Annals of Library and Information Studies
Vol. 64, September 2017, pp. 165-171

Finding the unfound: Recovery of missing URLs through Internet Archive

Vinay Kumar D[a] and B. T. Sampath Kumar[b]

[a]Lecturer, Department of Library and Information Science, Kuvempu University, Jnanasahyadri, Shankaraghatta, Karnataka, India, E-mail: vinay.86.kumar@gmail.com

[b]Professor, Department of Library and Information Science, Tumkur University, Tumakuru, Karnatka, India Email: sampathbt2001@gmail.com

Received: 23 February 2017; accepted: 22: September 2017

The study investigated the accessibility and permanency of citations containing URLs in the articles published in *DESIDOC Journal of Library and Information Technology* journal during 2006-2015. A total of 2133 URL citations were identified out of which 823 were found to be incorrect or missing. HTTP-404 was the most common error message associated with the missing URLs. The study also tried to recover the incorrect or URL citations using *Internet Archive* and recovered a total of 484 (58.81%) missing URL citations.

Keywords: URL citations; missing URLs; HTTP errors; Internet archive

Introduction

The outburst of e-content has made the internet the quickest and preferred medium for information exchange[1]. The adoption of web based technology for publishing information has quickened and increased the production of electronic information over the web and on the other hand its availability has influenced the large use of web resources in scholarly literature[2]**.** Searching and organizing information has never been easier before. Researchers prefer to access information sources on the web to support their research[3].

The increasing tendency to identify and use electronic resources is gradually replacing the use of traditional text based information resources in the scholarly literature[4]. The present day researchers have been increasingly citing Universal Resource Locators (URLs) in their scholarly publications. URL citations help to easily locate previous research works. However, web resources are subject to modifications and they can be overwritten[5]**.** Recently, a pertinent area of research that has started to receive the attention of researchers, academics and librarians concerns investigating the permanency of URLs of web resources cited in scholarly works[6-8].

Every researcher expects that the URL citations should remain stable and accessible. Preserving the web resources has become inevitable which assures the future researcher with regard to the accessibility to previous works. Web archiving initiatives such as *Internet Archive* are preserving the web based knowledge for posterity. The archivist at web archives collect and preserve web resources permanently. Thus they ensure long term accessibility to the URLs that are not accessible through the web browsers.

Previous studies[9-11] have documented the recovery of missing URLs through *Internet Archive*. Therefore, this study is an attempt to examine the issues such as, the trend of URL citations in a library and information science journal, missing URL citations and their recovery through *Internet Archive*.

Review of literature

Wu studied web references cited in two key Chinese academic journals. Out of 1637 web references, the study found 776 were inaccessible. Seventy eight inaccessible web references were randomly selected to know the rate of recovery. The author used a search engine and *Internet Archive* to

recover the missing web references. Search engines found 62.8%, and *Internet Archive* found 24.4% of missing web citations[12].

Moghaddam et al. studied the URLs of cited articles published in *Information Research* journal during the year 1995-2008 and found 1,761 cited URLs. Among these URLs, 1,290 (73%) were accessible and 471 (27%) were inaccessible. *Internet Archive* and Google search engine were used to retrieve inaccessible URLs. *Internet Archive* retrieved 28% of inaccessible URLs[9].

Sampath Kumar and Vinay Kumar studied two Indian LIS journals by extracting a total of 1290 URLs cited in 472 research articles published in Indian LIS journals spanning a period of 9 years (2002–2010). The study found that 39.84% of URL citations were not accessible. However, they used *Internet Archive* to recover the missing URLs which has increased the percentage of available URLs from 39.84% to 77.90%[13].

