Document Attrition in Web Corpora: an Exploration

Stephen Wattam[1], Paul Rayson[1] & Damon Berridge[2]

School of Computing and Communications[1], Mathematics and Statistics[2]
Lancaster University
{s.wattam, p.rayson, d.berridge}@lancaster.ac.uk

Abstract

Increases in the use of web data for corpus-building, coupled with the use of specialist, single-use corpora, make for an increasing reliance on language that changes quickly, affecting the long-term validity of studies based on these methods. This 'drift' through time affects both users of open-source corpora and those attempting to interpret the results of studies based on web data.

The attrition of documents online, also called link rot or document half-life, has been studied many times for the purposes of optimising search engine web crawlers, producing robust and reliable archival systems, and ensuring the integrity of distributed information stores, however, the affect that attrition has upon corpora of varying construction remains largely unknown.

This paper presents a preliminary investigation into the differences in attrition rate between corpora selected using different corpus construction methods. It represents the first step in a larger longitudinal analysis, and as such presents URI-based content clues, chosen to relate to studies from other areas. The ultimate goal of this larger study is to produce a detailed enumeration of the primary biases online, and identify sampling strategies which control and minimise unwanted effects of document attrition.

Keywords: Web-as-Corpus, Sampling, Corpora

1. Introduction

Those using corpus data are increasingly turning to the web as a source of language data. This is not surprising given the vast quantities of downloadable data that are readily available online. The Web as Corpus (WaC) paradigm(Kilgarriff and Grefenstette, 2003) has become popular for compilers of corpora for lexicographic use, replication of standard reference corpora as well as for studies of specific online varieties of language.

Two general models of WaC have emerged. Using the first model, 'browsing', corpus compilers collect data from a set of given domains and select whole or part texts online and incorporate them into their corpora (e.g. the BE06(Baker, 2009) corpus comprising material published in paper form but found on the web).

The second model, 'searching', sees the web through the lens of search engines, and is typically used to compile domain-specific corpora from a set of seed terms which are used to locate web pages for incorporation into a corpus (e.g. the BootCat & WebBootCat tools (Baroni and Bernardini, 2004)). In some cases, both approaches are combined, using searching for general language seed terms to produce reference corpora(Kilgarriff et al., 2010).

Collecting corpora online raises legal questions regarding redistribution rights. Consequently, many compilers choose to make data available only through a web interface (with restricted access for fair-use) or by distributing URI lists (known as open-source corpora(Sharoff, 2006b)).

Online content changes much faster due to the decentralised, open publishing model of the web, which may have a serious impact on two aspects of the WaC paradigm: availability and replicability.

If websites change, the URI lists need updating to reflect new locations. Worse, websites or pages may be completely removed, thus the corresponding part of a corpus is no longer available. This attrition of documents through time affects both users of open-source corpora and those attempting to interpret the results of studies using corpora that were built just a few years ago.

2. Background

Though many studies have looked at the life-cycle of web pages in general, these typically focus on the integrity of websites or specific repositories of information, rather than the documents and the language contained within.

Koehler (2002), through four years of weekly sampling, found that just 66% of their original seed URIs remain online after a year, with this proportion dropping to around 31% by the end of the study. Koehler started his study in the early days of the web (December 1996) using a relatively small sample of only 361 URIs. His analysis found differences in the type of page as well as variation across top-level domains (TLD).

Nelson and Allen (2002) found that only 3% of documents in digital libraries become unavailable in just over a year. This is perhaps unsurprising given the aim of such projects but serves to illustrate the degree of heterogeneity between types of document host.

Studies involving academic paper availability mirror open source corpora in that they use references in favour of original text, however, the centralised administration of academic repositories is notably in contrast to most web resources. Nevertheless, there has been much work into this area, some of which is comparatively reviewed in a paper by Sanderson et al. (2011). These studies, spanning years from 1993 to 2008, illustrate that even institutions charged with keeping an accessible record of information are still subject to rates of attrition in the region of 25-45% over five years.

Despite these enquiries, very little work has been carried out to estimate the effects that document attrition has upon corpus content. Sharoff (2006a) touches upon this in his WaC work , presenting a preliminary analysis of attrition within corpora generated by searching for 4-tuples. Although his studies lasted from one to five months, and contained modest sample sizes (1000), they indicate a rate of loss that is below that of other studies (just 20% per year in the month-long study), suggesting that the selection of documents may have significant effects upon document attrition rates.

