Modelling Information Persistence on the Web

Daniel Gomes
Universidade de Lisboa, Faculdade de Ciências
1749-016 Lisboa, Portugal
dcg@di.fc.ul.pt

Mário J. Silva
Universidade de Lisboa, Faculdade de Ciências
1749-016 Lisboa, Portugal
mjs@di.fc.ul.pt

ABSTRACT

Models of web data persistency are essential tools for the design of efficient information extraction systems that repeatedly collect and process the data. This study models the persistence of web data through the measurement of URL and content persistence across several snapshots of a national community web, collected for 3 years. We found that the lifetimes of URLs and contents are modelled by logarithmic functions. We gathered statistics on the structure of the web, identified reasons for URL death and characterized persistent URLs and contents. The lasting contents tend to be referenced by different URLs during their lifetime, while half of the contents referenced by persistent URLs do not change.[1]

Categories and Subject Descriptors

C.2.5 [**Computer-communication Networks**]: Local and Wide-Area Networks—*Internet*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*Collection*

General Terms

Management, Measurement, Experimentation

Keywords

URL persistence, Content persistence, tomba

1. INTRODUCTION

The design of a web data processing system reflects the characteristics of the web. Despite the ephemeral nature of web data, the persistency of some of the information should be considered in several engineering tasks:

- *Capacity planning*: A system is dimensioned based on the estimated amount of data to process. The identification and reuse of persistent data may reduce the storage requirements;
- *Scheduling maintenance operations*: A web data processing systems periodically refresh their data and prune stale information. These operations are costly and their scheduling should be optimized considering the persistence of web data;
- *Choice of algorithms*: Some web mining applications perform historical analysis of web documents. If URLs are commonly used as identifiers of web documents, the historical analysis becomes limited by the life span of the URLs.

However, the lack of models and up-to-date characterizations of the web frequently postpone important design decisions for a late development stage. For instance, a web data processing system could be designed to use a delta storage mechanism in order to save on storage space. However, its efficiency would be jeopardized in practice because delta storage mechanisms are built on the assumption of persistency of object identifiers and do not cope with web contents identified by short life URLs [10]. Despite the ephemeral nature of the web, there is persistent information. Web systems like search engines or archives that gather periodical snapshots of the web can benefit from the reuse of persistent data.

In this paper, we modelled the persistence of information on the web using two metrics: the persistence of URLs and the persistence of contents. We measured persistence across several archived snapshots of a national web and studied the characteristics of data that influence it. The main contributions of this study are the models for estimating the lifetime of URLs and contents, updated statistics on technological and structural characteristics of the web, and a characterization of persistent information.

This paper is organized as follows: in the next Section, we describe the data set analyzed in our study. Sections 3 and 4 model the persistence of URLs. Sections 5 and 6 model the persistence of contents. Section 7 analyzes the relation between these two metrics. We compare our results with related work in Section 8 and, in Section 9, we draw conclusions and propose future work.

2. 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 [12] and search engine collections focus on highly ranked pages [6]. We be-

[1]This study was partially supported by the FCT-Fundação para a Ciência e Tecnologia, under grants SFRH/BD/11062/2002 (scholarship).

ICWE'06, July 11-14, 2006, Palo Alto, California, USA.
ACM 1-59593-352-2/06/0007.

Crawl id.	Date	Size (GB)	# URLs (millions)	#Sites
1	2002-11-06	44	1.2	19,721
2	2003-04-07	129	3.5	51,208
3	2003-12-20	120	3.3	66,370
4	2004-07-06	170	4.4	75,367
5	2005-04-12	259	9.4	83,925
6	2005-05-28	212	7.3	81,294
7	2005-06-18	288	10	94,393
8	2005-07-21	299	10.2	106,841

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

lieve that a collection generated by exhaustive harvests of a national community web is representative of the persistence of information on the general web. It may differ in other aspects, such as language, but a national web includes a broad scope of sites that represent distinct genres (e.g. blogs, news, commercial sites). We modelled persistence 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 domains [11]. 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 sites successfully visited.

The home page of a site is referenced by an URL where the path component is empty or composed by a single '/'. Each new harvest of a crawl was seeded with the home pages of the sites successfully harvested in the previous one. The crawler iteratively downloaded and followed links to URLs, visiting at most once each one of them. Ideally, the crawls should be successively larger, tracking web growth. However, we found that 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.

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. 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. The crawls included contents from several media types 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 [12].

