Sogou-QCL: A New Dataset with Click Relevance Label

Yukun Zheng†, Zhen Fan†, Yiqun Liu*, Cheng Luo, Min Zhang, Shaoping Ma
Department of Computer Science and Technology,
Beijing National Research Center for Information Science and Technology,
Tsinghua University
Beijing, China
yiqunliu@tsinghua.edu.cn

ABSTRACT
Data is of vital importance in the development of machine learning technologies. Recently, within the information retrieval field, a number of neural ranking frameworks have been proposed to address the ad-hoc search. These models usually need a large amount of query-document relevance judgments for training. However, obtaining this kind of relevance judgments needs a lot of money and manual effort. To shed light on this problem, researchers seek to use implicit feedback from users of search engines to improve the ranking performance. In this paper, we present a new dataset, Sogou-QCL, which contains 537,366 queries and five kinds of weak relevance labels for over 12 million query-document pairs. We apply Sogou-QCL dataset to train recent neural ranking models and show its potential to serve as weak supervision for ranking. We believe that Sogou-QCL will have a broad impact on corresponding areas.

CCS CONCEPTS
• Information systems → Test collections; web log analysis; Relevance assessment; web crawling;

KEYWORDS
Test collection; document ranking; search evaluation

ACM Reference Format:

1 INTRODUCTION
Without data support, deep learning can’t achieve such rapid development today. The fast growth of data quantity has brought breakthroughs in a lot of machine learning problems, such as computer vision, speech recognition, etc. However, due to lack of high-quality data, there is a bottleneck of advancing the state-of-the-art technologies in many cases, such as document ranking in information retrieval (IR) [3, 12]. Document ranking is the central problem in IR, i.e. given a textual query and a set of candidate documents, the ranking model calculates a relevance score to represent the document’s degree of relevance with respect to the query, which determines the position of the document in the ranking list. Recently, a number of deep neural networks have been proposed to address document ranking problem. However, it’s very expensive and time-consuming to collect a large scale of query-document pairs with relevance labels for model training. Thus, data is a matter of concern for researchers in a lab-based environment to enhance the state-of-the-art approaches.

Several benchmarks have been released to examine the effectiveness of different retrieval models, such as TREC Web Tracks [2] and NTCIR We Want Web [11]. The relevance judgments of these datasets are available as a supervision signal for retrieval model training. However, these tasks usually have at most hundreds of queries. LETOR [8] is a package of benchmark datasets for research on learning to rank containing standard features, relevance judgments, several baselines, etc. The latest LETOR 4.0 [8] integrated 78,720 documents from Gov2 and 2,476 queries from Million Query track of TREC 2007 and TREC 2008. In terms of data size, all these datasets are inadequate compared to billions of information need from real search engine users.

Thus, researchers have begun to study the methods of automatic relevance annotation, such as click-through rates (CTR) deriving from search engines. Search engines can collect a number of query logs with click-through information automatically. However, raw query logs contain a lot of user privacy information, which is illegal and inappropriate to share. Several anonymous query log datasets have been published to promote IR studies, such as Sogou-Q [9], the WSDC series [15], MSN2006 [20] and AOL2006 [14]. In search engines, clicks are usually treated as implicit relevance feedback from users to improve the ranking list. However, there are limitations on adopting clicks as supervision signals to train neural ranking models. During the Web search, user clicks are often biased towards many aspects, such as the position and novelty of a document, the users’ attention to different vertical styles, etc. Thus, user clicks are biased and noisy [19]. A number of click models were proposed to estimate the click probability of a document from query logs by reducing the impacts of the biases and inferring its relevance to the query. This kind of relevance is named as “click model-based relevance” in previous studies[19].

In this paper, we employ click models to debias query logs sampled from Sogou.com and present a new dataset with various kinds of click model-based relevance labels, Sogou-QCL. This dataset contains 537,366 unique queries and 12 million unique query-document
pairs. The relevance labels are assessed by UBM [4], DBN [1], TCM [18], PSCM [16] and TACM [10] respectively. There are three advantages of Sogou-QCL:

- To our best of knowledge, Sogou-QCL is the first public dataset assessed with click model-based relevance. A large scale of query logs and five popular click models are integrated to generate weak relevance labels.
- Sogou-QCL provides abundant textual information, including queries and raw pages, titles and full-text contents of documents.
- Sogou-QCL protects user privacy and can be used in a wide range of research areas, such as ad-hoc retrieval, search evaluation, and etc.

