ClueWeb22: 10 Billion Web Documents with Rich Information

Arnold Overwijk
arnold.overwijk@microsoft.com
Microsoft

Chenyan Xiong
chenyan.xiong@microsoft.com
Microsoft

Jamie Callan
callan@cs.cmu.edu
Carnegie Mellon University

ABSTRACT

ClueWeb22, the newest iteration of the ClueWeb line of datasets, is the result of more than a year of collaboration between industry and academia. Its design is influenced by the research needs of the academic community and the real-world needs of large-scale industry systems. Compared with earlier ClueWeb datasets, the ClueWeb22 corpus is larger, more varied, and has higher-quality documents. Its core is raw HTML, but it includes clean text versions of documents to lower the barrier to entry. Several aspects of ClueWeb22 are available to the research community for the first time at this scale, for example, visual representations of rendered web pages, parsed structured information from the HTML document, and the alignment of document distributions (domains, languages, and topics) to commercial web search.

This talk shares the design and construction of ClueWeb22, and discusses its new features. We believe this newer, larger, and richer ClueWeb corpus will enable and support a broad range of research in IR, NLP, and deep learning.

ACM Reference Format:
Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. ClueWeb22: 10 Billion Web Documents with Rich Information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3477495.3536321

1 INTRODUCTION

Many important research topics in information retrieval are centered around web search. One of the most used AI applications, web search provides research challenges such as user understanding, search relevance, indexing and serving efficiency, to name a few. Many of these research challenges require a realistic web corpus to explore new ideas and benchmark new solutions. ClueWeb09 [11] and ClueWeb12 [12] have been the standard web corpora for much IR research during the last decade (e.g., [2, 3]), but they are becoming outdated. The web has evolved during the last decade, and our web datasets must evolve with it.

Much recent IR research has shifted to the MSMARCO benchmark [1], which has a large amount of relevance labels released from Bing, but only contains a smaller amount of web pages, fewer than 10 million, in comparison to 733 million in ClueWeb12. In addition, MSMARCO documents are selected from top search results of Bing QA queries, which do not align with the actual document distribution on the web, even in Bing.

Table 1: Size and sampling distribution of ClueWeb22.

Category	#Pages	Sampling Distribution
ClueWeb22-B	200M	Most Popular Web Pages
ClueWeb22-A	2B	Mainly Head Pages on Web Search
ClueWeb22-L	10B	Mixed Head-Tail Pages on Web Search

Another increasingly important usage of web corpora is to pretrain large scale language models. As the amount of language model parameters has grown from hundreds of millions to trillions in the past four years [4, 6], their pretraining also requires significantly more text data than well-processed corpora, e.g., from Wikipedia and news. Many recent large models sought more pretraining data from the web. One approach is to sift CommonCrawl data, e.g., the C4 dataset used to pretrain T5 [10], which provides sufficient quantity, but the quality quickly becomes a concern. For example, the cleaned CommonCrawl reflects a quite weird distribution of the web [5]. Language models pretrained on C4 often perform worse than models pretrained on higher quality corpora at the same scale.

With ClueWeb22, we aim to provide the web corpus for research in the near future. The design of ClueWeb22 emphasizes on these goals: 1) to reflect the *distribution* of the web in real scenarios; 2) to provide web pages at large *quantity* and also with high *quality*; 3) to enable new research directions by including information important in industry but previously not publicly available.

In the construction of ClueWeb22, we sample HTML web pages from the Bing index based on the distribution of estimated importance of pages on the web. Specifically, a page more likely to satisfy potential information needs from search engine users received a higher importance score, thus is assigned higher probability in our sampling distribution. Pages that are of lower quality are demoted in our sampling and spam pages are filtered. We also ensure the high coverage of the most important part of the web, often referred to as “head” in the web corpus. In total, we sampled 10 billion web pages from the indexed web pages in Bing and grouped them into three “categories” after spam and adult content filtering, as listed in Table 1. For all three categories, about half of them are in English and the rest are not. These categories follow the tradition of ClueWeb09 while also closely mimicking real scenarios used in web search.

(1) ClueWeb22-B (“Category B”) samples from the most frequently visited part of the web, e.g., main parts of Wikipedia, news websites, and other top internet domains. It represents the highest quality part of the web.
(2) ClueWeb22-A (“Category A”) approximates the general regularly visited part of the web through search. It includes two billion web pages and covers most notable URL domains.
(3) ClueWeb22-L (“Category L”) mixes in the “tail” part of the web into the collection, with a total of ten billion web pages.

SIGIR ’22, July 11–15, 2022, Madrid, Spain
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3536321

Table 2: Information of web pages included in ClueWeb22, obtained by state-of-the-art production-quality pipelines.

Information	Avail. in Category	Description
Raw HTML	A, B	The original, unprocessed HTML of the web page
Clean Text	A, B, L	Primary text content, i.e., without headers, footers, side bar, navigation panel, etc.
Semantic Annotations	A, B	Annotations of content structure: title, section headings, paragraphs, lists and tables
VDOM Features	A, B	The virtual representation features of the content, e.g., location and size of each text piece [13]
Topic Tag	A, B, L	The category classified by a supervised topic classifier
Language Tag	A, B, L	The language of the main content, detected by an updated version of BlingFire [9]
Image-Text Pair	A, B	The parsed image URL and their corresponding texts in the web page
Rendered Visual Page	A	The screen shot of the rendered web page as it appears to users

The three categories provide different trade-offs of quality and coverage. Each one is a subset of the larger ones, $B \subset A \subset L$, providing researchers three choices for different research focus. Together they form the largest available, spam-filtered, clean web corpus to date, designed to mimic the distribution of web pages in commercial search systems.