Prithviraj and Sampath Kumar conducted a study of three Indian LIS conference proceedings that yielded a total of 5698 web citations. The study showed that 50.9% of web citations (2854) were missing. They used *Internet Archive* to bring back the missing web citations and succeeded to retrieve 29.71% of missing web citations[14]. In the same year, Zhuo et al., conducted a study of web references cited in the articles published in Elsevier database. Out of 193955 links 70270 (36.2%) were found rotten. They used 'Memento Aggregator' that covers nine different web archives to recover the rotten links that resulted in the recovery of 43745 links (62.3%)[15]. Kumar and Kumar examined 5197 URL references of *Annals of Library and Information Studies* and *Webology* and found 417 references were missing[16].

The above said studies have used *Internet Archive* and Memento aggregator to locate the missing URL citations. These recovery tools have increased the rate of active URLs and subsequently decreased the percentage of missing URLs.

Objectives of the study

- To know the ratio of citations and URL citations in *DESIDOC Journal of Library and Information Technology*;
- To determine the URL accessibility and percentage of missing URLs;
- To identify the HTTP error message encountered for the missing URL citations;
- To know the URL domains and file formats associated with active, missing and recovered missing URL citations;
- To identify the co-relation between the path depth and URL citations decay; and
- To know the extent of recovered missing URLs through Internet Archive by year

Hypotheses

The hypotheses formulated for this study are:

- The growth in number of citations is positively correlated with the growth of URL citations.
- The percentage of missing URLs and the age of URLs are positively correlated.
- URL path depth and the missing URL citations were correlated.

Methodology

A preliminary browse showed that the research articles published in library and information science journals relied on web resources. *DESIDOC Journal of Library and Information Technology* has been selected as a representative journal to examine the URL citations' trend and their permanency. The 491 articles published in *DJLIT* during 2006 to 2015 were considered. A total of 7353 references appended to these articles examined and those contains URLs were extracted for further analysis.

Testing of URLs

Once the extraction of URLs from the reference list was completed, the duplicate URLs were identified and removed from the data set. Then, the unique URLs were tested for their accessibility using W3C Link Checker (http://validator.w3.org/checklink). This tool tests a submitted URL for broken or non-valid hypertext links. A useful feature of this application was that if a broken link was found, it brings out the exact HTTP error message.

Recovery of missing URLs

The missing URL citations were entered in the search box of *Internet Archive* (https://archive.org/web/) by copying the exact URL as it has appeared in the reference. If the URL was found in *Internet Archive*, the URL was recorded

'active'. If the missing URL was not found in *Internet Archive* and the Google search yielded no result, the inaccessible URL was considered as 'unrecovered'.

Results

Distribution of citations and URL citations

Table 1 shows that a total of 491 articles were published during the years 2006-2015. The highest average citations per article was in the year 2011 (19.07 citations per article), followed by 2006 (17.06 citations per article) and 2015 (16.44 citations per article). It is observed that the URL citations grown during the years 2006-2015.

Our assumption was that the growth in the number of citations is positively correlated with the growth of URL citations. The statistical analysis shows that the number of citations and URL citations are positively correlated and this correlation is statistically significant (r=.744, df=9, p=.008) (Fig. 1).

Authorship pattern

Table 2 shows the authorship pattern in the research articles published during the 2006-2015. The table reveals that the two-authored articles were more (202) followed by articles with one author (191). It was found that articles with more than five authors have highest average(19) of citations whereas, articles with three authors have highest average (4.65) of URL citations followed by article with two authors (4.65).

Table 3 gives the distribution of active and missing URLs by year. The result of the URL accessibility test in W3C link checker indicated that of the 2133 URL citations, 61.42% of URL citations were still

Table 1—Distribution of articles, citations and URL citations by year

Year	Total articles	Total citations	Average citation per paper	URL citations	Average URL citation/article
2006	18	307	17.06	50	2.78
2007	34	362	10.65	48	1.41
2008	49	721	14.71	250	5.10
2009	50	505	10.10	153	3.06
2010	47	692	14.72	204	4.34
2011	56	1068	19.07	185	3.30
2012	66	958	14.52	430	6.52
2013	59	952	16.14	196	3.32
2014	60	933	15.55	307	5.12
2015	52	855	16.44	310	5.96
Total	491	7353	14.98	2133	4.34

