Building a Web Test Collection using Social Media

Chia-Jung Lee
Center for Intelligent Information Retrieval
School of Computer Science
University of Massachusetts, Amherst
cjlee@cs.umass.edu

W. Bruce Croft
Center for Intelligent Information Retrieval
School of Computer Science
University of Massachusetts, Amherst
croft@cs.umass.edu

ABSTRACT

Community Question Answering (CQA) platforms contain a large number of questions and associated answers. Answerers sometimes include URLs as part of the answers to provide further information. This paper describes a novel way of building a test collection for web search by exploiting the link information from this type of social media data. We propose to build the test collection by regarding CQA questions as queries and the associated linked web pages as relevant documents. To evaluate this approach, we collect approximately ten thousand CQA queries, whose answers contained links to ClueWeb09 documents after spam filtering. Experimental results using this collection show that the relative effectiveness between different retrieval models on the ClueWeb-CQA query set is consistent with that on the TREC Web Track query sets, confirming the reliability of our test collection. Further analysis shows that the large number of queries generated through this approach compensates for the sparse relevance judgments in determining significant differences.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.m [**Information Storage and Retrieval**]: Miscellaneous—*Test Collections*

General Terms

Experimentation, Performance

Keywords

Test collection, social media, community question answering

1. INTRODUCTION

The most difficult part of building a test collection is perhaps creating a set of queries with associated relevance judgments. Click data can be used as a substitute in some cases, but this is not available for general use in academic environments. In this paper, we explore an approach to generating a set of queries and relevance judgments for a collection of web documents by exploiting the information in social media, and more specifically Community Question Answering services. CQA sites such as Yahoo! Answers[1] and Baidu Zhidao[2] provide social platforms for users to raise questions, to obtain useful information and to share potential answers. The collaborative nature of such platforms motivates interested users to voluntarily engage in providing useful answers to many topics. Answerers sometimes provide URLs as part of their answers. These links are used to provide additional information, to explain more complicated concepts that can not be detailed in short paragraphs, or to present reliable citations, etc. Figure 1 shows an example of an answer that incorporates several URLs.

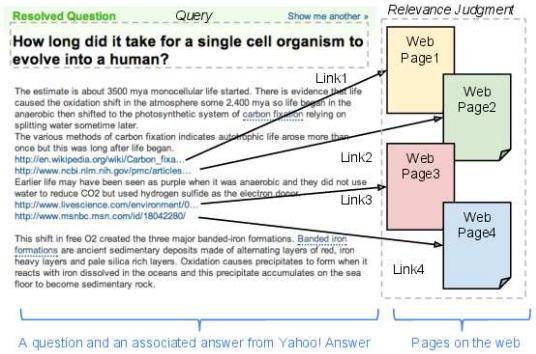

Figure 1: An example query with relevance judgments generated from social media data.

The central hypothesis of this paper is that these links can be regarded as relevant documents for the CQA question (or query). Although the relevance judgments obtained in this way are likely to be sparse, our expectation is that the large number of queries obtained will compensate for this.

In this paper, we collected a large number of question-answer pairs from existing and newly crawled CQA collections. We then reduced this set to include only questions whose answers contained links, where the links pointed to documents in the ClueWeb09 collection. The final ClueWeb-CQA (CW-CQA) set of queries and relevance judgments was produced after some additional filtering to deal with issues

[1] `http://answers.yahoo.com/`
[2] `http://zhidao.baidu.com/`

SIGIR '13, July 28 - August 01 2013, Dublin, Ireland

such as spam. To test the validity of this new test collection, the relative effectiveness of some well-known benchmark retrieval models is evaluated and compared. Our CW-CQA results show that a term dependency model significantly outperforms a bag of words model, and pseudo-relevance feedback techniques can be helpful in most cases. These findings are consistent with the results using standard TREC Web Track query sets. As expected, the relevance judgments are sparse and incomplete. Carterette et al. [4] demonstrated that, up to a point, evaluation over more queries with fewer judgments is as reliable as fewer queries with more judgments. Similarly, we show that evaluating using a sufficient number of queries shows significant differences between retrieval models despite the incomplete relevance judgments.