Rather than analysing web resources in isolation of their linguistic uses, we outline a preliminary analysis of what we term 'document attrition' relative to a number of corpora of differing construction. We do this by comparing each corpus' construction with a series of URI-based explanatory variables, as part of a larger longitudinal study that will go on to use full text features in order to identify linguistic influence and trends upon web-based corpora and document attrition as a whole.

3. Data and Methods

In order to measure document attrition across a number of linguistic sources we selected a series of corpora, chosen due to their differing constructions and ages, and downloaded them using a process that closely approximates an end user's view of the web. Statistics on the availability of these documents were then annotated with a series of URI-related variables for analysis.

3.1. Data Summary

Data were taken from four open-source corpora (outlined in Table 1), each of which consist of a sample of URIs referring to web resources. All of these corpora are cross-sectional, representing data from a short period, however, only the BootCat-based corpora are built using a script that is likely to sample quickly enough to count as a true point sample in the context of this study.

BE06 was built as a conventional, hand-selected corpus designed as an update to the LOB(Johansson, 1980) and FLOB(Hundt et al., 1998) corpora. It contains texts from sources published in 2006 but also available online.

The Delicious corpus represents a sample of links posted to the front page of delicious.com[2] during the whole of September 2009.

Sharoff's corpus is the same one used in his 2006 paper on open-source corpora, and is built using modified BootCat scripts from a series of 4-tuple seed terms selected from the British National Corpus.

The BootCat corpus is built for this study from 4-tuples that are themselves built from the same terms as Sharoff uses for his 2006 study, using updated versions of his original scripts[3].

3.2. Downloading Process

The process of recording the document's status was relatively simple: a small piece of custom software was written to download documents from an open-source corpus at regular intervals. This tool was configured to mimic requests made by common web browsing software in order to emulate a typical user's visit to the document. Handling SSL, cookies, and referrer links in a similar manner to a user following a bookmark allows us to assess more accurately the content, avoiding tricks that exploit search engine crawlers and other bots.

Taking this user perspective, the notion of document availability becomes slightly more complex. Redirect requests were followed up to a depth of 5[4]. Since we do not account for content changing in this study, failure was taken as receiving a HTTP status code other than 200, or a network timeout (60 seconds was the timeout used for DNS and TCP)[5].

Both metadata and original response details are stored by the download software. This study will focus on features of URIs, such as the presence of GET arguments[6] in the URI string, meaning that the resource is likely to be dynamic. These features have been chosen to indicate aspects of web hosting and affiliation that are likely to vary between users with both different reasons for uploading their content, and different degrees of technical expertise.

4. Preliminary Results

Table 2 describes the availability of documents within the corpora, as sampled on the 21st October 2011. This forms two data points, the former representing no attrition when the corpus was first compiled. Document lifetime statistics are calculated assuming exponential decay: both the half-life ($t_{1/2}$, the time it takes for only half of the original corpus to remain available) and the mean lifetime, τ, are provided.

Corpus	Age (yr.)	Loss	$t_{1/2}$ (yr.)	τ (yr.)
BE06	5.3	42%	6.5	15.8
Delicious	2.1	7%	17.8	42.4
Sharoff	5.3	34%	8.6	16.4
BootCat	.08	0.8%	4.8	20.4

Table 2: Loss from corpus inception to October 21st, 2011.

The large differences in corpus half-life are revealing — the Delicious corpus has significantly lower loss than the others. This is ostensibly owing to its construction: users are likely to bookmark resources that are useful (and hence are well-established, popular sites), in contrast to BootCat's uncritical selection or the deliberate *document*-seeking (rather than *information*-seeking) represented by BE06.

The difference in half-life between Sharoff's corpus and our own BootCat-derived one is harder to explain. Though both are large corpora built using similar methods, they were

[2] A social bookmarking site where users post and exchange links.

[3] As published online at `http://corpus.leeds.ac.uk/internet.html`

[4] As recommended in the HTTP specification and commonly implemented in browsers

[5] Timeout errors also occur stochastically due to routing policies, and are impossible to avoid entirely when downloading resources in bulk. The download process was tuned to minimise this source of error.

[6] Parameters appended to a URI string, typically used to control dynamic scripts

Corpus	Date	Size (URIs)	Sample Period	Construction
BE06	2006	473	1 year	Browsing
Delicious	Sep. 2009	630,476	1 month	Browsing
Sharoff	2006[1]	82,257	hours	Searching
BootCat	Sep. 2011	177,145	hours	Searching

Table 1: An overview of the corpora selected for study.

sampled years apart with heavy influence from search engines (which will have updated significantly in this period). It is also possible that 31 days is an insufficient period to achieve an estimate for attrition that is representative of a full year's loss, implying that a pattern may be evident due to external influences (such as hosting renewals around the commencement of the tax year).

