

University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Library Philosophy and Practice (e-journal)

Libraries at University of Nebraska-Lincoln

2021

AN ANALYSIS OF PERSISTENCE AND OBSOLESCENCE OF WEB CITATIONS OF PAVEMENT ENGINEERING LITERATURE

Shanthakumari K
shanthakrs@gmail.com

Keshava -
keshtut@gmail.com

Follow this and additional works at: <https://digitalcommons.unl.edu/libphilprac>



Part of the [Scholarly Communication Commons](#), and the [Scholarly Publishing Commons](#)

K, Shanthakumari and -, Keshava, "AN ANALYSIS OF PERSISTENCE AND OBSOLESCENCE OF WEB CITATIONS OF PAVEMENT ENGINEERING LITERATURE" (2021). *Library Philosophy and Practice (e-journal)*. 5923.

<https://digitalcommons.unl.edu/libphilprac/5923>

AN ANALYSIS OF PERSISTENCE AND OBSOLESCENCE OF WEB CITATIONS OF PAVEMENT ENGINEERING LITERATURE

ABSTRACT

This paper examines the persistent accessibility of cited URLs in International Journal of Pavement Engineering. Also, forsee to compute the half-life period of the cited web URLs in a journal article. A total of 901 URLs cited in 299 research articles published spanning ten years from 2010 to 2019 were extracted. W3C Link Checker is used to verify the accessibility of the cited URLs. Active and inactive URLs were listed. The inactive URLs are checked in the Internet Archive Wayback Machine to retrieve available URLs. The half-life period was computed using the formula “ $t(h) = [t \ln (0.5)] / [\ln W(t) - \ln W (0)]$ ”. The study's findings revealed that 299 (30.63%) out of 976 articles have URL citations and the percentage of URL citations increased from 2.28 to 5.07 percent from 2010 to 2018, and in 2019 decreased to 2.54 percent.

Keywords: Half-life, active links, error codes, URL decay, citation analysis, web references, pavement engineering

Introduction

In the present Internet era, everyone is directly or indirectly depending on internet for their routine works. With the wide applications of internet in all sectors it plays a very important role. The internet is becoming very popular source to access all kind of information. Today’s electronic publications play an important role in promoting electronic learning. In order to promote electronic learning the electronic resources such as e-journals, e-books, e-papers in all the streams plays an essential role. In this digital era electronic resources have more benefit than print resources (Bodomo et al., 2003). In this article an attempt has been made to know the persistence and decay of URL citations cited in the articles published in International Journal of Pavement Engineering of Pavement Engineering to recognize the usage of internet resources in the field of civil engineering for research communication.

Academic research writing obligates authors to cite existing research works that have been considered. Both in-text citations and references help readers locate where information, knowledge, facts, and figures were borrowed and find the source if they need further information. Citations and references indicate the relation of the previous study to the present study, accredit the author to validate their opinion, and credit other author’s ideas. Therefore, citing URL has become common in academic research writing whereby authors are now required to include those URLs as part of references (Sife & Lwoga (2017).

Objectives of the study

- To identify the total web citations cited in articles,
- To find out active and missing URL citations,
- To rectify HTTP errors linked with missing URL citations,

- To find the domains linked with missing URLs, and
- To calculate the half-life cycle of URL citations.

Scope and Limitation

Considering the unpredictable characteristics of Web resources, the study examined the extent of URLs vanishing in articles published in International Journal of Pavement Engineering. There are two reasons for selecting this journal: 1) Journal published by Taylor & Francis Online with IF 2.646 (2019) & 0.81SJR (2019) cover the latest research articles and case studies contributor from across the world. 2) Journal has got 5.50-year half-life and 0.22 immediacy index (both are journal level metrics). Journal is abstracted in EBSCO Databases; ISI Science Citation Index produced by Clarivate Analytics; Scopus; Construction and Buildings Abstract, and several others making it easily accessible.

