International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:11, No:10, 2017

Comparison of the H-Index of Researchers of Google Scholar and Scopus

Adian Fatchur Rochim, Abdul Muis, Riri Fitri Sari

Abstract**—**H-index has been widely used as a performance indicator of researchers around the world especially in Indonesia. The Government uses Scopus and Google scholar as indexing references in providing recognition and appreciation. However, those two indexing services yield to different H-index values. For that purpose, this paper evaluates the difference of the H-index from those services. Researchers indexed by Webometrics, are used as reference's data in this paper. Currently, Webometrics only uses H-index from Google Scholar. This paper observed and compared corresponding researchers' data from Scopus to get their H-index score. Subsequently, some researchers with huge differences in score are observed in more detail on their paper's publisher. This paper shows that the H-index of researchers in Google Scholar is approximately 2.45 times of their Scopus H-Index. Most difference exists due to the existence of uncertified publishers, which is considered in Google Scholar but not in Scopus.

Keywords**—**Google Scholar, H-index, Scopus.

I. Introduction

THE growth of scientific papers per year is an impact of information technology (IT) advancement. This technology allows institutions to store and publish their scientific papers much easier. Also, it makes researchers able to access other papers and giving communication facilities among researchers. Furthermore, the more paper stored as an open access documents, the more papers will be read and cited by others.

Each published paper can be cited by other papers. The number of other papers that cited a particular paper represents the number of citation. This citation shows how the published paper has been acknowledged by others. Here, researcher's performance can be stated as a quantitative representation index, calculated based on the number of published papers and the citations it has obtained. The well-known index is the H-Index. The H-Index indicator, currently, plays an important role in evaluating the quality of the researcher and hence every researcher needs to consider the bibliometric aspect.

Some institutions or countries use H-index as a tool to measure the productivity of researchers or lecturers [1]. Webometrics site by The Spanish National Research Council (CSIC) ranks researchers in some countries using H-index [2]. The research productivity indicators also use the H-index to evaluate performance in certain [3].

Indonesian government has started to use H-index and other bibliometrics from sources of the Google Scholar and Scopus to evaluate the performance of lecturers, researchers, publishers, and research institutions [4].

The institution/country for making the academic policy uses h-index of both indexing services. Therefore this paper intended to observe the difference of the H-index of the two services. For that purpose, 1,227 highly cited researchers with H-index > 100 by the webometrics site are considered [5]. To obtain those data, an application to collect data automatically from Scopus database has been developed. Data from Scopus were obtained through application programming interface (API).

This paper is divided into six chapters. The next chapter discusses about the work that has been done related to the comparison of h-index on some indexing services. Third chapter discusses about research methodology used. The fourth section explains data sources and data collection method. The fifth section discusses the data processing. Finally the conclusion is presented.

II. Related Works

Harzing compared some indicators of indexed journals on Google Scholar and ISI Thomson (CA). The results showed that Google's H-index is more accurate and comprehensive than ISI Thomson's (CA) [9].

Minasny, et al. conducted an H-index comparison study among the Google Scholar, Web of Science (CA) and Scopus of soil researchers. The research result showed that the Google Scholar came up with the highest H-index value among the three [10].

Lan et al. provided a statement that a database cannot replace other database. Google Scholar has the best total number of citation. CA is the best when considering the citations only in the journal, while Scopus is the best in displaying data from 1996 forward [11]. The study compared those databases and focused on data source of the databases.

III. Methodology

This paper was conducted based on the following methodology

1. Getting data of author name list of top highly cited researchers from the Webometrics site.
2. Observing valid researchers with corresponding researchers on Scopus based on name and institution.
3. Obtaining publication data from Scopus through API
4. Observing the different

Adian F. Rochim, Abdul Muis, and Riri Fitri Sari are with Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia,, Depok, 16242, Indonesia (e-mail: adian.fatchur@ui.ac.id, muis@ui.ac.id, riri@ui.ac.id).

5. Evaluating the result.

IV. Data Collection

The "Webometrics Ranking of Scientists World" is part of the webometrics rankings sites, an initiative of the Cybermetrics Lab, a research group belonging to the Consejo Superior de Investigaciones Científicas (CSIC), the largest public research body in Spain [12].

The webometrics site provided the ranking of highly cited researchers' data. The data of the researchers are based on H-index generated by Google Scholar. Google Scholar has been indexing journals, books, conference proceedings, and universities' or research's institutions repositories, since 2004. [7], [8].