3. LIFETIME OF URLS

The URLs are the identifiers of the resources available on the web and the basis of its structure. Hence, web data processing systems must be designed to manage URLs efficiently. A model for predicting the lifetime of URLs enables the definition of data structures and algorithms to manage them at an early stage of a system's development.

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 generate new URLs for each visit.

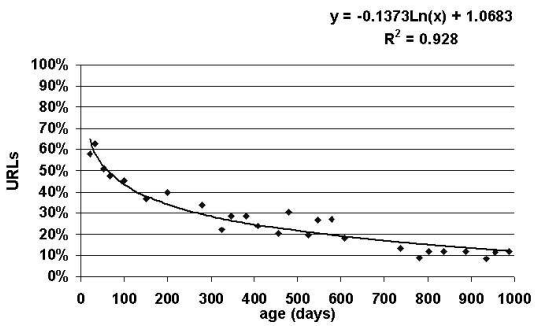

Figure 1: Lifetime of URLs.

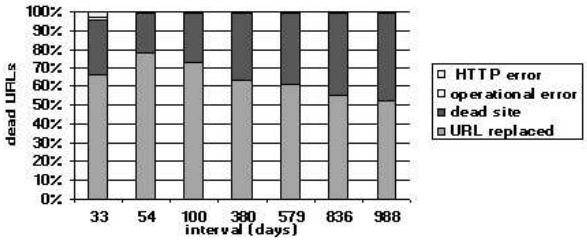

Figure 2: Reasons for URL death.

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.

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 1 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. 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. 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. The half-life of an URL is 61 days.

A previously proposed model to estimate the frequency of change of web pages under the assumption that URLs persist in time as identifers [5]. Our study complements that work by estimating the time span under which the assumption is valid.

3.1 URL death

We considered an URL dead if it was not referencing a content in the last crawl (8^{th}), but was successfully harvested previously. A site was considered dead if it did not provide at least one content. Figure 2 presents the main reasons we found for URL death. The *xx* axis represents the time elapsed in days between the pairs of crawls an-

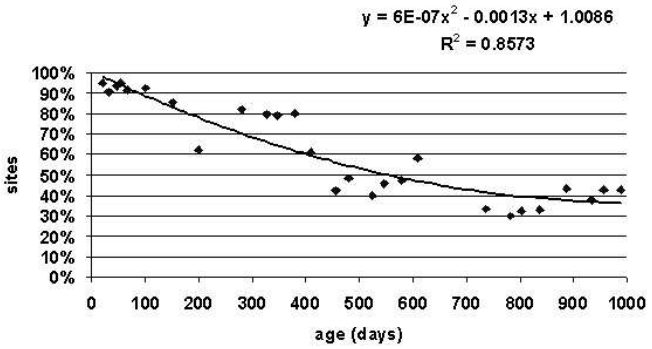

Figure 3: Lifetime of a site.

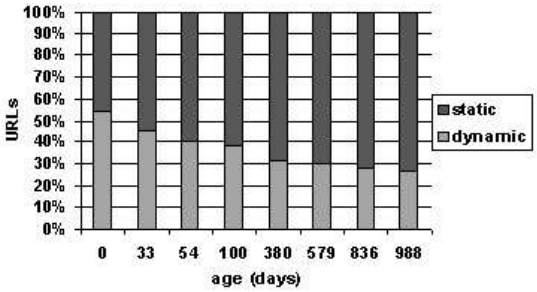

Figure 4: Distribution of dynamic URLs.

alyzed. 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 by new ones. For *21%* of the dead URLs, the corresponding sites were also found dead. The percentage of URL deaths due to site's disappearance increased with time. The linked URLs that could not be successfully harvested represent less than 1% of the dead URLs, except for the two crawls closest in time, which were executed with *33* days of interval. In these, we found a percentage of *4.4%*. While harvesting, operational problems like network failures may occur. 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 were *File Not Found* (404), *Internal Server Error* (500) and permanent or temporary redirections (301, 302). Notice that the crawler also visited the target URLs of the redirections. Spinellis studied the reasons of death among the URLs cited from research articles and obtained similar results [17]. Our conclusion is that the main causes of URL death are the frequent replacement of URLs and site death, independently from the source of citation.

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 visits to obtain the crawls. Figure 3 shows 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, which is significantly larger than the half-life of URLs. Hence, a web system can be designed to reuse information about a site although their URLs may disappear. For instance, consider a site that migrates to a new content management system, causing the replacement of most of its URLs. The information kept in a web directory system about the site, such as the description of its content, does not need to be updated, although most of the previous URLs of the site no longer reference contents.

