Usage Analysis of a Public Website Reconstruction Tool

Frank McCown
Harding University
Computer Science Department
Searcy, Arkansas, USA 72149
fmccown@harding.edu

Michael L. Nelson
Old Dominion University
Computer Science Department
Norfolk, Virginia, USA 23529
mln@cs.odu.edu

ABSTRACT

The Web is increasingly the medium by which information is published today, but due to its ephemeral nature, web pages and sometimes entire websites are often "lost" due to server crashes, viruses, hackers, run-ins with the law, bankruptcy and loss of interest. When a website is lost and backups are unavailable, an individual or third party can use Warrick to recover the website from several search engine caches and web archives (the Web Infrastructure). In this short paper, we present Warrick usage data obtained from Brass, a queueing system for Warrick hosted at Old Dominion University and made available to the public for free. Over the last six months, 520 individuals have reconstructed more than 700 websites with 800K resources from the Web Infrastructure. Sixty-two percent of the static web pages were recovered, and 41% of all website resources were recovered. The Internet Archive was the largest contributor of recovered resources (78%).

Categories and Subject Descriptors: H.3.5 [Information Storage and Retrieval] Online Information Services: Web-based services; H.3.7 [Information Storage and Retrieval] Digital Libraries: Collection

General Terms: Measurement

Keywords: digital preservation, search engine caches, web archiving

1. INTRODUCTION

Most Web users are very familiar with the 404 Not Found message; web pages often change location or disappear from the Web altogether. The same is true for entire websites. An individual may lose their website when their server's hard drive crashes or when their ISP goes out of business [7, 11]. Although a rigorous backup plan may prevent such loss, it is not uncommon for users to fail to backup their digital data [2]. Websites can also be lost when webmasters lose interest in maintaining their website or when a catastrophic event forces them to cease maintaining their websites (e.g., bankruptcy or death). In such cases, the website's users may be deprived of a valuable resource.

Noting that individuals have tried to download their lost websites from the Internet Archive [16] and have used search engine caches to retrieve lost web pages [17], we created Warrick, a *web-repository crawler* which can automatically reconstruct entire websites from search engine caches and web archives (collectively known as the Web Infrastructure or WI) [14]. The four web repositories currently being used are: Internet Archive (IA), Google, Live and Yahoo. Warrick begins with a starting URL, usually the root page of a website (e.g., `http://foo.org/`), and probes each web repository for a stored copy of the resource. The resource that is most recent or is in its canonical (original) format is saved, and HTML resources are mined for links to other missing resources. The reconstruction process continues until an attempt has been made to recover all discovered URLs.

Warrick was made available to the public in 2006, but it required some technical know-how to download, install and execute. In response to many inquiries for help, we created Brass [9], an online queueing system that runs on a network of computers at Old Dominion University[1]. Brass allows users to enter the URL of the website they want to recover (see the screenshot in Figure 1). After responding to a confirmation email (to make it more difficult to spam the system), the job is queued and later deployed to a free machine. Once the job completes, Brass emails the user a link to the files composing the reconstructed website.

Brass was made available to the public in July 2007 and has allowed them to reconstruct hundreds of lost websites. We have been collecting usage data to characterize the types of websites users want to recover and to determine how successful we are at reconstructing websites from the WI. In this paper we provide an analysis of this data. In the interest of preserving our users' privacy, we do not disclose individual statistics or URLs to websites that have been reconstructed.

2. BACKGROUND AND RELATED WORK

Work has been done on measuring (e.g., [5]) and preventing linkrot (e.g., [4]) and on archiving the Web in the hopes of keeping an historical record [3]. A wide variety of research has been performed on search engines, measuring their size, overlap and freshness (e.g., [6, 13]). Our work [14] was the first to suggest that web archives and search engine caches could be used collectively to preserve a wide-range of web-

[1]Available at `http://warrick.cs.odu.edu/`

JCDL'08, June 16–20, 2008, Pittsburgh, Pennsylvania, USA.

