STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Which Type of Research is Cited More Often in Wikipedia? A Case Study of PubMed Research

Tahereh Dehdarirad[*], Fereshteh Didegah[**] and Hajar Sotudeh[***]

[*]*tahereh.dehdarirad@chalmers.se.*
Department of Communication and Learning in Science, Chalmers University of Technology, Hörsalsvägen 2, Göteborg, SE-412 96 (Sweden)

[**]*fdidegah@sfu.ca.*
Scholarly Communication Lab, Simon Fraser University, Vancouver BC, V6B 5K3 (Canada)

[***]*sotudeh@shirazu.ac.ir.*
Department of Knowledge and Information Sciences, Shiraz University, Eram Campus, Shiraz ,71946-84471, (Iran)

Abstract

This study examines the characteristics of medical articles cited in Wikipedia and compares them with a sample of medical articles not cited in the platform. The aim is to determine the reasons why some articles are selected as reliable sources for Wikipedia and others are not. The characteristics studied are document type, open access status of article, article topic, article F1000 class and F1000 count, article tweet count, and article news count. The findings show a document type similarity for both cited and uncited sets of articles, with articles, reviews and editorial materials being more visible in both sets. While the articles cover a broad range of topics, the top three topics are the same in both sets. The results also reveal that Wikipedia favors OA articles, although a large number of cited articles are non-OA. Finally, significant, although weak correlations are found between Wiki citation counts and F1000, tweet and news counts. While F1000 and tweet counts correlate negatively with Wikipedia citation counts, news counts show a positive correlation, although the weakest compared to the other correlations.

Introduction

Wikipedia is a prominent source of general healthcare information, extensively used by the general public, students, and health care professionals (Kousha & Telwall, 2016). More than 155000 Wikipedia medical articles, written in different languages, were viewed more than 4.88 billion times in 2013, making it one of the most viewed medical and health care resources on the internet (Heilman & West, 2015). It is also frequently listed in search engine results for top health-related queries (Laurent & Vickers, 2009).

Given its popularity, it is important to ensure content quality of Wikipedia articles, which could be measured to an extent through articles' references. According to Wikipedia's verifiability policy, articles should cite external reliable research to confirm existing knowledge (Wikipedia, 2008). This research aims to study the characteristics of external sources cited in Wikipedia articles, in order to determine the reasons why some articles are selected as reliable sources for Wikipedia and others are not.

Previous research has studied some characteristics of research articles cited in Wikipedia. For example, articles from high impact factor journals were found to be cited more often in Wikipedia (Nielsen, 2007) and the majority of references citied in Wikipedia came from peer reviewed journals (Haigh, 2011). Moreover, previous studies have showed that Wikipedia articles favor open access articles (Didegah, 2017). Some studies also found that journal articles cited in Wikipedia have higher F1000 scores than the uncited ones (Evans & Krauthammer, 2011). The subject variation of medical articles cited in Wikipedia showed that *genetics* was the most frequent cited article topic (Evans & Krauthammer, 2011).
However, a large scale study, which considers several characteristics of medical articles cited in Wikipedia, and also compares cited with uncited articles, is missing from the literature. Thus, this paper aims to study the characteristics of a large sample of medical articles cited in Wikipedia and compare them with a sample of medical articles that are not cited in this platform. These characteristics include document type, open access status of article, article topic, article F1000 class and F1000 count, article tweet count, and article news count.
The findings will provide a comprehensive view of the type of medical research that is of interest to Wiki community. This is important, as Wiki editors aim to mediate between Wiki content and public interest (Thelwall, 2016) and transfer knowledge from academia to a broader community (Kousha & Thelwall, 2016).
To fulfill the research goals, the following questions will be addressed:

1. Which document types are cited more often in Wikipedia?
2. Are open access documents cited more than non-open access documents in Wikipedia? Which types of open access documents are favored?
3. Which Medical Subject Headings (Mesh) are cited more often in Wikipedia?
4. Which F1000 classes are cited more often in Wikipedia?
5. Are there significant correlations between Wiki citation counts and F1000 counts, news counts, and tweet counts?

