On URL and content persistence

Daniel Gomes
Mário J. Silva

DI–FCUL TR–05–21

Dezembro 2005

Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
Campo Grande, 1749–016 Lisboa
Portugal

On URL and content persistence

Daniel Gomes, Mário J. Silva
Departamento de Informática
Faculdade de Ciências, Universidade de Lisboa
1749-016 Lisboa, Portugal

{dcg, mjs}@di.fc.ul.pt

Dezembro 2005

Abstract

This report presents a study of URL and content persistence among 51 million pages from a national web harvested 8 times over almost 3 years. This study differs from previous ones because it describes the evolution of a large set of web pages for several years, studying in depth the characteristics of persistent data. We found that the persistence of URLs and contents follows a logarithmic distribution. We characterized persistent URLs and contents, and identified reasons for URL death. We found that lasting contents tend to be referenced by different URLs during their lifetime. On the other hand, half of the contents referenced by persistent URLs did not change.

1 Introduction

The Web is a privileged way of communication where a large amount of information is published every day. However, this information is ephemeral. Every web user has already found browser bookmarks or search engine results leading to faulty URLs. The ephemeral nature of URLs leads to accessibility problems and makes it difficult the design of efficient systems to process web data. For instance, off-the-shelf software components that use delta encoding could be considered to save on storage space [22]. However, these components can not be efficiently integrated in web mining systems because they assume the persistence of object identifiers and do not cope with web contents identified by short life URLs [15]. Uniform Resource Names (URNs) were intended to be persistent and location-independent resource identifiers [10]. The URNs should be registered and maintained by the holders of the domains, but their usage remains insignificant. National Libraries tried to maintain URNs to selected publications, but the human interventions required make it unbearable at web scale [25]. Even when URLs persist, citing a source of information on the web is problematic because the referenced contents change. Several initiatives have been working on the archival of web contents to extend their lifetime, but this is a daunting challenge and most contents published on the web are lost [11]. In general, the preservation of information does not concern publishers. Despite these problems, the web is a source of information too valuable to be ignored and many applications, such as search engines or proxies, must address the problem of the transience of web data.

In this report, we provide a thorough study on the persistence of URLs and contents. We used a set of 8 crawls kept in a web archive as basis for our experiments. We detail the followed methodology and compare the obtained results with previous studies. We observed that, despite the ephemeral nature of the web, there is information that persists for long periods of time. Although URLs and referenced contents can not be completely dissociated, we analyzed each one independently to determine the factors that influence their persistence. We derived models for estimating the lifetime of URLs, contents and sites. We characterized persistent information and identified the main reasons for URL death. Collaterally, we provide updated statistics on technological and structural characteristics of the web. This study, aims to contribute to the efficient modelling and design of systems that process historical web data. It is organized as follows: in the next Section, we present related work. In Section 3 we describe the data set analyzed in our study. In Section 4, we present a model for the lifetime of URLs and determine the main causes of URL death. Section 5, characterizes persistent URLs. The lifetime of contents is analyzed in Section 6 and the characteristics of persistent contents are presented in Section 7. Section 8 analyzes the relation between the persistence of URLs and contents. Finally, in Section 9, we draw conclusions and propose future work.

2 Related Work

The problem of URL persistence has been studied by the Digital Libraries community, motivated by the increasing number of citations to URLs on scientific publications [20, 28, 30]. The lifespan of resources related to scholarly pursuits should be permanent, but the referenced URLs have been proven to be ephemeral. Spinellis visited over 4375 URLs extracted from research articles gathered from the ACM and IEEE digital libraries [29]. The documents were harvested from February to May, 2000 and visited in June of the same year. Jonathan D. Wren studied the stability and persistence of 1630 URLs referenced in abstracts extracted from a digital library of life sciences and biomedical bibliographic information for one month, by visiting them almost daily [30]. The author observed that the URLs containing spelling or format errors could be corrected to make the referenced contents accessible. Markwell and Brooks monitored 515 web pages from distance learning courses for 24 months [21]. During this time, the authors witnessed that over 20% of the URLs became nonviable, moving without automatic forwarding or having their content changed. In the first year, they estimated a half-life of 55 months for the URLs population. Lawrence et al. analyzed the persistence of information on the web, looking at the percentage of invalid URLs contained in academic articles within the CiteSeer database [20]. In May, 2000 the authors found that a significant percentage of URLs became invalid, ranging from 23% for articles of 1999, to 53% for articles published in 1994. They studied the causes for the invalid URLs and proposed solutions for citing and generating URLs intended to improve citation persistence.

The frequency of change of web contents has a high variability. There are contents with a median inter-modification interval of approximately 3 hours [27] and entire sites that remain unchanged for months [9]. The frequency of change of web contents is important to search engines that tune harvesting rates to keep indexes up-to-date. Cho and Garcia-Molina studied the frequency of change of web pages by harvesting during 4 months a selection of 270 popular sites on a daily basis, summing a total of 720,000 pages [5]. They proposed estimators for the frequency of change of web pages and counted how many days each URL was accessible to derive its lifespan. Fetterly et al. studied the evolution of web pages by executing weekly crawls of a set of 150 million URLs gathered from the Yahoo! home page [13]. The study spanned 11 weeks in 2002. The authors focused on identifying characteristics that may determine the frequency and degree of change of a page. The contents referenced by each URL were compared to derive a measure of syntactic similarity. They found that most changes consisted of minor modifications, often of markup tags. In the end, the authors concluded that past changes are an excellent predictor of future changes. Ntoulas et al. studied the evolution of contents and link structure of 150 top-ranked sites picked from the Google directory [26]. They witnessed high levels of birth and death of URLs and concluded that the creation of new pages is a much more significant cause of change on the web than changes in existing pages.