The rest of paper is laid out as follows. Section 2 summarizes related work and Section 3 describes the methodology of building the CW-CQA test collection. We discuss the experimental results in Section 4. Section 5 makes closing remarks and discusses several future directions.

2. RELATED WORK

There have been several web test collections created for supporting reproducible experiments at TREC. Examples include .GOV2, WT2g, and WT10g as well as a recent larger collection ClueWeb09. For large collections, relevance judging is mostly done through pooling techniques [7]. Such techniques assemble and judge results from multiple searches and systems, with the assumption that most relevant documents will be found.

Even with the use of pooling methods, creating relevance judgments is often costly and the judged results can be biased and incomplete [2]. Buckley and Voorhees [1] proposed the metric *bpref* that is both highly correlated with existing measures when complete judgments are available and more robust to incomplete judgment sets. For the Million Query Track at TREC 2007, Carterette et al. [4] presented two document selection algorithms [3] to acquire relevance judgments. Their results suggested that, up to a point, evaluation over more queries with fewer judgments is more cost-effective and as reliable as fewer queries with more judgments.

3. METHODOLOGY

3.1 Test Collection

We build the CW-CQA test collection using large CQA datasets and the web collection ClueWeb09 [3]. We obtain a large number of question-answer pairs from the CQA corpora and harvest all links provided in answers. We then reduce this set to include only questions whose answers contained links pointing to ClueWeb09 documents. We obseve that some of the links contained in CQA answers can be considered to be spam pages. To ensure a reliable test collection, we filter the reduced question sets based on two spam-controlling parameters S_R and S_A. Cormack et al [5] proposed a content-based classifier that quantifies the "spamminess" of a document based on a scale of 0 to 100, where a lower score indicates that the page has a higher likelihood to be spam. Accordingly, S_A calculates the average spam score of all links L_Q extracted for a question Q and S_R records the ratio of spam links among L_Q [4]. Varying S_R and S_A affects the final number of queries and the proportion of spam links. After filtering, we can establish our final CW-CQA test collection by using the remaining questions and associated links as test queries and relevance judgments. We dicuss the parameter settings in Section 4.

[3] http://lemurproject.org/clueweb09/

3.2 Retrieval Models

We test four existing retrieval models including query likelihood model (QLM), relevance model (RM) [8], sequential dependency model (SDM) [9] and a query expansion model using latent concept expansion (LCE) [10]. We choose these models because they include both common baselines used in other papers and methods that are state-of-the-art in terms of effectiveness.

QLM computes the likelihood of generating query texts based on documents models, and can often be written as:

$$P(Q|D) \stackrel{rank}{=} \sum_{q_i \in Q} log(P(q_i|D))$$

RM ranks documents according to the odds of their being observed in the relevant class. $P(w|q_i \dots q_k)$ can be effectively used to approximate $P(w|R)$ with w as a word in the collection.

$$\frac{P(D|R)}{P(D|N)} \approx \prod_{w \in D} \frac{P(w|R)}{P(w|N)}$$

SDM is an effective instantiation of the Markov random field for information retrieval (MRF-IR) that makes the sequential dependence assumption. It ranks documents by:

$$\begin{aligned} P(D|Q) \stackrel{rank}{=} & \lambda_T \textstyle\sum_{q_i \in Q} f_T(q_i, D) \\ & + \lambda_O \textstyle\sum_{q_i \in Q} f_O(q_i, q_{i+1}, D) \\ & + \lambda_U \textstyle\sum_{q_i \in Q} f_U(q_i, q_{i+1}, D) \end{aligned}$$

LCE is a robust query expansion technique based on MRF-IR. This technique provides a mechanism for modeling term dependencies during expansion. The central idea is to compute a probability distribution over latent concepts using a set of pseudo-relevant documents in response to Q. Retrieval is then done by incorporating the top k latent concepts with the highest likelihood into original MRF model. Details of these models can be found in the appropriate papers.

