RECOUPING THE MISSING WEB CITATIONS IN LIBRARY HI-TECH JOURNAL

Dr. D. Vinay Kumar
Dr. B. T. Sampath Kumar

Dr. D. Vinay Kumar
Lecturer, Department of Library and Information Science, Kuvempu University, Jnanasahyadri, Shankaraghatta, Shivamogga, Karnataka, India
Email: vinay.86.kumar@gmail.com
Corresponding Author

And

Dr. B. T. Sampath Kumar
Professor, Department of Library and Information Science, Tumkur University, Tumakuru, Karnataka, India
Email: sampathbt_2001@gmail.com

This paper is an attempt to know the use of web citations, extent of missing web citations, and recovery of missing web citations using Time Travel tool. The study collected 3912 web citations cited in 568 research articles published in Library Hi-Tech Journal during the years 2006-2017. Of the total cited web citations 911 web citations (23.29%) are not accessible and 3001 (76.71%) are accessible. The attempt to recover 911 missing web citations through Time Travel tool has increased the active web citations from 3001 to 3595 that accounted for 91.90% of active web citations. The study also found that *Internet Archive* has recovered the highest number of missing web citations than other web archives as shown in the Time Travel platform.

Keywords: Web Citations, Missing web citations, Active web citations, Internet Archive, Time Travel

INTRODUCTION

The web is a familiar source of information among the research community that offers easy access to the abundant collection of scholarly literature (Casserly and Byrd, 2003). The web grants navigation among different web documents that eventually increase the use of web resources in scholarly literature. The increasing tendency of use of the electronic resources gradually increased the citation to web resources. The web citation links the researcher to a much larger community of related works. The Library and Information Science (LIS) research are not exempted from accessing and citing URLs of web resources. The increased use of web citations in scholarly literature posed a challenge to the researchers, academics, and librarians that promoted them to investigate the accessibility of web resources cited in scholarly works (Chen et al., 2014). Many early studies experienced the presence of the inaccessible URLs cited in LIS scholarly literature (McCown *et al.*, 2005; Sampath Kumar and Manoj Kumar, 2012; Gul et al, 2014). Subsequently, many studies tried to recover the inaccessible URLs using various tools such as web archives, search engines. In this study, an attempt has been made to know the use of URLs as citations in Library Hi-Tech journal articles and their accessibility. Further, the study tried to recuperate the missing URLs using Time Travel. It is a tool which yields the archived web resources from many web archive initiatives.

REVIEW OF LITERATURE

Previous studies witnessed the prevalence of URLs as citations in scholarly literature. Many studies also documented the trend of URLs and the problem of inaccessibility of URLs cited in research articles. Despite the problem of web decay exist, a number of studies focused on the recovery of missing URLs.

The use of URLs as citations and their inaccessibility in LIS scholarly literature are highly cited in several studies. A study by Sadat-Moosavi et al. (2012) reported an increase in the accessibility rate URLs from 64 to 95 percent by using the Wayback Machine and Google in four LIS scholarly journals. The use of Internet Archive was useful for Sampath Kumar and Vinay Kumar in the recovery of missing URLs. Their study in the year 2013 examined the trend of URLs in two Indian LIS Journals. They extracted a total of 1,290 URLs cited in 472 research articles published in two Indian LIS journals spanning a period of 9 years (2002–2010). The study documented that 39.84% of missing Web citations and remaining 60.15% of Web citations were still accessible. The Internet Achieve recovered 44.55% of missing URLs and increased the percentage of accessible URLs from 39.84% to 77.90% (Sampath Kumar and Vinay Kumar, 2013).