Douglis et al. investigated gateway and proxy traces collected over 17 days [12]. They recorded the ”Last-Modified” timestamp transmitted by the visited web servers and found that 16.5% of the resources (including HTML pages and other content, such as images) changed every time they were accessed. Brewington and Cybenko studied the change rate of web pages by recording the Last-Modified timestamp and the time of download of each page accessed by the users of a clipping service [2]. This analysis ignored those pages not relevant to the users' standing queries. The pages were observed over an average of 37 days, 4% of the pages changed on every repeat observation, while no change was observed for 56% of them.

Koehler examined the accessibility and content of 361 randomly chosen URLs for 6 years and concluded that once a collection has sufficiently aged, it tends to stabilize in the sense that its URLs become more durable [19]. The author witnessed the periodic resurrection of web pages and sites, sometimes after protracted periods of time. A reason for this situation is that site domains are resold and URLs resurrect referencing completely different and unrelated contents.

3 Data set

The representability of collected WWW samples has been a controversial issue. Should the samples include password protected contents, pages that do not receive any links or results of form submissions? Moreover, samples are biased towards the selection policy. Proxy or ISP traces are biased towards the pages visited by a limited set of users [1], web crawls are restricted to linked public pages [18] and search engine collections focus on highly ranked pages [7]. We assume that collections generated by exhaustive harvests of national webs are representative of the persistence of information on the general web, because they include sites with different scopes that represent distinct contexts of publication.

Crawl Id.	Date	Size (GB)	# URLs (millions)	#Sites
1	06-11-2002	44	1.2	19721
2	07-04-2003	129	3.5	51208
3	20-12-2003	120	3.3	66370
4	06-07-2004	170	4.4	75367
5	12-04-2005	259	9.4	83925
6	28-05-2005	212	7.3	81294
7	18-06-2005	288	10	94393
8	21-07-2005	299	10.2	106841

Table 1: Statistics of the crawls in the data set.

We studied the persistence of URLs and contents through the analysis of the data collected in the Tomba web archive (**tomba.tumba.pt**). Tomba's crawler has been periodically harvesting textual documents from the Portuguese web, which is broadly defined as the documents hosted on sites under the .PT domain, plus the documents written in the Portuguese language hosted in other sites [16]. Table 1 summarizes the 8 crawls that compose the analyzed data set. It presents the median date of harvest of the pages, the total size of the downloaded contents, the number of URLs and the number of sites successfully visited.

Each new harvest of crawl was seeded with the home pages of the sites successfully harvested in the previous one. Ideally, the crawls should be successively larger, tracking web growth. However, crawl *6* was stopped before it was finished due to hardware problems. The pairs of crawls that did not present an increasing number of contents were excluded from our analysis. There were also crawls obtained using slightly different versions or configurations of the software. To limit the effect of these changes, we carefully analyzed each crawl and corrected any data that would bias the results. For instance, the archive has a fingerprint for each content as meta-data. However, as the fingerprinting algorithm changed between software releases, we recomputed fingerprints to enable the comparison of contents between crawls.

Robustness measures against hazardous situations for harvesting, such as spider traps, were imposed. The crawler harvested at most 5,000 URLs per site, following links from the seeds until a maximum depth of 5 in a breadth-first mode, except for crawl 1 that was executed with a maximum depth of 3. The content sizes were limited to 2 MB and had to be downloaded within 1 minute. The length of the URLs was limited to a maximum of 200 characters. Related work showed that the constraints imposed do not significantly affect the coverage of the sites visited [4, 16].

The crawls included contents from several media types potentially convertible to text. We observed that 97% of the contents were HTML pages, which is not surprising since this is the dominant textual format on the web [18]. However, we also found that a significant amount of contents of other types were discarded during the harvest of the crawls because they could not be converted to text. Therefore, the results presented in this study should be interpreted as referring to HTML web pages.

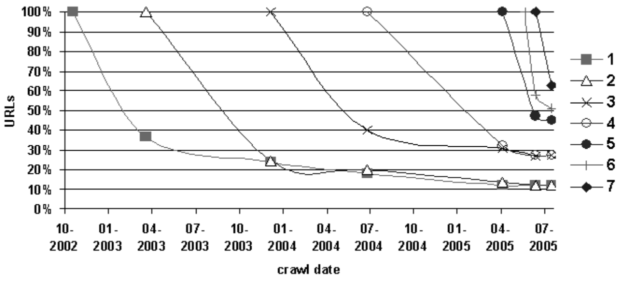

Figure 1: Persistence of URLs for each crawl.

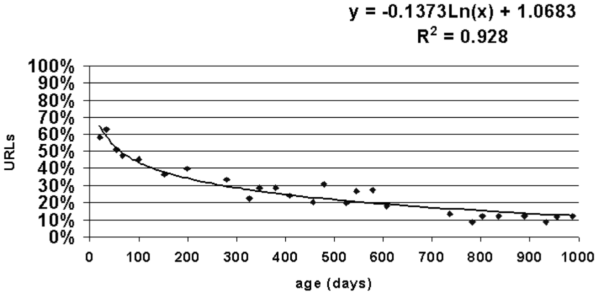

Figure 2: Lifetime of URLs.

4 Lifetime of URLs

There are several situations that lead to the bulk disappearance of URLs: webmasters migrate their servers to different technological platforms, entire sites are shut down and session identifiers embedded in URLs generate new URLs for each visit. We consider an URL persistent if the referenced content was successfully downloaded in two or more crawls, independently from content changes. The URLs that were not linked from any page could not be found by the crawler and would hardly be found by a web user through navigation. Thus, we assumed they died. For each crawl we computed the percentage of URLs that persisted on the following crawls. Figure 1 describes the obtained results. The marks on the lines represent the percentage of URLs from one crawl that persisted until a later crawl. For instance, crawl *1* (line with squares) was harvested in *November, 2002* and *36%* of its URLs were referencing a content in crawl *2*, harvested in *April, 2003*. Almost 3 years later, only *12%* of the URLs of crawl 1 persisted. We observed that most URLs have short lives and the death rate is higher in the first months. However, a minority of URLs persists for long periods of time.