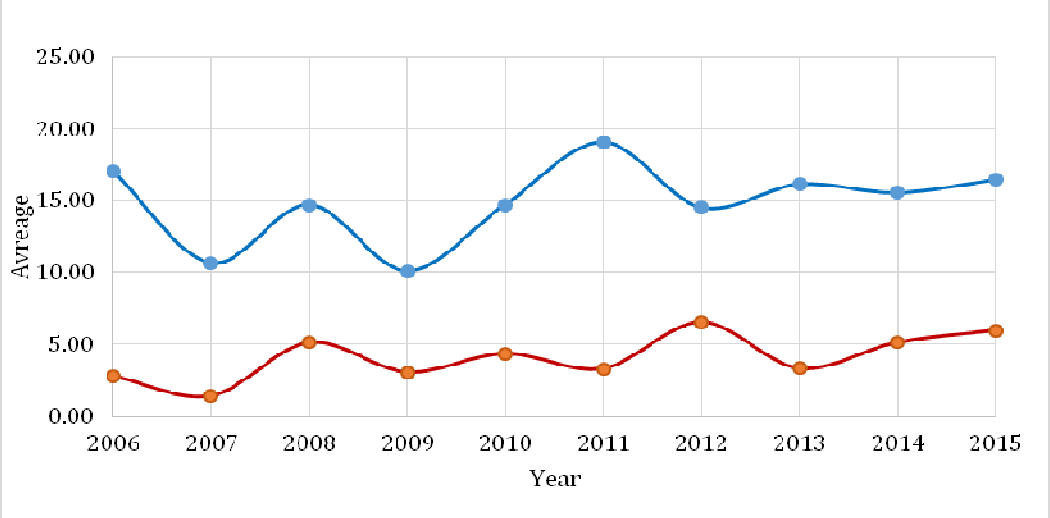

Fig. 1—Average total and URL citations per article

Table 2—Authorship pattern and citations

Authorship	Total articles	Total citations	Average citation per article	URL citations	Average URL per article	% of URL citations to total citations
Single	191	2762	14.46	802	4.20	29.04
Two	202	3145	15.57	908	4.50	28.87
Three	68	967	14.22	316	4.65	32.68
Four	21	308	14.67	78	3.71	25.32
Five and above	9	171	19.00	29	3.22	16.96
Total	491	7353	14.98	2133	4.34	29.01

Table 3—Distribution of active and missing URL citations by year

Year	URL citations	Active URL citations	%	Missing URL citations	%
2006	50	11	22.00	39	78.00
2007	48	21	43.75	27	56.25
2008	250	154	61.60	96	38.40
2009	153	86	56.21	67	43.79
2010	204	107	52.45	97	47.55
2011	185	102	55.14	83	44.86
2012	430	286	66.51	144	33.49
2013	196	91	46.43	105	53.57
2014	307	182	59.28	125	40.72
2015	310	270	87.10	40	12.90
Total	2133	1310	61.42	823	38.58

accessible while the remaining 38.58% of URL citations encountered access errors. The percentage of missing URL citations decreased significantly from 78% in the year 2006 to 12.90% in the year 2015.

It indicates that the percentage of missing URLs and age are positively correlated. The statistical analysis at the probability of 0.05 level shows that the percentage of missing URLs and their age are positively correlated and the correlation is statistically significant (r=.715, p=.013). The results of the study is consistent with the findings of earlier studies6, 13.

The summary of various HTTP error codes are presented in the Table 4. Of the 823 missing URL citations, the HTTP 404 error message- "page not found" was the overwhelming error message encountered and represented 78.13% percent of all HTTP error messages. It is followed by HTTP 500 (15.67%) and HTTP 403 (4.37%). The remaining missing URLs were associated with 7 other HTTP errors cumulatively accounted for 1.83% of the total.

