Five Hundred Most-Cited Papers in the Computer Sciences: Trends, Relationships and Common Factors

Phoey Lee Teh[1](✉) and Peter Heard[2]

[1] Department of Computing and Information Systems, School of Science and Technology, Sunway University, 47500 Sunway City, Malaysia
phoeyleet@sunway.edu.my

[2] Provost Office, Sunway University, 47500 Sunway City, Malaysia
pheard@sunway.edu.my

Abstract. This study reveals common factors among highly cited papers in the computer sciences. The 500 most cited papers in the computer sciences published between January 2013 and December 2017 were downloaded from the Web of Science (WoS). Data on the number of citations, number of authors, article length and subject sub-discipline were extracted and analyzed in order to identify trends, relationships and common features. Correlations between common factors were analyzed. The 500 papers were cited a total of 10,926 times: the average number of citations per paper was 21.82 citations. A correlation was found between author credibility (defined in terms of the QS University Ranking of the first named author's affiliation) and the number of citations. Authors from universities ranked 350 or higher were more cited than those from lower ranked universities. Relationships were also found between journal ranking and both the number of authors and the article length. Higher ranked journals tend to have a greater number of authors, but were of shorter length. The article length was also found to be correlated with the number of authors and the QS Subject Ranking of the first author's affiliation. The proportion of articles in higher ranked journals (journal quartile), the length of articles and the number of citations per page were all found to correlate to the sub-discipline area (Information Systems; Software Engineering; Artificial Intelligence; Interdisciplinary Applications; and Theory and Methods).

Keywords: Data search · Knowledge discovery · Citation · Trends · Common factors

1 Introduction

The Institute for Scientific Information (ISI) [1] is a bibliographic database of academic journals with citation indexing and analysis, which allows researchers to find out how many times a given article has been cited and by whom. Although opinion is divided on the merits of the metrics derived from such databases, the proliferation of similar abstract and database services, such as Scopus, Google Scholar and the more focused PubMed, is testament to the growth in their importance.

Á. Rocha et al. (Eds.): WorldCIST 2021, AISC 1366, pp. 13–22, 2021.
https://doi.org/10.1007/978-3-030-72651-5_2

The number of citations received by a given piece of scholarly work is an often-used proxy measure of the quality, importance and impact of the work: higher citation counts are assumed to be indicative of higher quality research, and greater impact. Citation measures, such as the average number of citations per paper, h-index, [2–4], i10-index [5] or g-index [6] may be used as part of academic recruitment, tenure and promotion exercises. At institutional-level citation metrics feed into the major league tables, such as the QS and Time Higher rankings. Citations per faculty (i.e. per member of academic staff) constitutes 20% to the total score in the QS world university rankings [7]; while in the THE world rankings citations per paper contributes 30% [8].

The number of times a piece of scholarly work is cited is thus of great importance to individuals, their academic department and their university. High citation rates are often used as an indicator of quality and impact, and may indicate to other researchers whether or not a particular article is worthy of reading [3] and citing, thus leading to further citations. Thelwall [9] noted that citation counts are used by researchers and research managers to assist in the evaluation of the quality or impact of published research, especially where it is impractical to employ peer judgements or where corroborating data is required.

Several other methodologies have been proposed to measure research output. Dorta-González et. al. [10] for example, proposed three dimensions, namely productivity (number of journal papers); impact (journal citations); and references (bibliographical sources). González-Betancor and Dorta-González [10] proposed an alternative citation impact indicator, based on the percentage of highly cited articles. The potential use of Google Scholar metrics as a feasible and reliable indicator of highly cited documents was examined by Martin-Martin [11], but it was found to be an unreliable method. Chang [12] conducted a study on high impact papers using ISI metrics for the 200 most highly cited journal in the sciences and social science. Results showed that the Sciences and Social Sciences are different in terms of the strength of the relationship of journal performance metrics, although the actual relationships were very similar.