2 DATASET

2.1 Data Preparation

Our dataset is based on the collection of query logs which contains 1.95 billion query sessions in a time span of 18 days. The query logs are collected by Sogou.com, the third largest commercial search engine in China. Each query session records the query, the URLs and vertical types of results in the SERPs, and the sequence of user clicks as well as timestamps. Besides, such query sessions contain a lot of user privacy information which is strictly protected by law and regulations. Therefore it’s impossible to release such kind of data even for research purposes. Sogou-QCL only contains weak labels derived from a large number of users’ clicks and hides individual’s behaviors.

Firstly, we remove the queries that appear less than 10 times and the query sessions with no click. The queries with low appearing frequency are likely to contain user privacy information, including the user’s address and phone number. Another consideration is that if the clicks are too sparse in a query session, it is insufficient for click models to calculate reliable click model-based relevance labels. For a large scale of URLs, we crawl their raw sources from the Web. Because a number of web pages are out-of-date or blocked in 2018, only about 60% of URLs’ resource pages are crawled successfully. Then, we conduct several cleaning processes on the dataset to make it more user-friendly to researchers:

1. We filter the pornographic queries.
2. We convert the encoding of web pages to UTF-8.
3. We extract the titles and full-text of crawled documents.

We then use the sequences of clicks to train five click models: UBM, DBN, TCM, PSCM and TACM, which are based on an open source implementation [16]. Due to time and computational resource constraints, all data is divided in 60 subsets to train 60 click models individually. We ensure that all the sessions with the same query are classified into the same subset. We randomly split the sessions per query in proportion to 4:1 for training and test respectively. We use the average perplexities on the test set to evaluate these click models. The performances of click models are listed in Table 1. It shows that TACM is the best-performed model in predicting the click probabilities of documents, followed by PSCM, while TCM performs worst among all five click models. Our experiment findings align with the results of click model in [10].

Table 1: The average perplexities of click models.

<table>
<thead>
<tr>
<th>Click Model</th>
<th>TACM</th>
<th>PSCM</th>
<th>UBM</th>
<th>DBN</th>
<th>TCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perplexity</td>
<td>1.375</td>
<td>1.490</td>
<td>1.577</td>
<td>1.608</td>
<td>3.181</td>
</tr>
</tbody>
</table>

2.2 Overview of Sogou-QCL

We will briefly introduce our dataset and present various statistics of Sogou-QCL. The contents of Sogou-QCL are listed in Figure 1. Table 2 shows the fundamental statistics of the dataset. In Sogou-QCL, we provide the entire list of URLs recorded in query logs as well as their weak relevance labels with respect to the query. For those reachable URLs, the titles and full-text contents are integrated into Sogou-QCL. Thus, we report numbers of total and crawled documents here. We segment all the textual data in Sogou-QCL using the Jieba toolkit and calculate the average query/doc length, i.e. the average number of words in queries/documents. Table 3 shows the statistics of TREC Web Track 09-14 datasets for ad-hoc retrieval and LETOR 4.0, two most popular datasets in ad-hoc retrieval studies. In terms of data size, Sogou-QCL is far larger than the two datasets, which is a main advantage to serve the training of deep neural networks.

Table 2: The statistics of Sogou-QCL dataset.