The details of the recovered URL citations through Internet Archive by year are presented in Table 5. It is evident from the table that of the 823 missing URL citations, Internet Archive recovered a total of 484 missing URL citations with an overall recovery rate of 58.81%. The highest percentage (82.09%) of recovery was achieved for the missing URLs cited in the year 2009 followed by 2006 (82.05%) and 2007 (77.78%).

Table 4—HTTP errors associated with missing URL citations

HTTP Error	Count	Percentage
HTTP 300	1	0.12
HTTP302	1	0.12
HTTP 400	6	0.73
HTTP 403	36	4.37
HTTP 404	643	78.13
HTTP 410	3	0.36
HTTP 415	1	0.12
HTTP 500	129	15.67
HTTP 502	1	0.12
HTTP 503	2	0.24
Total	823	100.00

Internet Archive has increased the percentage of active URLs from 61.42% (before recovery) to 84.11% (after recovery). The recovery of missing URLs through Internet Archive have increased the

Table 5—Distribution of recovered and unrecovered missing URLs by year

Year	Total URL citations	Missing URL citations	Recovered URL citations	%	Unrecovered URL citations	%	Active URL citations after recovery	%
2006	50	39	32	82.05	7	17.95	43	86.00
2007	48	27	21	77.78	6	22.22	42	87.50
2008	250	96	71	73.96	25	26.04	225	90.00
2009	153	67	55	82.09	12	17.91	141	92.16
2010	204	97	75	77.32	22	22.68	182	89.22
2011	185	83	39	46.99	44	53.01	141	76.22
2012	430	144	64	44.44	80	55.56	350	81.40
2013	196	105	53	50.48	52	49.52	144	73.47
2014	307	125	62	49.60	63	50.40	244	79.48
2015	310	40	12	30.00	28	70.00	282	90.97
Total	2133	823	484	58.81	339	41.19	1974	84.11

Table 6—Domains associated with total, missing and recovered URL citations

Domains	Total URL citations		Missing URL citations		Recovered URL citations	
	Number	%	Number	%	Number	%
CO/COM	535	25.08	175	21.26	93	19.21
EDU	221	10.36	107	13.00	73	15.08
ORG	665	31.18	213	25.88	138	28.51
AC	236	11.06	110	13.37	75	15.49
GOV	112	5.25	50	6.08	26	5.37
NIC	24	1.13	13	1.58	10	2.06
ERNET	9	0.42	2	0.24	2	0.41
NET	60	2.81	21	2.55	10	2.06
RES	13	0.61	3	0.36		0
INT	15	0.70	7	0.85	4	0.82
GEO DOMAINS	218	10.22	109	13.24	44	9.09
MIL	3	0.14	1	0.12	1	0.20
INFO	22	1.03	12	1.46	8	1.65
Total	2133	100.00	823	100.00	484	100

rate of active URLs by 22.69%. It is also evident from Table 5 that the number of unrecovered URL citations were varied during the year 2006-2015. Unrecovered URL citations were high (70%) for the year 2015, while the number of unrecovered URL citations were low for the year 2008 (17.91%).

URL Domains

The domains associated with cited URLs, missing URL citations and recovery of missing URLs are summarized in Table 6. Of the 2133 URL citations, the organizational domain is the leading domain accounted for 31.18% followed by commercial domain (25.08%). This result is consistent with the results found in the previous studies conducted by Wren in 2008, and Janakiramaiah & Doraswamy in 2011[17-18].

The study also examined the extent of missing URLs by domain. Of the 823 missing URL citations, the highest percentage of missing URLs were associated with organizational domain (25.88%) followed by commercial domain (21.26%) and academic domain (13.37%). The available data also indicated that a very less number of missing URLs were associated with the domains such as .mil (0.12%), .ernet (0.24), and .res (0.36%).