2 Credibility, Article Length, Number of Authors and Field of Study

Previous research has shown that citation rates vary with such parameters as author credibility, article length, number of authors and field of study. Author credibility refers to the credentials or other perceived qualities of the author. Perceived author credibility may be used as an indicator of whether or not their research is reliable, of high quality and thus a valuable source of reference. Measures of author credibility include the experience of the author, the ranking of the author's primary affiliation, and/or the number and prestige of awards received. Plomp [13] found that authors with a greater number of previously published outputs were more likely to receive a greater number of citations for subsequent work. Rodríguez-Navarro [14] noted that Nobel Prize-winning authors enjoy higher citation rates, and Bitetti [15] noted how the more influential a researcher is in a certain field, the greater their citation count. Akre [16] found that research originating from high-income European countries tends to have higher citation rates. In the context of this research, the primary affiliation of the first-named author – both at institutional-level and at departmental/subject-level – is used as a proxy measure of author credibility. The first named author was used, because, in the computer sciences, the first named author is usually the individual most identified with the work.

Previous research [17–20] has demonstrated a relationship between citation count and article length: articles of greater length attracted a high number of citations. The rationale for this observation is that longer papers have more content, which is of potential interest to others. Previous research [21] has also shown that papers with multiple authors are more highly cited than single author works. Gazni [22] showed that single-authored papers received, on average one citation, while multi-authored papers received an average of 2.12 citations per paper. Tahanam [19] also concluded that multiple authorship is a contributory factor to higher citation rates.

Oakleaf [23] found that monodisciplinary papers received higher citation rates than multidisciplinary papers, and Akneses and Rip [21] concluded that the majority of citations come from other researchers in the same field. In contrast, [19] showed that research focused on a single field of study had lower citation rates. Other research [10] has shown that citation practices differ across scientific fields, and found no evidence to support the hypothesis that multidisciplinary papers were cited differently to single-discipline outputs. Thus, there remains no consistent view on the impact of multi- versus single-disciplinary work. Much research in the computer sciences is multidisciplinary and we were interested to explore any relationship between the field of study and citation rates.

Building on the aforementioned research we postulate that academics with higher levels of peer esteem will, on average, be cited more often than other academics. We thus suggest that publications from researchers affiliated with more highly ranked universities will, on average, be more highly cited: likewise, of authors from highly ranked departments. We also, consider whether authors from higher ranked universities and departments will have proportionately more papers in higher ranked journals. Review articles offer a broader perspective on an area of research, summarising previous work and drawing out more general conclusions. Reviews are therefore a valuable resource to researchers, and we suggest that reviews will, on average, be more highly cited than primary papers. Since reviews collate and summarise multiple studies, we anticipated that reviews will be longer than primary publications. Compiling all the suggestions above, this study seeks to establish if: (1) there is a correlation between article length and citation rates; (2) citation rates correlate with the number of authors; or (3) citation rates vary by sub-discipline.

3 Method

The 500 most cited papers in the computer sciences published over the 5 years period January 2013 to December 2017 were downloaded from the Web of Science on 14th October 2018. Parameters extracted directly from the ISI database or the QS rankings organisation included: year of publication; sub-discipline (as defined by Web of Science); number of citations; article length; number of authors; University Ranking (QS) and Subject Ranking (QS).

Table 1 shows papers that published earlier generally had more citations. The raw citation data were thus normalised to allow for comparison. Data were normalised by

Table 1. Average citations vs. number of years since publication

Number of years since publication	Average
5	30.6
4	23.3
3	16.2
2	15.8
1	17

determining the mean citation count for each year and then converting raw counts to fractions of the yearly average. The number of citations received was compared in a pair-wise fashion to see if any common features emerged. Papers were a mix of primary articles and reviews articles. Unless otherwise indicated, results given are for the analysis of both reviews and primary publications.