We calculated an approximate age for the persistent URLs found in each pair of crawls given by the difference in days between their dates. Figure 2 shows the relation between the percentage of persistent URLs and their age. For instance, crawl *1* was executed in November 2002 and crawl *3* was executed in December 2003. Hence, the URLs of crawl 1 that persisted until crawl 3 were *409* days old and *24*% of these URLs persisted between the two crawls. The lifetime of URLs follows a logarithmic function with an R-squared value of *0.928*. The function estimates the probability of an URL being available given its age. It suggests that the half-life of an URL is 61 days. There were proposed mathematical models to estimate the frequency of change of web pages under the assumption that URLs persist in time as identifers [6]. Our results complement this work by estimating the time span under which the assumption is valid.

We compared our work with previous studies that presented results on the persistence of URLs referenced from web pages and digital libraries. Cho and Garcia-Molina witnessed that 10% of the URLs had a lifespan of less than 1 week, 20% between 1 week and 1 month, 34% between 1 month and 4 months, and 36% lived more than 4 months [5]. Our results suggest that a larger fraction of URLs (20%) live less than 1 week and a smaller fraction has a lifespan between 1 and 4 months (19%). Fetterly et al. observed that 88% of the URLs in their data set were still available after 11 weeks, representing an URL persistence much higher than the 47% we estimated for the same interval of time [13]. Ntoulas et al. reported that after 1 year only 20% of the URLs were still accessible, which is consistent with our results that suggest that 26% of the URLs are available after 1 year. Koehler's collection of randomly collected URLs remained in a fairly 'steady-state' for two years after it lost approximately two-thirds of its population over a 4 year period [19]. Our results suggest a quicker decay of URLs but they strengthen Koehler's conclusion that once a collection has aged sufficiently, their URLs become more durable in time.

Nelson and Allen found that only 97% of the objects placed in digital libraries accessible via the Web were available after 1 year [24]. Spinellis witnessed that 1 year after the publication of research articles, 80% of the cited URLs were accessible, but this number decreased to 50% after 4 years [29]. Rumsey addressed the legal literature and ran a test on URL viability in mid 2001 [28]. She reported that 63% of the citations to URLs dated 2000 and 30% of the citations with approximately 4 years old persisted. Although, the persistence of URLs varies according to the analyzed document collection, we conclude that URLs referent to contents kept in online digital libraries or cited from scholar articles are more persistent than those linked from web pages in general.

4.1 URL death

We considered an URL dead if it was not referencing a content in the last crawl (8^{th}), but it was successfully harvested in a previous one. A site was considered dead if it did not provide at least one content. Figure 3 presents the main reasons we found for URL death. The *xx* axis represents the time elapsed in days between the pairs of crawls analyzed. We observed that most of the disappeared URLs did not receive any link. Considering an interval of *54* days between crawls, we observed that for *78*% of the dead URLs, the corresponding site was alive but did not link to them. This suggests that the URLs were replaced. For *21*% of the dead URLs, the corresponding sites were also found dead. The percentage of URL death due to site's decease increased with time: *47*% of the URLs died after *988* days. The linked URLs that could not be successfully harvested represent less than 1% of the dead URLs, except for the two closest crawls in time, which

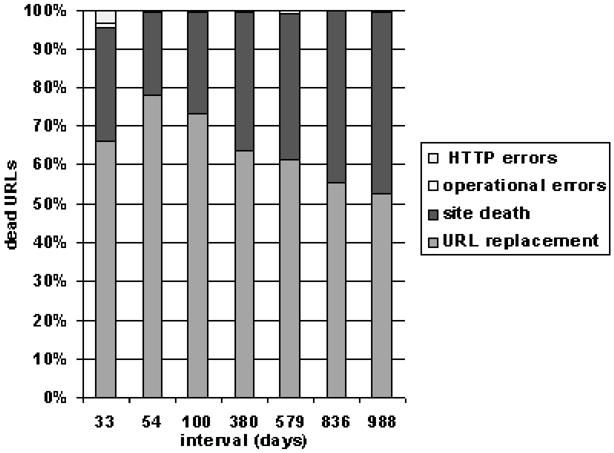

Figure 3: Reasons for URL death.

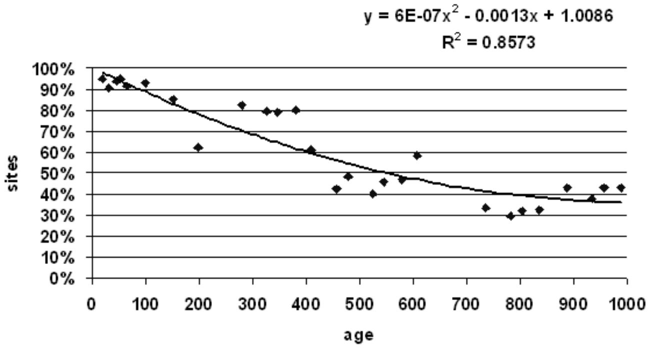

Figure 4: Lifetime of a site.

were executed with *33* days of interval, where we found a percentage of *4.4%*. While harvesting, operational problems like network failures are frequent. On average, only 0.4% of the URLs were considered dead due to these problems, most of them because the referenced content could not be downloaded within 1 minute. URL unavailability identified through HTTP errors represents on average 0.8% of the causes of URL death, but these errors become more visible in shorter intervals, *3.5%* of the URLs in crawl 7 presented HTTP errors in crawl 8 (*33* days of interval). The most common HTTP errors are *File Not Found* (404), *Internal Server Error* (500) and permanent or temporary redirections (301, 302). Notice that the crawler visited the target URLs of the redirections. Our results contrast with the ones presented by Spinellis, which reported that most of the failures were due to 404 errors (60%), invalid hostnames (22%) and network problems (8%) [29]. However, the author used research articles as URL referrers and not web pages. Our conclusion is that the main causes of URL death are the frequent replacement of URLs and site death.