Table 6 also indicates the recovery of missing URL citations with respect to their domains. The URLs

Table 7—File formats associated with total, missing and recovered URL citations

File formats	Total URL citations		Missing URL citations		Recovered URL citations	
	Number	%	Number	Number	%	Number
HTML/HTM	1553	72.81	533	34.32	314	58.91
PDF	382	17.91	205	53.66	116	56.59
ASP	62	2.91	32	51.61	19	59.38
NSF	4	0.19	3	75.00	3	100.00
DOC	11	0.52	8	72.73	6	75.00
PPT	6	0.28	4	66.67	4	100.00
TXT	4	0.19	2	50.00	1	50.00
JSP	14	0.66	4	28.57	0	0.00
PHP	67	3.14	19	28.36	12	63.16
CFM	19	0.89	9	47.37	8	88.89
CGI	11	0.52	4	36.36	1	25.00
Total	2133	100.00	823	38.58	484	58.81

Table 8—Path depth associated with total, missing and recovered URL citations

Path depth	Total URL citations		Missing URL citations		Recovered URL citations	
	Number	%	Number	%	Number	%
PD=0	185	8.67	27	14.59	19	70.37
PD=1	393	18.42	128	32.57	75	58.59
PD=2	567	26.58	209	36.86	117	55.98
PD=3	418	19.60	201	48.09	117	58.21
PD=4	309	14.49	132	42.72	83	62.88
PD=5	159	7.45	79	49.69	45	56.96
PD=6	55	2.58	27	49.09	13	48.15
PD=7	28	1.31	13	46.43	10	76.92
PD=8	8	0.38	3	37.50	2	66.67
PD=9	8	0.38	3	37.50	2	66.67
PD≥10	3	0.14	1	33.33	1	100.00
Total	2133	100.00	823	38.58	484	58.81

associated with .ernet and .mil were the fully recovered (100%) followed by .nic (76.92%), .edu (68.22%), and .ac (68.18%). It is also observed from the Table 6 that no missing URL citations belonged to .res (30%) domain have recovered.

URL File formats

Data presented in Table 7 shows that the highest percentage of URLs belonged to .html files [1553 (72.81%)]. This was followed by 382 PDF files (17.91%) URLs and PHP files [67 (3.14%)] came third. Remaining file formats of URLs cover only 6.16% of the total URL citations which is negligible. The highest rate of HTML file formats was also documented by earlier studies[13&19].

It is evident from Table 7 that the highest percentage of missing URLs were found among the URLs with .nsf files (75%), followed by .doc files (72.73%) and .ppt files (66.67%). The data presented in Table 7 also shows the recovery of missing URLs though *Internet Archive* by file formats.

Out of 484 missing URLs recovered through *Internet Archive*, the highest percentage (100%) of recovery was achieved for the missing URLs with .nsf and .ppt followed by .cfm file format (88.89%) and .doc file format (75%). A low percentage of recovery was achieved for the missing URLs with .cgi file format (25%) and no missing URLs with .jsp file format were recovered.

Path depth

The data presented in Table 8 reveals that of the 2133 URL citations, highest number of cited URLs belonged to the path depth level '2' (26.58%)

followed by path depth level '3' (19.60%). Very less number of URL citations with the path depth level 7 and above have been cited.

One of the objectives of this study was to understand the relationship between the path depth and missing URL citations. The highest percentage of missing URLs (49.69%) were found in the missing URL citations associated with Path depth level 5 followed by Path depth level 6 (49.09%) and Path depth level 7 (46.43%). Goh and Ng in 2007 have found that the lengthy path depth of the URL could cause link failure[20]. Therefore, an attempt was made to know the relationship between the path depth of URLs and the number of missing URLs. The correlation between the path depth and missing URLs was measured and found that there is a negative correlation between these two and the correlation is statistically significant (r=-0.6506, df=10, p=0.022).

However, the data presented in Table 8 indicates that the highest percentage of recovery was achieved for the missing URLs with the path depth level ≥10 (100%) followed by path depth level 7 (76.92).

Conclusion

The study concludes that the *Internet Archive* is capable of recovering timeworn URL citations. *Internet Archive* could be an indispensable asset for the authors of LIS research.