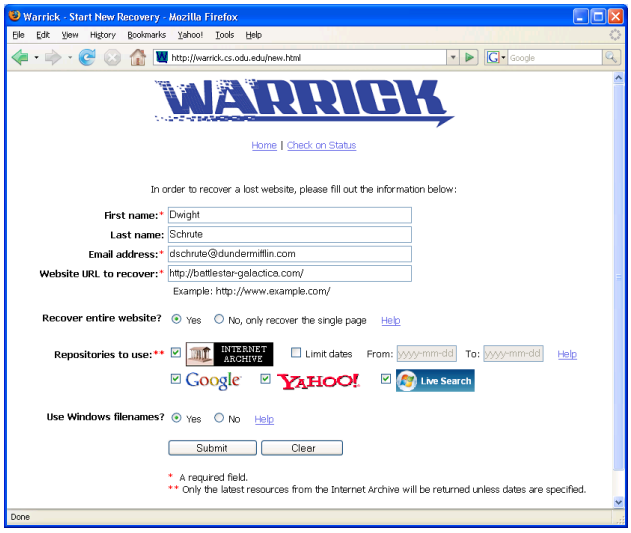

Figure 1: Screenshot of Brass's submit job interface.

sites of unknown importance. We call this preservation service *lazy preservation*. Lazy preservation requires no work from the content producer, but it offers no quality-of-service guarantees.

We have used Warrick to recover a number of lost websites. For example, we reconstructed an academic conference website that was lost due to a fire [12]. On behalf of the Library of Congress, we reconstructed a Congressman's website when it was suddenly shutdown due to allegations of misconduct [8]. Others have used Warrick to reconstruct personal websites they have lost, informational websites about popular entertainers and medical information, and websites used for marketing purposes. We have investigated some of these loses by surveying and interviewing known Warrick users in order to better understand the reasons why websites are lost in the first place [7, 11].

We have also focused on measuring the success of lazy preservation in general. In previous experiments [10, 14], we measured the effectiveness of reconstructing websites from the WI using randomly selected *live* websites. Live websites were crawled and then reconstructed, and comparisons between the crawls and reconstructions were performed to measure how successfully reconstruction would be *if* the site were to suddenly become lost. The experiments revealed that 61% of a typical website (77% textual resources, 42% images and 32% other) could be recovered from the WI. The problem with these experiments is that they were not performed on truly lost websites; since the websites were still live, they likely had a much higher footprint in the search engine caches than a truly inaccessible website.

By capturing reconstruction data from sites that are actually lost, we are able to better understand what types of sites the public is interested in recovering and provide more accurate recovery statistics. We can also determine which web repositories are most helpful for reconstructing the websites.

3. RECONSTRUCTION MECHANICS

Brass runs Warrick processes on a series of machines with their own IP address since search engine APIs limit the number of daily queries that can be issued from the same IP address on the same day. Seven different hosts are currently deployed, allowing seven reconstructions to be running concurrently.

Warrick is executed using the Exhaustive policy [12] which means all four web repositories are initially asked to list all URLs they have stored for a website (we call these *lister queries*). This discovery process is somewhat limited for large websites since search engines will only reveal at most 1000 URLs they have cached. Once these initial lister queries are complete, the recovery of individual resources begins.

Warrick will attempt to recover any URL it discovers from the lister queries or from any pages recovered from the WI. While search engines normally have only recent pages cached, IA may have many pages archived that are no longer a part of the lost website. We recover these pages anyway (unless the user specifies to only recover resources from more recent dates) since they could potentially be used by the user to reconstruct resources which could not be recovered. Because we do not ask the user to evaluate the effectiveness of the reconstructions, we are unable to determine which resources were ultimately useful or not.

Since we do not have access to the actual website that was lost, only its footprint within the WI, we are unable to say precisely what was lost and what was found. However, we can infer that if a URL to a resource was found among the recovered resources and the URL was not found in the WI, it is lost. This may not be necessarily true since the "lost" resource may be a duplicate of a resource that was found at another URL. The link may also not have resolved when the website was live.

Table 1: Brass usage statistics from 2007.