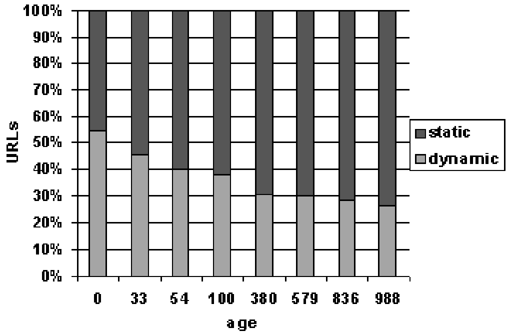

Figure 5: Distribution of dynamic URLs.

The previous results raised our interest on the lifetime of sites. For each crawl we computed the percentage of sites that were still alive in subsequent crawls. The age of a site is the difference between the dates of the crawls in which they were visited. Figure 4 presents the results obtained. We can observe that over *90%* of the sites younger than *100* days were alive, but this percentage decreased to *30-40%* among those older than *700* days. The closest trend we found to model the lifetime of sites was a polynomial function with an R-squared value of *0.8573*. We estimate that the half-life of a site is 556 days.

5 Characteristics of persistent URLs

We used crawl 8 as baseline to characterize persistent URLs. We identified the URLs in the baseline that persisted from previous crawls and compared feature distributions. The age of an URL is the difference in days between the date of the crawl and the date of the baseline (which has age 0).

5.1 Dynamic URLs

URLs containing embedded parameters are commonly generated on-the-fly by the referrer page to contain application specific information (e.g. session identifiers). These URLs are frequently used just once. We defined the URLs containing embedded parameters as dynamic and the remaining as static. Figure 5 describes the distribution of static and dynamic URLs. The first column identified with *age 0* shows that *55%* of the URLs in the baseline were dynamic and *45%* were static. The second column presents the distribution of static and dynamic URLs that persisted from crawl 7 until the baseline. As the date of the baseline was 21-07-2005 and the date of crawl 7 was 18-06-2005, the persistent URLs have an age of *33* days. The URLs were extracted from links in web pages, so we did not consider dynamic URLs resultant from the input of values in forms. Notice that a single form can be used to generate an infinite number of URLs. We observed that the presence of dynamic URLs decreases smoothly as they grow older: *46%* of the URLs *33* days old were dynamic, but this percentage decreased to *26%* among URLs *988* days old. We conclude that

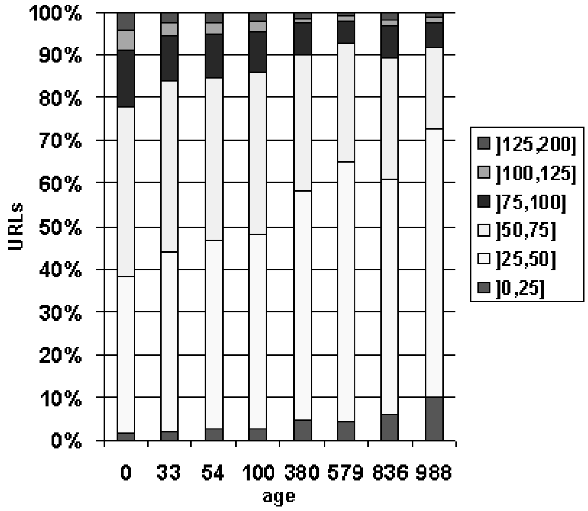

Figure 6: Distribution of URL length (number of characters).

static URLs are more persistent than dynamic URLs.

5.2 URL length

Spinellis concluded that deep path hierarchies, which correspond to longer URLs, are linked to failures because the elements of the path are sensitive to organizational changes and URLs with a short path are more likely to be cited, therefore site administrators try to keep them alive [29]. We studied the relation between the length and persistence of URLs. Figure 6 presents the obtained results. The results show that URLs shorter than 50 characters are more persistent than longer ones. This observation is consistent with the results of the previous subsection, because dynamic URLs were longer (average 77.1 characters) and less persistent than static URLs (average 49.2 characters). We observed that very long URLs are commonly used in poorly designed web sites that end up being significantly remodelled or deactivated.

5.3 Depth

The depth of an URL is the minimum number of links followed from the home page of the site until the URL. The URLs at lower depths are usually the most visited. So, they should be more persistent because broken links are easier to detect. Figure 7 describes the distribution of the URLs per depth. Surprisingly, we witnessed that depth did not influence URL persistence. We analyzed a sample of persistent URLs and observed that they can be found at different levels of depth according to the structure of the site. There are sites presenting a deep tree structure, while others have a shallow and wide structure. So, an URL with depth 3 may be deep in one site but not in another. Moreover, there are sites that maintain URLs for long periods of time updating the referenced

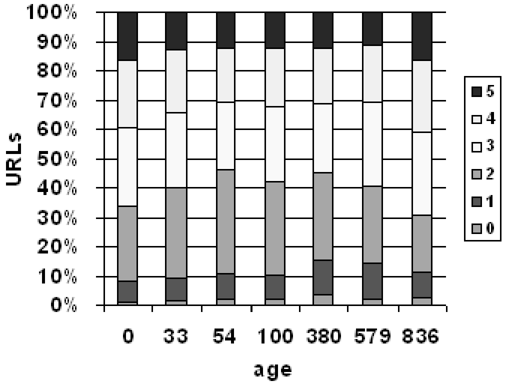

Figure 7: Distribution of URL depths.