Although we have omitted an analysis of content here, the half-lives of our samples are above those stated in other, more general, studies (Koehler reports a half-life of 2.4 years, for example). This may be the result of bias introduced when deliberately omitting non-document portions of the web, such as navigation pages or images. Another influence is the age of many attrition studies, as it is possible that, with reduction in the price of web hosting, resources simply remain online for longer.

The relatively high rate of attrition in BE06 is surprising given that it only features documents that were already in print, which ostensibly reside in archives or the websites of large institutions (shown to be relatively nonvolatile in other work). One possible counter argument is that BE06's sampling policy was to take documents *published* in 2006, rather than merely being available, such that these samples have witnessed the initial steep descent on the document survival curve.

The diversity of status codes returned varies significantly between corpora, with older ones showing more intricate and descriptive modes of failure (such as code 410 Gone). Delicious exhibits differences to the other corpora, exhibiting codes that are presented by WebDAV and similar systems unlikely to be crawled by search engines.

Each of the corpora exhibit similar distributions of each top level domain (TLD), though the large differences in sample size make formal comparison difficult. The overall distribution of the more popular domains is provided in Figure 1. This indicates that .com dominates the selection across corpora, with .org following. Only BE06 differs from this distribution in its selection of .uk addresses, however, it has been deliberately biased this way so as to represent British English.

Other studies have identified statistically significant differences in the rate of attrition between the major TLDs. Though each corpus exhibits a dependence between these TLD groups (chosen to represent the vast majority of each corpus' content) in a χ^2 indepdence test ($\chi^2 > 33.8; p < 0.01$ in all cases), generalised linear models reveal that the nature of this dependence varies greatly between corpora, indicating that this estimate is far too simplistic to represent the real causes of attrition.

The shape of the empirical distribution for path length of the URI is shown in Figure 2. Delicious.com users may be expected to bookmark top-level domains with relative frequency, but the difference between the two BootCat samples is more subtle, perhaps an effect of search engine changes. The preponderance of introductory or 'launch' pages in the Delicious data set may also go some way to explaining the longevity of its content — it seems reasonable to presume that top-level pages remain online for longer (though also perhaps that they change more often).

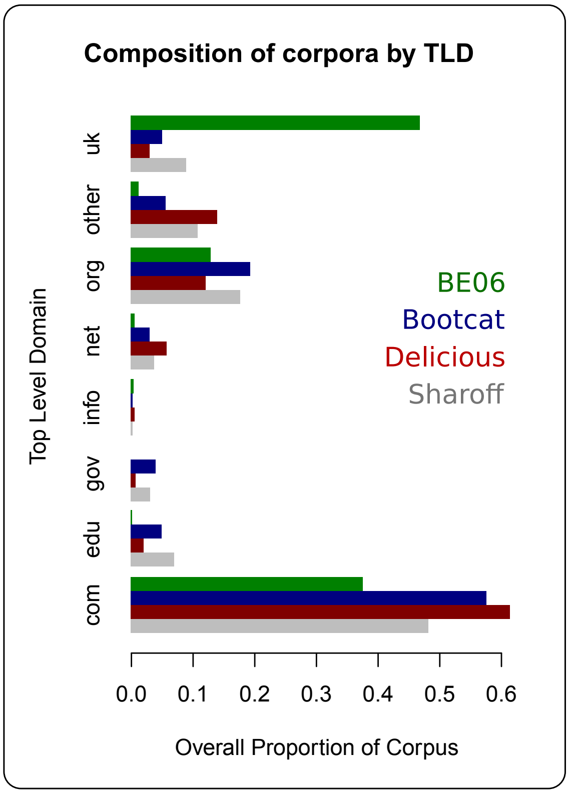

Figure 1: Overall distribution of top level domains.

Taking the presence of GET-arguments in a URI as an indicator of a page being dynamic, a number of effects may be seen across the corpora. The BE06 corpus had 24% of all links dynamic, exceeding Sharoff's at 17% and Delicious & BootCat at 8% and 6% respectively. This difference is probably due to the selection of published documents, since the compiler was seeking specific materials within sites rather than attempting to retain the location of a resource (as with Delicious) or sampling randomly from URI-space (as with BootCat). Differences between the two BootCat-based corpora may reflect changing weights within search engine algorithms.

