LWAC: Longitudinal Web-as-Corpus Sampling

Stephen Wattam[1] **Paul Rayson**[1] **Damon Berridge**[2]

Lancaster University
School of Computing and Communications[1], Mathematics and Statistics[2]
f.lastname@lancs.ac.uk

1 Sampling

Many sampling efforts for linguistic data on the web are heavily focused on producing results comparable to conventional corpora. These typically take two forms: those based on URI lists (e.g. from search results, as in (Sharoff, 2006), BE06 (Baker, 2009), BootCat (Baroni and Bernardini, 2004)), and those formed through crawling (e.g. enTenTen (TLCME), UKWaC (Ferraresi et al., 2008)).

Though initial efforts in web-as-corpus (WaC) first focused on the former method many projects are now focused on constructing supercorpora, which may themselves be searched with greater precision than the 'raw' web, in line with Kilgarriff's vision of linguistic search engines (Kilgarriff, 2003). This has led to the proliferation of crawlers such as those used in (Schäfer and Bildhauer, 2012) and (Remus, 2008).

This approach, with its base in a continually-growing supercorpus, parallels the strategy of a monitor corpus (Sinclair, 1982), and is applicable to linguistic inquiry concerned with diachronic properties (Koehler, 2006). Indeed, we could conclude that the web is mature enough to require date-based lookup in casual use, and such tools are increasingly being included in consumer search engines such as Google.

This crawling approach to repeated sampling tells us about the state of language change online in a manner that is comparable to other diachronic corpora, however, it omits subtler technical aspects that govern consumption of data online, most notably the user's impression of its location, as defined by the URI. Low publishing costs online, paired with increasing corporate oversight and reputation management, (both personal (Meyerdee and Smith, 2010) and professional (Malaga, 2001)) , leads to a situation where this content is being revised frequently, often without users even noticing.

This URI-oriented change has been studied from a technical perspective by those interested in managing and maintaining network infrastructure, compiling digital libraries (Tyler and McNeil, 2003), and optimising the maintenance of search engine databases (Koehler, 2004). The needs of these parties are quite aside from those of corpus researchers, however, since they focus around a best-effort database of information, rather than a dependable longitudinal sample with known margins for error.

We present here a tool, LWAC, for this form of longitudinal sampling, designed to maximise the comparability of documents downloaded in each sample in terms of their URI rather than content. To accomplish this, we use a batch-mode sampling strategy, as illustrated in Figure 1, to get full coverage over a list of URIs, at the expense of sampling new content.

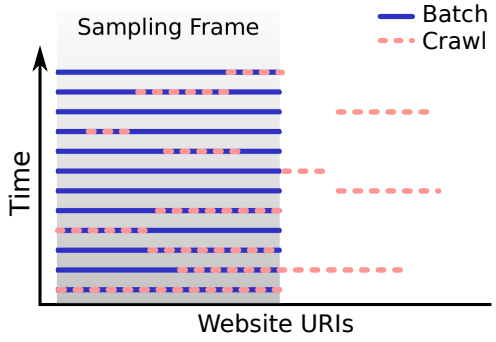

Figure 1: URI coverage for batch and crawl.

2 Applications

This strategy allows us to investigate how language may change in relation to technical and social events in a way that mimics the experience of the end user, and offers a useful perspective on many epistemic problems of WaC methods, to determine:

- The portions of web pages that typically change as main content regions;
- The impact of social feedback and user generated content on page content;
- How censorship, redaction and revision affect website contents;
- Website resource persistence and its relation to linguistic content (link rot/document attrition);
- How institutions' publishing policies affect reporting of current events.

In order to maximise its coverage of these topics, LWAC is designed to construct longitudinal samples from arbitrary URI lists, using commodity hardware, in a way that mimics the user's experience of a website.

