Indian Journal of Information Sources and Services
ISSN: 2231-6094 Vol.11 No.1, 2021, pp.16-21

Recovery of Web Citations using Time Travel in Journal of Informetrics

B. Niveditha[1] and Mallinath Kumbar[2]
[1]UGC-Senior Research Fellow, [2]Professor,
[1&2]Department of Library and Information Science, University of Mysore, Mysore, Karnataka, India
E-mail: niveditha.jb@gmail.com

Abstract **- The present paper explores the accessibility and recovery of web citations using Time Travel in the Journal of Informetrics during the year 2008-2017. A total of 647 articles were downloaded and 24901 references were extracted. Out of 3546 web citations2084 references contained URLs, DOIs were found in 1298 references and 164 references contained arXiv, WOS article identifier, etc. It was found that 3163 web citations were accessible and the remaining 383 web citations were missing. The study also investigated the characteristic features of display and destination URLs like the path depth, URL length, file format, and top-level domain.**
Keywords: **References, Web Citations, URLs, DOIs, Accessibility, Recovery, Time Travel**

I. INTRODUCTION

The Internet which is used for gaining access to all kinds of information online is emerging as a powerful medium for academicians. Electronic or online resources are now supplementary to printed information sources. They endorse efficient information dissemination and a norm for academicians for conducting research. The major advantage of using electronic resources is the ease of access to the needed information. With the electronic resources becoming a vital tool in conducting research, references to these resources have also increased significantly. Though there is increased use of web citations, the decay of the web resources has also emerged. It is important to make the constructive use of these resources, so that the current and future researchers have access to the electronic information sources. The need of the hour is to address the issue of accessibility. The present study has thus sought to explore the effect of accessibility and recovery of web citationsusing Time Travel in Journal of Informetrics during the period 2008-2017.

II. OBJECTIVES OF STUDY

This study aims to investigate the accessibility and recovery of web citations in "Journal of Informetrics" during the year 2008-2017. This study addresses the following objectives

1. To know the proportion of web citations in the Journal of Informetrics.
2. To determine the percentage of URLs and DOIs in the Journal of Informetrics.
3. To know the percentage of inaccessible web citations.
4. To identify the file format, path depth, top-level domain, character length of web citations.
5. To recover the inaccessible web citations through Time Travel.

III. METHODOLOGY

For the present study, data was drawn from "Journal of Informetrics." The journal was selected based on its high impact factor of 3.484 as per Clarivate Analytics' 2018 "Journal Citation Report." All the research articles published during 2008-2017 were taken up for the study. Editorial notes, book reviews, short communication were excluded. The references that were adjoined at the end of each article were considered for the study. A total of 24901 references were selected from 647 articles published in "Journal of Informetrics." The references that contained web links and DOIs were extracted as the study deals with the accessibility and recovery of the web citations. A total of 3546 web citations were extracted and were checked for their accessibility in the web browser. The W3C link checker (http://validator.w3.org/checklink) was used to report the HTTP error message for inaccessible URLs. Further, the features of the web citations, such as their length, top-level domain, file format and path depth were found for the display as well as for destination URLs.

The study used Time Travel (http://timetravel. mementoweb.org/) to find whether the web citations were archived or not. The Time travel recovers the inaccessible web citations that are archived in Internet Archive, Library of Congress Web Archive, Archive-it, Perma-cc, etc. The study also made an attempt to find whether the inaccessible URLs were available by searching through their title present in their respective reference. The web citations that were not archived or not found through title search were considered as missing web citations.

IV. DATA ANALYSIS AND INTERPRETATION

A. Year-Wise Distribution of Web Citations

Table I shows that the 647 research articles published in "Journal of Informetrics" during the years 2008-2017 consisted of 24901citations. The percentage of web citations varied from a low of 5.61% during the year 2008 to a high of 22.99% in the year 2015.

TABLE I DISTRIBUTION OF ARTICLES, REFERENCES AND WEB CITATIONS IN JOI

Year	Total articles	Total references	Average reference per article	Total web citations	Percentage
2008	31	891	28.74	50	5.61
2009	30	1249	41.63	85	6.81
2010	62	2033	32.79	172	8.46
2011	55	1963	35.69	180	9.17
2012	64	2128	33.25	177	8.32
2013	87	3197	36.75	263	8.23
2014	78	2816	36.10	296	10.51
2015	80	3353	41.91	771	22.99
2016	79	3558	45.04	734	20.63
2017	81	3713	45.84	818	22.03
Total	647	24901	38.49	3546	14.24

B. Distribution of URLs and DOIs

The use of DOIs in place of URLs has increased in the recent years to prevent the deterioration of web citations. The DOI is a character string used to identify intellectual property in the digital environment. Table II shows the distribution of URL and DOI in Journal of Informetrics. It was found that out of the total 3546web citations, 1332 were URL links, 2050 were DOIs and 164 were arXiv identifier and WOS article identifier.

