SogouT-16: A New Web Corpus to Embrace IR Research

Cheng Luo
luochengleo@gmail.com
DCST, Tsinghua University
Beijing, China

Yukun Zheng
zhengyk11@mails.thu.edu.cn
DCST, Tsinghua University
Beijing, China

Yiqun Liu*
yiqunliu@tsinghua.edu.cn
DCST, Tsinghua University
Beijing, China

Xiaochuan Wang
wxc@sogou-inc.com
Sogou Inc.
Beijing, China

Jingfang Xu
xujingfang@sogou-inc.com
Sogou Inc.
Beijing, China

Min Zhang
z-m@tsinghua.edu.cn
DCST, Tsinghua University
Beijing, China

Shaoping Ma
msp@tsinghua.edu.cn
DCST, Tsinghua University
Beijing, China

ABSTRACT

Web collection is essential for many Web based researches such as Web Information Retrieval (IR), Web data mining, Corpus linguistics and so on. However, it is usually expensive and time-consuming to collect a large scale of Web pages in lab-based environment and public-available collection becomes a necessity for these researches. In this study, we present a Chinese Web collection, SogouT-16, which is the largest free-of-charge public Chinese Web collection so far. We provide a variety of descriptive characteristics of SogouT-16 and discuss its adoption in a newly-designed ad-hoc retrieval task in NTCIR-13, *We Want Web*. SogouT-16 also provides online retrieval service and contains a number of auxiliary resources including hyperlink structure graph, query logs, word embedding, and etc. We believe that SogouT-16 will provide new opportunities for novel investigations and applications in IR and other related communities.

CCS CONCEPTS

• **Information systems → Test collections**; *Web searching and information discovery*; *Web mining*;

KEYWORDS

test collection; Web corpus; search evaluation

1 INTRODUCTION

Web corpus is one of the central requirements in a number of researches, for example, Web structure analysis, information retrieval, and human language related studies. However, advancing the state-of-the-art technologies in a lab-based environment is usually difficult. One of the challenges is that it is expensive and time-consuming to collect a large-scale of pages which are representative of the size and diversity of the Web. In this paper, we present a new Web corpus, SogouT-16, which is the largest free Chinese corpus so far.

One of the commonest IR system evaluation methodologies is the test collection-based method, referred to as the Cranfield framework [6]. A typical test collection consists of a set of queries/requests, a collection of documents, and relevance judgments for query-document pairs. Test collection-based method has the advantage that once the relevance judgments are collected, they can be easily reused to compare the performance of multiple ranking systems with effectiveness metrics. In Web search studies, the documents are usually pages crawled from the Web.

A number of test collections have already provided much support for corresponding research tasks, such as WT10g [4], VLC2 [7], Gov2 [5], ClueWeb09, and ClueWeb12 [1]. These datasets are mainly in English and they are not necessarily adequate for researches in the context of Chinese.

Several Chinese-centric test collections have also been proposed in the past few years. ClueWeb09 has a Chinese subset (denoted as ClueWeb09-zh)[1], which consists of about 177M Web pages collected in 2009. CWP200T [2] which is released in 2016 contains about 7B Chinese pages collected from 2002 to 2015[2]. We have not found a lot of researches which are based on CWP200T yet. A potential reason is that the dataset is too large (200TB) and it is not a trivial task to process the dataset in laboratory. SogouT-08 and SogouT-12[3] are two public Web corpora, which were released in 2008 and 2012 respectively. They have already supported a variety of researches, for example, spam detection [8], intent discovery [3], Chinese linguistic analysis [12], word embeddings [11], and etc. They also served as the Web corpus in several NTCIR tasks: INTENT-1/2 and IMine.

[*] Corresponding Author

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5022-8/17/08…$15.00
http://dx.doi.org/10.1145/3077136.3080694

[1] The ClueWeb12 consists of only English Web pages.

[2] The dataset is free for research purpose. However, the organization who distributes it charges a commission fee (about $435)

[3] http://www.sogou.com/labs/resource/t.php

SogouT-16 is a new version of SogouT-08 and SogouT-12. Similar to its previous versions, the pages of SogouT-16 were sampled from the indexed documents of *Sogou.com*, which is the third largest commercial search engine in China. SogouT-16 will first be applied in an ad-hoc retrieval task, NTCIR-13 *We Want Web* (NTCIR-WWW). The organizers of NTCIR-WWW believe that to provide a testbed for ad-hoc Web search is still of utmost practical importance for academia.