Gul et al. (2014) analyzed 1184 web citations cited in 73 articles published in Ariadne-web magazine. This study confirmed the loss of 7.85% of cited web resources. The authors indicated that the use of WebCite and LOCKSS could prevent web decay. A study by Prithvi raj and Sampath Kumar (2014) conducted on three Indian LIS conference proceedings yielded 5,698 web citations. The study showed that 50.9% of web citations (2,854) were missing. They used Internet Archive to recover the missing web citations and succeeded to retrieve 29.71% of missing web citations. The study found that the use of Internet Archive could retrieve the missing web citations which resulted in an increase in the percentage of active URLs from 49.91% to 79.08% (Prithviraj and Kumar. In the same year, Klein et al. (2014) observed the reference rot after studying over one million web resources cited in over 3.5 million articles belonged to Science, Technology and Medicine disciplines. They found that the one in five articles suffered from reference rot. Another study by Burnhill et al. (2015) examined 46,000 URIs appended to 6,400 e-theses downloaded from institutional repositories of five US institutions. This study witnessed 36.7% of URIs are not available on the live web and they tried to recover them from the Time Travel. The study found that 18.3% of rotten links were preserved in web archives.

Recently, Sife and Lwoga (2017) extracted 574 web citations cited in 822 articles published in four East African health science online journals during 2001-2015. The authors found 253 (44.1%) web citations were inaccessible. This study reported that only 36 (6.3%) web citations recovered using the Wayback Machine. Similarly, Tajeddini et al. (2017) found that there were 4562 online citations cited in 1109 articles published in six LIS scholarly journals. The rate of inaccessible URLs was 34% that reduced to 5% after adopting various techniques. They used Internet Archive which extracted 11% of the inaccessible URLs.

Vinay Kumar and Sushmitha (2019) studied 1105 URLs cited in 342 research articles published in Annals of Library and Information

Studies during 2006-2015. The study found that 43.44% of URLs were missing. They used Time Travel to recover the missing URLs and succeeded in the recovery of more than 50% of missing URLs. Another recent study by Krol (2019) studied the link rot in websites of rural tourism facility. He collected 919 websites and found that 464 sites have broken links. Also approximately 65% of them have few broken links. This paper opined that the age of website relates to the link rot phenomenon. The previous studies have elucidated the susceptible nature of web resources cited in LIS scholarly literature. An extensive volume of literature affirmed the existence of the problem of missing URLs cited in scholarly literature. The previous literature documented the use of recovery tools such as web archives, search engines or Time Travel and recouped around 11% to 59% of missing URLs.

RESEARCH QUESTIONS

1. What percentage of URLs used as citations in articles published in Library Hi-Tech journal?
2. What is the extent of missing web citations?
3. What percentage of missing web citations can be recovered through Time Travel.

METHODOLOGY

The study has located 568 research articles published in Library Hi-Tech journal during 2006-2017. Library Hi-Tech journal has been published by Emerald Publications which was taken as a representative LIS journal to know the trend of URL citations in LIS research journal as well as the accessibility of cited URLs. The journal has the impact factor of 1.014. The study considered 14,962 references cited in 568 research articles published in Library Hi-Tech during the years 2006-2017. Of the 14,962 references, 3912 were web citations. The study also attempted to check the accessibility of cited URLs. The web citations were separated from the reference list and once the extraction is completed, the URLs of web citation were tested for their accessibility using *W3C Link Checker* (http://validator.w3.org/checklink). This tool tests a submitted URL for broken hypertext links and reports the HTTP error code matching with the missing URL. The web citations with HTTP error codes were considered as missing and the accessible web citations were considered as active.The study further attempted to recover missing web citations using Time Travel. Therefore, the missing URLs of web resources were submitted to the search box of Time Travel and the 'Find' button was clicked. The recovered URLs were recorded as active and the rest were considered as unrecovered.

Time Travel

Time Travel is a free service that helps to find the web pages that existed currently or in the past. According to Time Travel, the user can access the web pages archived in more than twenty web archives like Bibliotheca Alexandrina Web Archive, Internet Archive, Library of Congress Web Archive, UK Web Archive, WebCite etc. When the URL of a webpage/website is entered in the search box of Time Travel (figure-1), it locates the web page archived in more than 23 web archives and bring back at least one result per archive. The Time Travel tries to recover the web page closer to the time requested by the user. The incompatible scripts or the robots.txt-enabled web pages hinder the process of retrieving exact webpage requested through the Time Travel (Time Travel, 2016).