<table>
<thead>
<tr>
<th>#Query</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>#Doc_total</td>
<td>9,046,737</td>
</tr>
<tr>
<td>#Doc_crawled</td>
<td>5,480,860</td>
</tr>
<tr>
<td>#Query-#DocPair</td>
<td>12,238,726</td>
</tr>
<tr>
<td>#Domain of URLs</td>
<td>4,298,597</td>
</tr>
<tr>
<td>Avg. Query Length</td>
<td>1.46 words</td>
</tr>
<tr>
<td>Avg. Doc Length</td>
<td>1,060.7 words</td>
</tr>
<tr>
<td>Sampling Date</td>
<td>April 1st-18th, 2015</td>
</tr>
<tr>
<td>Language</td>
<td>Chinese</td>
</tr>
</tbody>
</table>

Table 3: The statistics of several datasets for ranking.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Query</th>
<th>#Doc</th>
<th>#Pair</th>
<th>Collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC 09-14</td>
<td>298</td>
<td>111,909</td>
<td>113,272</td>
<td>ClueWeb 09 &amp; 12</td>
</tr>
<tr>
<td>LETOR 4.0</td>
<td>2,476</td>
<td>78,720</td>
<td>84,834</td>
<td>Gov2</td>
</tr>
</tbody>
</table>

The characteristics of a dataset have a great impact on neural model training. Here, we will present four of our most concerned aspects of Sogou-QCL, i.e. the query/document length, the number of documents per query, the distribution of relevance labels and the quality of web pages.

Query/document length. In most neural ranking models, there is a limit on the maximum length of queries and documents, which is a tradeoff between the computational cost and the effectiveness of text representation. We look into the distributions of query length

https://github.com/fxsjy/jieba
and document length in Sogou-QCL, which are shown in Figures 2(a) and 2(b) respectively. More than 90% of queries are within six words, and about 85% of documents are less than 1,200 words. Although these results will change when we use other word segmentation methods, they are still instructive for model training.

Number of documents per query. Intuitively, the neural ranker can be trained better with more documents covering different contents under a query. The average number of documents per query in the TREC series and LETOR 4.0 are 380.1 and 34.3, while it’s 22.8 in Sogou-QCL. Figure 3 shows the distribution of queries with different number of document. Each record of a query session in our query logs mostly contains about 10 results in the first SERP. However, these results were changing during collection. Therefore, as the log data accumulates, more than 70% of queries contain 10 to 25 documents with about 14 documents successfully crawled on average.

The distribution of relevance labels. In the pairwise training process, pairs of positive and negative documents are generated according to their relevance labels, whose quality significantly determines the performance of a ranking model. Figure 4 presents the distribution of relevance scores estimated by click models, where the x-axis is the relevance scores ranging from 0 to 1 and the y-axis is the number of corresponding documents in a logarithmic scale. With diverse assumptions in these click models, the distributions of their click model-based relevance scores are different. Interestingly, in the range of [0, 1], DBN tends to predict lower scores, while PSCM and TACM tend to give higher ones. For the other two click models, UBM and TCM, the distributions of their relevance scores are relatively uniform.

Quality of web pages. To investigate the quality of web pages, we calculated the average PageRank values of all the URLs’ domains in Sogou-QCL. The top five most frequent domains are listed in Table 4 with their appearing frequency, values and ranks of PageRank. Among five domains, there are three Q&A websites, while the other two are popular BBS and document resource websites in China. All these domains have relatively high PageRank values among all domains in Sogou-QCL dataset. On the other hand, since all documents are retrieved at the top ranks in the SERPs by Sogou search engine, we can affirm that most of them are of high quality and relevant to the queries.

3 APPLICATION

In the following part, we will describe our experiment of training several recent neural ranking models on Sogou-QCL dataset. We randomly split Sogou-QCL dataset into two parts, in which 1000 queries for validation and the rest of the others for training. The textual data of queries and documents’ titles are used as training data with CTR and TACM-based relevance as the supervision labels. We evaluate rankers on two test sets [17], Test-same and Test-diff, which are sampled from the same query log data as Sogou-QCL and assessed by CTR and TACM respectively. For the sake of convenience, we rename the two test sets as Test-CTR and Test-TACM using the types of their relevance. We train the word embedding on documents’ titles in the training set using word2vec [13] and select 300 as the embedding size. We also adopt the same TREC evaluation toolkit in [17] to make our results comparable with theirs.