Comparing to previous test collections and datasets, SogouT-16 has several advantages: 1) SogouT-16 is the most recent large-scale Web collection in Chinese. Although the aforementioned datasets (ClueWeb09-zh, SogouT-08/12, and etc.) have already gained much success in supporting corresponding researches, they are somewhat dated and small, given the scale of today's Web. SogouT-16 contains 1.17B Web pages from 2.03M domains. We believe that it is representative of the size and diversity of Chinese Web environment. 2) From NTCIR-INTENT (2010) to NTCIR-WWW (2017), SogouT-08/12/16 have provided stable support for IR related researches. In NTCIR-WWW, we are going to provide relevance judgments and anonymized behavior information (query logs from a commercial search engine) for the queries in the topic set. To help researchers utilize this corpus, we provide Web-based retrieval service for free. 3) SogouT-16 provides several related datasets, such as link structure, word embeddings, and query log of NTCIR-WWW topic set. These datasets together with SogouT-16 allow researchers to investigate a wide range of topics in information science and language technologies.

In the remainder of this paper, we describe how we collect the Web corpus and present a variety of descriptive characteristics of SogouT-16. Then we briefly introduce how we apply it in NTCIR-WWW.

2 DATA COLLECTION

2.1 Sampling Process

The Web pages are sampled from the indexed Web pages of a commercial search engine, *Sogou.com*. We design the sampling strategy based on the following considerations: 1) Diversity: the Web pages should be sampled from a wide range of sites to cover various topics. 2) Quality: Web spams have been obstacles in usersâĂŹ information acquisition process. It was estimated by Wang et al. [10] that approximately one seventh of English Web pages were spam in 2006. We assume that the majority of SogouT-16's users might be researchers from academia. It can be very labor intensive for them to filter the pages of poor quality. On the other hand, it is reasonable that valuable information is more likely to appear on the pages of relatively high quality. We design a sampling algorithm as shown in Algorithm 1.

We can see that the algorithm will iteratively select pages from all the indexed URLs. For a particular URL u, two factors are taken into consideration. The first one is the PageRank value of u [9], which is one of the most important features of a Web page, indicating its relative importance within the World Wide Web. A parameter ω is designed to control the probability that a page will be selected. A positive constant, λ, can make sure that even a page with low PageRank value can be selected with a relatively small probability. The second factor which would have an impact on the probability is the number of selected pages under the same site. For a specific site, as the number of selected pages grows, the probability that a further page will be selected will be penalized. For SogouT-16, we set $\omega = 0.9$, $\lambda = 0.1$.

Algorithm 1 Sampling Algorithm

```
procedure SAMPLE($U$)                                   ▷ $U$: all the indexed URLs
    shuffle $U$
    initial $U_s$                                       ▷ $U_s$: initial the selected URLs
    initial $loop = 0$
    while $loop < 10$ do
        for $u \in U$ do
            if $u \in U_s$ then
                continue
            end if
            $d \leftarrow$ site name of $u$             ▷ get the site name of $u$
            $c_d \leftarrow$ #selected pages of $d$
            $pr_u \leftarrow$ PageRank of $u$           ▷ get PageRank of $u$
            $r \leftarrow rand()$                       ▷ generate random number$\in (0, 1)$
            if $r < \min\left(\omega * \left(pr_u{}^2 + \lambda\right), \frac{1}{\log c_d + 1}\right)$ then
                $U_s$ add $u$                           ▷ select $u$
                $c_d \leftarrow c_d + 1$
            end if
        end for
        $loop \leftarrow loop + 1$
    end while
    return $U_s$
end procedure
```

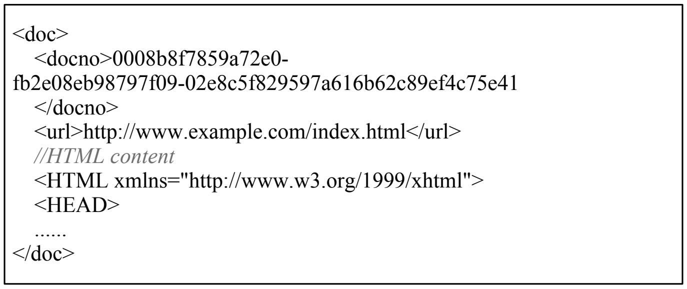
After sampling, we conduct several cleaning process on the Web pages to make the corpus more friendly to researchers: (1) We filter the illegal Web pages including: pornography contents, malware, phishing, spyware and virus-infected pages. (2) We filter the Web pages which are not in Chinese. (3) We detect the encoding of the Html content. All the Html files are converted to UTF-8 if they are originally encoded in other character sets (ASCII, GBK, GB2312, UTF-8, ISO-8895-1).