4 Result

The 500 papers were cited a total of 10,926 times over the period: the average was 21.82 citations per paper; median 13; mode 8. The majority of the papers, (74%) received 20 or fewer citations; only 3% had more than 80 citations. Mean citation rates are relatively flat for the first three years post publication, only rising significantly after that.

Table 2. Average normalised citation counts compared with QS University World Ranking

QS University World Ranking (2019)	Average normalised citation count	QS University World Ranking (2019)	Average normalised citation count
1–50	1.086	451–500	0.615
51–100	1.232	501–550	0.594
101–150	0.982	551–600	0.704
151–200	1.032	601–650	0.693
201–250	1.036	651–700	0.558
251–300	1.412	701–750	1.291
301–350	1.121	751–800	0.842
351–400	0.644	801–1000	0.745
401–450	1.052	Unranked	0.999

Tables 2 and 3 show the relationships between ranking and citations. Data indicate small differences between citation rates for authors from higher ranked universities compared to lower ranked ones. Not all universities and departments (subjects) have a ranking: such universities/subjects are grouped under the heading "unranked". The group of

Table 3. Average normalised citations compared with QS subject ranking.

QS world subject ranking (2019)	Average normalised citation count	QS world subject ranking (2019)	Average normalised citation count
1–50	1.342	301–350	0.637
51–100	0.807	351–400	1.486
101–150	0.731	401–450	0.921
151–200	1.207	451–500	0.720
201–250	1.019	Unranked	0.936
251–300	0.874		

unranked institutions includes government and industry laboratories. Data reveal that the average citation rates for publications emanating from universities ranked 350 or higher is greater than and those for lower ranked universities. The difference is not large, but is statistically significant (P = 0.03): papers originating from higher ranked universities receive, on average, about 25% more citations than those from lower ranked universities. Examination of Table 3, reveals no statistically significant relationship between citation rates and subject ranking.

Table 4. Percentage of publications by journal quartile

Journal quartile	Q1	Q2	Q3	Q4	Q5
Percentage of outputs (%)	57.6	23.9	14.0	3.7	0.7

Table 4 shows the percentage of papers that are published in Q1, Q2, Q3, Q4 and unranked journals (designated Q5), respectively. Overall approximately 58% of the papers are in Q1 journals; 81.5% of papers are in Q1 or Q2 journals.

As depicted in Fig. 1, our data indicate no relationship between ranking (university ranking or subject ranking) and the percentage of papers in higher quartile journals. Questions 3 and 4 are thus shown to be proven false. One possible rationale for such a finding might be the efficacy of the peer review process: peer reviews showing no bias towards authors from higher ranked universities or departments.

The number of review articles in the sample was relatively small: 35 papers (7%). Nevertheless statistically significant differences in article length and citation rates were observed. The average length of review articles was found to be 1.5 times longer ($p < 0.01$), whilst the average number of citations was found to be 2.5 times greater ($p < 0.01$). For review papers, the number of citations also appears to be correlated with the number of authors: citations increase gradually with number of authors, up to 5 authors, then flattens (Slope = 1.06; Rˆ2 = 0.94).

Since journal rankings are based on citations [13, 24], it is to be expected that journal rankings and average citations correlate. The mean normalized citation rates are:

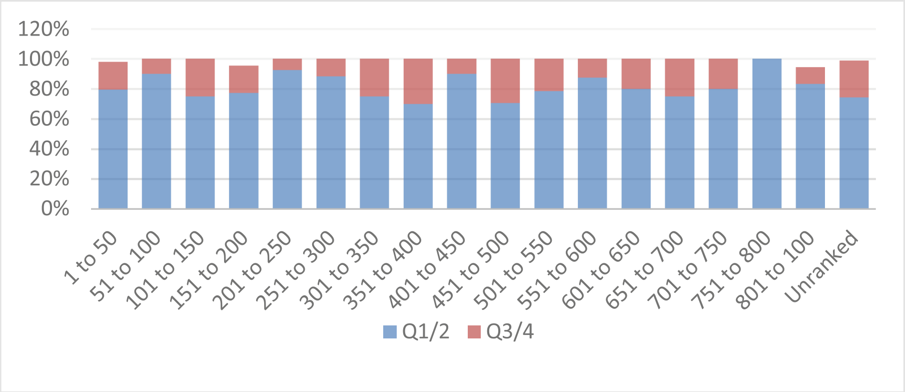

Fig. 1. Proportion of Q1/Q2, and Q3/Q4 papers as a function of QS World University Ranking