TABLE II YEAR-WISE DISTRIBUTION OF URLS AND DOIS

Year	URL		DOI		Others		Total web citations
	Number	%	Number	%	Number	%	
2008	42	84.00	6	12.00	2	4.00	50
2009	72	84.71	5	5.88	8	9.41	85
2010	106	61.63	39	22.67	27	15.70	172
2011	85	47.22	87	48.33	8	4.44	180
2012	107	60.45	65	36.72	5	2.82	177
2013	138	52.47	116	44.11	9	3.42	263
2014	91	30.74	199	67.23	6	2.03	296
2015	199	25.81	526	68.22	46	5.97	771
2016	272	37.06	444	60.49	18	2.45	734
2017	220	26.89	563	68.83	35	4.28	818
Total	1332	37.56	2050	57.81	164	4.62	3546

C. Distribution of Accessible and Inaccessible Web Citations

Table III gives the distribution of accessible and inaccessible web citation by year. The web citations were checked in the web browser by clicking directly on the URLs.

For this all the DOIs and arXiv were resolved to URLs using the syntax https://doi.org/ and https://arxiv.org/ respectively.

The result of the accessibility check indicated that of the 3163 web citations, 89.20% were accessible while the remaining 10.80% encountered accessibility error.

D. File Format of Web Citations

The data illustrated in Table IV indicates that the greatest number of cited web resources were .html files. Out of the total 3546 web citations, 3153 were .html files, followed by 303.pdf files, 34.php file, 24 .asp files, 21 .cfm file and 11 .jsp files.

TABLE III DISTRIBUTION OF ACCESSIBLE AND INACCESSIBLEWEB CITATIONS

Year	Total web citations	Accessible web citations	Percentage	Inaccessible web citations	Percentage
2008	50	37	74.00	13	26.00
2009	85	60	70.59	25	29.41
2010	172	141	81.98	31	18.02
2011	180	146	81.11	34	18.89
2012	177	148	83.62	29	16.38
2013	263	225	85.55	38	14.45
2014	296	272	91.89	24	8.11
2015	771	716	92.87	55	7.13
2016	734	671	91.42	63	8.58
2017	818	747	91.32	71	8.68
Total	3546	3163	89.20	383	10.80

TABLE IV FILE FORMAT OF WEB CITATIONS

File format	Number	Percentage
.html	3153	88.92
.cfm	21	0.59
.pdf	303	8.54
.php	34	0.96
.asp	24	0.68
.jsp	11	0.31
Total	3546	100.00

E. Path Depth of Display and Destination URL

In this study, an attempt has been made to distinguish the characteristics between the display and destination URL. Display URL is the URL displayed to the user. It is the web address or the DOI found at the end of the reference which is given by a central server regardless of the article's physical location. The destination URL is the URL where the user is eventually taken after multiple redirects are involved. It is the article landing page or where the article resides. The article landing page is under the control of the publisher.

The distribution of display URL by path depth is shown in table V. Out of 3546 web citations, display URLs with path depth 2 (2452) were frequently cited, followed by 384 URLs with depth of 3. A total of 227 URLs had path depth 4 and 1, 89URLs with path depth 1, 80 URLs with depth of 1, 80 URLs had path depth 5,41 URLs with path depth 6 and 24 URLs with path depth 7, 22 URLs had path depth more than 7. Unlike the display URL, almost 1112 destination URLs have a path depth of 2, followed by 765 URLs having a path depth of 5,707 URLs have a path depth of 4, 461 have a path depth of 1, 121 have a path depth of 6, 33 URLs have a path depth of 7, 25 URLs had path depth more than 7 and 2 URLs had path depth 0.

TABLE V PATH DEPTH OF DISPLAY AND DESTINATION URL

Path Depth	Display URL	Percentage	Destination URL	Percentage
PD = 0	89	2.51	2	0.06
PD = 1	227	6.40	461	13.00
PD = 2	2452	69.15	1112	31.36
PD = 3	384	10.83	707	19.94
PD = 4	227	6.40	765	21.57
PD = 5	80	2.26	320	9.02
PD = 6	41	1.16	121	3.41
PD = 7	24	0.68	33	0.93
PD>7	22	0.62	25	0.71
Total	3546	100.00	3546	100.00

F. Character Length of Display and Destination URL

Table VI shows the URL length and it can be found that a total of 1662 display URLs had length 41-50, 944 URLs had length of 31-40, 259 URLs had a length of 51-60. Unlike the display URL, only 375 destination URL had length 41-50. The highest number of destination URLs (939) had length of 61-70, followed by 731 destination URLs having length of 51-60.

