. First, we compared $K\text{-}NRM_{click}$ with $Duet_{BM25}$ and then with BM25 and chose the most preferred one among the three models, K-NRM, to compare with the cascade ranking framework. In each comparison between two rankers, we invited seven people to annotate their preference in seven-level criteria (+3 to -3), indicating how much the left page is better than the right page. There were 14 annotators in total. We calculated the average score of users' preferences for each query as the final preference score. Table 6 shows win/tie/loss of preferred query numbers of $K\text{-}NRM_{click}$ compared to $Duet_{BM25}$, BM25, and cascade ranking framework. Our results show that $K\text{-}NRM_{click}$ is preferred

Table 5
*The Performances of Cascade Ranking Models on Test-NTCIR-Top20BM25-Top10Duet, Which is first Ranked by BM25 and Then Re-ranked by DuetBM25 with Ranking Ranges 20 and 10. (△ Indicates Statistical Significance over the same Models Evaluated on Test-NTCIR with $p \le 0.05$. * Indicates Statistical significance over DuetBM25 with $p \le 0.05$.)*

Model	nDCG@1	nDCG@3	nDCG@10
Cascade	$0.6747^{\triangle *}$	$0.6710^{\triangle *}$	$0.6532^{\triangle *}$
BM25	0.6109	0.6196	0.6386

Table 6
Preference Test on K-NRM$_{click}$ and Other Three Opponent Rankers. (Win/Tie/Loss are the Number of Queries Where Annotators Regard the Opponent Ranker Better, Equally, or Worse Than K-NRM$_{click}$.)

Comparison	Win	Tie	Loss
Duet$_{BM25}$ vs. K-NRM$_{click}$	24	16	60
BM25 vs. K-NRM$_{click}$	38	16	46
Cascade vs. K-NRM$_{click}$	42	21	37

by users on more queries than Duet$_{BM25}$ and BM25, while our cascade ranking framework outperforms single model K-NRM$_{click}$, which are consistent with experimental results of nDCG metrics.

6 Conclusions

In this paper, we investigate the difference between BM25-based relevance and click model-based relevance in training neural ranking models. Extensive experiments are conducted to show the effectiveness of click model-based relevance in training neural ranking models when tested on human-assessed test set. Our results demonstrate that the click label can improve the rankings of highly relevant documents, while the BM25 label can help rankers capture more exact matching signals. We also propose a cascade ranking framework to fuse the two kinds of weak labels, which significantly improves the performances of neural rankers and wins in the preference test comparing with other single ranking models. Our work provides a novel and feasible solution, cascade ranking framework, to train data-driven ranking models. For the future work, we would like to design a single neural ranking model to jointly take advantage of BM25 relevance and click relevance.

Acknowledgments: This work is supported by the National Key Research and Development Program of China (2017YFB0202204) and Natural Science Foundation of China (Grant Nos. 61622208, 61732008, 61532011).

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Ghemawat, S.(2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*. Retrieved from https://arxiv.org/pdf/1603.04467.pdf

Agichtein, E., Brill, E., Dumais, S., & Ragno, R. (2006,). Learning user interaction models for predicting web search result preferences. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3-10. doi:10.1145/1148170.1148175

Ai, Q., Bi, K., Guo, J., & Croft, W. B. (2018). Learning a deep listwise context model for ranking refinement. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 135-144. doi:10.1145/3209978.3209985

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics, 18*(4), 467-479.

Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., & Hullender, G. N. (2005). Learning to rank using gradient descent. *Proceedings of the 22nd International Conference on Machine learning (ICML-05)*, 89-96. doi:10.1145/1102351.1102363

Chapelle, O., & Zhang, Y. (2009). A dynamic bayesian network click model for web search ranking. *Proceedings of the 18th International Conference on World Wide Web*, 1-10. doi:10.1145/1526709.1526711

Chuklin, A., Markov, I., & Rijke, M. D. (2015). Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 7*(3), 1-115.

Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008). An experimental comparison of click position-bias models. *Proceedings of the 2008 international conference on web search and data mining*, 87-94. doi:10.1145/1341531.1341545

Dehghani, M., Zamani, H., Severyn, A., Kamps, J., & Croft, W. B. (2017). Neural ranking models with weak supervision. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 65-74. doi:10.1145/3077136.3080832

Dupret, G., & Liao, C. (2010). A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 181-190. doi:10.1145/1718487.1718510

Dupret, G. E., & Piwowarski, B. (2008): A user browsing model to predict search engine click data from past observations. *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 331-338. doi:10.1145/1390334.1390392.