4. CHARACTERISTICS OF PERSISTENT URLS

The characteristics of an URL can predict its persistence. We used crawl 8 as baseline to characterize persistent URLs in the evaluated data set. 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 is 0 days old).

4.1 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 4). The URLs were extracted from links in web pages, so we did not consider dynamic URLs resultant from the input of values in forms. 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 2005-07-21 and the date of crawl 7 was 2005-06-18, the persistent URLs have an age of *33* days. 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 static URLs are more persistent than dynamic URLs, although there are dynamic URLs that persist for years.

4.2 URL length

We studied the relation between the length and the persistence of URLs. Figure 5 shows 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 tend to be used in poorly designed web sites that are quickly remodelled or deactivated.

4.3 Depth

The depth of an URL is the minimum number of links followed from the home page of the site to the URL. The URLs at lower depths are usually the most visited. We hypothesized that they should be more persistent because broken

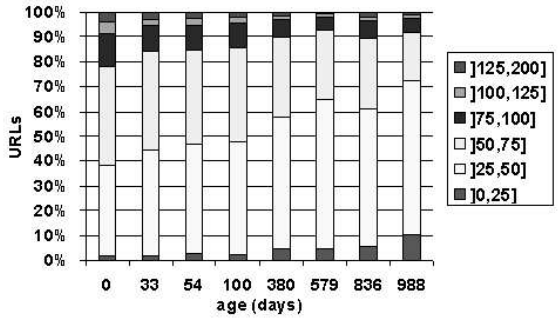

Figure 5: Distribution of URL length (number of characters).

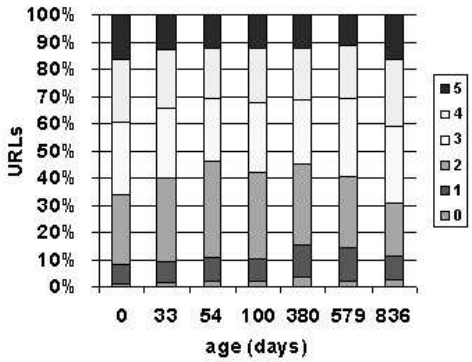

Figure 6: Distribution of URL depths.

links are easier to detect. Figure 6 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.

4.4 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 7 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 the owners of popular URLs take special care in preserving them because of their commercial value.

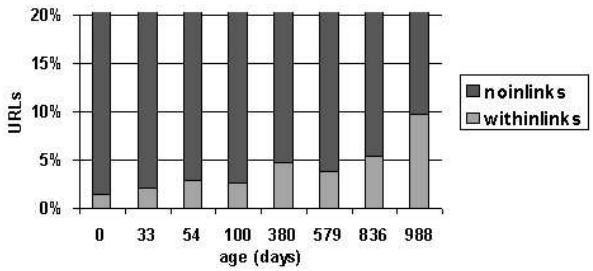

Figure 7: Distribution of linked URLs.

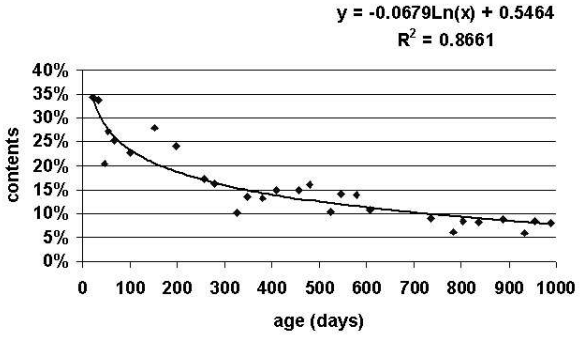

Figure 8: Lifetime of contents.

5. LIFETIME OF A CONTENT

A web data processing system needs to periodically update the information collected from the web. A model that determines the lifetime of a content enables measuring the freshness of the information kept and schedule refreshment operations. Fetterly et al. observed a set of web pages for 11 weeks and observed that the age of a content is a good predictor of its future persistence [8]. In this study, we analyzed persistence for a longer period of time to model persistence for older contents. The definition of a boundary for deciding if a content has changed enough to imply its refreshment from the web is highly subjective. A change in the number that shows the total of visits to a page may be negligible but the correction of a number on the date of a historical event seems important. We assumed that any change in a page generates a new content and studied the persistence of contents on the web.