The Table 1 indicates the distribution of articles, citations, and web citations appeared in Library Hi-Tech journal during 2006-2017. While, 568 articles consists of 14,962 citations (26.34 citations per article) published in a period of 12 years. The table also shows that an inconsistent growth in the percentage of citations and a gradual increase after the year 2010. Meanwhile, the average web citations per article are 6.89 during this period. The study also recorded inconsistency for the average web citations that ranged between a low of 4.56 web citations per

ANALYSIS AND INTERPRETATION

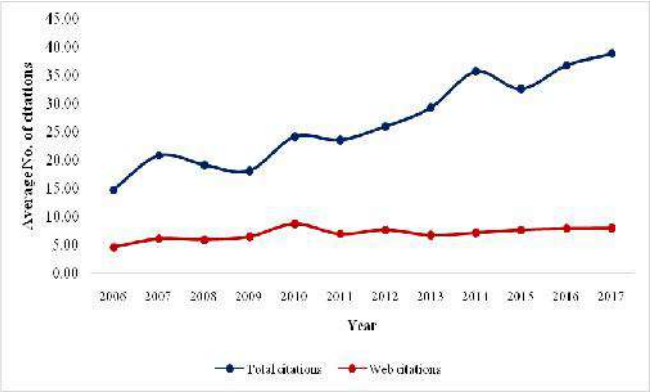
Table 1: Distribution of articles, citations and web citations

Year	Total Articles	Percentage	Total citations	Average citation per article	Web citations	Average web citation per article
2006	50	8.80	737	14.74	228	4.56
2007	49	8.63	1020	20.82	295	6.02
2008	51	8.98	978	19.18	299	5.86
2009	48	8.45	870	18.13	307	6.40
2010	47	8.27	1136	24.17	407	8.66
2011	51	8.98	1202	23.57	351	6.88
2012	48	8.45	1247	25.98	365	7.60
2013	46	8.10	1347	29.28	306	6.65
2014	46	8.10	1642	35.70	326	7.09
2015	41	7.22	1339	32.66	312	7.61
2016	46	8.10	1694	36.83	360	7.83
2017	45	7.92	1750	38.89	356	7.91
Total	568	100.00	14962	26.34	3912	6.89

Figure-2: Growth of Average citations/w

article in the year 2006 to a high of 8.66 citations per article in the year 2010.

The correlation test indicates that the number of citations and the average number of web citations is positively correlated (r=0.6902, p=0.009). Hence, it is arguable that the increase in the total number of citations duly increases the number of web citations.



The Table 2 indicates the distribution of active and vanished URLs by year. The W3C link checker was used to identify the accessibility of 3912 web citations and found that 3001 (76.71%) are accessible, while the remaining web citations (23.29%) encountered access errors. The percentage of missing web citations have been decreased from 41.36% in the year 2007 to 10.11% in the year 2017.

The statistical analysis shows that the percentage of missing URLs and their age are positively correlated (r=0.927, p=.000) and the correlation is statistically significant. The result of the study is consistent with the findings of the previous study (Sampth Kumar and Vinay Kumar, 2013). The analysis of the data also confirmed that the web citations cease to exist over time.