Methodology

The current study is based on a random sample of publications from PubMed proportionally gathered from 1996 to 2017, which accounted for 3,905,323 records. PMID and MESH subject headings for each record were obtained from PubMed. In this paper, we refer to these headings as topics. Using PMID, a search was made in *Altmetric.com* (October 2017 version) *for the Wikipedia citations* of the corresponding documents. From this, 384,394 (~10%) PMIDs were cited at least once in Wikipedia, while the rest of PMIDs (3,520,929) were not cited. For comparison purposes, a random sample of uncited documents was also selected proportionally from 1996 to 2017, which accounted for 371,521 documents. All types of documents were taken into account for this study. Types of documents for both sets of cited and uncited publications in Wikipedia were extracted from *Web of Science*. Open access status of publications was obtained from *Unpaywall.org*. F1000, news and tweet counts were obtained via Altmetric.com for both collections of cited and uncited publications. To answer question 4, documents were classified into six F1000 classes, including new finding, confirmation, technical advance, controversial, novel drug target and good for teaching.

Statistical procedures

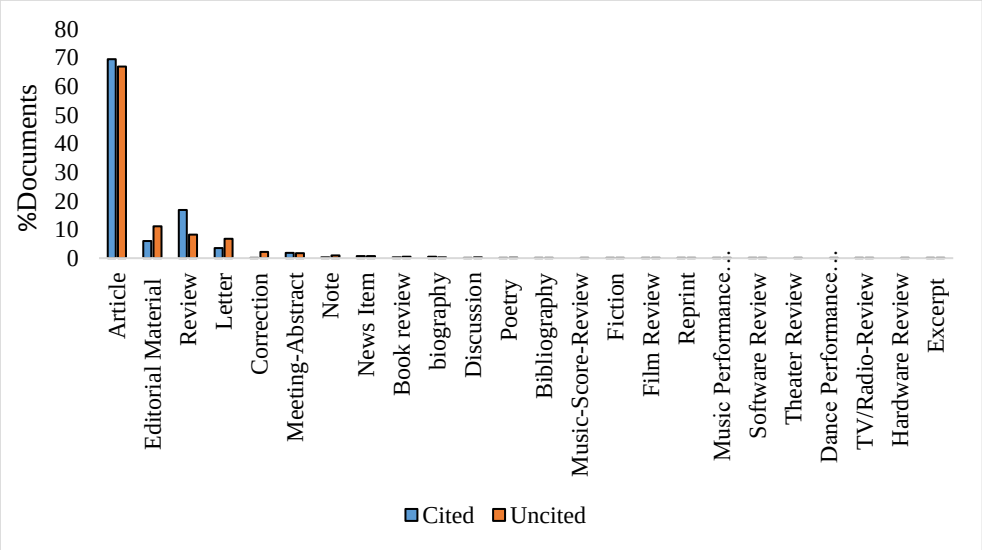
To compare the percentage of cited OA documents in both cited and uncited sets in Wikipedia, a two-sample proportion test was used. Similarly, to compare the percentage of F1000 classes between cited and uncited sets, a two-sample proportion test was used. Spearman correlations have been used to study the relationship between Wikipedia citation counts, F1000 counts, news counts, and tweet counts for the entire collection of cited and uncited documents in Wikipedia.

Results and discussions

Question 1. Which document types are cited more often in Wikipedia?

In both the cited and uncited sets of documents in Wikipedia, editorial materials, reviews and letters are the top document types. However, the percentage of articles and reviews is slightly higher in the cited document set than the uncited set (though non-significant). In a relevant study on drug-related Wikipedia articles, Koppen, Phillips, and Papageorgiou (2015) found that Wikipedia's most commonly cited documents were journal articles (49.2%) and news articles (12.0%).

Figure 1: Document types for cited and uncited sets of documents in Wikipedia.



Question 2. Are open access documents cited more than non-open access documents in Wikipedia? Which types of open access documents are favored?

The open access status and type of cited and uncited documents were examined. The results show that whilst around 30% of cited documents in Wikipedia are open access, less than 20% of uncited documents were found to be open access. The result of a two-sample proportion test also shows that the percentage of cited OA documents is significantly higher than that of the uncited set [P<0.0001]. As with previous studies, Wikipedia acts as an "amplifier" for the already freely available OA literature (Teplitskiy et al., 2017). The results show the "amplifying role" of open social media, especially social web, in strengthening and widening the visibility and impact of open access documents, which could provide a potential citation advantage over their non-OA peers (Hajjem, Harnad & Gingras, 2006; Sotudeh, Ghasempour & Yaghtin, 2015; Sotudeh & Estakhr, 2018). Previous studies confirm that citing OA content has both quality and visibility advantages (Sotudeh & Estakhr, 2018). Moreover, availability via multiple platforms was found to play an important role in increasing citation counts for OA documents (Xia et al., 2011).

However, more than 70% of documents in both sets are not open access (Figure 2). This was expected, as the number of OA documents is smaller than non-OA ones (Sotudeh & Estakhr, 2018; Björk & Paetau, 2012; Laakso, 2014).