In addition to raw HTML data, we also include a variety of information produced by document analysis. Much of this information provides rich signals for web search production systems, but is not previously available for the research community. Table 2 shows the information available for each web page. The additional information facilitate a variety of research directions. For example, the Clean Text is the parsed primary contents of the web page, which can significantly reduce the engineering effort required for academics to work with web documents. More importantly, many of these document features enable research in directions that are important in real world practice but previously nearly impossible without available public data. For example, the semi-structured contents, such as tables and lists, are critical parts for modern QA systems, but most previous research is restricted to Wikipedia due to limited data availability [8]. the VDOM information is widely used in document understanding in industry but there were no such data at this scale available publicly [13]. We are also excited to see more research on the multi-modality front with the Image-Text pairs and Rendered Visual Pages.

ClueWeb22 is also a multi-lingual corpus, with more than one hundred languages discovered in the web. The language distribution closely follows their frequencies in our samples of the web. To the best of our knowledge, this is the first widely-available multi-lingual corpus at this scale and quality. We believe it will encourage more research on multi-lingual scenarios, rather than concentrating research attention on the most popular language.

ClueWeb22 is scheduled to be available in summer 2022, following the distribution patterns and research license of previous ClueWeb datasets. We also have plans in building benchmarks for varied IR tasks upon this valuable corpus and are engaging with the research community for further developments.

2 GOALS OF THIS TALK

The main goal of having a talk about ClueWeb22 at SIGIP is to share more details and potential usages of ClueWeb22, and to gather feedback from community on further improvements. For example, we would like to share our approaches to ensure ClueWeb22 reflects the technology needs from industry and also the future research we envisioned. We will share various existing and potential use cases of ClueWeb (and the corresponding part of real production data) in both academic research and industry applications.

We also plan to gather feedback using this talk and through SIGIP. We sought feedback from the community when constructing ClueWeb22. Its current form reflects the beliefs of us and several academic experts about the web corpus needed by current and future research. With this talk we would like gather feedback from a larger group of members in our community on what they would like us to potentially update, modify, or preprocess in ClueWeb22. Having an early understanding of what tasks people are interested to work on with ClueWeb will also be beneficial for our follow up efforts in building benchmarks with ClueWeb.

3 COMPANY PORTRAIT

Microsoft is one of the leaders in information retrieval, both in industry products and academic research. We cover perhaps the most diverse search scenarios and provide industry-leading solutions in nearly all fronts, for example, in Bing, Office search, Azure search, Edge recommendation, to name a few. Members of Microsoft, both from production teams and research labs, appear frequently at IR venues and have contributed significantly to IR technologies.

One big differentiator of Microsoft compared to other members of the industry is the openness of the company to academia. Besides publications, Microsoft has released various datasets and benchmarks that have significantly impacted the research and development of the community. A recent notable example is MSMARCO, which has becomes the go-to benchmark and training corpus for neural information retrieval [1]. ClueWeb22 marks another step forward in the tight collaboration between Microsoft and academia.

4 PRESENTER BIO

Arnold Overwijk is leading the Document Understanding area at Microsoft. His team is responsible for developing state-of-the-art representation learning models, such as ANCE [14] and SEED-Encoder [7], as well as document layout understanding for structured information extraction and NLP fundamentals like extreme classification for topic modeling, key phrase extraction [13] and fast tokenization [9]. These models are powering a wide range of product scenarios within Microsoft, including: web, multimedia and enterprise search, question answering, content recommendation, entity pane, people also ask and more. Prior to Microsoft, Arnold got a graduate degree from Carnegie Mellon University and his undergrad from the University of Twente in the Netherlands.

ACKNOWLEDGMENTS

This research was supported in part by NSF grants CNS-1822975 and CNS-1822986. Any opinions, findings, and conclusions in this paper are the authors’ and do not necessarily reflect those of the sponsors.

REFERENCES

[1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).

[2] Charles L Clarke, Nick Craswell, and Ian Soboroff. 2009. *Overview of the TREC 2009 Web Track.* Technical Report. NIST.

[3] Charles L Clarke, Nick Craswell, and Ellen M Voorhees. 2012. *Overview of the TREC 2012 Web Track.* Technical Report. NIST.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2019.* 4171–4186.

[5] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758* (2021).

[6] William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv preprint arXiv:2101.03961* (2021).

[7] Shuqi Lu, Di He, Chenyan Xiong, Guolin Ke, Waleed Malik, Zhicheng Dou, Paul Bennett, Tieyan Liu, and Arnold Overwijk. 2021. Less is More: Pretrain a Strong Siamese Encoder for Dense Text Retrieval Using a Weak Decoder. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021.*

[8] Kaixin Ma, Hao Cheng, Xiaodong Liu, Eric Nyberg, and Jianfeng Gao. 2021. Open Domain Question Answering over Virtual Documents: A Unified Approach for Data and Text. *arXiv preprint arXiv:2110.08417* (2021).

[9] Microsoft. 2019. BlingFire. https://github.com/microsoft/BlingFire

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[11] Carnegie Mellon University. 2009. ClueWeb09. http://lemurproject.org/clueweb09/

[12] Carnegie Mellon University. 2012. ClueWeb12. http://lemurproject.org/clueweb12/

[13] Lee Xiong, Chuan Hu, Chenyan Xiong, Daniel Campos, and Arnold Overwijk. 2019. Open Domain Web Keyphrase Extraction Beyond Language Modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP 2019.* http://arxiv.org/abs/1911.02671

[14] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *International Conference on Learning Representations, ICLR 2021.*