Table 2: Distribution of active and missing web citations

Year	Web citations	Active web citations	Percentage of active web citations	Missing web citations	Percentage of missing web citations
2006	228	134	58.77	94	41.23
2007	295	173	58.64	122	41.36
2008	299	189	63.21	110	36.79
2009	307	206	67.10	101	32.90
2010	407	286	70.27	121	29.73
2011	351	266	75.78	85	24.22
2012	365	310	84.93	55	15.07
2013	306	257	83.99	49	16.01
2014	326	280	85.89	46	14.11
2015	312	263	84.29	49	15.71
2016	360	317	88.06	43	11.94
2017	356	320	89.89	36	10.11
Total	**3912**	**3001**	**76.71**	**911**	**23.29**

web citations

Table 3: Number of missing web citations recovered by different web archives as shown in Time Travel

Year	No. of Missing web citations	Internet Archive	Library of Congress web archive	Bib Alexandria	WebCite	archive.is	Archive-it	Portughese Archive	Web Harvest	UK Web archive	Proni.
2006	94	63	22	11	7	6	2	9	4	3	1
2007	122	76	27	16	7	7	3	10	3	3	1
2008	110	65	25	13	4	3	0	5	1	2	0
2009	101	62	26	15	4	7	2	6	2	1	0
2010	121	72	27	18	5	4	1	5	2	0	0
2011	85	51	19	12	2	1	0	4	1	3	0
2012	55	40	12	7	3	5	2	3	2	1	1
2013	49	29	12	6	1	1	0	2	0	1	0
2014	46	28	11	7	1	4	1	3	1	1	1
2015	49	25	7	6	0	1	1	4	1	1	0
2016	43	25	9	6	4	1	4	2	2	1	1
2017	36	19	4	2	1	0	1	1	0	1	0
Total	**911**	**555**	**201**	**119**	**39**	**40**	**17**	**54**	**19**	**18**	**5**
% of recovery		**60.92**	**22.06**	**13.06**	**4.28**	**4.39**	**1.87**	**5.93**	**2.09**	**1.98**	**0.55**

Table 4: Distribution of missing, recovered, and unrecovered web citations

Year	Total web citations	Missing web citations	Percentage of missing web citations	Recovered missing web citations	Active web citations before recovery	Active web citations after recovery	Percentage of active web citations after recovery	Unrecovered missing web citations	Percentage of missing web citations after recovery
2006	228	94	41.23	72	134	206	90.35	22	9.65
2007	295	122	41.36	82	173	255	86.44	40	13.56
2008	299	110	36.79	67	189	256	85.62	43	14.38
2009	307	101	32.90	68	206	274	89.25	33	10.75
2010	407	121	29.73	76	286	362	88.94	45	11.06
2011	351	85	24.22	54	266	320	91.17	31	8.83
2012	365	55	15.07	42	310	352	96.44	13	3.56
2013	306	49	16.01	30	257	287	93.79	19	6.21
2014	326	46	14.11	31	280	311	95.40	15	4.60
2015	312	49	15.71	25	263	288	92.31	24	7.69
2016	360	43	11.94	28	317	345	95.83	15	4.17
2017	356	36	10.11	19	320	339	95.22	17	4.78
Total	3912	911	23.29	594	3001	3595	91.90	317	8.10

An attempt to recover 911 missing web citations through the Time Travel shows that the Wayback Machine recovered the highest percentage of missing web citations (60.92%) followed by Library of Congress Web Archive (22.06%) and Bib Alexandria (13.06%). The rest of the web archives recovered a meager percentage of missing URLs.

The Table 4 shows that of the 911 missing web citations before recovery, the Time Travel brought back 594 web citations which has increased the number of active web citations from 3001 (before recovery) to 3595. However, it is also evident that 317 web citations (8.10%) remain missing and unrecovered from Time Travel. It is clear from t-test that there is a significant difference between the number of missing web citation before and after recovery of missing web citations (t=4.9325, p=0.000).

DISCUSSION AND CONCLUSION

The study found that of the 26.34 citations per article, 6.89 citations belonged to web resources. This could be due to the extensive use of the web as an essential source of information. Nevertheless, free access to the digital information endorsed the LIS authors to cite more web resources in their scholarly publications.