4. EXPERIMENTS

4.1 Building The Collection

CQA datasets. We used two large CQA datasets, Yahoo Webscope L6 (Y6) [5] and a recently crawled Yahoo! Answers dataset (YA). Corpus Y6 provides a 10/25/2007 Yahoo! Answers dump. We additionally crawled the YA corpus by using the Yahoo! Answers API. Specifically, we collected up to 10,000 questions for each of the 26 Yahoo root categories as well as their corresponding answers [6]. Table 1 shows the number of questions N_Q, the average number of associated answers per question N_{Aavg}, and the average number of links per question N_{Lavg} in first row.

[4] We consider a page with spam score below 60 to be spam.

[5] http://webscope.sandbox.yahoo.com/

[6] The collection contains approximately one month Yahoo! Answers data starting from 7/31/12 to 9/5/12.

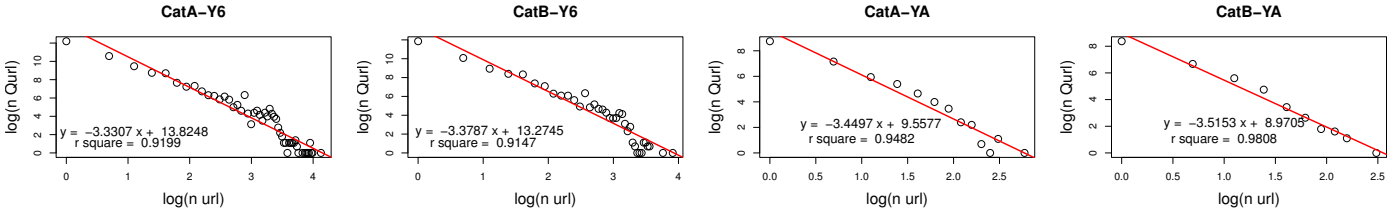

Figure 2: Linear fit of the logarithms of links per question (n_{url}) and frequency of these questions ($n_{Q_{url}}$).

		Y6	YA
CQA	N_Q	4,483,032	216,474
	N_{Aavg}	7.11	3.42
	N_{Lavg}	1.95	1.92
CW−CQA (CatA)	N_Q	272,619	8,386
	N_{Lavg}	1.74	1.44
CW−CQA (CatB)	N_Q	186,651	5,567
	N_{Lavg}	1.64	1.33

Table 1: Dataset statistics.

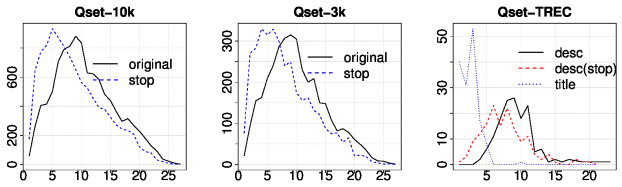

Figure 3: Query length (x-axis) and the frequency of queries (y-axis) for query sets 10k, 3k and TREC.

Connecting CQA and ClueWeb. To find connections, we then compared the CQA links with two subsets of ClueWeb09 pages, namely Category A (CatA) and Category B (CatB), which contain approximately 500M and 50M English documents, respectively. The second and third rows in Table 1 summarize the number of questions whose answers contained links to the ClueWeb data N_Q and their corresponding N_{Lavg} [7]. Figure 2 shows the relation between the number of links each question has (n_{url}) and the frequency of questions with n_{url} links ($n_{Q_{url}}$). Since this is a log-log plot, Figure 2 shows that n_{url} and $n_{Q_{url}}$ follow a power law distribution; that is, questions with few links occupy a significant portion of entire population. R^2 statistics show the goodness of the fit. The connection distributions for CatA and CatB resemble each other; in the following, we focus on evaluation using the ClueWeb09 CatB connections and searching on CatB for computational efficiency.

Queries and Relevance Judgments. To build the final test collection, we aggregate questions from Y6 and YA constrained by $S_R \leq 0.1$, $S_A \geq 90$ and $n_{url} > 1$. The constraint $n_{url} > 1$ is used to avoid the extremes in the power law distributions shown in Figure 2. A final set of 9988 questions are selected as the CW-CQA test queries and the associated links are regarded as relevance judgments. This set is denoted as the *10k* set. We construct an additional query set by selecting queries from 10k that have at least one relevant document returned by any of the four retrieval models described in Section 3, resulting in the query set *3k* containing 3440 questions. The maximum n_{url} in query sets 10k and 3k are respectively 22 and 19, and the minimum for both is 2. The relation of n_{url} and $n_{Q_{url}}$ remains as power law distributions for both query sets. The nature of CQA questions can make the queries quite long. We apply stop structure removal techniques [6] to the CW-CQA queries and Figure 3 compares the query length distributions.