Month	Completed Jobs	Resources Recovered
Jul	118	129,884
Aug	75	84,697
Sep	140	191,186
Oct	129	146,336
Nov	118	161,721
Dec	128	125,294
Average	118	139,853

4. USAGE STATISTICS

4.1 Reconstructed Websites and Resources

Brass usage statistics have been kept since its deployment in early July 2007 and is summarized in Table 1. In the final six months of 2007, Warrick has been used by 520 individuals to recover a total of 708 websites composed of 839,118 recovered resources. For 7% of the jobs submitted, only a single resource was requested for recovery instead of a complete website. Of those completed jobs that recovered at least one resource (84%), 18% were not picked-up by the user. In these cases, the users may have been able to recover the lost websites another way or lost interest in recovering the websites. It is also possible the websites were merely spam URLs submitted by sophisticated spambots or that the user was just satisfying their curiosity by starting the reconstructions.

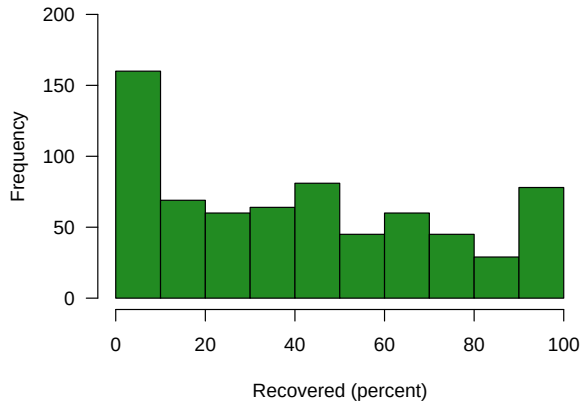

Figure 2: Distribution of websites by percentage of recovered resources.

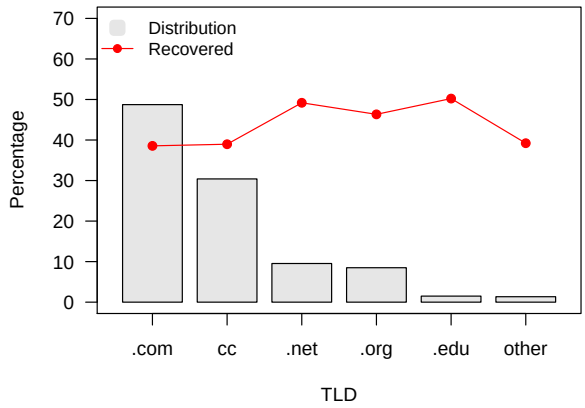

Figure 3: Brass recovered websites by TLD.

Table 2: Brass recovery summary.

	Mean	Median	Max	Std
URLs discovered	3632.4	206	196,448	13,181.6
URLs recovered	1214.4	67	26,682	3470.3
Days to complete	1.2	1	46	4.3

Table 2 gives descriptive statistics for the reconstructed websites. URLs discovered is the number of URLs that were found by mining recovered HTML pages for links, and URLs recovered were URLs where a resource was recovered. The table also summarizes the number of days that it took for Warrick to complete the submitted jobs. Most websites were rather small in number of URLs, but a few were quite large (website sizes have a power-law distribution on the Web [1]).

On average, 41% of the websites' resources were recovered (this is obtained by averaging the percentage of URLs recovered for each website). The percent jumps to 49% if only considering websites with at least one recovered resource. The distribution of recovered resources per website is shown in Figure 2. For 16% of the jobs, none of the resources could be recovered. Less than 10% of the resources were recovered for a quarter of the jobs (24%), and a small minority (14%) of jobs have an 80% recovery rate or better.

4.2 Reconstructions by TLD

The distribution of reconstructed websites by top level domain (TLD) is shown in Figure 3 along with the percent of recovered resources for each TLD. Websites with .com TLDs make up approximately half of all reconstructions with country code TLDs (ccTLD) coming in second, making up a quarter of all jobs. Recovery success differs some by TLD. Websites with a .net or .edu TLD experienced nearly a 50% recovery rate compared to only 40% for the .com and ccTLD sites that were reconstructed.

