# {{citation needed}}: Filling in Wikipedia's Citation Shaped Holes

Kris Jack[1], Pablo López-García[2], Maya Hristakeva[1], and Roman Kern[2]

[1] Mendeley Ltd - 144a Clerkenwell Road, London EC1R 5DF (United Kingdom)
[2] Know-Center GmbH - Inffeldgasse 13/6, A-8010 Graz (Austria)

**Abstract.** Wikipedia authors cite external references to support claims made in their articles in order to increase their validity. A large number of claims, however, do not have supporting citations, putting them in question. In this paper, we describe a study in which we attempt to retrieve relevant citations for claims using a variety of information retrieval algorithms. These algorithms are inspired by bibliometric and altmetric insights that exploit readership data from Mendeley's community and rerank results using a Bradfordising approach. The results of the small scale study indicate that both of these approaches can improve upon basic keyword-based search, typically used in digital libraries, in order to return relevant documents for unsupported claims.

**Keywords:** Digital Libraries, Bibliometrics, Altmetrics, Wikipedia

## 1 Introduction

Wikipedia has shown to be of extreme value not only for the general public but also in the academic world. Research has shown an increasing number of scholarly publications citing Wikipedia and that academic institutions are one of Wikipedia's major consumers [7]. Given the crowdsourced nature of Wikipedia and its associated coordination problems, however, Wikipedia is often sceptically viewed by experts with respect to information quality [2, 1].

Similar to scholarly literature, Wikipedia authors are encouraged to cite reliable and verifiable[1] information sources that back up their claims and therefore strengthen the information quality of articles. This practice has been widely adopted, resulting in a rich set of Wikipedia articles with numerous references.

Experts, however, might still find two serious concerns regarding information quality when it comes to Wikipedia claims and their associated references. On the one hand, despite the large number of already existing references, it is still commonplace for co-authors or readers to encounter unsupported claims in Wikipedia articles. These claims needing corresponding evidence are marked with a {{citation needed}} tag[2]. It is then up to Wikipedia contributors to find the corresponding references that support such claims.

[1] `http://en.wikipedia.org/wiki/Wikipedia:Verifiability`
[2] `http://en.wikipedia.org/wiki/Wikipedia:Citation_needed`

In order to increase the number of cited claims in Wikipedia, it would be useful to develop a tool that can automatically suggest articles that back them up. Wikipedia contributors could then check if the articles indeed back up the claims and choose to insert relevant ones into Wikipedia pages.

In this paper, (i) we investigate the distribution of citation sources in Wikipedia to better understand what authors are currently using as verifiable evidence and the extent to which citations are missing, and (ii) we explore if findings in the field of bibliometrics can be exploited in developing a system that can automatically retrieve articles that back up claims. The second study is particularly focussed on comparing the well established technique of Bradfordising [12] to a new technique of biasing retrieval ranking based on signals from digital communities of researchers. As this study was carried out in Mendeley, the digital community chosen is that of Mendeley's social network, whose readership data has previously shown to be useful in understanding research trends [3, 4].

## 2 Wikipedia Citations

While many claims in Wikipedia articles are backed up with citations, there is also a considerable amount which is not. It can be difficult to quantify how many claims made in Wikipedia articles lack citations but estimates can be made based on the number of missing citations that have been explicitly signposted using the {{citation needed}} tag.

In order to estimate the number of claims that are not backed up with citations we downloaded a copy of the English Wikipedia from Academic Torrents[3]. The download contained 4.4 million articles written in English in an XML format. These articles were parsed using the SAXParser from the WikiXMLJ project[4], to find all citation tags that appear within the collection, following Wikipedia's citation conventions[5]. All parsing was done on a laptop with 8GB RAM and took less than 30 minutes to run.

Around one million Wikipedia articles contained at least one citation. In total, just over nine million citations were parsed out. We focused on the top five types of citations, which account for 93.68% of all citations made (see Table 1). Citations can cite different types of objects. The most popular type of citation is the {{cite web}} tag which indicates that a web page is being cited. This type accounted for over half of all citations made. Citations for news articles, books and journals accounted for 17.08%, 11.00%, and 8.22% of all citations made, respectively. The fifth most frequent citation type was {{citation needed}}, the tag used to indicate that a claim is missing a citation.