In terms of OA article type, Figure 3 shows that for cited documents, more green type documents are found, whereas for non-cited documents, there are more gold type documents. This could be due to the PubMed Central policy of archiving green versions, or due to the fact

that open access full-texts of research documents may not necessarily be the official published version of the document.

Figure 2: OA status for cited and uncited sets of documents in Wikipedia.

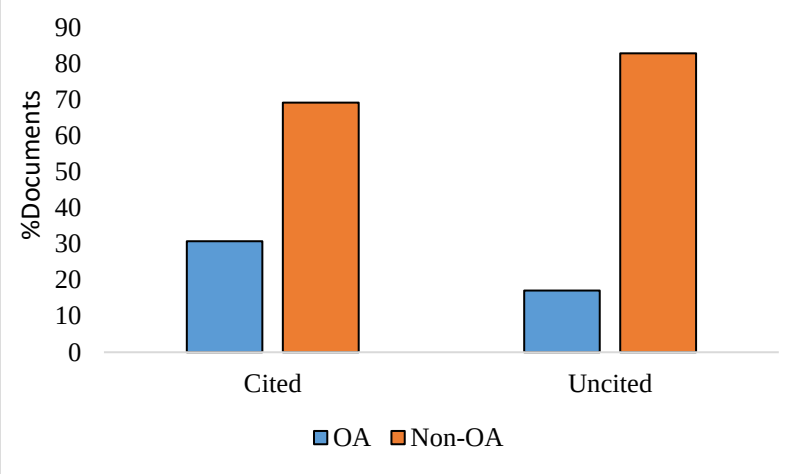
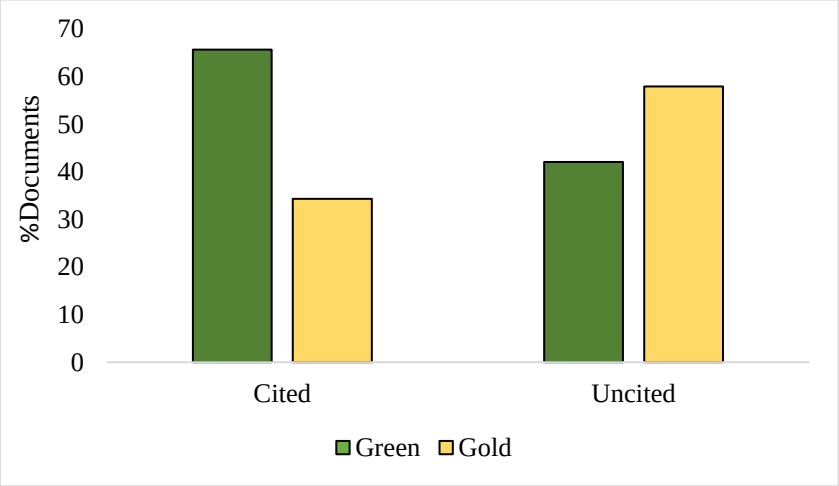


Figure 3: OA models for cited and uncited sets of documents in Wikipedia.



Question 3. Which Medical Subject Headings (Mesh) are cited more often in Wikipedia?

Cited documents from the sample were classified into 15,852 topics, and the uncited documents were classified into 10,289 topics. Neoplasms, Tuberculosis and Disease are the top three topics in both sets. However, the top 10 topics for cited documents are not the same as that for the uncited set (Table 1). In a sample of PubMed documents cited in Wikipedia pages (published before 2010), Evans and Krauthammer (2011) found that a quarter of documents' MESH headings are related to genetics. This shows that topic interest of Wikipedia citations has changed over time, as our current study finds that disease-related documents are of more interest to the Wikipedia community.

Table 1. Top 10 topics and their corresponding percentages for cited and uncited sets.