3 Architecture

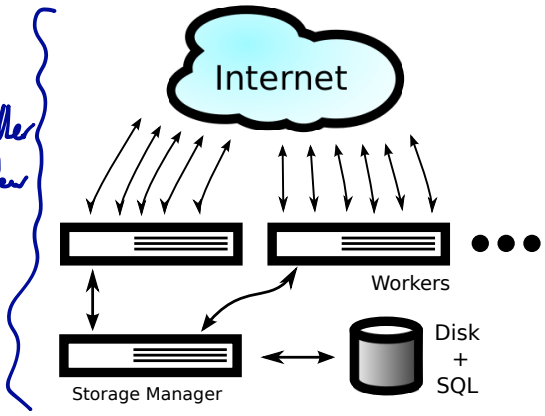

Figure 2: System Architecture

In order to form a useful longitudinal sample, each data point should be as cross-sectional as possible. As such, a highly parallel, distributed architecture was selected (Figure 2). This yields technical benefits in terms of throughput (especially where the internet connection is a bottleneck), flexibility, and the ability to differentiate between websites that are blocked for a given area of the internet and those that are offline 'proper'.

Data storage in the system is split between metadata, stored in an SQLite database, and website sample data itself, which is stored as raw HTTP response data in a versioned structure, compressed using hard links[1]. The storage format is optimised for large samples, and is nested in order to avoid common filesystem limits.

The download process itself is managed by a central server, which co-ordinates storage and metadata access in order to provide full atomicity and recording of sample data. This central server distributes batch jobs, according to policies governing reliability and throughput, to worker servers, which compete for the opportunity to download websites.

In order to avoid search engine optimisation tricks, workers imitate the behaviour of end users' browsers: they retain cookies and present typical user-agent and referrer strings in their request headers.

4 Performance

In order to obtain the most simultaneous samples, the system was designed to maximise the parallel number of connections on each client. This is limited by the underlying OS, which in our tests showed a practical maximum of 120 simultaneous downloads[2].

In practice, throughput is limited by several factors, among them the available bandwidth, number of worker servers, speed of DNS lookups, and the proportion of links which are destined to time out during connection. In testing, each worker proved capable of downloading at a sustained 8-12MBps, depending on job size and remote server properties.

5 Conclusion

The LWAC sampling tool, available online[3], offers an easy and rigorous way to compile longitudinal web corpora from arbitrary URI lists. We believe it has particular utility to investigation of challenges that face WaC methods, as well as fine-grained sampling of language linked to current events and other fast-moving phenomena.

References

[Baker2009] Paul Baker. 2009. The BE06 corpus of British English and recent language change. *International journal of corpus linguistics*, 14(3):312–337.

[Baroni and Bernardini2004] Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora

[1] Linked in a similar manner to `rsync`'s `-H` option.

[2] Using the Linux 2.8 kernel

[3] `http://stephenwattam.com/projects/LWAC`

and terms from the web. In *Proceedings of LREC*, volume 4.

[Ferraresi et al.2008] Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating UKWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

[Kehoe2006] Andrew Kehoe. 2006. Diachronic linguistic analysis on the web with WebCorp. *Language and Computers*, 55(1):297–307.

[Kilgarriff2003] Adam Kilgarriff. 2003. Linguistic search engine. In *proceedings of Workshop on Shallow Processing of Large Corpora (SProLaC 2003)*, pages 53–58.

[Koehler2004] Wallace Koehler. 2004. A longitudinal study of web pages continued: a consideration of document persistence. *Information Research*, 9(2).

[Madden and Smith2010] Mary Madden and Aaron Smith. 2010. Reputation management and social media.

[Malaga2001] RossA. Malaga. 2001. Web-based reputation management systems: Problems and suggested solutions. *Electronic Commerce Research*, 1:403–417.

[Renouf2003] Antoinette Renouf. 2003. WebCorp: providing a renewable data source for corpus linguists. *Language and Computers*, 48(1):39–58.

[Schäfer and Bildhauer2012] Roland Schäfer and Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of LREC*, volume 8.

[Sharoff2006] Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. *WaCky Working papers on the Web as Corpus*, pages 63–98.

[Sinclair1982] John Sinclair. 1982. Reflection on computer corpora in English language research. *Computer Corpora in English Language Research*, pages 1–6.

[Tyler and McNeil2003] David C Tyler and Beth McNeil. 2003. Librarians and link rot: a comparative analysis with some methodological considerations. *portal: Libraries and the Academy*, 3(4):615–632.

→ "LREC Proc. Volume 4."

Does the template specify full first names? If not, reduce to initial to save space

→ connect to footnote to website link

save one line by shortening

→ "LREC Proc. Volume 8"

Does template allow citations to be numbered brackets rather than names eg [7]
If so, this will save space.
The bracketed numbers should also appear in the running text thus saving more space.