This study determined the extent of Web references cited in the journal articles, represented the percentage of missing web citation links, identified error messages related to inaccessible URLs. Indicated top domain levels of decayed URLs, calculated the URL citations' half-life and determined the portion of recovered URLs.

Related work

With the emergence of the World Wide Web, the potential to retrieve and access information rapidly and suitably, the study and use of URL citations have been considered by the researchers. Consequently, many studies have been conducting in this area of research and related works. Jalalifard & Isfandyari (2013) examined “few related studies on web citations on scholarly communication.” Bansal and Parmar (2020) did a research on “accessibility, deterioration and half-life of URLs of web documents cited in Current Science Journal published during 2015 -2016, analysed 1724 URLs cited in the 1564 articles and found that 56.67% of URLs were accessible and the remaining 43.33% of URLs were not accessible”. Moreover, the study also revealed that the average half-life period is URLs is 1.76 years and 59.03% of error messages were HTTP 404 (page not found). Another study conducted by Parmar and Pateria (2019) to “identify the prevailing citation trends, durability of web citations or URLs, decay and half-life of URLs by analyzing articles citations of Indian Journal of Agricultural Library and Information Services published during 2012 -2016.” After analyzing 980 URL citations appended in 94 journal articles, it was found that 33.16 percent was having web citations. The study examined that of URL citations, 62.15% of URLs were accessible and the remaining 37.85 % were not accessible.

Further, indicates that out all HTTP error message 51.22% error messages are HTTP 404 (page not found) and average half-life of URLs is 4.62 years. Loan & Shah (2020) analyzed “web references of the Journal of Informetrics and found 32.12% web references were inactive; majority of the error were associated with HTTP 404 ‘File not found’ (60%) followed by 400 ‘Bad request error’(35.65%); domains .gov, .edu and .com contains more number of missing URLs.” Prithvi Raj & Sampath Kumar (2013) examined “accessibility, corrosion and half-life of URLs cited in the articles of Indian LIS conference proceedings published from 2001 to 2010 and found 5,698 URLs cited in 1700 articles.”

Research Methodology

Archives for table of content (TOC) of the journal entries were accessed for the years 2010 to 2019. The study was carried out from February to June 2020 in the COVID-19 period. Only research articles termed as ‘Article/Original Article’ were considered and references cited at the end of an article/Original Articles considered as citations. Editorial notes, technical notes, corrigendum were not considered and not even counted and tested (Sampath Kumar, & Prithvi Raj (2012)). A total of 56 issues were considered for the study and details of articles recorded using MS Excel to analyze the citations. Also, the volume and issue number were recorded. A total of 901 URL citations were found in 976 journal articles. Each article reviewed to get the total number of references, references that included URL, and access date. Web citations on the basis of domain and file type were determined. The web citation is retrieved considered as persistent otherwise as non-persistent. Checked all the web citations with the W3C Link checker to check their accessibility of cited URL. Missing URLs were retrieved by using the application of internet archive (<https://archive.org/web/>) Wayback Machine.

Findings and Results

Table 1: Year-wise Distribution of Articles, Citations and URLs.

Year of Publications	Total no. of articles	Number of Citations	Average Citations Per Article	Number of articles with URL citations	Number of URL Citations	Average URL Per Article	Percentage of URL Citations (%)
2010	49	1182	24.12	12	27	0.55	2.28
2011	45	1584	35.20	18	61	1.36	3.85
2012	48	1237	25.77	18	53	1.10	4.28
2013	64	1522	23.78	26	64	1.00	4.20
2014	74	2138	28.89	32	70	0.95	3.27
2015	64	2351	36.73	30	124	1.94	5.27
2016	73	2159	29.58	27	61	0.84	2.83
2017	338	2947	8.72	39	164	0.49	5.56
2018	90	3256	36.18	46	165	1.83	5.07
2019	131	4402	33.60	51	112	0.85	2.54
Total	976	22778	28.26	299	901	1.09	3.92