Fan, Y., Pang, L., Hou, J., Guo, J., Lan, Y., & Cheng, X. (2017). Matchzoo: A toolkit for deep text matching. *arXiv preprint*

arXiv:1707.07270. Retrieved from https://arxiv.org/pdf/1707.07270.pdf

Frénay, B., & Verleysen, M. (2014). Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems, 25*(5), 845-869. doi:10.1109/TNNLS.2013.2292894

Freund, Y., Iyer, R., Schapire, R. E., & Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of machine learning research, 4*(6), 933-969. doi:10.1162/1532443041827916

Guo, J., Fan, Y., Ai, Q., & Croft, W. B. (2016). A deep relevance matching model for ad-hoc retrieval. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 55-64. doi:10.1145/2983323.2983769

Hu, B., Lu, Z., Li, H., & Chen, Q. (2014). Convolutional neural network architectures for matching natural language sentences. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, & K.Q. Weinberger(Eds.). *Advances in neural information processing systems* (pp. 2042-2050). Cambridge, Massachusetts, USA: MIT Press.

Huang, P. S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2013). Learning deep structured semantic models for web search using clickthrough data. *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 2333-2338. doi:10.1145/2505515.2505665

Joachims, T., Granka, L. A., Pan, B., Hembrooke, H., & Gay, G. (2017). Accurately interpreting clickthrough data as implicit feedback. *ACM SIGIR Forum*, 51(1), 4-11. doi:10.1145/3130332.3130334

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. Retrieved from https://arxiv.org/pdf/1412.6980.pdf

Lee, D. H. (2013, June). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *The 30th International Conference on Machine Learning* (Vol. 3, pp. 2-7). New York, USA: ACM.

Liu, Y., Xie, X., Wang, C., Nie, J. Y., Zhang, M., & Ma, S. (2017). Time-aware click model. *ACM Transactions on Information Systems (TOIS)*, 35(3), 16.

Luo, C., Sakai, T., Liu, Y., Dou, Z., Xiong, C., & Xu, J. (2017a). Overview of the NTCIR-13 we want web task. *Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies*. 40-49.

Luo, C., Zheng, Y., Liu, Y., Wang, X., Xu, J., Zhang, M., & Ma, S. (2017b). SogouT-16: a new web corpus to embrace IR research. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1233-1236. doi:10.1145/3077136.3080694

MacAvaney, S., Hui, K., & Yates, A. (2017). An approach for weakly-supervised deep information retrieval. *arXiv preprint arXiv:1707.00189*. Retrieved from https://arxiv.org/pdf/1707.00189.pdf

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* (pp. 3111-3119). Cambridge, Massachusetts, USA: MIT Press.

Mitra, B., Diaz, F., & Craswell, N. (2017). Learning to match using local and distributed representations of text for web search. *Proceedings of the 26th International Conference on World Wide Web*, 1291-1299. doi:10.1145/3038912.3052579

Pang, L., Lan, Y., Guo, J., Xu, J., Wan, S., & Cheng, X. (2016,). Text matching as image recognition. *Thirtieth AAAI Conference on Artificial Intelligence*, 2793-2799.

Pang, L., Lan, Y., Guo, J., Xu, J., & Cheng, X. (2017a). A deep investigation of deep ir models. *arXiv preprint arXiv:1707.07700*. Retrieved from https://arxiv.org/pdf/1707.07700.pdf

Pang, L., Lan, Y., Guo, J., Xu, J., Xu, J., & Cheng, X. (2017b). DeepRank: A new deep architecture for relevance ranking in information retrieval. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 257-266. doi:10.1145/3132847.3132914

Qin, T., & Liu, T. Y. (2013). Introducing LETOR 4.0 datasets. *arXiv preprint arXiv:1306.2597*. Retrieved from https://arxiv.org/ftp/arxiv/papers/1306/1306.2597.pdf

Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014). Learning semantic representations using convolutional neural networks for web search. *Proceedings of the 23rd International Conference on World Wide Web*, 373-374. doi:10.1145/2567948.2577348

Wang, C., Liu, Y., Zhang, M., Ma, S., Zheng, M., Qian, J., & Zhang, K. (2013). Incorporating vertical results into search click models. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 503-512. doi:10.1145/2484028.2484036

Wang, C., Liu, Y., Wang, M., Zhou, K., Nie, J. Y., & Ma, S. (2015). Incorporating non-sequential behavior into click models. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 283-292. doi:10.1145/2766462.2767712

Wu, Q., Burges, C. J., Svore, K. M., & Gao, J. (2010). Adapting boosting for information retrieval measures. *Information Retrieval, 13*(3), 254-270. doi:10.1007/s10791-009-9112-1

Xiong, C., Dai, Z., Callan, J., Liu, Z., & Power, R. (2017). End-to-end neural ad-hoc ranking with kernel pooling. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 55-64. doi:10.1145/3077136.3080809

Yin, D., Hu, Y., Tang, J., Daly, T., Zhou, M., Ouyang, H., ... & Langlois, J. M. (2016). Ranking relevance in yahoo search. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 323-332. doi:10.1145/2939672.2939677

Zhang, Y., Chen, W., Wang, D., & Yang, Q. (2011). User-click modeling for understanding and predicting search-behavior. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1388-1396. doi:10.1145/2020408.2020613

Zheng, Y., Fan, Z., Liu, Y., Luo, C., Zhang, M., & Ma, S. (2018). Sogou-qcl: A new dataset with click relevance label. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1117-1120. doi:10.1145/3209978.3210092