Q1 (0.32) > Q2 (0.21): $p < 0.001$; Q2 > Q3 (0.16): $p < 0.05$. No significant differences exist in the rates of citation between Q3 and Q4 or unranked journals.

Table 5. Journal quartile and the average number of pages from the top 500 papers

Journal quartile	Number of papers	Average number of pages
1	264	15.898
2	132	18.152
3	77	20.896
4	22	23.545

Table 5 reveals that article length is inversely and linearly correlated with journal ranking: higher ranked journal articles are shorter ($R^2 = 0.998$).

The Web of Science categorises computer science outputs into five different sub-disciplines: Information Systems; Software Engineering; Artificial Intelligence; Interdisciplinary Applications; and Theory and Methods. The highest proportion of papers in Q1/Q2 journals is found for the sub-discipline of Interdisciplinary Application; papers in this sub-discipline are also found to be shorter on average than those of other sub-disciplines. In contrast, papers in the sub-discipline areas of Software, and Theory and Methods are much less likely to be in Q1/Q2 journals, and are on average longer: particularly for Theory and Methods papers. Determining whether pressure for space in journals drives shorter papers, or whether Interdisciplinary Applications papers are naturally shorter and at the same time considered closer to the cutting-edge and thence more publishable in Q1 and Q2 journals is an interesting, but open question.

The proportion of papers in Q1/Q2, and Q3/Q4 journals with different numbers of authors is shown in Fig. 2. The proportion of Q1 and Q2 journals generally increases with the number of authors: this increase is statistically significant ($R^2 = 0.80$; $p < 0.05$).

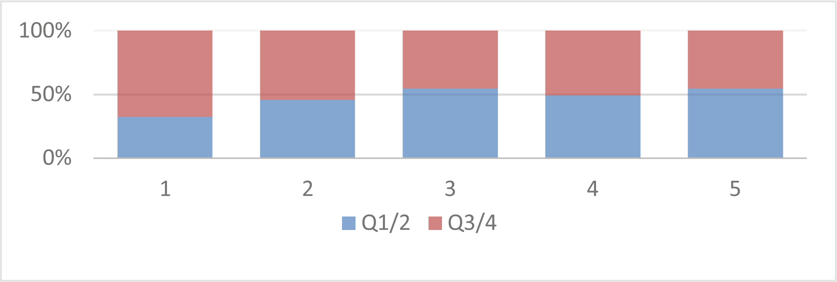

Fig. 2. Proportion of papers in Q1/Q2 and Q3/Q4 journals versus the number of authors

It may simply be that doing high quality work requires input from a greater number of investigators.