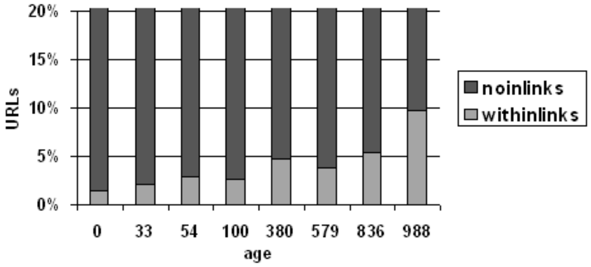

Figure 8: Distribution of linked URLs.

contents, while others generate a new URL for every new content independently from depth.

5.4 Links

Authors use links to reference information related to their publications. The number of links that an URL receives from external sites represents a metric of importance, while links internal to the site are navigational. Figure 8 describes the distribution of the URLs that received at least one link from another site. We found that *98.5%* of the URLs in the baseline did not receive any link. However, the presence of linked URLs among persistent URLs slightly increased with time. It raised from *1.5%* among URLs aged *33* days to *9.6%* among URLs 988 days old. We found two explanations for this fact. Firstly, persistent URLs are more likely to accumulate links during their lifetime. Second, the number of links to an URL increases its measure of popularity in search engines and popular URLs have high commercial value [8, 17]. Hence, the owners of popular URLs take special care in preserving them.

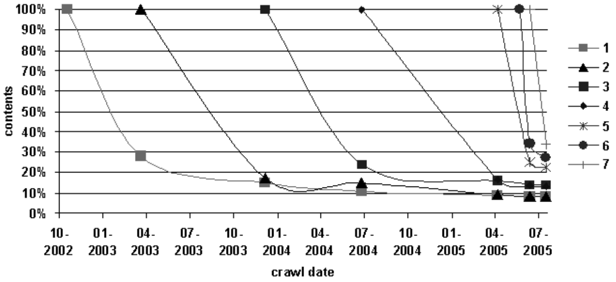

Figure 9: Persistence of contents for each crawl.

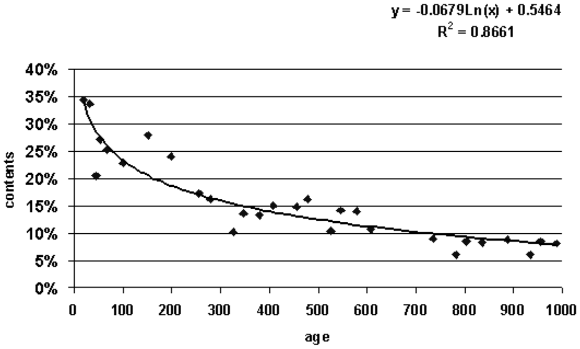

Figure 10: Lifetime of contents.

6 Lifetime of a content

There are pages that present small changes over time, while others suffer complete renewals. The definition of a boundary that determines if the content of a page has changed enough to be considered a new one is controversial. A change in the number that shows the total of visits to the page may be negligible. On the other hand, the change of a single number on the date of a historical event seems important. We assumed that any change in a page generates a new content. Hence, a content dies if it changes or if it becomes unavailable. We identified persistent contents by comparing their fingerprints between crawls, independently from the URLs that referenced them. For each crawl, we computed the percentage of contents that were still available on the following crawls. Figure 9 presents the obtained results. We can observe that crawl *1* (line with squares) was harvested in *November, 2002* and 5 months later only *28%* of its contents were available. However, *8%* of the contents of crawl *1* were still available after 2 years and 8 months.

Figure 10 summarizes the percentages of persistent contents according to their age. We can observe that just *34%* of the contents *33* days old persisted, but *13%* of the contents lived approximately 1 year. The lifetime of contents matches a logarithmic function with an R-squared value of *0.8661*. This function enables the estimation of the contents percentage that remain unchanged within a collection of pages given its age. Our results suggest that the half-life of a content is 2 days. We conclude that most contents have short lives, but there is a small percentage that persists for long periods of time. The age of a content is a good predictor of its future persistence on the web.

Cho and Garcia-Molina observed that 77% of contents persist for at least 1 day [5]. Our results suggest a figure of 55% for the same interval of time. Fetterly et al. analyzed the amount of change in a page between two successful downloads with less than 1 week of interval and witnessed that 65% did not differ at all [13]. We observed that after 1 week, only 41% of the contents remain unchanged. In our data set, approximately 15% of the contents persisted for 1 year, a result close to the 10% witnessed by Ntoulas et al. [26]. Brewington and Cybenko estimated that about half of the web's content is younger than 3 months and the older half has a very long tail: 25% are older than 1 year [2]. Our results suggest that 77% of the contents are younger than 100 days and 15% are older than 1 year. The results presented in previous studies suggest a higher presence of contents younger than 1 year than we observed but the figures presented for older contents are close to those we found.

We suspect that the function derived from our results may underestimate the presence of younger contents, because the analyzed data set was composed by crawls harvested with relatively long time intervals among them. On the other hand, the results presented in related works are biased towards highly ranked contents. The contents used in our study were gathered using breadth-first harvesting, a strategy known to be biased towards pages with high PageRank [23]. However, these contents were gathered from exhaustive harvests of the Portuguese web, including a larger fraction of less popular contents than those studied in previous works, where the most popular contents at world level were chosen. Given that the popularity of web pages follows a Zipf distribution [3], studies restricted to the most popular contents are not a representative sample of all the information available on the web. We speculate that popular contents are more persistent in short or medium periods of time (several weeks) but they are not maintained for long intervals. For instance, daily newspapers maintain articles available for several weeks after their publication to receive the readers that follow links to them from other sites but a historical archive of all the news published is not publicly accessible.

7 Characteristics of persistent contents

In this section we analyze a set of features characterizing persistent contents. We used crawl 8 as baseline and derived feature distributions among persistent contents.