Further, study of missing URLs and their recovery provides an in-depth understanding of the present day dynamics of the web. Also the existence of the recovery tools such as *Internet Archive*, Memento, and other web archives allows the researcher to re-locate the missing web resources. The knowledge related to the dynamic web are essential for the library and information science professionals to plan things in a scientific way to locate, harvest, preserve and disseminate the web resources for longer durations.

References

1. Riahinia N, Zandian F and Azimi A, Web citation persistence over time: a retrospective study, *The Electronic Library*, 29(5) (2011) 609-620.
2. Spinellis D, The decay and failures of web references, *Communications of the ACM*, 46(1) (2003) 71–77.
3. Thorp A W and Brown L, Accessibility of Internet references in Annals of Emergency Medicine: is it time to require archiving? *Annals of Emergency Medicine*, 50(2)(2007)188-192.
4. Sellitto C, A study of missing web-cites in scholarly articles: towards an evaluation framework, *Journal of Information Science*, 30(6) (2004) 484-495.
5. Casserly MF and Bird JE, Web citation availability: analysis and implications for scholarship, *College and Research Libraries*, 64(4) (2003) 300-317.
6. Dimitrova DV and Bugeja M, Raising the dead: recovery of decayed online citations, *American Communication Journal*, 9(2) (2007) 2.
7. Nagaraja A, Joseph SA, Polen HH and Clauson KA, Disappearing act: Persistence and attrition of uniform resource locators (URLs) in an open access medical journal, *Program: Electronic Library and Information Systems*, 45(1) (2011) 98–106.
8. Gul S, Mahajan I and Ali A, The growth and decay of URLs citation: A case of an online Library and Information Science Journal, *Malaysian Journal of Library and Information Science,* 19(3) (2014) 27-39.
9. Moghaddam AI, Saberi MK and Esmaeel SM, Availability and half-life of web references cited in Information Research Journal: a citation study, *International Journal of Information Science and Management (IJISM)*,8(2) (2010) 57-75.
10. Tajeddini O, Azimi A, Sadatmoosavi A and Sharif-Moghaddam H, Death of web citations: a serious alarm for authors, *Malaysian Journal of Library and Information Science*, 16(3) (2011) 17-29.
11. Kumar S B T and Kumar M K, Persistence and half-life of URL citations cited in LIS open access journals,*Aslib Proceedings: New Information Perspectives*, 64(4) (2012) 405-422.
12. Wu Z, An empirical study of the accessibility of web references in two Chinese academic journals, *Scientometrics*, 78(3) (2009) 481-503.
13. Kumar S B T and Kumar V D, HTTP 404-page (not) found: Recovery of decayed URL citations, *Journal of Informetrics*, 7(1) (2013) 145-157.
14. Raj PKR and Kumar SBT, Web Citation trends in Indian LIS Journals: A Citations Analysis, *COLLNET Journal of scientometrics and Information Management*, 9(2) (2015) 298-310.
15. Zhao D Z and Logan E, Citation analysis using scientific publication on the web as data source: a case study in the XML research area, *Scientometrics,* 5(3) (2002) 449-72.
16. Kumar D V and Kumar B T S, Recovery of vanished URLs: Comparing the efficiency of Internet and Google, *Malaysian Journal of Library and Information Studies*, 22(2) (2017) 31-43.
17. Wren J D, 404 not found: the stability and persistence of URLs published in MEDLINE, *Bioinformatics*, 20(5) (2004) 668-672.
18. Janakiramaiah M and Doraswamy M, Measuring impact of web resources in conference proceedings: A citation analysis *In CALIBER 2011,* (2011) 541–549.
19. McCown F, Chan S, Nelson ML and Bollen J, The availability and persistence of Web citations in D-Lib Magazine, (2005) Available from http://iwaweuroparchive-org/05/papers/iwaw05-mccown1pdf.
20. Goh DHL and Ng PK, Link decay in leading information science journals, *Journal of the American Society for Information Science and Technology*, 58(1) (2007) 15–24.