Article

Relationship of Google Scholar Versions and Paper Citations

Marjan Mohammadjafari 1*, Hadi Salehi 2, Kaveh Bakhtiyari 3,4, Nader Ale Ebrahim 5, Mahmoud Danaee 6, Masoud Shakiba 7, Masoomeh Shahbazi-Moghadam 8 and Azam Zavvari 9

1 Department of Industrial Engineering, Kerman Branch, Islamic Azad University, Kerman, Iran
2 Faculty of Literature and Humanities, Najafabad Branch, Islamic Azad University, Najafabad, Isfahan, Iran
3 Department of Computer & Cognitive Science, Faculty of Engineering, University of Duisburg-Essen, Duisburg 47057, Germany
4 Department of Electrical, Electronics and Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi 43600, Malaysia
5 Centre of Research Services, Institute of Research Management and Monitoring (IPPP), University of Malaya, 50603 Kuala Lumpur, Malaysia; aleebrahim@um.edu.my
6 Academic Enhancement and Leadership Development Centre (ADeC), University of Malaya, 50603 Kuala Lumpur, Malaysia
7 Manukau Institute of Technology, Auckland 2023, New Zealand
8 Perdana School of Science, Technology and Innovation Policy, Universiti Teknologi, 81310 Skudai, Johor, Malaysia
9 Center for Software Technology and Management, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

* Correspondence: marjan_mohamadjafari@yahoo.com; Tel.: +60-379-676-289

Abstract: The number of citations that a paper has received is the most commonly used indicator to measure the quality of research. Researchers, journals, and universities want to receive more citations for their scholarly publications to increase their h-index, impact factor, and ranking respectively. In this paper, we tried to analyses the effect of the number of available Google Scholar versions of a paper on citations count. We analyzed 10,162 papers which are published in Scopus database in year 2010 by Malaysian top five universities. Then we developed a software to collect the number of citations and versions of each paper from Google Scholar automatically. The result of spearman correlation coefficient revealed that there is positive significant association between the number of Google Scholar versions of a paper and the number of times a paper has been cited.

Keywords: h-index; citations; published version; Scopus database; highly cited paper; bibliometrics

1. Introduction

Jorge E. Hirsch in [1] proposed the Hirsch-index which is commonly abbreviated as the H-index. The h-index is an index that attempts to measure the productivity and cumulative impact of a researcher’s published work by looking at the distribution of citations the work has received [2, 3]. Another common indicator which measuring the reputation and academic standard of a journal is so-called ‘Impact Factor’, which, with some qualifications, is the average number of citations for papers published in a particular journal [4]. Impact Factor is obtained as the ratio of the total number of citations received by the papers published in the journal to the number of papers published in the journal [5, 6]. A majority of world university ranking are also rely on paper citations. So, receiving more citations is very important for authors, journals, and university to get high h-index, impact factor, and world ranking [7, 8]. In this research, we analyses the effect of the number of available Google Scholar versions of a paper from the web on receiving citations of the paper. We targeted to analyze all of published papers in year 2010 related to five top university of Malaysia which appear...
in the Scopus database. To achieve on this purpose, 10,162 papers which are published in Scopus database in year 2010 are selected. Then we developed software to collect the number of citations and versions of each paper from Google Scholar automatically.

2. Definition of Citation

A bibliographic citation is a reference to a book, article, web page, or other published item. Citations should supply detail to identify the item uniquely [9, 10]. Citation is a reference to a published or unpublished source. Citing sources points the way for other scholars [11].

*Important Purposes Citation:*

- To avoiding plagiarism or support of academic honesty [12].
- To attribute prior or unoriginal work and ideas to the correct sources [13].
- To allow the reader to determine independently whether the referenced material supports the author’s argument in the claimed way.
- To help the reader gauge the strength and validity of the material the author has used.

3. Number of Versions for a Paper

Publishing a research paper in a scholarly journal is necessary but not sufficient for receiving citations in the future. We need to ensure that the paper is visible to the relevant users and authors. When the authors published a paper, the publisher will put the published version of the paper to own website and repository. This means we have a product, and we also have one shop then if somebody wants to have our product must come into our shop and buy it. But if we have so many versions then we can find more customers. For example, somebody made a pen and put on the one shop to sell it, beside somebody else made another pen and put on the 20 shops to sell it. It is appeared that the pen in the 20 shops is more visible for customers and then this pen will sell more.