7.1 Dynamic contents

Dynamic contents are generated on-the-fly when a web server receives a request. They became popular because they enable the management of transient information in databases independently from publishing formats (e.g. online shops). The *static* contents do not have to be regenerated in each visit. However, identifying dynamic contents is not trivial. The strings within an URL that

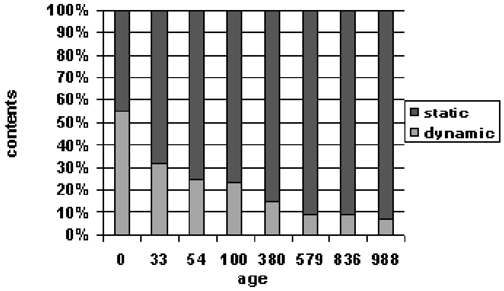

Figure 11: Distribution of dynamically generated contents.

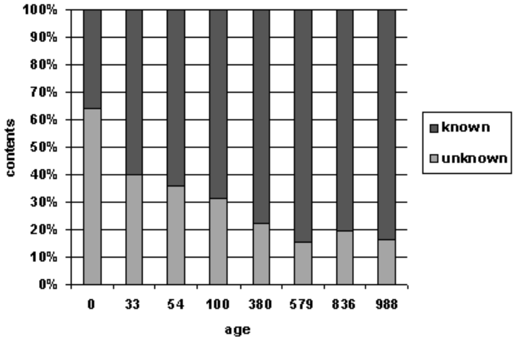

Figure 12: Distribution of contents with known Last-Modified dates.

suggest that the content was dynamically generated (e.g. .asp, .php, /cgi) are sometimes hidden as a security measure. We assume that the URLs containing embedded parameters referenced dynamic contents and define the remaining as static. Nevertheless, there are dynamically generated contents referenced from URLs that do not contain any parameter (e.g. www.somesite.com/index.asp). Figure 11 describes the distribution of static and dynamic contents. It shows that 55% of the contents in the baseline (age *0*) were dynamically generated, a number superior to the *34%* witnessed by Castillo [4]. The presence of dynamic contents decreased to *32%* among contents *33* days old and to less than *9%* among contents older than *579* days. We conclude that static contents are more persistent than dynamic contents.

7.2 Last-Modified date

The Last-Modified header field contained in HTTP responses provides the date of the last modification of the content referenced by an URL. However, web servers returned unknown values for *64%* of the contents in the baseline (Figures 12), a problem also witnessed in previous studies but with a smaller extent [1, 2, 12, 16]. We observed that contents with a known Last-Modified date tend to be more persistent. For *84%* of the contents *988* days old, the web servers

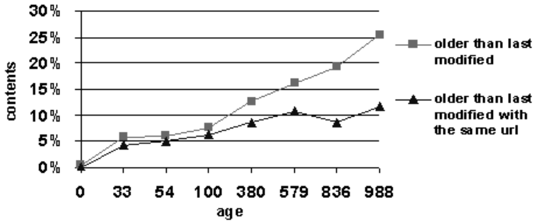

Figure 13: Contents that present underestimated ages due to erroneous Last-Modified dates.

returned the last date of modification. Our results strengthen the observation by Brewington and Cybenko that the absence of the Last-Modified date indicates a more volatile resource [2]. We noticed that webmasters are encouraged to disable the Last-Modified field for pages that change frequently [14].

The Last-Modified header field can provide an erroneous date because the server's clock is not correctly set or the file date was updated with no associated change to its content. We compared the ages of the contents derived from the Last-Modified dates with the ages calculated from the dates of harvest to measure the presence of Last-Modified dates that underestimated the longevity of contents on the web. Figure 13 depicts the obtained results. The line with squares shows that the number of contents older than the Last-Modified date increased with age. *6%* of the contents with approximately *33* days old had an associated Last-Modified date that underestimated their age. This percentage increased to *26%* among contents that persisted for *988* days. The reason we found for these results is that older contents are more subjective to experience site reorganizations that move them to different locations and update timestamps without causing changes in the contents. These operations commonly cause changes in the URLs. Hence, we recomputed the ages of the contents that maintained the same URL (line with triangles). We witnessed that the number of erroneous Last-Modified dates dropped significantly for contents older than 100 days. We conclude that contents with an associated Last-Modified date are more persistent, but its usage has been decreasing in the past years. The presence of inaccurate Last-Modified dates increases among elder contents but it is less visible among contents that maintain the same URL.

7.3 Content length

Figure 14 presents the size distribution of contents. We can observe that the presence of small contents increased with age, *27%* of the contents in the baseline were smaller than *10* KB but this percentage increased to *74%* among the contents *988* days old. We conclude that small contents are more persistent than bigger ones. Our results are consistent with the observation by Fetterly et al. that large pages change more often than smaller ones [13].

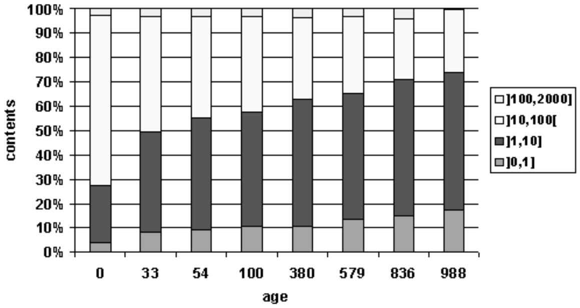

Figure 14: Distribution of content size (KB).

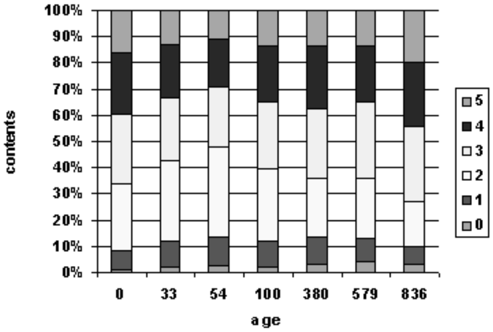

Figure 15: Distribution of content depth.