The study witnessed that 23.29% of web citations are missing. The study also found that more than forty percent of missing web citations have been cited a decade ago. Hence, it is evident that the longevity of web citation is questionable and it obstructs the present day readers from accessing scholarly content. The study has attempted to recover missing URLs using the Time Travel and recouped 60% of missing web citations. Hence, the Time Travel is found to be useful Internet tool for researchers, webmasters, and LIS researchers to recuperate the missing web citation and eventually obtain access to increased number of web resources.

Even though Time Travel is a noteworthy tool to recover missing URLs, the problem of missing web citations still persists. The study witnessed 317(8.10%) web citations are inaccessible and the use of Time Travel also unable to recover them. Earlier, Sampath Kumar and Vinay Kumar in 2013 opined that authors lack appropriate mechanisms to preserve their scholarly content permanently over the web. This may be due to the vastness and dynamicity of the web. At this juncture, the webmasters who manage the web-based scholarly content can voluntarily submit URLs to Web Archives. One such tool is the "Webrecorder" (https://webrecorder.io) that captures the submitted URL of a website or webpage. Meanwhile, in the ambient of Internet Archive, webmasters can archive their Web pages in the Wayback Machine (http://www.archive.org/) by embedding the following link.

```
<div id="wb404"/>
<scriptsrc="https://archive.org/web/wb404.js"></script>
```

If the webpage disappears, the embedded link helps to bring back the missing webpage from Internet Archive's memory. Further, it is also essential that the webmasters need not disallow the website from being crawled by web

archives so that the permanent preservation of web resource is possible.

REFERENCES

1. Burnhill, P., Mewissen, M., & Wincewicz, R. (2015). Reference rot in scholarly statement: threat and remedy. *Insights, 28*(2).

2. Casserly, M., & Byrd, J. (2003). Web citation availability: analysis and implications for scholarship.*College and Research Libraries. 64*(4), 300–317.

3. Gul, S., Mahajan, I., & Ali, A. (2014). The growth and decay of URLs citation: A case of an online Library & Information Science journal. *Malaysian Journal of Library & Information Science*, *19*(3).

4. Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & Tobin, R. (2014). Scholarly context not found: one in five articles suffers from reference rot. *PloS one*, *9*(12), e115253.

5. Król, K. (2019). The link rot phenomenon and its influence on the quality of the websites of rural tourism facilities in Poland. *Economic and Regional Studies/ Studia Ekonomiczne i Regionalne*, *12*(1), 68-79.

6. McCown, F., Chan, S., Nelson, M. L., & Bollen, J. (2005). The availability and persistence of web references in D-Lib Magazine. *arXiv preprint cs/0511077*.

7. Prithviraj, K. R., & Sampath Kumar, B. T. (2014). Corrosion of URLs: Implications for electronic publishing. *IFLA journal*, *40*(1), 35-47.

8. Sadat-Moosavi, A., Isfandyari-Moghaddam, A., & Tajeddini, O. (2012). Accessibility of online resources cited in scholarly LIS journals: A study of Emerald ISI-ranked journals. *Aslib Proceedings, 64*(2), 178-192.

9. Sampath Kumar, B. T., & Manoj Kumar, K. S. (2012). Persistence and half-life of URL citations cited in LIS open access journals. *Aslib Proceedings*, *64*(4), 405-422.

10. Sampath Kumar, B. T., & Vinay Kumar, D. (2013). HTTP 404-page (not) found: Recovery of decayed URL citations. *Journal of Informetrics*, *7*(1), 145-157.

11. Sife, A. S., & Lwoga, E. T. (2017). Retrieving vanished Web references in health science journals in East Africa. *Information and Learning Science*, *118*(7/ 8), 385-392.

12. Tajeddini, O., Azimi, A., & Moghaddam, H. S. (2017). Death of web citations: a serious alarm for authors. *Malaysian Journal of Library & Information Science*, *16*(3), 17-29.

13. Vinay Kumar, D., & Sushmitha, M. (2019). Recovery of missing URLs cited in Annals of Library and Information Studies: a study of Time Travel. *Annals of Library and Information Studies, 66*(1), 24-32.

❖❖❖