The most popular ccTLD was .jp by far, making up 47% of all ccTLDs. Although at least one missing resource was found for 81% of the .jp websites, only 21% of these reconstructions were picked-up by the user (compare this with an 82% pick-up rate overall). This suggests that many of the .jp reconstructions were likely spam URLs.

4.3 Reconstructions by Resource Type

We wanted to know the content makeup of the websites being reconstructed and which types of resources were most recoverable. In order to tell what types of resource were recovered, the resources' MIME types were examined when extracting the resources from the repositories. Because search engines convert Microsoft Office and PDF documents into HTML, we counted HTML resources that ended with an Office or .pdf extension to be of their original type. For example, if a Word document with a URL ending in .doc was recovered from Google with a `text/html` MIME type, we classified the resource as an Office document.

For those resources which could not be recovered, we examined the file extension or URL ending to classify the type. If the resource ended in a forward slash (/), .html, .htm, .shtml, .php, .asp, .aspx or .jsp, we considered the resource to be of type HTML. URLs that used a query string were also considered to be HTML resources.

Figure 4 shows the distribution of resources grouped by type and the percent of resources recovered in each group. HTML resources are divided into static resources (those with no query string) and dynamic resources (those with a query string). PDF and Microsoft Office documents are grouped together under the "text doc" category.

Static HTML pages made up a majority of the websites (43%) and were the most recoverable (62%). Dynamic pages were only slightly less prevalent than images, making up 23% of the resources; they were only recovered 32% of the time on average. This may be because some repositories prefer not to crawl and/or cache resources with query strings.

4.4 Repository Contributions

All four repositories were used in a majority of the reconstructions (Table 3), but IA contributed significantly more resources per reconstruction than the other three repositories. The search engines combined only contributed 22% of the recovered resources. In 20% of the reconstructions, IA is the *only* contributor. This is likely due to two reasons: 1) the websites being reconstructed have been lost for long enough of time for their resources to disappear from the search engine caches, and 2) IA has archived many resources which are no longer part of the website that was lost; since the Exhaustive policy is used (see Section 3), all these archived resources are recovered.

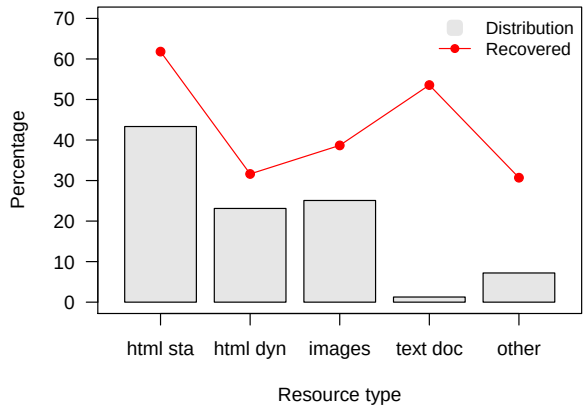

Figure 4: Brass recovered resources by type.

Table 3: Repository use, contributions and requests.

	Used in recons	Contribution	Requests per recon (ave)
IA	99.3%	77.7%	2614.2
Google	96.0%	9.6%	1018.1
Live	94.9%	6.4%	660.9
Yahoo	95.1%	6.3%	881.0

5. DISCUSSION

In previous experiments [10], the percentage of recovered resources was significantly higher (61%) than the 41% reported in Section 4.1. There are several likely reasons for this. In [10], websites were sampled from the ODP which means they were well-connected in the web graph and were more likely to be discovered by search engines when crawling; perhaps the websites being reconstructed by Brass were not as well-connected. Also in [10], URLs that were blocked by the robots exclusion protocol were not considered, but the real usage data presented here does not account for these blocked URLs which are avoided by web repositories; this may account for a significant number of missing resources. And, as mentioned in Section 2, websites reconstructed in earlier experiments were not yet lost, so they likely had a much larger footprint in the WI, especially in the search engine caches.