2.2 Overview of SogouT-16 Dataset

After sampling and data cleaning, SogouT-16 has about 1.17B Web pages from 2.03M domains. The corpus is stored in 10 folders and each folder contains 4,096 zipped files. In each file, the Web pages are concatenated together with two additional fields: DocId and URL. An example of a Web page is shown in Figure 1.

We then compare SogouT-16 with several similar datasets in Table 1. It can be seen that SogouT-16 is almost 9 times larger than its previous version (SogouT-12). It is almost as large as ClueWeb09 (1.04B). It should be noted that CWP200T was collected from 2002 to 2015 and its advantage lies that it contains the histories of Web pages, i.e. different versions of a URL would be treated as different pages. Although CWP200T contains much more pages than SogouT-16, at this moment, we have no idea how many unique URLs it contains. The other two datasets, WT10g and Gov2 are somewhat dated and small given the rapid growth of the Web.

Among all the above Web corpora, SogouT-16/12 and CWP200T consist of mainly Chinese Web pages. ClueWeb09 contains Web



```
<doc>
   <docno>0008b8f7859a72e0-
fb2e08eb98797f09-02e8c5f829597a616b62c89ef4c75e41
   </docno>
   <url>http://www.example.com/index.html</url>
   //HTML content
   <HTML xmlns="http://www.w3.org/1999/xhtml">
   <HEAD>
   ......
</doc>
```

Figure 1: An example of a Web page.

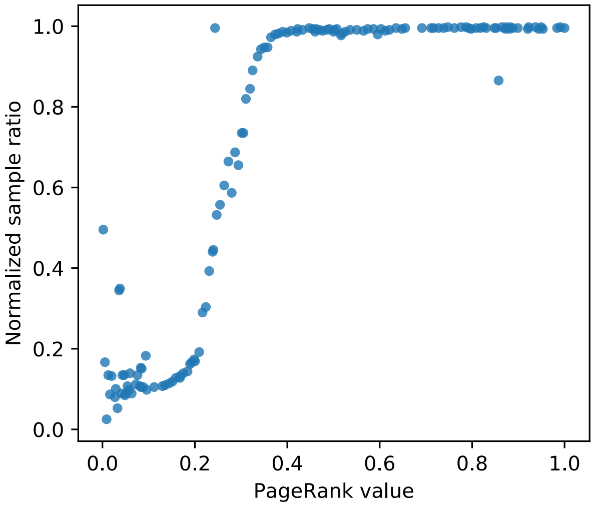


Figure 2: Sample ratio v.s. PageRank value on 150 sites which are randomly selected from SogouT-16.

pages in 10 most popular languages: English (503M), Chinese (177M), and etc. ClueWeb12 intentionally filtered non-English pages.

To investigate whether SogouT-16 meets our proposed considerations about quality and diversity, we then conduct several analyses on the Web pages.

First, we are wondering how the PageRank value would impact the sampling process. 150 sites are randomly selected from SogouT-16. We consider the sample ratios of the corresponding sites, i.e. the ratio of selected pages to all the pages of a particular site. The sample ratios of these sites are then normalized to 0-1 range. The relationship between sample ratio and average PageRank value of the site's pages is illustrated in Figure 2. We can see that when the PageRank value is relatively small (0.0 to 0.4), the pages from corresponding sites are less likely to be selected by our sampling algorithm. The PageRank value has little impact on the sample ratio when it is larger than 0.4, indicating that our algorithm will get a stable proportion of pages from the domains with relatively high PageRank values (0.4 to 1.0).

We then investigate the sample ratios on a larger set of domains. 1200 domains are sampled from SogouT-16. The relationship between the sample ratio and the number of pages under a particular site is illustrated in Figure 3. We can see that the sample ratios tend to converge as the size of the sites grow.