It's reasonable to assume that not all claims lacking citations in Wikipedia articles have been explicitly tagged with {{citation needed}}. As a result, the 402,347 {{citation needed}} tags are likely to cover only a subset of the actual

[3] Wikipedia English Official Offline Edition (version 20130805) [Xprt] - `http://academictorrents.com/details/30ac2ef27829b1b5a7d0644097f55f335ca5241b`

[4] `https://code.google.com/p/wikixmlj/`

[5] `http://en.wikipedia.org/wiki/Wikipedia:Citation_templates`

| Type of Wiki Citation Tag | Citation Count | Distribution(%) |
|---|---|---|
| {{cite web}} | 4,796,157 | 52.94% |
| {{cite news}} | 1,547,056 | 17.08% |
| {{cite book}} | 996,433 | 11.00% |
| {{cite journal}} | 744,866 | 8.22% |
| {{citation needed}} | 402,347 | 4.44% |
| **Total** | **8,486,859** | **93.68%** |

**Table 1.** Top 5 citation links that appear in English Wikipedia articles.

claims that require citations. The number of missing citations in Wikipedia articles indicates the need for a tool that can help people to retrieve articles that back up claims.

## 3 Approaches to Finding Citations

In this study, we were interested in applying some insights from bibliometrics and altmetrics to inform the design of a tool that can help retrieve articles that support natural language claims. The behaviour of three algorithms was investigated. The first is Bradfordising, a technique that has been shown to improve the ranking of research article results in search engines [5]. The second is to bias search results based on how often they are read in Mendeley's community. The third is to combine both approaches.

All algorithms were investigated using the popular search engine Lucene[6]. The article metadata (e.g. title, authors, year of publication, abstract) of Mendeley's 100 million research articles were indexed. For each claim, the text of the claim plus the title of the Wikipedia page were entered into the search engine and the results were reanked based on either Bradfordising, readership or a combination of Bradfording and readership.

The standard Bradfordising approach was followed, applying it to the first 100 results. That is, the first 100 results were reranked so that the articles from the most frequent publication venue, appearing in the first 100 results, were ranked above the articles appearing in less frequent publication venues, from the first 100 results.

In order to exploit Mendeley's readership information, a new query handler was written in Lucene, that extended the basic keyword-based search with a weighted boost. The weighted boost is based on a logarithmic function of the number of readers that an article has, as follows;

$score * \log_{10}(number\_of\_readers + 1)$

The final score given to each result was based on its original keyword-based score plus the boosting given by the logarithmic function. As a result, articles that had more readers should have higher ranked positions.

Finally, the third algorithm combines both approaches, first applying the readership bias and then Bradfordising.

[6] `https://lucene.apache.org/`

## 4 Experimental Setup

As a case study, a small scale gold standard data set was manually created by selecting 10 Wikipedia pages with claims citing scholarly articles in them. A claim was randomly chosen from each Wikipedia page and associated with the scholarly article that it cited. It is recognised that more than one scholarly article may support the claim made. This approach, despite this limitation, is common practice in the information retrieval community, as it provides enough information to fairly compare different algorithms and can be fully automated for large scale testing. The task of the algorithms is, given a claim, to retrieve articles that can be used to back it up. We employed two baseline systems: (i) Google Scholar and (ii) Mendeley's catalogue search. Both cover a broad range of research disciplines and are two of the world's largest research collection repositories. These baselines were compared to versions of Mendeley's search engine enhanced separately using Bradfordising and readership biases, as described in the previous section.

Five of the 10 selected claims are provided as examples (Table 2). These claims are made using natural language sentences that paraphrase and/or summarise findings from research articles. The 10 claims cross multiple disciplines of research, just as Google Scholar and Mendeley's collections do.

| **Claim Summary** | **Full claim** | **Readers** |
|---|---|---|
| Quiet Revolution | The Quiet Revolution is called such because it was not a "big bang" revolution; rather, it happened and is continuing to happen gradually | 109 |
| Opting Out and In | They are passed up for promotions because of the possibility that they may leave, and are in some cases placed in positions with little opportunity for upward mobility to begin with based on these same stereotypes | 31 |
| WWW | One study, for example, found five user patterns: exploratory surfing, window surfing, evolved surfing, bounded navigation and targeted navigation | 8 |
| Statistical Learning | There is an ongoing debate about the relevance and validity of statistical approaches in AI, exemplified in part by exchanges between Peter Norvig and Noam Chomsky | 22 |
| Cognitive Robotics | Within developmental robotics, developmental learning approaches were elaborated for lifelong cumulative acquisition of repertoires of novel skills by a robot, through autonomous self-exploration and social interaction with human teachers, and using guidance mechanisms such as active learning, maturation, motor synergies, and imitation | 148 |

**Table 2.** Claims citing a corresponding research article available in both Google Scholar and Mendeley. The last column shows the number of readers in the Mendeley catalogue.