Table 1 shows that a total of 22778 citations were appended to 976 articles, on an average of 28.26 citations per article of which 901 (3.95%) were URL citations. On an average 1.09 URL citations per article were found. This result appears to be significantly similar result documented by previous study by Jalalifard et. al in 2013 March (0.16 URL per paper). Between 2010 to 2018 the percentage of URL citations in scholarly articles increased substantially from 2.28 to 5.07 and slightly decreased 2.54 (2019). This result appears to

be similar result documented by previous study by Sampath Kumar & Prithvi Raj in 2012 “the web citations as a percentage of all citation also significantly increased from 2001 to 2008.” Average number of URLs was 1.09 per article from 2010 to 2019. This journal suggests authors to follow Harvard B reference citation style but does not insist to cite URLs with DOI.

Table 2: Summary of Active and Missing Web Citations

Year	Total number of URL citations	Active URL Citations	Percentage (%)	Missing URL citations	Percentage (%)
2010	27	26	96.30	1	3.70
2011	61	55	90.16	6	9.84
2012	53	46	86.79	7	13.21
2013	64	59	92.19	5	7.81
2014	70	63	90.00	7	10.00
2015	124	110	88.71	14	11.29
2016	61	56	91.80	5	8.20
2017	164	151	92.07	13	7.93
2018	165	146	88.48	19	11.52
2019	112	104	92.86	8	7.14
Total	901	816	90.56	85	9.44
DOI	116(12.87%)	104	89.65	12	10.35

Table 2. for the URL accessibility test indicates out of the 901 (include 116 with DOI) extracted URL citations, 816 (90.56%) were accessible. In contrast, the remaining (85, include 12 citations with DOI) URL citations (9.44%) were inaccessible and encountered access errors, as shown in Table 2. The ratio of missing URLs between 2010 to 2019 has fluctuated between 3.70% to 13.21%.

“The Digital Object Identifier (DOI) initiative was launched in October 1977 at the Frankfurt Book Fair in order to develop a common mechanism to enable intellectual content management to be integrated with internet technology (Jue Wang, 2007)”. There are inaccessible (missing) (10.35 per cent) DOI. This is due to DOI errors belong to Prefix-type error or Suffix-type error or Other-type error (Xu, S. et al., 2019) that required further discussion.

Table 3: HTTP errors linked with missing URL citations.

Year	Total Number of URL Citations	Missing URL Citations	HTTP 404 (Not Found)	HTTP 500 (Internet Server error)	HTTP 403 (Forbidden)
2010	27	1	0	1	0
2011	61	6	1	5	0
2012	53	7	5	2	0
2013	64	5	3	1	1
2014	70	7	5	1	1
2015	124	14	8	6	0
2016	61	5	3	2	0
2017	164	13	9	4	0
2018	165	19	9	5	5
2019	112	8	3	5	0
Total	901	85	46	32	7
Percentage		9.44	54.12	37.65	8.23

As mentioned in Table 3, most error messages shown are HTTP 404 (page not found) 54.12% of all missing inactive URLs followed by HTTP 500- Internet server error 37.6% and the least HTTP 403 – Forbidden 8.23%. This study's result is significantly similar to earlier studies by Sife & Lwoga (2017) and many previous studies. HTTP 404 is the most common error reported in similar studies. These findings indicate the web's mercurial nature as a publishing platform where electronic resources can easily remove from their original location. Paskin (1999) stated that “researchers need to take care while typing the URLs, whereas the editors need to check the URLs before publishing carefully”. An author should prefer to cite DOIs instead of URLs.

Table 4: Distribution of URL citations by domain type.