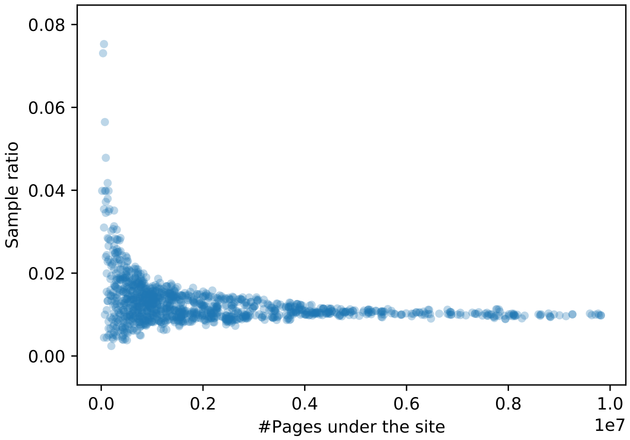


Figure 3: Sample ratio v.s. the number of the site's pages on 1200 sites which are randomly selected from SogouT-16.

The sampling algorithm is actually a trade-off between diversity and quality. The large sites which comprise a lot of pages are usually assumed to provide content of relatively high quality. If these sites get too much benefit during the sampling process, the diversity of the Web corpus may be harmed. According to the above results shown in Figure 2 and Figure 3, SogouT-16 is basically a uniform sampling among the sites of relatively high quality.

2.3 Subset of SogouT-16

Considering that processing large scale of Web pages usually needs a lot of computational resources and efforts. ClueWeb09 and ClueWeb 12 both provided subsets of corresponding datasets for researchers not yet ready to scale up to 1 billion documents. These subsets are referred to as the "Category B" versions of the ClueWeb datasets. They have already supported a number of search tasks, for example, the Web Track and Million Query Track of TREC, and the INTENT-1/2 of NTCIR.

SogouT-16 needs about 81TB before compression. There might be some difficulties for researchers to handling such a large corpus. Following the ClueWeb datasets, we continue the tradition started by TREC evaluations of to provide a smaller subset of SogouT-16 to make it more convenient for some research groups. The ClueWeb09 Category B consists of crawl seeds (2.5M), English Wikipedia (6.0M) and crawled pages (41.8M) while ClueWeb12 is a 5% sample of the full dataset. The subset of SogouT-16, denoted as SogouT-16-B, comprises 177,936,163 unique URLs from 818,182 domains. The compressed file needs about only 1.5TB. This dataset will be applied in the NTCIR-WWW task.

3 WE WANT WEB IN NTCIR-13

Information access tasks have diversified: currently there are various novel tracks/tasks at NTCIR, TREC, CLEF etc. This is in sharp contrast to the early TRECs where there were only a few tracks, where the ad hoc track (a set of new topics run against a static document collection) was at the core. The organizers of We Want Web (WWW) [4] believe that ad hoc Web search, is still of utmost

[4] http://www.thuir.cn/ntcirwww

Table 1: Comparison between SogouT-16 and several similar datasets

Dataset	SogouT-16	SogouT-12	ClueWeb12	ClueWeb09	CWP200T	WT10g	Gov2
#Page	1.17B	0.13B	0.73B	1.04B	7B	1.7M	25M
Language	CHN	CHN	ENG	Multiple	CHN	ENG	ENG
Latest Crawl Date	2016	2012	2012	2009	2015	1997	2004

practical importance. From the participants' view, they will have a stable environment in which they can evaluate ad hoc web search and monitor progress across rounds. From the organizers' view, statistically-motivated methods can be established for designing and maintaining test collections and for quantifying progress.

The WWW task is a traditional ad hoc task and have already been accepted as a formal task by NTCIR-13. This round of WWW will have English and Chinese subtask and each of them has a topic set of 100 queries covering different intents (informational and navigational queries). The Chinese queries are sampled from a commercial search engine's query log. For English queries, 60 of them are sampled from query log and 40 of them are translated from Chinese topic set, which allow further analysis on cross-language IR topics. One common subset of the topic set in WWW will be "frozen", i.e. the relevance assessments will be disclosed only after NTCIR-15 (after three rounds of WWW). This will enable studies about the longevity of test collection and standardization factors.

By the end of February 2017, NTCIR-WWW has 18 participants and it is the second largest NTCIR-13 task.