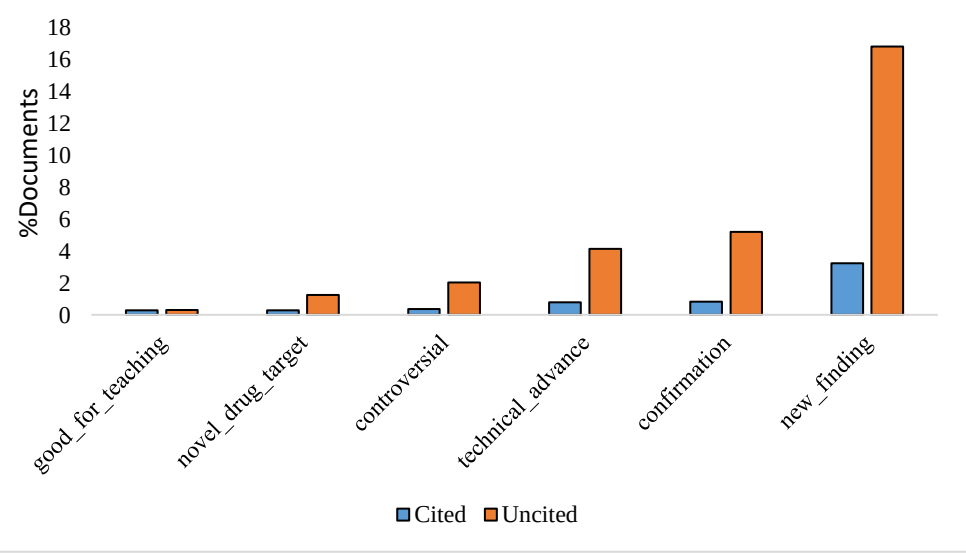
Cited	**Number (%)**	**Uncited**	**Number (%)**
Neoplasms	3631 (0.94)	Neoplasms	3503(0.94)
Tuberculosis	3139 (0.82)	Tuberculosis	3123 (0.84)
Disease	2737 (0.71)	Disease	2591 (0.70)
Mutation	2362 (0.61)	Medicine	1560 (0.42)
Biological Evolution	1958 (0.51)	Biometry	777 (0.21)
Phylogeny	1833 (0.48)	Intestines	768 (0.21)
Medicine	1651 (0.43)	Blood	766 (0.21)
Evolution. Molecular	1466 (0.38)	Brain	759 (0.20)
Gene Expression Regulation	1379 (0.36)	Anesthesia	733 (0.20)
Signal Transduction	1354 (0.35)	Tooth	730 (0.20)

Question 4. Which F1000 classes are cited more often in Wikipedia?

The findings show that majority of documents in both cited and uncited sets of documents are classified into the '*new finding*' class of F1000. However, the proportion of uncited documents in this class (~16%), is significantly higher than that of cited documents (~4%; P<0.0001). '*Confirmation*' and '*technical advance*' classes are respectively the second and third classes in both cited and uncited document sets (Figure 4).

This result is in line with Wikipedia's policies on *identifying reliable sources (*Wikipedia, 2018*).* The policy requires that editors rely on secondary sources, which accurately reflects *current* and *up-to-date* medical knowledge and can be found in *recent*, authoritative review articles, statements and practice guidelines, issued by scientific and widely respected health authorities.

Figure 4: F1000 classes for cited and uncited sets.



Question 5. Are there significant correlations between Wikipedia citation counts and F1000 counts, news counts, and tweet counts?

Spearman correlations have been used to study the relationship between Wikipedia citation counts, F1000 counts, news counts, and tweet counts for the entire collection of cited and uncited documents in Wikipedia.
Whilst a significant negative correlation is found between Wikipedia citation counts and F1000 and tweet counts, a very weak positive correlation is found between Wikipedia citation counts and news counts (Table 2). The results show that whilst 9.71% of documents cited in Wikipedia are mentioned in news outlets, only 7.13% of uncited documents are mentioned in news outlets. This finding concurs with previous studies confirming an inter-relation between altmetric indicators (Priem et al., 2012). The negative correlation between Wikipedia citation count and the F1000 and tweet counts, and the positive correlation found between Wikipedia citation count and news count, may signify different interests across social web communities, and also content (mis)alignment. While Wikipedia and news communities have slightly common interests, F1000 and Twitter communities do not show any alignment with the Wikipedia community. This may imply that Wikipedia editors' focus is neither ad-hoc – to be of interest to specialists - nor very common – to be of interest only to the general public.

Table 2. Spearman's rho correlation coefficients for the relationship between Wiki citation counts, F1000, news, and tweet counts.