7.4 Depth

The contents kept at low depths are the most reachable. Empirically, they should change often to include new advertisements or navigational links within the site. The contents kept deep in the sites are frequently archives of persistent information. We analyzed the distribution of contents per depth. Figure 15 depicts the obtained results. We observed that the depth distribution is maintained regardless of the contents' age. We observed that some sites permanently change their contents (e.g. online auctions), while others keep them unchanged (e.g. online digital libraries), regardless of depth. Based on our observations, depth is not a predictor of content persistence.

7.5 Site size

The size of a site is the number of contents that it hosts. One may argue that only large sites, such as digital archives, maintain contents online for long periods of time. In this case, there would be a prevalence of large sites among persistent contents. Figure 16 describes the distribution of the site sizes. We observe that *28%* of the sites in the baseline hosted a single content. This percentage was of *26%* among contents *33* days old and *35%* among contents

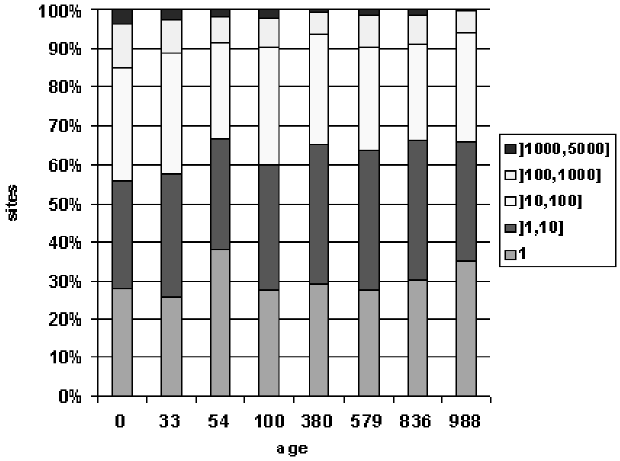

Figure 16: Distribution of site size.

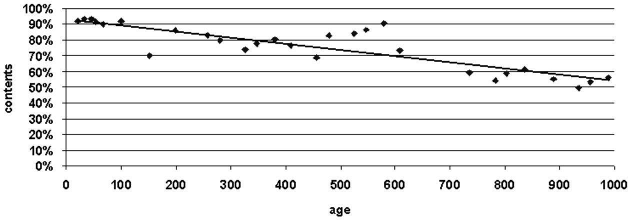

Figure 17: Persistent contents that maintained the same URL.

988 days old. These contents were mainly home pages of sites under construction that were never finished. We observed that the percentage of sites that hold a large number of persistent contents tends to slightly decrease with time but the general distribution of the site sizes does not significantly change. We conclude that the distribution of the number of persistent contents per site is similar to the one we can find on a snapshot of the web.

8 Relation between URL and content persistence

Previous work on the study of the evolution of the web focused on the change of contents under the same URL, assuming that the unavailability of an URL implied the death of the referenced content [6, 13]. However, a simple change of a site's domain name modifies all the correspondent URLs without implying changes on the referenced contents. Lawrence et al. witnessed that for almost all invalid URLs found in academic articles it was possible to locate the information in an alternate location [20]. In Figure 17 we quantify the presence of persistent contents that maintained the same URL. We observed that over *90%* of the persistent contents younger than *100* days maintained the same URL. However,

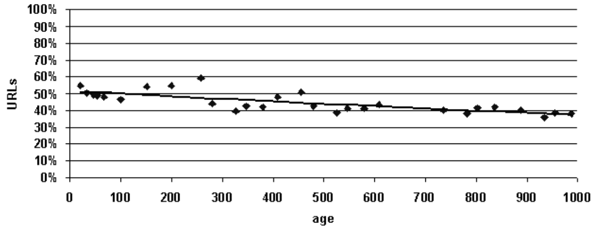

Figure 18: Persistent URLs that maintained the content.

this relation tends to decrease as contents grow older, on average only *58%* of the contents older than *700* days maintained the same URL. These results show that the assumption that the death of an URL implies the death of the referenced content is inadequate in long-term analysis.

The permanent change of contents on the web may lead us to believe that most URLs reference several different contents during their lifetime. Figure 18 depicts the relation between persistent URLs and persistent contents. We can observe that *55%* of the URLs *33* days old referenced the same content during their lifetime. This percentage does not vary much as URLs grow older. On average, 45% of the persistent URLs referenced persistent contents. Ntoulas et al. observed that 50% of the URLs that lived for 1 year referenced the same content during their lifetime [26]. Our results show that this relation between URL and content persistence is extensible to longer periods of time.

9 Conclusion and Future Work

We studied the persistence of URLs and contents over almost 3 years through the analysis of a set of 8 crawls harvested from the Portuguese web. This data differs from those in previous studies, as it was built from exhaustive harvests of a partition of the web regardless page importance or the selection bias of documents kept by topic-specific digital libraries. We found that the lifetime of URLs follows a logarithmic distribution. Most URLs have short lives and the death rate is higher in the first months but there is a minority that persists for long periods of time. Our results contrast with previous work and evidence a quicker decay of URLs. The half-life of an URL was 2 months and the main causes of death were the replacement of URLs and the deactivation of sites. We estimated that the half-life of a site is 556 days. We concluded that persistent URLs are static, short and tend to be linked from other sites. We witnessed that depth did not influence URL persistence.

We concluded that the lifetime of contents follows a logarithmic distribution. Hence, most contents have short lives but some persist for long periods of time. We estimated an half-life of just 2 days for web contents. Typically, persistent contents are not dynamically generated, have an associated Last-Modified date and are small. We witnessed that inaccurate Last-Modified dates increased among elder contents but they were less visible among contents that maintained the same URL. Persistent contents were not related to depth and were not particularly distributed among sites. Most contents younger than 100 days maintained the same URL, but this proportion decreased as they grow older. On the other hand, around 45% of the persistent URLs referenced the same content during their lifetime.