TABLE VI CHARACTER LENGTH OF DISPLAY AND DESTINATION URL

Character length	Display URL	Percentage	Destination URL	Percentage
<20	29	0.82	108	3.05
21-30	185	5.22	139	3.92
31-40	944	26.62	396	11.17
41-50	1662	46.87	375	10.58
51-60	259	7.30	731	20.61
61-70	172	4.85	939	26.48
71-80	93	2.62	283	7.98
81-90	66	1.86	312	8.80
91-100	54	1.52	48	1.35
>100	82	2.31	215	6.06
Total	3546	100.00	3546	100.00

G. Top-Level Domain of Display and Destination URL

The top-level domain associated with the display and destination URL is summarized in table VII. It can be seen that a total of 2571 display URLs had the organizational top-level domain, followed by 430 having the commercial top-level domain. On the other hand, a total of 2003 destination URL have commercial top-level domain followed by 1002 organization domain.

TABLE VII TOP-LEVEL DOMAIN OF DISPLAY AND DESTINATION URL

Top-level Domain	Display URL	Percentage	Destination URL	Percentage
.com	430	12.13	2003	56.49
.edu	107	3.02	120	3.38
.gov	58	1.64	53	1.49
.info	5	0.14	5	0.14
.net	46	1.30	34	0.96
.org	2573	72.56	1002	28.26
Others	327	9.22	329	9.28
Total	3546	100.00	3546	100.00

H. HTTP Errors Associated with Missing Web Citations

TABLE VIII HTTP ERRORS ASSOCIATED WITH INACCESSIBLE WEB CITATIONS

Error Type	Total	Percentage
400	2	0.52
403	65	16.97
404	239	62.40
406	1	0.26
408	1	0.26
410	1	0.26
500	72	18.80
503	2	0.52
Total	383	100

The missing or inaccessible URLs were checked in W3C link checker (http://validator.w3.org/checklink) to report the HTTP error codes. Table VIII shows that the HTTP 404 error message that is “Page not Found” error was the error message that mostly occurred and represented 62.40% of all the HTTP error messages. It is followed by HTTP 500 (18.80%), HTTP 403 (16.97%) error messages and HTTP 406,408 410 with 0.26%.

I. Distribution of Recovered Web Citations

The study intended to recover the inaccessible web citations through Time Travel tool. The inaccessible web citations were entered in the search box of “Time Travel.” Table IX depicts that out of the 383 inaccessible web citations, 227 were recovered and the remaining 156 have not been recovered.

TABLE IX DISTRIBUTION OF RECOVERED WEB CITATIONS

Year	Total number of inaccessible web citations	Number of inaccessible web citations recovered through Time Travel	Percentage	Number of inaccessible web citations not recovered	Percentage
2008	13	4	30.77	9	69.23
2009	24	8	33.33	16	66.67
2010	31	4	12.90	27	87.10
2011	34	5	14.71	29	85.29
2012	29	29	100.00	0	0.00
2013	38	34	89.47	4	10.53
2014	24	15	62.50	9	37.50
2015	55	42	76.36	13	23.64
2016	63	57	90.48	6	9.52
2017	72	29	40.28	43	59.72
Total	383	227	59.27	156	40.73

J. Distribution of Recovered Web Citations in Web Archives

Table X shows the distribution of archived web citations in various web archives. The Internet Archive recovered the highest percentage of inaccessible web citations (59.39%), followed by Library of Congress web archive (9.90) and Web citation memento (9.02%).

TABLE X DISTRIBUTION OF ARCHIVED WEB CITATIONS IN TIME TRAVEL

Year	Total number of inaccessible web citations	Internet Archive	LOC	Archive it	Perma.cc	Archive.is	Arquivo.pt	Stanford Web Archive	Icelandic Web Archive	UK Web Archive	Web Citation Memento	Bibliotheca Alexandrina	Canadian Archive Memento	UK Government Web archive
2008	13	4	0	2	1	2	1	0	0	1	1	2	0	0
2009	24	6	4	2	1	4	3	3	1	1	0	1	1	0
2010	31	3	2	0	0	0	0	0	0	0	1	0	0	0
2011	34	5	2	0	0	3	1	0	0	0	2	0	0	0
2012	29	4	0	0	0	1	29	0	0	0	0	0	0	0
2013	38	13	0	1	0	5	34	0	0	1	1	0	0	0
2014	24	7	0	0	0	3	13	0	0	0	3	0	0	0
2015	55	10	2	2	0	6	42	0	0	1	1	0	0	0
2016	63	20	2	1	0	5	56	0	0	0	1	0	0	0
2017	72	22	3	1	2	2	9	2	0	0	2	0	0	1
Total	383	94	15	9	4	31	188	5	1	4	12	3	1	1

K. Title-Wise Search for Inaccessible Web Citations

An attempt was made in this study to search the inaccessible web citations by their title present in their respective cited reference. It is found from table XI that out of the 383web citations, 175web citations were found through title search. The percentage of web citations found through title search varied a low of 1.75% cited during the year 2008 to a high of 6.24% cited in the year 2016.