4 RELATED SERVICES AND DATASETS

To help researchers use SogouT-16 more easily, we also provide some services. Similar to the ClueWeb datasets, we will provide online retrieval service for individual query and batch of queries. The retrieval system is built based on an open-source search platform, `Apache Solr`[5]. For each query, the retrieval system could return a retrieved list which comprises at most 1000 documents and their scores. We also plan to provide a "DocId to Document" in the near future. To the best of our knowledge, this is the first Chinese Web corpus which provide online retrieval service. We believe these service would help SogouT-16's users find relevant information in the corpus more conveniently.

Together with SogouT-16, we also release several related datasets to accelerate the pace of researches.

- SogouT-16 Link: This dataset is the link structure of the Web pages in SogouT-16. For each page, we extract all the outlinks using a Python library, `BeautifulSoup`[6]. The mapping between URLs and DocIds is also provided. This dataset could potentially support researches such as web graph analysis, spam detection and etc.
- SogouT-16 Embeddings: We release a Chinese word embedding trained on a corpus which is much larger than SogouT-16. It can power many aspects of research, such as sentiment analysis, recommendation, and etc.
- SogouT-16 WWW: For NTCIR-WWW, we will release the relevance judgments for the queries which are not in the "frozen" dataset. This data will get ready after all the participants submit their runs. This data may support researches about ranking and search evaluation.
- SogouT-16 Qlog: For the queries in NTCIR-WWW's topic set, we will release anonymized query log. Actual users' behavior is often unavailable in academia. We believe this behavior dataset will encourage a lot of researches.

5 CONCLUSION AND FUTURE WORK

In this paper, we present a new Chinese Web corpus, SogouT-16, which is sampled from a search engines' indexed documents by considering both quality and diversity. Together with SogouT-16, we also provide several online services and related datasets to support a wide range of research topics. It will be applied in an ad hoc search task, NTCIR-WWW. In the future work, we will continue to investigate how to use SogouT-16 to encourage innovations in corresponding research topics.

6 ACKNOWLEDGMENTS

This work was supported by Natural Science Foundation (61622208, 61532011, 61472206) of China and National Key Basic Research Program (2015CB358700).

REFERENCES

[1] 2012. The ClueWeb12 Dataset – The Lemur Project. http://www.lemurproject.org/clueweb12.php. (2012). Online; Accessed: 2017-02-01.

[2] 2016. CWP200T Dataset. http://www.ccf.org.cn/sites/ccf/xhdtnry.jsp?contentId=2937064120111. (2016). Online; Accessed: 2017-02-01.

[3] John A Akinyemi and Charles LA Clarke. 2011. UWaterloo at NTCIR-9: Intent discovery with anchor text.. In *NTCIR*.

[4] Peter Bailey, Nick Craswell, and David Hawking. 2003. Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing & Management* 39, 6 (2003), 853–871.

[5] Chris Buckley, Darrin Dimmick, Ian Soboroff, and Ellen Voorhees. 2007. Bias and the limits of pooling for large collections. *Information Retrieval* 10, 6 (2007), 491–508.

[6] Cyril W Cleverdon and Michael Keen. 1966. Aslib Cranfield research project-Factors determining the performance of indexing systems; Volume 2, Test results. (1966).

[7] David Hawking, Ellen Voorhees, Nick Craswell, and Peter Bailey. 1999. Overview of the TREC-8 web track. In *TREC*.

[8] Yiqun Liu, Fei Chen, Weize Kong, Huijia Yu, Min Zhang, Shaoping Ma, and Liyun Ru. 2012. Identifying web spam with the wisdom of the crowds. *ACM Transactions on the Web (TWEB)* 6, 1 (2012), 2.

[9] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

[10] Yi-Min Wang, Ming Ma, Yuan Niu, and Hao Chen. Spam double-funnel: Connecting web spammers with advertisers. In *WWW '07*.

[11] Qi Zhang, Jihua Kang, Jin Qian, and Xuanjing Huang. Continuous word embeddings for detecting local text reuses at the semantic level. In *SIGIR '14*.

[12] Yu Zhao and Maosong Sun. 2013. Exploiting Lexicalized Statistical Patterns in Chinese Linguistic Analysis. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Springer, 238–246.

[5] http://lucene.apache.org/solr/

[6] https://www.crummy.com/software/BeautifulSoup/