We choose ARC-I [7], DRMM [6] and K-NRM [17] in the experiment. These models can cover two categories of network architectures [6]. ARC-I belongs to the representation-focused model, while K-NRM and DRMM are classified to the interaction-focused approach. In Table 4, we list the top five domains and their frequency, average PageRank value, and rank from best to worst. Additionally, we also show the number of documents crawled in Sogou-QCL. For example, the domain "wenwen.sogou.com" has the highest frequency and the best average PageRank value, which is 482,028 with a rank of #277,563. The domain "tieba.baidu.com" has the second-highest frequency and the second-best average PageRank value, which is 487,577 with a rank of #257,000. The domain "wenku.baidu.com" has the third-highest frequency and the third-best average PageRank value, which is 277,563 with a rank of #482,028.
Table 5: The performances of ranking models on Test-CTR and Test-TACM. (* and ** indicate statistical significance over BM25 with \( p \leq 0.05 \) and \( p \leq 0.01 \) respectively.)

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>Test-CTR</th>
<th>Test-TACM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>nDCG@1</td>
<td>nDCG@3</td>
</tr>
<tr>
<td>CTR</td>
<td>ARC-I</td>
<td>0.1476</td>
<td>0.1926**</td>
</tr>
<tr>
<td></td>
<td>DRMN</td>
<td>0.1485</td>
<td>0.1951**</td>
</tr>
<tr>
<td></td>
<td>K-NRM</td>
<td>0.1511*</td>
<td>0.2057**</td>
</tr>
<tr>
<td>TACM</td>
<td>ARC-I</td>
<td>0.1413</td>
<td>0.1852**</td>
</tr>
<tr>
<td></td>
<td>DRMN</td>
<td>0.1578**</td>
<td>0.2053**</td>
</tr>
<tr>
<td></td>
<td>K-NRM</td>
<td>0.1664**</td>
<td>0.2147**</td>
</tr>
<tr>
<td></td>
<td>BM25</td>
<td>0.1261</td>
<td>0.1583</td>
</tr>
</tbody>
</table>

model. All models are implemented using MatchZoo [5] based on tensorflow. We employ cross entropy loss with softmax as the loss function in the pairwise training process. We tune all hyperparameters of models based on the validation set. In all training processes with learning rate equals to 0.001, we adopt adam as the gradient descent optimisation algorithms. The student’s t-test are employed to examine the significance of ranking models’ performances over the baseline, BM25.

Table 5 shows the performances of ranking models on Test-CTR and Test-TACM. The K-NRM_{TACM} achieves the best performances on both test sets and outperforms BM25 on all nDCG metrics. All the models trained with CTR and TACM-based relevance perform better than BM25 on Test-CTR, while the performances of ARC-I and DRM are worse than BM25 when tested on Test-TACM. This application of Sogou-QCL dataset shows its usability and effectiveness in training neural ranking models.

4 DISCUSSION

To address the lack of training data, a number of weakly supervised methods have been proposed in document ranking studies. Most of these methods focus on heuristics like BM25 to use exact matching scores as weak relevance labels [3, 12]. However, Guo et al. [6] suggested that the exact matching can’t represent relevance matching because it ignores the semantic matching signals. Different from BM25, click-through behaviors consist of abundant click preferences of users. Meanwhile, the sequence of documents that a user clicked can imply the user’s intent in the search session. Thus, we believe that the weak relevance derived from click-through information using click models can serve as a weak supervision signal to train neural ranking models. In addition, Sogou-QCL as a high-quality document collection can also be applied in other IR and language computing related studies.

5 CONCLUSIONS

In this paper, we publish a novel dataset named Sogou-QCL, which is the first public dataset with weak relevance labels in the IR community. Besides five kinds of weak relevance labels estimated by popular click models, it also contains queries and multiple kinds of textual data of documents. Our dataset is far larger than existing datasets for ranking. To examine different aspects of Sogou-QCL dataset, we make a detailed investigation of the dataset. Furthermore, we present an application of Sogou-QCL dataset by training neural ranking models on queries and document titles with CTR and TACM-based relevance labels as supervision. Our experimental results show Sogou-QCL’s potential to serve as training data for neural ranking models. We believe that this dataset will provide more opportunities for researchers to advance the development of technologies in IR and related communities.

REFERENCES