Year	Total number of URL Citations	Missing URL Citations	Domain association with missing URL citations						
			.com/.co	.edu	.ac	.gov	.org	.net	Others
2010	27	1	1	0	0	0	0	0	0
2011	61	6	0	0	0	1	4	0	1
2012	53	7	0	0	0	3	3	0	1
2013	64	5	1	2	0	1	0	0	1
2014	70	7	2	1	0	0	3	0	1
2015	124	14	2	3	0	2	4	0	3
2016	61	5	2	1	1	0	0	0	1
2017	164	13	1	3	0	0	6	1	2
2018	165	19	2	0	0	5	8	2	2
2019	112	8	1	2	0	1	0	0	4
Total	901	85	12	12	1	13	28	3	16
Percentage (%)			14.12	14.12	1.18	15.29	32.94	3.53	18.82

Table 4 summarized the top-level domains are associated with missing web citations. A total 85 cited URLs missing, top level domain is .org in highest with 32.94% followed by others 18.82%, .gov

15.29%, .edu, .com/.co with 14.12%, .net 3.53% and .ac is least missing URLs with 1.18%. This clearly indicates that the URLs of .gov and .com/.co and not persistent for a longer period of time.

Table 5: Half-life period of URL Citations.

Year	(t)	Total number of URL citations (W(0))	Active URL citations (W(t))	Half-life (t(h))
2010	1	104	13	-0.51
2011	2	146	48	-1.59
2012	3	151	24	-1.99
2013	4	56	32	20.76
2014	5	110	43	-7.67
2015	6	63	82	6.14
2016	7	59	42	145.55
2017	8	46	131	4.36
2018	9	55	133	5.68
2019	10	26	83	4.75
Total		901	816	17.55

The study employed the method used by Koehler (1999) to “calculate half-life of URL citations cited in the Journal.” The measurement of half-life period of URL citations cited in journal articles determines the URLs’ persistence over time. The half-life of URL citations also indicates the percentage of permanent URL citations that are accessible.

The formula used to calculate the half-life of URLs is: $W(t) = W(0) e^{\alpha t}$. Where $W(0)$ is the number of working URLs at the time of publication; $W(t)$ is the number of working URL citations at some later time(t); and α is constant that can be calculated from the available data.

The half-life of URL is: $t(h) = [t \ln (0.5)] / [\ln W(t) - \ln W(0)]$

Table 5 indicates the average half-life of the cited URL is estimated as 17.55 years, which means it will take about 17 years and five months to vanish half of the cited URLs in this journal. This average half-life is higher than those results reported by Sife & Lwoga (2017); Koehler (2002) (two years); Markwell and Brooks (2003) (4.6 years); Moghaddam et al. (2010) (14.94 years). The half-life of URL citation is likely to increase when observing for a longer period.

Major findings and limitations of the study

The following are the important findings and the limitations of the study

- The study was conducted on articles published in International Journal of Pavement Engineering. The study results may not be generalized for all journals in Engineering in general and Pavement Engineering in particular.

- The main findings included that very less number (1.09%) of web citation in this journal articles. It indicates that the editorial board of the journal may insist the civil engineering authors and research scholars to cite electronic resources with URLs.
- The primary reason for missing web citation links is HTTP 404 error (file not found).
- Very less number (12.87 percent) DOI cited in this journal articles.

Discussion

Very less number (1.09 per cent) of web citations found in the journal articles. Only 12.87 per cent of articles cited DOI. 10.35 per cent of inactive DOIs found. The DOI is a mechanism for improved access of electronic contents through an automated digital environment. The DOI enables the publisher, users and librarians to overcome the encumbrance of missing URLs. A library can retrieve URLs with DOIs to make enduring links to full-text intellectual property online and for increasing usage of acquired resources and enhanced localized linking at no cost. Further, "a library can integrate DOI to access its locally licensed materials" (Jue Wang, 2007) for end user.