The question is that, due to copyright roles how we can publish our paper in more than one journal to get more websites for advertise our paper. Actually, no need to publish in more than one journal but the authors can use some tools that help in enhancing the visibility and readership of research papers. Effective use of these tools can result in increased citations and, thus, improve the h-index of the author and journal impact factor. Here is a sample of tools to increase the visibility of one’s published papers.

4. Strategies for Enhancing the Impact of Research Dissemination

- Submit the manuscript to a digital subject repository.
- Submit the manuscript to an institutional repository.
- Set up a web site devoted to the research project and post manuscripts of publications and conference abstracts [14].
- Take advantage of SEO (search engine optimization).
- Present preliminary research findings at a meeting or conference.
- Follow up preliminary research findings presented at a meeting or conference with a published manuscript [15].
- Consider submitting the same article to a journal in a different language as a “secondary publication.”
- Start a blog devoted to the research project [16].
- Contribute to Wikipedia.
- Contribute to a social network [17].
5. Methodology

In this research, five research universities of Malaysia namely University of Malaya (UM), Universiti Putra Malaysia (UPM), Universiti Kebangsaan Malaysia (UKM), Universiti Sains Malaysia (USM) and Universiti Teknologi Malaysia (UTM) were selected to conduct this research. We collected 10,162 papers related to year 2010 from the Scopus database, and the extraction process for collecting these papers was done in 13 July 2013 11:00 AM (UTC +8:00) for 2 hours. The process of data collection is shown in the Figures 1-7.

**Figure 1:** Process of data collection from Scopus database

```
Start
Open Scopus Search tab / Affiliation Search
Search The University Name
Select Correct Name from Affiliation results
Select on the Show documents button
Refine the result and limit to publication year 2010
Select all Papers and Export them
End
```

**Figure 2:** The capture of affiliation search in Scopus database
Figure 3: The capture of affiliation search in Scopus database

Figure 4: The capture of affiliation search result in Scopus database

Figure 5: The capture of refine affiliation to show documents in Scopus database
To collect the number of citations and versions of these articles, Google Scholar search engine was used. We decided to focus on this tool because of its popularity and ability to provide a simple way to find the citations of articles. Also, Google Scholar database covers more resources and it reflects more versions and citations in comparison with the other databases such as ISI Thomson Reuters or Scopus. Therefore, we developed a software to collect the number of citations and versions of each paper from Google Scholar automatically.

All records should have been processed for the number of citations and versions in only a day. Because each day, new citations and versions might be created which results the incompetency in our data analysis. In order to overcome this issue, a server-based software application was developed to retrieve citations and versions. ASP.NET platform was selected to software development, and launched on a high-speed and band-width server to be able to process all these 10,000+ records in few hours.

**Software Algorithm**

It searched every single title in Google Scholar in 2 times. The first time with quotation marks (“”) and the second time without quotation marks. In the resulted page of Google Scholar, the titles and description might be included with some HTML tags as below:

- `<b>`: For the keywords match the search query, they would be bold to show the matched title with the keywords.
- `<i>`: This tag was also found in few titles in Google Scholar search results.
<sup> </sup> <sub> </sub>: Those titles with the superscripts and subscripts (e.g. Chemical formulas) consist of these tags to show the titles properly.

In order to extract and find the correct matching title in Google Scholar, all these HTML tags were removed from the titles. But still there was another challenge, and it was the different spacing. Some of the titles extracted from Scopus were 1 or 2 spaces different with those indexed in Google Scholar. So, after removal of all tags from titles, all spaces were also removed to find the correct match of the paper in Google Scholar results. In some cases, there were more than two items matched with the full title, and then in this case, the year and the authors’ names of the published data were matched to find the relevant record.