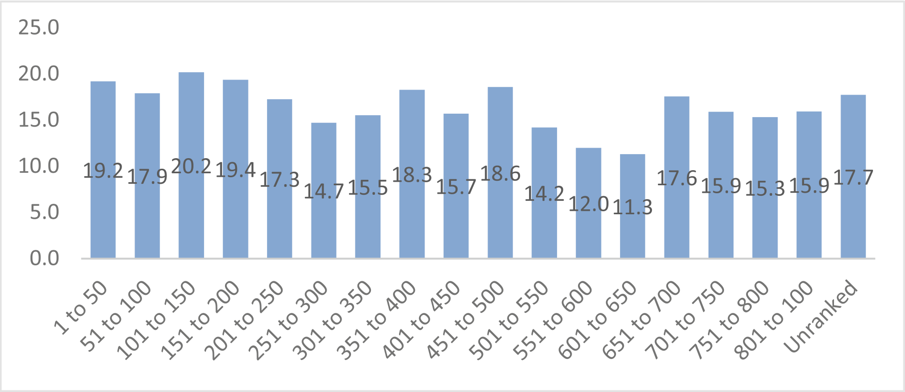

Fig. 3. Average article length vs. QS World University Ranking

The overall percentage of single authored papers is 14%, which compares to a global average (across all disciplines) of about 11% [25]. In line with previous findings [26] the proportion of single authored papers varies by sub-discipline: Theory and Methods publications are most likely to single author (17%); Artificial Intelligence and Software papers are least likely to be single author works (6% and 5%, respectively).

Data also show that authors from higher ranked universities or departments write marginally longer papers than authors from lower ranked universities and departments (Fig. 3). For example, articles emanating from top 100 ranked departments are on average just over 2.5 pages longer than articles emanating from departments outside of the top 100: average page length = 19.67 and 17.02, respectively (P = 0.03). The relationship between University Ranking and article length is less pronounced, and of marginal statistical significance (P = 0.055).

We observed a statistically significant relationship between the sub-discipline area and both the average number of pages and the average (normalised) citation rates (Table 6). In particular, Theory and Methods papers were found to be much longer, because of

Table 6. Number of articles, average page counts and normalised citation counts by sub-discipline

Subject sub-discipline	Number of articles	Average number of pages	Average normalised citation counts per paper
Artificial intelligent	96	16.32	0.57
Information system	113	16.96	1.03
Inter-disciplinary	203	14.49	1.05
Software engineering	42	19.95	0.81
Theory & methods	46	32.63	0.98

substantial space given over to mathematical argument. The mean number of citations per page was 0.71, compared to 0.57 (for AI) to 1.05 (interdisciplinary studies). The reason for the lower impact of AI journal articles is not obvious, but, interestingly, [4] found that there was a greater occurrence of papers being retracted in this sub-discipline over a similar timeframe (2013 to 2017). Whether there is any link between the occurrence of retractions and lower impact remains an open, but interesting question.

5 Conclusion

This paper presents an analysis of the 500 most cited papers in the computer sciences over the five-year period 2013 to 2017. Seventy-four percent of papers received 20 or fewer citations, with only 3% receiving more than 80 citations; the average was 21.82 citations per paper. A correlation was found between citation rates and author credibility: authors from universities ranked 350 or higher were more cited than those from lower ranked universities. Relationships were also found between journal ranking and the number of authors, and the article length: higher ranked journals tend to have a greater number of authors, but are shorter in length. The article length was also found to be correlated with the number of authors and the QS Subject Ranking of the first author's affiliation. The proportion of articles in higher ranked journals, the length of articles and the number of citations per page were all found to depend on the sub-discipline area: the greatest impact, measured in terms of citations per page was found to be in Interdisciplinary Applications, with the lowest in Artificial Intelligence, and Theory and Methods.

References

1. Garfield, E.: Citation analysis as a tool in journal evaluation. Serial Librarian **178**(4060), 471–479 (1972)
2. Hirsch, J.E.: hα: an index to quantify an individual's scientific leadership. Scientometrics **118**, 673–686 (2019)
3. Cormode, G., Ma, Q., Muthukrishnan, S., Thompson, B.: Socializing the h-index. J. Informetrics **7**(3), 718–721 (2013)