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 ones. Figure 8 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 con-

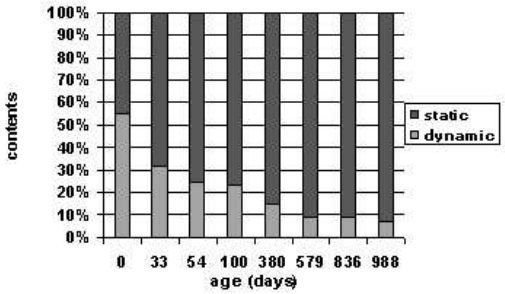

Figure 9: Distribution of dynamically generated contents.

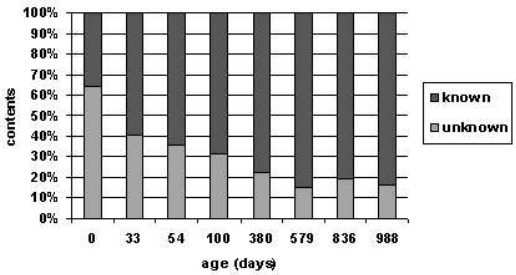

Figure 10: Distribution of contents with known Last-Modified dates.

tent is 2 days. We conclude that most contents have short lives, but there is a small percentage that persists for long periods of time.

6. CHARACTERISTICS OF PERSISTENT CONTENTS

The characteristics of contents influence their persistence. Modelling persistent contents allows the design of adaptive web data processing systems, which may be tuned to the characteristics of the data collections processed. For instance, a model on the persistence of web data can be used to tune the refresh policy of a caching proxy, according to the lifetime and characteristics of the cached information. We analyzed a set of 5 features that characterize persistent contents. We used crawl 8 as baseline and derived feature distributions. Each of these features will be dissected in the rest of this section.

6.1 Dynamic contents

Dynamic contents are generated on-the-fly when a web server receives a request. They became popular because they enable the efficient management of transient information in databases independently from publishing formats (e.g. online shops). So, empirically the contents dynamically generated are not persistent. On the other hand, the *static* contents are not generated in each visit and should be more persistent. A system can benefit from applying different storage policies for static and dynamically generated contents gathered from the web. We assumed that the URLs containing embedded parameters referenced dynamic contents and the remaining were static. Figure 9 shows that 55% of the contents in the baseline, harvested in July, 2005, were dynamically generated, a figure superior to the *34*% witnessed by Castillo in May, 2004 [3]. 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 at long term the static contents are more persistent than dynamic contents.

6.2 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. The Last-Modified date allows to detect if a content has changed since a previous visit without having to download it. This meta-data can be used to implement refresh policies in web data processing systems. For instance, cache entries in proxies can be invalidated based on this information.

Author	Date of the sample	%Unknown
Douglis [7]	1997-?	21%
Brewington [2]	1999-03	35%
Bent [1]	2003-07	44%
Baseline	2005-07	64%

Table 2: Evolution of the presence of unknown dates in the Last-Modified header field.

Webmasters are encouraged to disable the Last-Modified field for pages that change frequently [9]. So, the simple presence of this information for a content can be an indicator of its persistence. Figure 10 shows that the contents with an associated Last-modified date are significantly more persistent than those with an unknown date of last modification. Web servers returned unknown values for *64*% of the contents in the baseline. Table 2 presents the results obtained in previous works and shows that the usage of the Last-modified header field tends to decrease.

Web data processing systems should not rely blindly on the Last-Modified date because it can provide erroneous values. The web server's clock may not be correctly set or the file date may have been updated with no associated change to its content. We compared the ages of the contents derived from the Last-Modified header field 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 11 depicts the obtained results. The line connecting the squares shows that the number of contents older than the Last-Modified date increased with age. One reason for this result might be that older contents are more liable 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) and 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. The presence of inaccurate Last-Modified dates increases among elder contents but it is less visible among contents that maintain the same URL. Hence, a web data processing systems should be designed

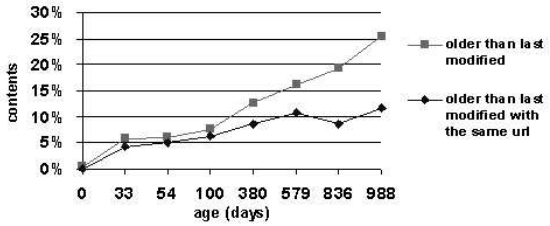

Figure 11: Contents that present underestimated ages due to erroneous Last-Modified dates.