If the title, were found, then the number of citations and versions were extracted from the page and it was updated in the database; and if it was not found, it would be marked as “not found” in the database. The whole extraction process was done in 15 July 2013 12:00 AM (UTC +8:00) for 4 hours. After completion of the data extraction, those not found records, were also checked manually to make sure that system and data analysis have minimum incomplete data and no record has been missed on Google Scholar. The structural procedure is visualized in the Figure 8.

![Figure 8: The structure of developed software](image-url)
6. Data Analysis

As both number of citations and number of versions were not distributed normally, a non-parametric method, was applied to test the differences among top five universities. Table 1 shows the collected data of five top Malaysian research universities for year 2010 from Scopus database. The result of spearman correlation coefficient revealed that there was positive significant association between number of citation and number of versions for different universities publication. The overall correlation coefficient was a moderate and positive ($r = 0.431$, $p < 0.01$). Relationship between number of citation and number of version shows in Table 2.

### Table 1: Total number of publications and citations of five top universities of Malaysia

<table>
<thead>
<tr>
<th>University</th>
<th>Number of Publication</th>
<th>Number of Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>UKM</td>
<td>1961</td>
<td>7243</td>
</tr>
<tr>
<td>UM</td>
<td>2276</td>
<td>11118</td>
</tr>
<tr>
<td>UPM</td>
<td>2061</td>
<td>9251</td>
</tr>
<tr>
<td>USM</td>
<td>2480</td>
<td>12575</td>
</tr>
<tr>
<td>UTM</td>
<td>1384</td>
<td>5081</td>
</tr>
<tr>
<td>Total</td>
<td>10162</td>
<td>45268</td>
</tr>
</tbody>
</table>

**Table 2: Correlation between number of citations and number of versions**

<table>
<thead>
<tr>
<th>Number of Versions</th>
<th>UKM</th>
<th>UM</th>
<th>UPM</th>
<th>USM</th>
<th>UTM</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Citation</td>
<td>.413**</td>
<td>.478**</td>
<td>.403**</td>
<td>.427**</td>
<td>.405**</td>
<td>.431**</td>
</tr>
</tbody>
</table>

**Note:** Correlation is significant at the 0.01 level (2-tailed).

6.1. Comparison among 5 Malaysian top universities for number of citations and number of versions

As both number of citations and number of versions were not distributed normally, Kruskal Wallis test, a non-parametric method, was applied to test the differences among these universities. The results revealed that there were significant differences among these 5 universities for both number of citation and version (Table 3).

### Table 3: The results of Kruskal Wallis Test for comparison among universities

<table>
<thead>
<tr>
<th></th>
<th>Number of citation</th>
<th>Number of versions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>51.786</td>
<td>21.816</td>
</tr>
<tr>
<td>df</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>P value</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

Figures 9 and 10 showed the differences for the average number of citation and version among universities. The highest mean number of citation per article observed for USM (M= 5.07) and UM (M= 4.88) respectively while UTM (M=3.69) and UKM (M=3.67) had the lowest mean number of citation per article.
6.2. Relationship between type of document and type of university for number of publication

The frequency of different type of document in each university was calculated. Table 4 showed the pattern of publication in each university based on the document type. To study the relationship between type of document and universities a chi square test was applied the result of Fisher exact test revealed that there was a significant relationship between type of publication and university (Table 5).
Table 4: Number of publication Relationship between type of document and type of university