Scopus has been indexing journals, books, bibliographic data of authors and publishers. Scopus has indexed more than 22,748 peer-reviewed journals and over 3,643 journals with full open access types [6]. Scopus has indexed a total of 7.7 million conference papers from 97,100 conferences. In addition, patents have been incorporated in the Scopus database. There are 28 million patents from patent offices has been indexing, i.e. US Patent & Trademark Office, European Patent Office, Japan Patent Office, World Intellectual Property Organization, and UK Intellectual Property Office [6].

Scopus provides data access services through two ways as:

1. Accessing the web with authenticated user password at http://scopus.com
2. Accessing via an API by token authentication and internet protocol (IP) listed by the Scopus, via url address http://api.elsevier.com.

It can be accessed from registered internet protocol (IP) gateway by Scopus.

This work used the API access, since the website access needs a long time to retrieve citation data set. The menu of the web can only be accessed manually. A data set of researcher can take 3-5 minutes to be retrieved, whereas the data taken in this study were about 1,720 researchers' dataset. Accessing data by API service is the best alternative to get data set from Scopus. There are some provisions of the Scopus API service as follows:

1) Each user can have 10 Scopus API service keys.
2) Each query is limited up to 50,000 for seven days.
3) Each query is limited to 200 results.
4) Each citation data takes 2-3 seconds per citation per paper.

For the purpose of the data for this research and considering the requirements and limitations given by the Scopus, it is necessary to make a method that satisfies such provisions. Fig. 1 shows step by step access data from the Scopus by API.

1. Querying Scopus ID based on name and affiliation of researchers.
2. Querying H-index score based on the Scopus ID.
3. Tabulating the data.
4. Comparing H-indexes from Scopus and Google Scholar using statistical analysis.
5. Using Peirce's Method to eliminate the outlier data from comparison result.
6. Analyzing and representing the result in graph and table.

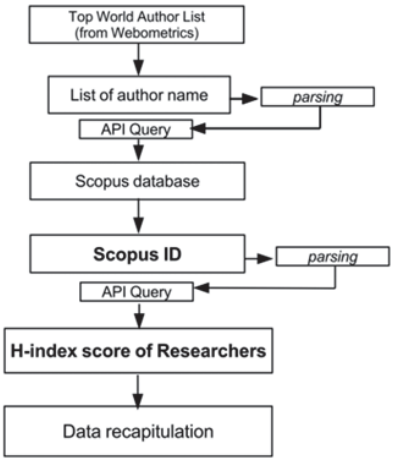

Fig. 1 The flowchart of data retrieval by Scopus API

```
  document_count              name
0            784  Ronald C. Kessler
1              7  Ronald C. Kessler
2              3  Ronald C. Kessler
[{'affiliation': u'Harvard Medical School', 'author_id': u'7202074046', 'name':
u'Ronald C. Kessler', 'document_count': 784}, {'affiliation': u'Harvard Medical
School', 'author_id': u'56606710700', 'name': u'Ronald C. Kessler', 'document_count':
7}, {'affiliation': u'Uniformed Services University of the Health Sciences', 'author_id':
u'55951054200', 'name': u'Ronald C. Kessler', 'document_count': 3}]
A total number of  0  records for the query.
Empty DataFrame
Columns: []
Index: []
[]
A total number of  0  records for the query.
Empty DataFrame
Columns: []
Index: []
[]
A total number of  1  records for the query.
  affiliation   author_id  document_count           name
0  Collège de  55911103200              66  Pierre Bourdieu
[{'affiliation': u'Coll\xe8ge de', 'author_id': u'55911103200', 'name': u'Pierre
Bourdieu', 'document_count': 66}]
A total number of  1  records for the query.
                    affiliation   author_id  document_count  \
0  The Johns Hopkins School of Medicine  35406091300            1246
```

Fig. 2 Result sample obtained of second step (before parsed)

Fig. 2 shows an example of data obtained as the query result to the Scopus database using input data containing the researcher's name and affiliation. This resulted in output data containing the researcher's name, affiliation, Scopus ID number and amount of data owned. The data must be parsed to get the clear data. Table II shows the result of first stage. Some of the omitted data are as follows:

- ✓ Header
- ✓ No ID Scopus Data,
- ✓ Researcher names and affiliations.