TABLE XI TITLE-WISE SEARCH FOR INACCESSIBLE WEB CITATIONS

Year	Total number of inaccessible web citations	Number of inaccessible web citations recovered through Title search	Percentage	Number of inaccessible web citations not recovered through Title search	Percentage
2008	13	9	1.75	4	0.08
2009	24	12	2.34	12	0.23
2010	31	20	3.90	11	0.21
2011	34	15	2.92	19	0.37
2012	29	16	3.12	13	0.25
2013	38	13	2.53	25	0.49
2014	24	12	2.34	12	0.23
2015	55	27	5.26	28	0.55
2016	63	32	6.24	31	0.60
2017	72	19	3.70	53	1.03
Total	383	175	34.11	208	4.05

V. CONCLUSION

The digital revolt, which involves WWW and the Internet, has made a huge revolution in research. The extensive use of electronic resources has made a positive impact on the research process. Because of this widespread utilization of electronic resources by the academicians, the use of web citations has also increased in the recent past. The present study confirms the use of web citations in the references cited in the Journal of Informetrics during the year 2008-2017. Nevertheless, the accessibility of the web resources is of a major concern, as they tend to decay after a specific period of time. The findings of the study showed that the DOI was used in recent years as they permanently identify an article or a document in the digital environment to overcome the problem of inaccessibility. Further, it was found from the study that the changes in path depth and character length of a display and destination URL is because the DOI server translates the DOI-based link into its actual URL on the publication server or where the article resides. If the article is moved to a different location, the DOI-based URL should redirect to the new location. It is thus the responsibility of the publisher to ensure that the current information is registered for each DOI. Moreover, to avoid the use of inaccessible web references in publications, the authors should check for their availability before using them in their references as well as the publishers should also verify them before publications. It is important for the web citations to be archived by the authors as well as publishers. Hence, the publishers, editors, and authors should systematically check the web citations before publication to ensure that the cited web citations are available for the academicians to accomplish their research activity.

REFERENCES

[1] Isfandyari-Moghaddam A., Saberi M.K. and Mohammad Esmaeel S. (2010). Availability and Half-life of Web References Cited in Information Research Journal: A Citation Study. *International Journal of Information Science and Management*, 8(2), 57-75.

[2] Mardani A. (2012). An investigation of the web citations in Iran's chemistry articles in SCI. *Library Review, 61*(1), 18-29.

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

[4] Niveditha, B. and Mallinath Kumbar. (2020). Accessibility and Characteristics of Web Citations in Journal of Computer-Mediated Communication during 2008-2017. *Journal of Indian Library Association, 56*(2), 39-50.

[5] Niveditha, B. and Mallinath Kumbar. (2020). Availability of Web Citations in Scholarly Library and Information Science Journals: A Study. *PEARL - A Journal of Library and Information Science, 14*(2), 202-208.

[6] Prithvi Raj, K. R. and Sampath Kumar, B. T. (2015). Web Citation Trends in Indian LIS Journals: A Citation Analysis. *COLLNET Journal of Scientometrics and Information Management, 9*(2), 295–310.

[7] Sampath Kumar, B. T. and Vinay Kumar, D. (2013). HTTP 404-page (not) found: recovery of decayed URLs. *Journal of Informetrics, 7*(1), 145-157.

[8] Spinellis, D. (2003). The Decay and Failures of Web References. *Communications of the ACM, 46*(1), 71-77.

[9] Vinay Kumar, D. and Sampath Kumar, B. T. (2017). Prevalence of URLs in Library and Information Science (LIS) Literature: A Citation Analysis. *COLLNET Journal of Scientometrics and Information Management, 11*(2), 287-297.

[10] Vinay Kumar, D. and 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.

[11] Wu, Z. (2008). An empirical study of the accessibility of web references in two Chinese academic journals. *Scientometrics, 78*(3), 481-503.

[12] Zhang, Y. (2007). The Effect of Open Access on Citation Impact: A Comparison Study Based on Web Citation Analysis. *Libri, 56*(3), 145-156.