In previous experiments [10, 14], HTML resources have always been the most recoverable. This is also true in our usage data when we exclude URLs with query strings. Images and other media types are generally less recoverable because these are types that most web repositories do not favor.

Previous experiments with live websites also showed IA to be far less useful in reconstructions, contributing only 23% of the recovered resources. In reconstructing truly lost websites, however, IA is by far the largest contributor (78%). If the websites are recovered immediately after being lost, we would expect the search engines to be more useful. However, Brass users are likely reconstructing websites that have been lost for quite some time. And since *all* archived resources from IA are recovered, this boosts IA's contribution rate.

6. CONCLUSIONS

Warrick is a unique service that has attracted wide-spread use and endorsement from the Internet Archive. In the first six months of operation, Warrick (through Brass) has been used to reconstruct more than 700 lost websites, almost half from the .com TLD. Warrick was able to recover at least one resource for 84% of all websites with a recovery rate of 41%. Static HTML resources proved to be the most recoverable (62%). The Internet Archive was the largest contributor to the reconstructions, contributing 78% of all resources.

From the large number of .jp URLs that are submitted to Brass, it is apparent that more stringent measures are necessary to prevent spamming. We are currently investigating using reCAPTCHA [15].

7. ACKNOWLEDGMENTS

This work is supported in part by NSF Grant IIS-0610841.

8. REFERENCES

[1] R. Baeza-Yates, C. Castillo, and E. N. Efthimiadis. Characterization of national web domains. *ACM Transactions on Internet Technology*, 7(2):9, 2007.

[2] A. Cantrell. Data backup no big deal to many, until... *CNNMoney.com*, 2006. `http://money.cnn.com/2006/06/07/technology/data_loss/index.htm`.

[3] M. Day. Preserving the fabric of our lives: A survey of web preservation initiatives. *Research and Advanced Technology for Digital Libraries*, pages 461–472, 2003.

[4] T. L. Harrison and M. L. Nelson. Just-in-time recovery of missing web pages. In *Proceedings of HYPERTEXT '06*, pages 145–156, Aug. 2006.

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

[6] S. Lawrence and C. L. Giles. Searching the world-wide web. *Science*, 280(4):98–100, 1998.

[7] C. Marshall, F. McCown, and M. L. Nelson. Evaluating personal archiving strategies for Internet-based information. In *Proceedings of IS&T Archiving 2007*, pages 151–156, May 2007. arXiv:0704.3647v1.

[8] F. McCown. Mark Foley websites - reconstructed, 2006. `http://www.cs.odu.edu/~fmccown/foley/`.

[9] F. McCown, A. Benjelloun, and M. L. Nelson. Brass: A queueing manager for Warrick. In *Proceedings of IWAW '07*, June 2007.

[10] F. McCown, N. Diawara, and M. L. Nelson. Factors affecting website reconstruction from the web infrastructure. In *Proceedings of JCDL '07*, pages 39–48, June 2007.

[11] F. McCown, C. C. Marshall, and M. L. Nelson. Why websites are lost (and how they're sometimes found). *Communications of the ACM*, 2008. To appear.

[12] F. McCown and M. L. Nelson. Evaluation of crawling policies for a web-repository crawler. In *Proceedings of HYPERTEXT '06*, pages 145–156, 2006.

[13] F. McCown and M. L. Nelson. Characterization of search engine caches. In *Proceedings of IS&T Archiving 2007*, pages 48–52, May 2007. arXiv:cs/0703083v2.

[14] F. McCown, J. A. Smith, M. L. Nelson, and J. Bollen. Lazy preservation: Reconstructing websites by crawling the crawlers. In *Proceedings of WIDM '06*, pages 67–74, 2006.

[15] reCAPTCHA. `http://recaptcha.net/`.

[16] A. Ross. Internet Archive forums: Web forum posting. Oct. 2004. `http://www.archive.org/iathreads/post-view.php?id=23121`.

[17] J. Symons. How the Google cache can save your a$$, Dec. 2005. `http://www.smartmoneydaily.com/Business/How-the-Google-Cache-can-Save-You.aspx`.