Based on the claim and the title of the Wikipedia article page, a query was constructed. The query contained all the words that appeared in the claim with the Wikipedia article page's title concatenated to it. This query was used to evaluate each of the approaches: (i) Google Scholar (Google Sch.), (ii) Basic Mendeley Keyword Search (Men), (iii) Mendeley + Readers (Men+R), (iv) Mendeley + Bradfordising (Men+B), and (v) Mendeley + Readers + Bradfordising (Men+R+B). The first 100 results lists from each tool were gathered and the position of the cited article in each list was recorded. The closer the article's position to the start of the results list, the better the approach performs.

## 5 Results

Table 3 shows the results of the five tested algorithms to retrieve the corresponding citations to the claims in Table 2.

| **Claim Summary** | **Google Sch.** | **Men** | **Men+R** | **Men+B** | **Men+B+R** |
|---|---|---|---|---|---|
| Quiet Revolution | >100 | 73 | **2** | 76 | 29 |
| Opting Out and In | >100 | >100 | >100 | >100 | >100 |
| WWW | >100 | 67 | 66 | **1** | 71 |
| Statistical Learning | >100 | 69 | **24** | 73 | 37 |
| Cognitive Robotics | >100 | 10 | **2** | **2** | 9 |
| Benford's Law | >100 | **2** | 8 | 9 | **2** |
| DNA Sequencing | >100 | 38 | **6** | 64 | 69 |
| Naturalism | >100 | >100 | >100 | >100 | >100 |
| Mathematics Definition | >100 | 47 | 38 | **3** | 8 |
| Core Samples | >100 | **2** | **2** | 34 | 52 |
| **Totals** | **1,000 (0)** | **508 (2)** | **348 (5)** | **462 (3)** | **477 (1)** |

**Table 3.** Rank of the citations for the two baseline and three modified approaches. A lower number indicates the approach performs better. If the target article did not appear in the first 100 results returned, then >100 appears. Best ranks appear in bold. The score in parentheses is the count of the number of best placed ranks that the algorithm achieved.

Mendeley's basic keyword search outperformed Google Scholar in retrieving the correct target document (i.e. the document actually cited in the Wikipedia article) in the first 100 results in 8 cases. Google Scholar failed to retrieve the target documents in the first 100 results for all queries. When considering only the top 10 results returned, 2 of the sample queries retrieved the correct target citations in the top 10 results using Mendeley's basic keyword search.

Three algorithms were tested based on bibliometrics and altmetrics insights. The use of readership counts ranked the target articles higher, on average, than the use of Bradfordising and the combination of readership and Bradfordising, resulting in 5, 4, and 3 top 10 hits respectively. When considering the number of cases in which an algorithm ranked the target document highest, Mendeley

+ Readership provided the highest ranked results in 5 of the cases. Mendeley's basic keyword search ranked the target article higher in 2 cases compared to Mendeley + Readership + Bradfordising's single case. The results suggest the combination of the 2 algorithmic enhancements, readership boosting and Bradfordising, appear to produce worse results than using either of these algorithms alone.

There were 2 cases when all algorithms tested failed to retrieve the target article in the top 100 results. In both of these cases, the claims did not contain the keywords present in the metadata of the articles.

## 6 Discussion

There is evidence that scholarly articles are increasingly citing Wikipedia. One study showed that Wikipedia had been cited 3,679 times within a reference data set taken from the Web of Science (WoS) and Elsevier's Scopus databases [7]. Regarding the information quality of Wikipedia articles, in 2007 Nielsen studied the relationship between a journal citation in Wikipedia and the impact factor of the journal, and a correlation between them could be observed, especially for high impact journals [6]. In 2012 Priem et al. sampled a number of scholarly articles and found that about 5% were cited by the English Wikipedia [9]. These results suggest that Bradfordising also applies in Wikipedia articles.

When it comes to the influence of readership, incorporating the readership count or popularity into a information retrieval system has been studied by many research groups. Researchers proposed that the readership count can be seen as an indicator for the quality of the retrieved articles and the to rerank the results accordingly. They found that among multiple quality metrics, the popularity contributed significantly to the improvement of the results [13], in good agreement with our findings. In our case, using keyword-based search with Mendeley and readership boosting retrieved the target citation in the top 10 results in 5 out of the 10 queries ran. Comparisons of download and citation data from Scopus with readership data from Mendeley have shown a medium to high correlation between downloads and readership and downloads and citations, while there is a medium-sized correlation between readership and citations. These results suggest some difference between the different usage features [11, 10].