A total sample of 1,260 H-index data of top cited researchers by Google Scholar is obtained after filtered. The things that led to the difference in the amount of the number of total individual h-index are:

1. The application cannot found the name of researchers in the Scopus.
2. The names of the researcher have no first and family name.
3. Researchers have no Scopus ID, but only Google Scholar ID.

TABLE I
A SAMPLE OF THE DATA AFTER PARSED

No.	Name	Scopus ID	Doc. Number
1	researcher 1	7202074046	171
2	researcher 2	55911103200	10
3	researcher 3	35406091300	106
4	researcher 4	55978052600	1
5	researcher 5	36077704000	148
…		7402409226	173
1259	researcher 1259	16151582900	195
1260	researcher 1260	35463345800	195

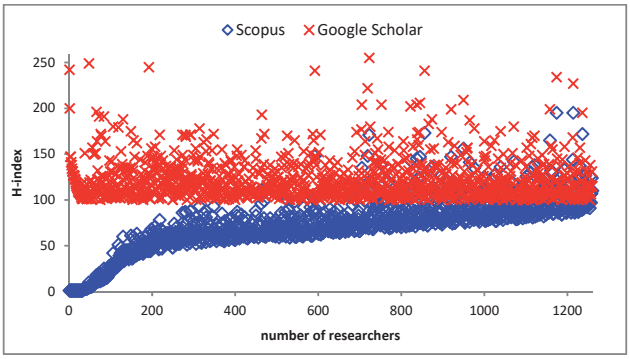

Fig. 3 H-index distribution of Scopus vs. H-index of Google Scholar based on 1,260 top scientist webometrics version (Webometrics August 2017)

TABLE II
H-INDEX DISTRIBUTION DATA OF GOOGLE SCHOLAR AND SCOPUS BASED ON 1,260 TOP SCIENTIST WEBOMETRICS VERSION (WEBOMETRICS, AUGUST 2017)

Database	Mean	Median	Minimum	Maximum	H-index range
Google Scholar	120.91	245	100	255	100-255
Scopus	72.12	106	0	195	0-195

Table II shows that several H-index based on Scopus have zero value.

V. RESULT AND DISCUSSION

A. Preprocessing Data

The query data get 1,260 individual H-index scores of Scopus and Google Scholar. The data obtained from the retrieval are citation data sets of 2,558 world top scientist of webometrics version based on Google Scholar H-index. Fig. 3 shows H-index comparison of Google Scholar and Scopus by scatter graph. The figure shows that H-index of Google Scholar is higher than the H-index of Scopus. Several data show the difference between both H-indexes. To explain the difference of the H-index, we analyze the data using statistical analysis.

The data of the top scientist based on the Google Scholar has range of 100-255. Table II shows the comparison of statistical analysis of the H-index between Google Scholar and Scopus.

The data source of Scopus is different from Google. To make a comparison of the data, we need to clear the outlier data. To calculate and clear the data from the outlier data, we used the Peirce's Method. To prepare the data for clearing the outlier data, first of all we find the H-index multiplier factor of Scopus and Google Scholar. Table III shows distribution data of H-index multiplier factor of Google Scholar and Scopus. The equation to find the multiplier factor is:

$$mf = (Hindex\ of\ GS)/(Hindex\ of\ Scopus) \quad (1)$$

For example: A researcher has H-index 10 in Google Scholar, and 5 in the Scopus. The multiplier factor is = 10/5 = 2.

TABLE III
MULTIPLIER FACTOR DISTRIBUTION DATA FROM H-INDEX OF SCOPUS TO GOOGLE SCHOLAR BEFORE OUTLIERS DATA REMOVED

Mean	Median	Minimum	Maximum	Range
5.66	1.59	1.05	242	1.05-242

Table III shows that the data are not in normal distribution. To make the normal distribution of the data we consider for eliminating the "outliers" data [13], [14]. Using the Peirce's method we need 3 times loops to eliminate the outliers. Algorithm 1 shows the algorithm of the Peirce's Method. Table IV shows the steps of the calculation for removing outliers.