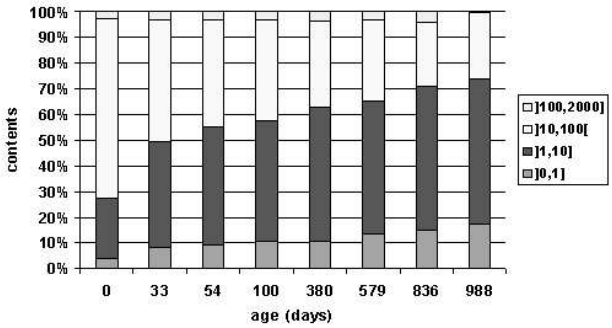

Figure 12: Distribution of content size (KB).

assuming that the contents with a known Last-Modified date are more persistent but the supplied date is not a reliable source of information for the elder contents.

6.3 Content length

Figure 12 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 [8].

6.4 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 13 shows 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. We conclude that depth is not a predictor of content persistence.

6.5 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 14 describes the distribution of the

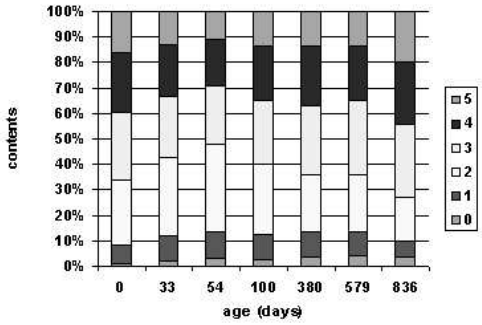

Figure 13: Distribution of content depth.

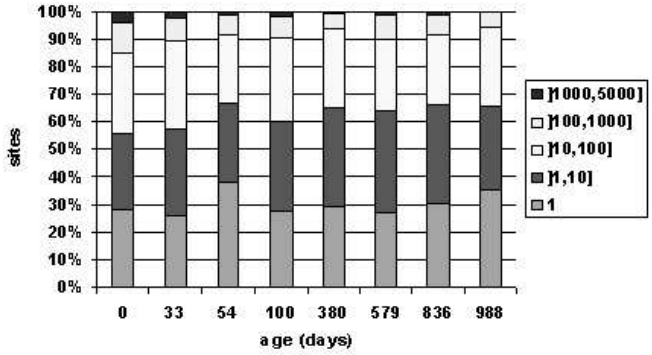

Figure 14: Distribution of site size.

site sizes. The sites that hosted a single content were mainly home pages of sites under construction that were never finished. The percentage of sites that hold more than 100 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.

7. 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 [5, 8]. However, a 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 [14]. In Figure 15 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, 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.

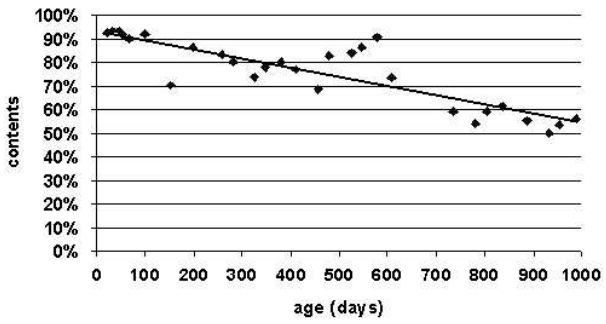

Figure 15: Persistent contents that maintained the same URL.

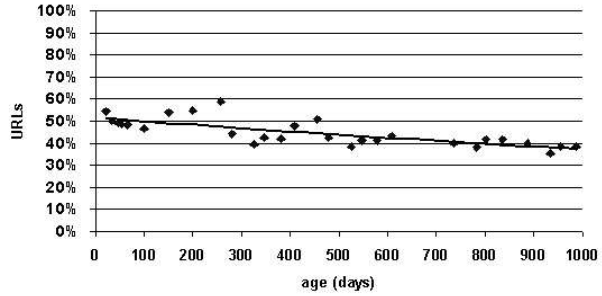

Figure 16: Persistent URLs that maintained the content.

The permanent change of contents on the web may lead us to believe that most URLs reference several different contents during their lifetime. Figure 16 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. We conclude that persistent URLs tend to reference persistent contents.

8. RELATED WORK

We compared our work with previous studies that presented results on the persistence of URLs and contents cited from web pages and articles in Digital Libraries.

