Vision

Empirical investigation of how use of the Web is changing provides crucial input to decisions about its future. Common Crawl (CC) has played a key role in the development of Large Language Models (LLMs) and the Generative AI applications such as ChatGPT based on them (GenAI). This project will study how recent CC datasets reflect the shift away from human-authored content on the Web.

Importance: As Tim Berners-Lee once said, “We have to study the Web so we don’t break it by mistake”. There is a lot of low-level empirical data about Internet traffic, and a fair amount about Web traffic volume on a per-host basis, but much less is known about how the pages on the Web themselves are evolving. Addressing this deficit has recently become much more important as the Web, and in particular very large-scale Web archives, are the primary sources of language data for training LLM-based AI systems. As more and more of the Web is itself generated by such systems, there is an existential risk of a spiralling degradation of quality.

Methodological advance: By giving concrete examples of what can be achieved by applying High Performance Computing resources to very large-scale web archive data, and by explaining how this can be replicated and why it can be trusted, we aim to enable and encourage study of the way Web content is changing.

Policy influence: Regulation of GenAI is needed to protect its own future. Our work aims to produce evidence of the growing impact of GenAI on the Web, which will incentivise and inform such regulation.

Scientific importance: Applying the methodology we have developed for empirical measurements of the reliability of results derived from Common Crawl datasets to the problem of detecting GenAI outputs is crucial for supporting the argument for regulation.

Scope: Involve 5 MSc students with a project within the proposed research, training them on the use of HPC tooling for work with very-large-scale Web datasets

Context: The proposed research builds on our previous work in this area, in particular our completed projects **LURID2: Assessing the validity of Common Crawl**, EPSRC Access to HPC Award, as reported in [Thompson 2024]* and **LURID3: Longitudinal studies of the World Wide Web**, UKRI APP39557.

We want to exploit the augmented versions of the Common Crawl datasets produced by that work, using the computational resource Cirrus provides, in order to quantify the extent to which GenAI outputs are appearing on the Web.

Common Crawl (CC) is a very-large-scale web archive, totalling over 120 billion web pages. Collection began in 2008 with annual archives, expanding steadily to the point that recent archives contain over 3x10^9 pages, (about 75 Terabytes (compressed). We have local copies of 10 months of CC in a petabyte store attached to Cirrus.

The very large size of CC's constituent archives makes using it for research, particularly longitudinal studies, which necessarily involve multiple archives, very expensive in terms of compute time and storage space and/or web bandwidth. The proposed work will build on the results of **LURID2** and **LURID3** (see R4RI for details) to address this problem. Our work began to quantify the extent to which Web content is automatically generated “just in time” from data, reducing the availability of high-quality human-authored information.

Overall goal: The enormous success of GenAI is based on Web content in general, and CC in particular. However, their very success poses a significant risk to the value of CC and other very large-scale archives in the future. We need a sound empirical baseline to enable the detection of the rapidly growing amount of auto-generated content on the Web, in order to avoid a vicious cycle of the degradation which would be caused by models being trained on their own output.

*See “Applicant and Team capability” section for references

Approach

Objectives

1. Expand the empirical basis for longitudinal study of CC

We have demonstrated in **LURID2** and **LURID3** that augmenting the index of a CC dataset with Last-Modified timestamps enables fine-grained longitudinal research of Web content. We propose to create similarly augmented indices for further recent CC datasets, so that there are indices for datasets every six months from 2022 to date.

2. Ameliorate the decline of high-quality human-authored content in CC

CC's original policy of truncating each individual result to 1MB renders much of its already limited PDF content (around 1% of the whole) completely unusable. This is problematic because we found some evidence that PDF content is more likely to use persistent identifiers for linking than HTML content, consistent with the use of PDF for publishing scholarly content. We have already re-fetched all the truncated PDFs from one previous CC dataset: we propose to re-fetch the truncated PDFs for further CC datasets, some before and some after the launch of ChatGPT in 2022, and to take advantage of the increase in maximum size of PDFs in CC datasets since early 2025.

3. Studies based on (1) and (2)

Between January and May, when MSc projects start, I will build the additional resources (augmented indices and PDFs) described above.

In the May and June I will introduce the students to the Cirrus environment and the tooling we have built up to date, using small pilot projects:

- Replicate our published study of URI length using more data;
- Explore the evolution of primary natural language detected;
- Explore the evolution of declared versus detected media type;
- Document the rate of 'bit-rot' incurred in re-fetching older PDFs;
- Document and compare the growth in URI length of vs. within retrieved PDFs.

The students will then each undertake their own projects, each based on a different existing approach to detecting GenAI output, comparing their performance on material from different dates and/or in different formats (HTML vs PDF).

After project submission in August, I'll refine and extend the most promising work and undertake the impact work listed below.

Environment and methods

Cirrus provides the resources, both in terms of compute power and disk storage, to conduct work on the scale required. The tooling and experience we have built up while using Cirrus for our previous work mean that we have everything necessary to implement the above studies with a minimum of overhead in terms of designing new experimental pipelines or writing new code.

Impact

We will publish the additional augmented indices via the Edinburgh International Data Facility and publicise their availability via the CC online forum and a paper submitted to WebSci27 presenting the best results from the MSc work and my follow-up.

I will propose and organise a one-day meeting at the Turing Institute on the topic of the existential risk posed by GenAI by training on its own outputs.

We will also offer a workshop at WebSci27 on how our techniques and published resources can facilitate use of CC for a range of research projects and applications.

We will prepare and submit a Response-mode grant application to EPSRC to use a combination of our outputs, compared to forthcoming CC datasets, to explore unsupervised learning approaches to detecting AI-generated Web content.