Conclusion

The present study explored the citation behaviour of authors in the field of engineering in general and pavement engineering in particular. The studies on URL persistence in other branches of engineering field and further multidisciplinary comparisons are recommended. These kinds of study could depict the ability to perceive clearly web citation behaviour among research scholars of different disciplines. Moreover, extending the URL citation study by recommending helpful and novel methods can be studied. For instance, two or more than two URL referencing can be surveyed. In this study, collaboration, the author affiliation, cross national analysis and more specifically URL citation behaviour of authors can be done. The method may be very comprehensive but citation persistence may be increased. Further, it is hope that this study as an addition to the existing literature, can help researchers and academics to extend the horizon of the study.

References

- Bansal, Sonia, and Seema Parmar. "Decay of URLs Citation: A Case Study of Current Science." *Library Philosophy & Practice* (2020).
- Bodomo, Adams, Mei-ling Lam, and Carmen Lee. "Some students still read books in the 21st century: A study of user preferences for print and electronic libraries." *The Reading Matrix* 3, no. 3 (2003).
- Goh, Dion Hoe-Lian, and Peng Kin Ng. "Link decay in leading information science journals." *Journal of the American Society for Information Science and Technology* 58, no. 1 (2007): 15-24.
- Jalalifard, Mariam, Yaghoub Norouzi, and Alireza Isfandyari-Moghaddam. "Analyzing web citations availability and half-life in medical journals." In *Aslib Proceedings*. Emerald Group Publishing Limited, 2013.

- Koehler, W. (2002), "Web page change and persistence: a four-year longitudinal study." *Journal of the American Society for Information Science and Technology*, Vol. 53 No. 2, pp. 162-171, DOI: 10.1002/asi.10018.
- Loan, Fayaz Ahmad, and Ufaira Yaseen Shah. "The decay and persistence of web references." *Digital Library Perspectives* (2020).
- Markwell, John, and David W. Brooks. "'Link rot' limits the usefulness of web-based educational materials in biochemistry and molecular biology." *Biochemistry and Molecular Biology Education* 31, no. 1 (2003): 69-72.
- Moghaddam, A. Isfandyari, M. K. Saberi, and S. Mohammad Esmaeel. "Availability and half-life of web references cited in information research journal: A citation study." *International Journal of Information Science and Management (IJISM)* 8, no. 2 (2012): 57-75.
- Nagaraja, Aragudige, Shine A. Joseph, Hyla H. Polen, and Kevin A. Clauson. "Disappearing act: Persistence and attrition of uniform resource locators (URLs) in an open access medical journal." *Program: electronic library & information systems* 45, no. 1 (2011): 98-106.
- Parmar, Seema, and Rajive K. Pateria. "Web Citations and Decay of URLs: A Case Study of Indian Journal of Agricultural Library and Information Services." (2019).
- Paskin, Norman. "The digital object identifier system: digital technology meets content management." *Interlending & document supply* (1999).
- Prithvi Raj, K. R., and B. T. Sampath Kumar. "URLs as references in Indian LIS conference papers: an Analysis." *Annals of Library and Information Studies (ALIS)* 60, no. 4 (2014): 284-295.
- Prithviraj, K. R., and BT Sampath Kumar. "Corrosion of URLs: Implications for electronic publishing." *IFLA journal* 40, no. 1 (2014): 35-47.
- Riahinia, Nosrat, Fatemeh Zandian, and Ali Azimi. "Web citation persistence over time: a retrospective study." *The Electronic Library* (2011).
- Sampath Kumar, B. T., and K. R. Prithvi Raj. "Availability and persistence of web citations in Indian LIS literature." *Electronic library* 30, no. 1 (2012): 19-32.
- Sife, Alfred Said, and Edda Tandi Lwoga. "Retrieving vanished Web references in health science journals in East Africa." *Information and Learning Science* (2017).
- Wang, Jue. "Digital object identifiers and their use in libraries." *Serials review* 33, no. 3 (2007): 161-164.
- Xu, Shuo, Liyuan Hao, Xin An, Dongsheng Zhai, and Hongshen Pang. "Types of DOI errors of cited references in Web of Science with a cleaning method." *Scientometrics* 120, no. 3 (2019): 1427-1437..