Koehler examined the accessibility of a collection of 361 URLs randomly selected from a web crawl during 4 years and concluded that once a collection has sufficiently aged, it tends to stabilize [13]. 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. Cho and Garcia-Molina studied the frequency of change of web pages by harvesting a selection of 270 popular sites during 4 months on a daily basis, summing a total of 720,000 pages [4]. 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 [8]. The study

Author	Age	URL persistence	Our estimation
Koehler[13]	1.9 years	50%	17%
Cho[4]	1 month	70%	60%
Fetterly[8]	2.8 months	88%	47%
Ntoulas[16]	1 year	20%	26%
Digital Libs.			
Spinellis[17]	1 year	80%	26%
Markwell[15]	4.7 years	50%	5%
Lawrence[14]	1 year	77%	26%

Table 3: Comparison of URL persistence.

Author	Age	Content persistence	Our estimation
Brewington[2]	100 days	50%	23%
Cho[4]	1 day	77%	55%
Fetterly[8]	7 days	65%	41%
Ntoulas[16]	1 year	10%	15%

Table 4: Comparison of content persistence.

spanned 11 weeks in 2002. The authors focused on identifying characteristics that may determine the frequency and degree of change of a page. They found that most changes consisted of minor modifications, often of markup tags. Ntoulas et al. studied the evolution of contents and link structure of 150 top-ranked sites picked from the Google directory for 1 year [16]. 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. 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.

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 [14]. The methodology used was to extract citations to URLs from archive documents over several years and verify if they were still available. Changes to contents were not evaluated, because the digital libraries did not keep copies of the cited documents. Spinellis visited URLs extracted from research articles gathered from the ACM and IEEE digital libraries [17]. The author witnessed that 1 year after the publication of the research articles, 80% of the cited URLs were accessible, but this number decreased to 50% after 4 years [17]. Markwell and Brooks monitored 515 web pages from distance learning courses for 24 months [15]. During this time, the authors witnessed that over 20% of the URLs became nonviable, moving without automatic forwarding or having their content changed. Lawrence et al. analyzed the persistence of information on the web, looking at the percentage of invalid URLs contained in academic articles published since 1993 within the CiteSeer database [14] and validated them in May, 2000. They studied the causes for the invalid URLs and proposed solutions for citing and generating URLs intended to improve citation persistence.

Table 3 and Table 4 summarize the degree of persistence of URLs and contents found in previous works and compare these values with our estimations for the same ages. These

comparisons must be taken with a grain of salt because the presented results were derived using different methodologies, the experiments were executed in different dates and the main scope of some of the presented works was not the study of the persistence of information. Nonetheless, our results suggest a quicker decay of URLs and contents on the web than previous studies, but they strengthen Koehler's conclusion that once a collection has aged sufficiently, it becomes more durable in time.

9. CONCLUSION AND FUTURE WORK

In this study we modelled the persistence of information on the web, analyzing the lifetime of URLs and contents, and the characteristics of web data that influence them. Web data models help on important design decisions in the initial phases of web data processing systems implementation projects. We modelled the persistence of information on the web through the analysis of a set of 51 million pages harvested from a national community web. This data differs from those in previous studies, as it was built from exhaustive harvests of a partition of the web during several years, regardless of 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. 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 also witnessed that depth did not influence URL persistence. Our results contrast with previous work and evidence a quicker decay of URLs. In particular, the lifetime of URLs cited from documents in digital libraries is at least 3 times longer than the one we can find on the web.

We concluded that the lifetime of contents follows a logarithmic distribution with an estimated half-life of just 2 days. The comparison of our results with previous works suggests that the lifetime of contents is decreasing. Typically, persistent contents are not dynamically generated, are small and have an associated Last-Modified date. However, the presence of contents with known Last-Modified dates has been decreasing in the past years, which is consistent with our conclusion that the lifetime of contents is getting shorter. Inaccurate Last-Modified dates increased among elder contents but they were less visible among those that maintained the same URL. Persistent contents were not related to depth and were not particularly distributed among sites. About half of the persistent URLs referenced the same content during their lifetime.

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 have higher persistency than textual contents. So, another important direction would be extending this study to other media types.

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

[4] 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*, pages 200–209, September 2000.

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

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

[7] 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.

[8] 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.

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

[10] 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.

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

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

[13] W. Koehler. Web page change and persistence–a four-year longitudinal study. *J. Am. Soc. Inf. Sci. Technol.*, 53(2):162–171, 2002.

[14] 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.

[15] 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.

[16] 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.

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