4. Ayaz, S., Masood, N., Islam, M.A.: Predicting scientific impact based on h-index. Scientometrics **114**(3), 993–1010 (2018)
5. Noruzi, A.: Impact factor, h-index, i10-index and i20-index of webology. Webology **13**(1), 1–4 (2016)
6. De Visscher, A.: What does the g-index really measure? J. Am. Soc. Inf. Sci. Technol. **62**(11), 2290–2293 (2011)
7. QS World University Rankings – Methodology|Top Universities. https://www.topuniversities.com/qs-world-university-rankings/methodology. Accessed 19 June 2019
8. World University Rankings 2019: methodology | Times Higher Education (THE). https://www.timeshighereducation.com/world-university-rankings/methodology-world-university-rankings-2019. Accessed 19 June 2019
9. Thelwall, M.: Dimensions: a competitor to Scopus and the web of science? J. Informetrics **12**(2), 430–435 (2018)
10. Dorta-González, P., Dorta-González, M.I., Suárez-Vega, R.: An approach to the author citation potential: measures of scientific performance which are invariant across scientific fields. Scientometrics **102**(2), 1467–1496 (2014)
11. Martin-Martin, A., Orduna-Malea, E., Harzing, A., Lopez-Cozar, E.D.: Can we use Google Scholar to identify highly-cited documents? J. Informetrics **11**(1), 152–163 (2017)
12. Chang, C.L., McAleer, M., Oxley, L.: Coercive journal self citations, impact factor, journal influence and article influence, mathematics and computers in simulation. Int. Assoc. Math. Comput. Simul. (IMACS) **93**, 190–197 (2013)
13. Plomp, R.: The highly cited papers of professors as and indicator of a research group's scientific performance. Scientometrics **29**(3), 377–393 (1994)
14. Rodríguez-Navarro, A.: A simple index for the high-citation tail of citation distribution to quantify research performance in countries and institutions. PLoS ONE **6**(5), e20510 (2011)
15. Bitetti, M.S.D., Ferreras, J.A.: Publish (in English) or perish: the effect on citation rate of using languages other than English in scientific publications. Ambio **46**(1), 121–127 (2017)
16. Akre, O., Barone-Adesi, F., Pattersson, A., Pearce, N., Merletti, F., Richiardi, L.: Differences in citation rates by country of origin for papers published in top-ranked medical journals: do they reflect inequalities in access to publication? J. Epidemiol. Community Health **65**(2), 119–123 (2011)
17. Hamrick, T.A., Fricker, R.D., Brown, G.G.: Assessing what distinguishes highly cited from less-cited papers published in interfaces. Interfaces **40**(6), 454–464 (2010)
18. Coupé, T.: Peer review versus citations - an analysis of best paper prizes. Res. Policy **42**(1), 295–301 (2013)
19. Tahamtan, I., Safipour Afshar, A., Ahamdzadeh, K.: Factors affecting number of citations: a comprehensive review of the literature. Scientometrics **107**(3), 1195–1225 (2016). https://doi.org/10.1007/s11192-016-1889-2
20. Fox, C.W., Paine, C.E.T., Sauterey, B.: Citations increase with manuscript length, author number, and references cited in ecology journals. Ecol. Evol. **6**(21), 7717–7726 (2016). https://doi.org/10.1002/ece3.2505
21. Aksnes, D.W., Rip, A.: Researchers' perceptions of citations'. Res. Policy **38**(6), 895–905 (2009)
22. Gazni, A., Didegah, F.: Investigating different types of research collaboration and citation impact: a case study of Harvard University's publications. Scientometrics **87**(2), 251–265 (2011)
23. Oakleaf, M.: Writing information literacy assessment plans: a guide to best practice. Commun. Inf. Literacy **3**(2), 80–90 (2009)
24. Petersen, C.G., Aase, G.R., Heiser, D.R.: Journal ranking analyses of operations management research. Int. J. Oper. Prod. Manag. **31**(4), 405–422 (2011)

25. Baker, S.: Authorship: are the days of the lone research ranger limited? Times Higher Education. https://www.timeshighereducation.com/news/authorship-are-days-lone-research-ranger-numbered. Accessed 03 July 2019
26. Al-Hidabi, M.D., The, P.L.: Multiple publications: the main reason for the retraction of papers in computer science. In: Arai, K., Kapoor, S., Bhatia, R. (eds.) Advances in Information and Communication Networks. FICC 2018. Advances in Intelligent Systems and Computing, vol. 886, pp. 551–526. Springer, Cham (2019)