For comparison, we use 148 standard TREC web track 2009, 2010 and 2011 title and description (desc) queries. We search these TREC queries on ClueWeb09 CatB and evaluate the results using standard TREC relevance judgments. Figure 3 suggests that the CW-CQA queries are more similar to TREC description queries in terms of query length.

Retrieval Setup. Indri [8] is used for indexing and searching. We use Dirichlet smoothing with $\mu = 2500$ for all runs without tuning. We apply spam filtering to all retrieval runs based on [5]. We evaluate using the top 1000 documents and report mean average precision (MAP), precision at 10/100 (P@10, P@100), mean reciprocal rank (MRR) and bpref.

4.2 Retrieval Performance

Table 2 shows the retrieval performance of the CW-CQA query sets 10k and 3k where the top performing runs are underlined. We perform paired t-tests on pairs of retrieval models (QLM, SDM), (QLM, RM) and (SDM, LCE). Specifically, RM and SDM are marked † if p-value < 0.05 compared to QLM. LCE is marked $_\star$ if p-value < 0.05 compared to SDM. We observe that SDM significantly outperforms QLM in both query sets for every metric. RM can significantly improve QLM for most metrics, showing the utility of pseudo-relevance feedback. LCE seems to improve SDM, but the significant difference is only observed for metrics P@100 and bpref. In general, models SDM and LCE are the most effective compared to others. The performance of the 10k and 3k query sets show similar trends. For query set 3k, MRR shows that on average all models rank the first known relevant document above rank 20.

Table 3 shows the retrieval performance of TREC title and desc queries. Similar to CW-CQA results, SDM significantly outperforms QLM in all cases. For title queries, unlike CW-CQA results, pseudo-relevance feedback techniques such as RM and LCE sometimes can hurt performance of QLM and SDM, respectively. The utility of pseudo-relevance feedback is more evident for desc queries. The similarity of the query length for CW-CQA queries and TREC descriptions provides a possible explanation for the higher level of consistency between their results.

In general, the relative effectiveness between retrieval models is similar for CW-CQA and TREC queries. The improvements based on term dependency modeling are significant

[7]Multiple appearances of the same URL for a question is considered only once.

[8]`http://www.lemurproject.org/indri/`

	Model	MAP	P@10	P@100	MRR	bpref
10k	QLM	.0107	.0047	.0018	.0195	.1815
	SDM	.0114†	.0051†	.0019†	.0208†	.1866†
	RM	.0114†	.0050†	.0019†	.0204	.1942†
	LCE	.0114	.0051	.0020$_{\star}$	.0203	.2014$_{\star}$
3k	QLM	.0312	.0137	.0051	.0566	.5271
	SDM	.0331†	.0149†	.0054†	.0605†	.5417†
	RM	.0330†	.0144†	.0055†	.0593	.5639†
	LCE	.0331	.0149	.0057$_{\star}$	.0590	.5849$_{\star}$

Table 2: Retrieval results for CW-CQA query sets 10k and 3k.

	Model	MAP	P@10	P@100	MRR	bpref
title	QLM	.1804	.3628	.1853	.4860	.2715
	SDM	.1989†	.3831	.1928†	.5171†	.2877†
	RM	.1810	.3622	.1848	.4808	.2747
	LCE	.2037$_{\star}$	.3830	.1926	.4910	.2926
desc	QLM	.1309	.2892	.1147	.4559	.2953
	SDM	.1471†	.2932	.1184†	.4611	.3030†
	RM	.1365	.2896	.1168†	.4482	.2975
	LCE	.1463	.3000	.1214	.4537	.3049

Table 3: Retrieval results for TREC query sets.

for all query sets.