Our study contributes to the characterization of the web evolution and of its current state. It will help in the design of efficient web mining systems and algorithms. The presented results validate some of the conclusions presented in previous works and shed new discussions on the characterization of persistent information on the web.

In future work, we intend to study the influence of popularity in the persistence of information. It would be interesting to combine multiple features that influence persistence in a weighted model, sensitive to the peculiar characteristics of web collections. We suspect that images are more persistent than textual contents. So, another important direction would be extending this study to other media types.

10 Acknowledgements

We would like to thank Andreas Wichert, Francisco Couto and Nuno Cardoso for their comments. This study was partially supported by the FCCN-Fundação para a Computação Científica Nacional and FCT-Fundação para a Ciência e Tecnologia, under grants POSI/SRI/40193/2001 (XMLBase project) and SFRH/BD/11062/2002 (scholarship).

References

[1] L. Bent, M. Rabinovich, G. M. Voelker, and Z. Xiao. Characterization of a large web site population with implications for content delivery. In *Proceedings of the 13th international conference on World Wide Web*, pages 522–533. ACM Press, 2004.

[2] B. E. Brewington and G. Cybenko. How dynamic is the Web? *Computer Networks (Amsterdam, Netherlands: 1999)*, 33(1–6):257–276, 2000.

[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proceedings of the 9th International World Wide Web Conference on Computer networks*, pages 309–320. North-Holland Publishing Co., 2000.

[4] C. Castillo. *Effective Web Crawling.* PhD thesis, University of Chile, November 2004.

[5] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14*, pages 200–209, September 2000.

[6] J. Cho and H. Garcia-Molina. Estimating frequency of change. *ACM Trans. Inter. Tech.*, 3(3):256–290, 2003.

[7] J. Cho, H. García-Molina, and L. Page. Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1–7):161–172, 1998.

[8] J. Cho and S. Roy. Impact of search engines on page popularity. In *Proceedings of the 13th international conference on World Wide Web*, pages 20–29. ACM Press, 2004.

[9] P. D. Corporation. Perseus blog survey. September 2004.

[10] L. Daigle, D. van Gulik, R. Iannella, and P. Faltstrom. *Uniform Resource Names (URN) Namespace Definition Mechanisms*, October 2002.

[11] M. Day. Collecting and preserving the world wide web. `http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf`, 2003.

[12] F. Douglis, A. Feldmann, B. Krishnamurthy, and J. C. Mogul. Rate of change and other metrics: a live study of the world wide web. In *USENIX Symposium on Internet Technologies and Systems*, 1997.

[13] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 669–678, New York, NY, USA, 2003. ACM Press.

[14] T. A. S. Foundation. *Apache HTTP Server Version 1.3: Module mod_include*, November 2004.

[15] D. Gomes, A. L. Santos, and M. J. Silva. Managing duplicates in a web archive. In L. M. Liebrock, editor, *Proceedings of the 21th Annual ACM Symposium on Applied Computing (ACM-SAC-06)*, Dijon, France, April 2006. accepted for publication.

[16] D. Gomes and M. J. Silva. Characterizing a national community web. *ACM Trans. Inter. Tech.*, 5(3):508–531, 2005.

[17] Z. Gyongyi and H. Garcia-Molina. Web spam taxonomy. Technical report, Stanford Database Group, April 2004.

[18] A. Heydon and M. Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2(4):219–229, 1999.

[19] W. Koehler. A longitudinal study of web pages continued: a report after six years. *Information Research*, 9(2):paper 174, January 2004.

[20] S. Lawrence, F. Coetzee, E. Glover, G. Flake, D. Pennock, B. Krovetz, F. Nielsen, A. Kruger, and L. Giles. Persistence of information on the web: analyzing citations contained in research articles. In *CIKM '00: Proceedings of the ninth international conference on Information and knowledge management*, pages 235–242, New York, NY, USA, 2000. ACM Press.

[21] J. Markwell and D. W. Brooks. 'link rot' limits the usefulness of web-based educational materials in biochemistry and molecular biology. *Biochemistry and Molecular Biology Education*, 31(1):69–72, 2003.

[22] P. Miller. Aegis is only for software, isn't it? http://aegis.sourceforge.net/auug96.pdf, 1996.

[23] M. Najork and J. L. Wiener. Breadth-first crawling yields high-quality pages. In *Proceedings of the 10th International World Wide Web Conference*, pages 114–118, Hong Kong, May 2001. Elsevier Science.

[24] M. L. Nelson and B. D. Allen. Object persistence and availability in digital libraries. *D-Lib Magazine*, 8(1), January 2002.

[25] N. Noronha, J. P. Campos, D. Gomes, M. J. Silva, and J. Borbinha. A deposit for digital collections. In *Proc. 5td European Conf. Research and Advanced Technology for Digital Libraries, ECDL*, pages 200–212. Springer-Verlag, 2001.

[26] A. Ntoulas, J. Cho, and C. Olston. What's new on the web?: the evolution of the web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web*, pages 1–12. ACM Press, 2004.

[27] V. N. Padmanabhan and L. Qiu. The content and access dynamics of a busy web site: findings and implications. In *SIGCOMM '00: Proceedings of the conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, pages 111–123, New York, NY, USA, 2000. ACM Press.

[28] M. Rumsey. Runaway train: Problems of permanence, accessibility, and stability in the use of web sources in law review citations. *Law Library Journal*, 94(1):27–39, 2002.

[29] D. Spinellis. The decay and failures of web references. *Communications of the ACM*, 46(1):71–77, 2003.

[30] J. D. Wren. 404 not found: the stability and persistence of urls published in medline. *Bioinformatics*, 20(5):668–672, 2004.