None of the algorithms tested managed to retrieve the target article in their top 100 results in 2 of the tests. In considering the 2 queries, it appears that they do not share enough keywords in common with the target article's metadata. This points to the need for an alternative representation of articles beyond metadata and possibly an alternative representation of the query itself. Including the full text could possibly prove beneficial in such scenarios as suggested by [8]. Furthermore, a deeper linguistic representation such as the semantics revealed through topic modelling is worth considering.

## 7 Conclusions and Future Work

This paper reports on the results of two studies. First, the contents of the English Wikipedia were parsed in order to find what type of objects tend to be cited and the extent to which missing citations were present. This work confirmed that most citations are made to webpages, news articles, books and journals, and that there is a sizable number of claims that have been explicitly marked as needing a citation (over 400,000). Second, a small scale study was conducted in which Google Scholar was compared with Mendeley's search engine, and three modified versions of Mendeley's search engine to test how well they could retrieve articles based on natural language claims. The results show that reranking through Bradfordising and boosting readership scores both improve upon keyword-based search.

This study has been conducted as part of the EEXCESS project[7]. In its current form, it requires some manual interventions preventing us from scaling it up. In the future, we will automate the entire process so that we can run large scale tests and generate statistically significant results. We will also explore how to return not just scholarly articles, but objects that appear in cultural repositories (e.g. Europeana[8]). As seen by the types of citations made in Wikipedia articles, we need to go beyond scholarly articles in order to meet this community's needs.

We also intend to take all citations to journal articles in the English Wikipedia and automatically build a data set of citation claims paired with articles that they cite and have been deduplicated against Mendeley's catalogue. This data set will serve as a training and testing data set for a large scale evaluation of how well the target citations can be retrieved given the claims as input and will allow us to compare the attributes of different information retrieval algorithms in more detail.

## Acknowledgements

The presented work was developed within the EEXCESS project funded by the European Union Seventh Framework Programme FP7/2007-2013 under grant agreement number 600601. The Know-Center GmbH is funded within the Austrian COMET Program Competence Centers for Excellent Technologies of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency (FFG).

[7] http://eexcess.eu/
[8] http://www.europeana.eu/

# Bibliography


[1] Kittur, A., Kraut, R.E.: Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In: Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work. pp. 37–46. CSCW '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1460563.1460572

[2] Kittur, A., Suh, B., Pendleton, B.A., Chi, E.H.: He says, she says: Conflict and coordination in wikipedia. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 453–462. CHI '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1240624.1240698

[3] Kraker, P., Körner, C., Jack, K., Granitzer, M.: Harnessing User Library Statistics for Research Evaluation and Knowledge Domain Visualization, p. 1017. ACM (2012), http://know-center.tugraz.at/download_extern/papers/user_library_statistics.pdf

[4] Kraker, P., Trattner, C., Jack, K., Lindstaedt, S., Schloegl, C.: Head Start : Improving Academic Literature Search with Overview Visualizations based on Readership Statistics (2013)

[5] Mayr, P.: Relevance distributions across bradford zones: Can bradfordizing improve search? arXiv preprint arXiv:1305.0357 (2013)

[6] Nielsen, F.Å.: Scientific citations in wikipedia. arXiv preprint arXiv:0705.2106 (2007)

[7] Park, T.: The visibility of wikipedia in scholarly publications. First Monday 16(8) (2011), http://pear.accc.uic.edu/ojs/index.php/fm/article/view/3492

[8] Peacock, P.J., Peters, T.J., Peacock, J.L.: How well do structured abstracts reflect the articles they summarize. European Science Editing 35(1), 3–6 (2009)

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

[10] Schloegl, C., Gorraiz, J., Gumpenberger, C., Jack, K., Kraker, P.: Are downloads and readership data a substitute for citations? The case of a scholarly journal. In: LIDA (2014)

[11] Schloegl, C., Gorraiz, J., Gumpendorfer, C., Jack, K., Kraker, P.: Download vs. Citation vs. Readership Data: The Case of an Information Systems Journal (2013), http://know-center.tugraz.at/download_extern/papers/issi2013_schloegletal.pdf

[12] White, H.D.: 'bradfordizing'search output: how it would help online users. Online Information Review 5(1), 47–54 (1981)

[13] Zhu, X., Gauch, S.: Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In: ACM SIGIR Conference. pp. 288–295. SIGIR '00, ACM, New York, NY, USA (2000), http://doi.acm.org/10.1145/345508.345602