The absolute retrieval performance in Table 2 is rather low compared to Table 3. This is to be expected given the sparseness of relevance judgments. Carterette et al. [4] suggested that evaluation over more queries with fewer judgments can be reliable. Similarly, an interesting question from our perspective is: *how many queries do we need to confirm the existence of significant differences between retrieval models?* To this end, we compute the p-value between QLM and SDM using different number of CW-CQA queries. Specifically, from the 3k set, we randomly sample k queries where k ranges from 100 to 3400 in steps of 100. We perform 20 random samples at each k and report the average p-value in Figure 4. All metrics share a tendency that using more queries results in smaller p-values. For metrics such as MAP, MRR and P@10, stable significance (i.e., p-value < 0.05) is reached when the sample size grows beyond 2100. For other metrics such as P@100 and bpref, a sample size of more than 1000 queries is sufficient to confirm a significant difference. These observations support [4] in a sense that evaluating using a sufficient number of CW-CQA queries distinguishes retrieval model effectiveness despite incomplete judgments.

5. CONCLUSIONS

We proposed a novel way of building a test collection for web search by considering CQA questions as queries and the associated URLs as relevant documents. This approach has the advantage that a large number of queries and relevance judgments can be gathered automatically and efficiently. We filtered CW-CQA queries based on the spam scores of their links. Experimental results on the CW-CQA query sets show that the relative effectiveness between different retrieval models is consistent with previous findings using the TREC queries, showing the reliability of the test collection. The relevance judgments for the CW-CQA queries are incomplete and the absolute retrieval performance is relatively low. However, we demonstrated that evaluation using a sufficient number of queries ensures that significant differences can be found.

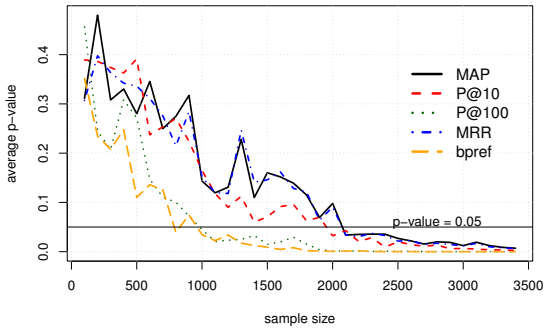

Figure 4: Average p-value of 20 times of sampling at each sample size.

These initial experimental results indicate several directions for future work. Validating consistency with manual relevance judgments will be important for our study. We plan to select a small set of queries for human assessors to judge, and compare the results with the automated approach. In addition, we will evaluate the same method using the newly constructed ClueWeb12 collection. The CW-CQA query sets for both ClueWeb09 and ClueWeb12 will be distributed through the Lemur project.

Acknowledgements

We thank Samuel Huston for his professional suggestions. This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF IIS-1160894. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

6. REFERENCES

[1] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. of SIGIR*, SIGIR '04, pages 25–32, 2004.

[2] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proc. of SIGIR*, SIGIR '07, pages 63–70, 2007.

[3] B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In *Proc. of SIGIR*, SIGIR '06, pages 268–275, 2006.

[4] B. Carterette, V. Pavlu, E. Kanoulas, J. A. Aslam, and J. Allan. Evaluation over thousands of queries. In *Proc. of SIGIR*, SIGIR '08, pages 651–658, 2008.

[5] G. V. Cormack, M. D. Smucker, and C. L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *CoRR*, abs/1004.5168, 2010.

[6] S. Huston and W. B. Croft. Evaluating verbose query processing techniques. In *Proc. of SIGIR*, SIGIR '10, pages 291–298, 2010.

[7] K. Jones, C. Van Rijsbergen, B. L. Research, and D. Dept. *Report on the Need for and Provision of an Ideal Information Retrieval Test Collection*. British Library Research and Development reports. 1975.

[8] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proc. of SIGIR*, SIGIR '01, pages 120–127, 2001.

[9] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proc. of SIGIR*, SIGIR '05, pages 472–479, 2005.

[10] D. Metzler and W. B. Croft. Latent concept expansion using markov random fields. In *Proc. of SIGIR*, SIGIR '07, pages 311–318, 2007.