<table>
<thead>
<tr>
<th>Type of document</th>
<th>UKM Count</th>
<th>%</th>
<th>UM Count</th>
<th>%</th>
<th>UPM Count</th>
<th>%</th>
<th>USM Count</th>
<th>%</th>
<th>UTM Count</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article</td>
<td>1260</td>
<td>64.30%</td>
<td>1767</td>
<td>77.60%</td>
<td>1673</td>
<td>81.20%</td>
<td>1915</td>
<td>77.20%</td>
<td>632</td>
<td>45.70%</td>
</tr>
<tr>
<td>Article in Press</td>
<td>5</td>
<td>0.30%</td>
<td>6</td>
<td>0.30%</td>
<td>4</td>
<td>0.20%</td>
<td>5</td>
<td>0.20%</td>
<td>1</td>
<td>0.10%</td>
</tr>
<tr>
<td>Conference Paper</td>
<td>644</td>
<td>32.80%</td>
<td>395</td>
<td>17.40%</td>
<td>306</td>
<td>14.80%</td>
<td>463</td>
<td>18.70%</td>
<td>728</td>
<td>52.60%</td>
</tr>
<tr>
<td>Editorial</td>
<td>1</td>
<td>0.10%</td>
<td>14</td>
<td>0.60%</td>
<td>7</td>
<td>0.30%</td>
<td>10</td>
<td>0.40%</td>
<td>8</td>
<td>0.60%</td>
</tr>
<tr>
<td>Erratum</td>
<td>1</td>
<td>0.10%</td>
<td>3</td>
<td>0.10%</td>
<td>2</td>
<td>0.10%</td>
<td>3</td>
<td>0.10%</td>
<td>1</td>
<td>0.10%</td>
</tr>
<tr>
<td>Letter</td>
<td>11</td>
<td>0.60%</td>
<td>21</td>
<td>0.90%</td>
<td>3</td>
<td>0.10%</td>
<td>12</td>
<td>0.50%</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Note</td>
<td>3</td>
<td>0.20%</td>
<td>8</td>
<td>0.40%</td>
<td>1</td>
<td>0.00%</td>
<td>3</td>
<td>0.10%</td>
<td>1</td>
<td>0.10%</td>
</tr>
<tr>
<td>Review</td>
<td>36</td>
<td>1.80%</td>
<td>61</td>
<td>2.70%</td>
<td>65</td>
<td>3.20%</td>
<td>67</td>
<td>2.70%</td>
<td>11</td>
<td>0.80%</td>
</tr>
<tr>
<td>Short Survey</td>
<td>0</td>
<td>0.00%</td>
<td>1</td>
<td>0.005%</td>
<td>0</td>
<td>0.00%</td>
<td>2</td>
<td>0.10%</td>
<td>2</td>
<td>0.10%</td>
</tr>
</tbody>
</table>

Table 5: Type of publication and university

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>df</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Chi-Square</td>
<td>915.163</td>
<td>32</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Fisher’s Exact Test</td>
<td>853.92</td>
<td>--</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

6.3. Comparison among different types of publication for number of citation and number of versions

Kruskal Wallis test was applied to test the differences for number of citation and number of versions among different types of publication. The results revealed that there were significant differences for both number of citation and version (Table 6).

Table 6: The results of Kruskal Wallis Test for comparison among different types of publication

<table>
<thead>
<tr>
<th></th>
<th>Number of citations</th>
<th>Number of versions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>691.304</td>
<td>584.03</td>
</tr>
<tr>
<td>df</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>P value</td>
<td>&lt;0.01</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

Figure 11 shows the differences for the average number of citations and versions among universities the highest mean for citations per article observed for review article (M=16.07) and the lowest was for editorial (M=0.45).
7. Conclusion

In this paper, we analyses the effect of the number of available Google Scholar versions of a paper from the web on receiving citations for the paper. We targeted to analyze all of published papers in year 2010 related to five top university of Malaysia which appear in the Scopus database. To achieve on this purpose, 10,162 papers which are published in Scopus database in year 2010 are selected. Then we developed a software to collect the number of citations and versions of each paper from Google Scholar automatically. Since there is a strong association between the number of Google Scholar versions of a paper and the number of times a paper has been cited, we encourage the researchers to increase the number of paper versions by depositing in different open access repositories.

Author Contributions: All authors checked manually on Google Scholar for “not found records” which returned by the software. Nader Ale Ebrahim did the data collection and final touch up of the paper. While, Marjan Mohammadjafari; wrote the initial draft of the paper, Hadi Salehi; revised it, and Mahmoud Danaee analyzed the data. Kaveh Bakhtiyari was the software programmer. Masoud Shakiba, Azam Zavvari, and Masoomeh Shahrzadi-Mohammad contributed to clean the data and wrote some parts of the paper.

Conflicts of Interest: The authors declare no conflict of interest.
References


13. Rubio, C.C., Colombia’s Poetic World of Authors’ Moral Rights: Considerations on Imprisoning a Professor for Plagiarism. 2012.