5. Conclusions

These preliminary results indicate that the process of corpus compilation, by introducing deliberate bias into the con-

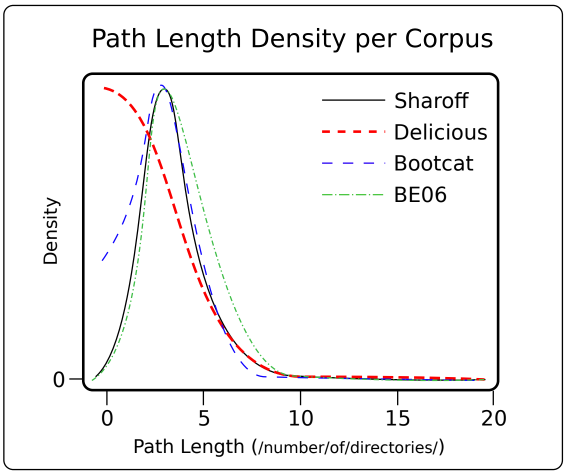

Figure 2: Differences in the empirical distribution of URI path length.

tent (the selection of full documents, filtering of navigation pages, etc.), impacts the observed document attrition rate. These biases have been evidenced by the URI features alone, raising interesting questions about the effect that collection strategies have upon corpus integrity — should the tendencies of different groups of web publishers be factored into sampling strategies for open-source or subject-specific corpora?

The ramifications of these biases for the WaC availability sampling strategy remain an open question — does 'searching' for links imply a minimum age, and hence a pre-existing skew towards certain content?

There are indications that sampling a cross-section of production, rather than consumption, observes the initial steep decline in document availability that is inherent in most survival distributions. It is possible that these effects are minimised by the WaC approach, and are actually more pronounced in conventional, offline, corpora: search-based sampling may compensate for this effect by weighting reliable and established websites through the algorithms used to rank relevance, though further work is needed to establish the degree to which this occurs. Another possible effect is the disproportionate availability of archived, out of copyright, documents.

5.1. Further Work

Further sampling and analysis is necessary to confirm the issues highlighted above. This paper comprises a preliminary look at data sampled in a longitudinal study, which will go on to relate the influences of extrinsic document features (which may be used to inform sampling strategies) to their linguistic content.

This will involve, primarily, identifying the extent to which document attrition applies bias on linguistic content, rather than technical features, and how this varies through the sampling period. Issues of particular interest include:

- The distribution through time of documents online — does sampling using search engines apply particular topical bias as they respond to time-critical events?;
- The propensity of document contents to change in meaningful ways (rather than simply updating boilerplate and navigation page code);
- Whether similar documents or websites display similar levels of loss;
- The characteristics of compilation strategies with regards to their effects on attrition — is it possible to control biases in web-derived corpora through stratified or two-phase sampling techniques.

6. Acknowledgements

The authors would like to thank Paul Baker and Serge Sharoff for assistance in providing details of their corpora. This work is part of an ESRC funded PhD at Lancaster University.

7. References

P. Baker. 2009. The be06 corpus of british english and recent language change. *International journal of corpus linguistics*, 14(3):312–337.

M. Baroni and S. Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proceedings of LREC*, volume 4. Citeseer.

M. Hundt, A. Sand, and R. Siemund. 1998. *Manual of information to accompany the Freiburg-LOB Corpus of British English ('FLOB')*. Albert-Ludwigs-Universität Freiburg.

S. Johansson. 1980. The lob corpus of british english texts: presentation and comments. *ALLC journal*, 1(1):25–36.

A. Kilgarriff and G. Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347.

A. Kilgarriff, S. Reddy, J. Pomikálek, and A. Pvs. 2010. A corpus factory for many languages. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC-2011), Valletta, Malta*. Citeseer.

W. Koehler. 2002. Web page change and persistence — a four-year longitudinal study. *Journal of the American Society for Information Science and Technology*, 53(2):162–171.

M.L. Nelson and B.D. Allen. 2002. Object persistence and availability in digital libraries. *D-Lib Magazine*, 8(1):1082–9873.

R. Sanderson, M. Phillips, and H. Van de Sompel. 2011. Analyzing the persistence of referenced web resources with memento. *Arxiv preprint arXiv:1105.3459*.

S. Sharoff. 2006a. Creating general-purpose corpora using automated search engine queries. *WaCky*, pages 63–98.

S. Sharoff. 2006b. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.