Variable	F1000 post count	News post count	Tweet post count
Wiki citation count	-0.26*	0.07*	-0.35*

* p <0.0001

Conclusion

This study investigates different characteristics of cited medical articles in Wikipedia versus uncited articles. The findings show a document type similarity for both cited and uncited sets of documents, with the articles, reviews and editorial materials being more visible. Whilst the documents cover a broad range of topics, the top three topics are similar between the two sets. This implies a global significance of these topics, particularly as they are also rated as top or highly important in the Wikipedia rating system, and are listed as subjects for which Wikipedia should provide high-quality articles (Shafee et al., 2017). The open access status of documents indicates that Wikipedia favors OA documents, although a large number of cited documents are non-OA. Regarding the F1000 classes, the majority of both cited and uncited documents are categorized as "new finding".
Finally, our findings show significant, although weak correlations between Wiki citation counts and F1000, tweet and news counts. Whilst F1000 and tweet counts correlate negatively with Wikipedia citation counts, the news counts have a positive correlation.
Overall, according to Teplitskiy et al. (2017), the editors of English Wikipedia in medicine act as "distillers" of quality science. They interpret and distribute open/closed access knowledge to a broad public audience via different document types, whilst focusing on new findings and current medical knowledge. Moreover, it seems that Wikipedia's focus is neither specialized, nor generalized, but it is something of a rather "general scientific" nature.

References

Björk, B. C., & Paetau, P. (2012). Open access to the scientific journal literature–status and challenges for the information systems community. Bulletin of the Association for Information Science and Technology, 38(5), 39-44.

Didegah, F. (2017). Factors associated with Wikipedia citations vs. traditional citations to research articles. WikiCite Conference. Vienna, Austria, 23 May.

Evans, P., & Krauthammer, M. (2011). Exploring the Use of Social Media to Measure Journal Article Impact. *AMIA Annual Symposium Proceedings, 2011,* 374-381.

Hajjem, C., Harnad, S., & Gingras, Y. (2006). Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact. arXiv preprint cs/0606079.

Haigh, C. A. (2011). Wikipedia as an evidence source for nursing and healthcare students. *Nurse Educ Today, 31*(2), 135-139.

Heilman, J. M., & West, A. G. (2015). Wikipedia and medicine: quantifying readership, editors, and the significance of natural language. *J Med Internet Res, 17*(3), e62.

Koppen, L., Phillips, J., & Papageorgiou, R. (2015). Analysis of reference sources used in drug-related Wikipedia articles. *Journal of the Medical Library Association: JMLA, 103*(3), 140–144.

Kousha, K., & Thelwall, M. (2017). Are wikipedia citations important evidence of the impact of scholarly articles and books? *Journal of the Association for Information Science and Technology, 68*(3), 762-779.

Laakso, M. (2014). Green open access policies of scholarly journal publishers: a study of what, when, and where self-archiving is allowed. Scientometrics, 99(2), 475–494. Retrieved from http://link.spri nger.com/article/10.1007%2Fs11192-013-1205-3.

Laurent, M. R., & Vickers, T. J. (2009). Seeking Health Information Online: Does Wikipedia Matter? *Journal of the American Medical Informatics Association: JAMIA, 16*(4), 471-479.

Nielsen, F. A. (2007). Scientific citations in Wikipedia. *First Monday; Volume 12, Number 8 - 6 August 2007*.

Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). Altmetrics in the wild: Using social media to explore scholarly impact. arXiv preprint arXiv:1203.4745.

Shafee, T., Masukume, G., Kipersztok, L., Das, D., Haggstrom, M., & Heilman, J. (2017). Evolution of Wikipedia's medical content: past, present and future. *J Epidemiol Community Health, 71*(11), 1122-1129.

Sotudeh, H., & Estakhr, Z. (2018). Sustainability of open access citation advantage: the case of Elsevier's author-pays hybrid open access journals. Scientometrics, 1-14.

Sotudeh, H., Ghasempour, Z., & Yaghtin, M. (2015). The citation advantage of author-pays model: The case of Springer and Elsevier OA journals. Scientometrics, 104(2), 581-608.

Teplitskiy, M., Lu, G., & Duede, E. (2017). Amplifying the impact of open access: Wikipedia and the diffusion of science. *Journal of the Association for Information Science and Technology*, *68*(9), 2116-2127.

Thelwall, M. (2016). Does astronomy research become too dated for the public? Wikipedia citations to astronomy and astrophysics journal articles 1996–2014. *El Profesional de la Información*, *25*(6), 893–900.

Unpaywall (2018). Retrieved from http://unpaywall.org.

Xia, J. F., Myers, R. L. & Wilhoite, S. K. (2011). Multiple open access availability and citation impact. Journal of Information Science, 37(1), 19-28.

Wikipedia (2008). Wikipedia: verifiability. Wikimedia Foundation. Retrieved 3 April 2018, from http://en.wikipedia.org/wiki/Wikipedia:Verifiability.

Wikipedia (2018). Wikipedia: Identifying reliable sources (medicine). Wikimedia Foundation. Retrieved 5 April 2018, from https://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources_(medicine).