TABLE IV
ITERATION STEP-BY STEP FOR REMOVING OUTLIERS DATA

No	Iteration-n	Data	Outliers removed
1.	0	1260	0
2.	1	1228	32
3.	2	1227	1
4.	3	1227	0
Total Outliers removed			33

The algorithm for removing the outliers is as follows:

1. Calculate the mean and the sample standard deviation (SD) of the complete data set.
2. Obtain R corresponding to the number of measurements taken from Peirce's table.
3. Calculate the maximum allowable deviation:|xi – xm| max
4. For any suspicious data measurements, obtain |xi – xm|
5. Eliminate the suspicious measurements if:

$$|x_i - x_n| > |x_i - x_m|max \quad (2)$$

6. If this result is within the rejection of one measurement, go to step 8.
7. If more than one measurement is rejected in the above test go to step 9.
8. Repeat the above calculations (steps 2 – 5),
9. Now obtain the new value of the mean and sample
10. Saving the new SD.

B. Comparing H-Index of Google Scholar and Scopus

After filtering, we have only 1,227 H-index data ready to compare based on Google Scholar and Scopus H-index. To compare the H-index of Google Scholar and Scopus, we describe the multiplier factor of Google Scholar and Scopus using normal distribution.

Fig. 4 shows the normal distribution of the multiplier factor data of H-index Google Scholar and Scopus. After removing

the outliers, we analyze data for comparing the H-index of Google Scholar vs. Scopus. Table V shows the result of statistical analysis.

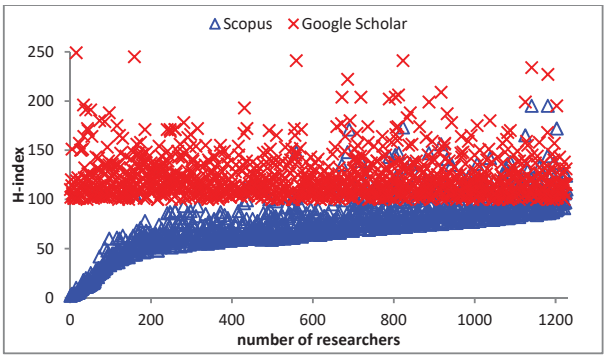

Fig. 4 H-index distribution data of Scopus vs. H-index of Google Scholar after outlier's data removed

In this work, we compared H-index of the Google Scholar and Scopus, and removed the outlier of the multiplier factor, and we found that H-index of the Google Scholar is 2.45 times the average of the H-index of Scopus. Table VII shows samples the differences significantly of H-index of Google Scholar and Scopus.

Fig. 5 shows that H-index of Google Scholar and Scopus significantly difference at quartile range 25% to 50%. It causes the correlation test low. Correlation test results with Pearson's method get 0.46 results indicating that H-index Scopus has no correlation with Google Scholar. Fig. 6 shows test correlation of H-index of the Scopus and Google Scholar by Pearson's Method. Table VIII shows the quartiles of the data H-index of Google Scholar and Scopus.

Table VI shows the multiplier factor statistical analysis of the comparison H-index between Google Scholar and Scopus.

Fig. 5 shows the quartiles from distribution of H-index data of Google Scholar and Scopus. The next study will explore the other different databases to be explored such as: Microsoft Academic Search (MAS) and Connecting Repositories (CORE).

TABLE V
H-INDEX OF GOOGLE SCHOLAR AND SCOPUS DISTRIBUTION DATA AFTER REMOVING THE OUTLIERS

Database	Mean	Median	Minimum	Maximum	H-index range
Google Scholar	120.80	114	100	255	100-255
Scopus	73.92	74	2	195	2-74

TABLE VI
DISTRIBUTION DATA OF THE MULTIPLIER FACTOR FROM H-INDEX OF SCOPUS TO GOOGLE SCHOLAR, AFTER REMOVING THE OUTLIERS

Mean	Median	Minimum	Maximum	SD	Range
2.45	1.57	1.05	55	4.04	1.05-55

VI. CONCLUSION

The result of deeply analysis from 1,227 H-index of Scopus and Google Scholar concluded that the H-index of Google Scholar has 2.45 times of their H-index in Scopus, and the standard deviation of the multiplier factor is 4.04. The Spearman correlation coefficient value of 0.46, there appears to be weak correlation between H-index of Google Scholar and Scopus.

TABLE VII
SAMPLE OF RESEARCHERS H-INDEX OF GOOGLE SCHOLAR AND SCOPUS THAT HAS SIGNIFICANTLY DIFFERENCE

Researcher	Scopus			Google Scholar	
	number of citation	number of papers	H-index	number of citation	H-index
1	4	1	1	790615	242
2	66	1	1	218078	200
3	0	3	1	130666	147
4	68	1	1	80900	141
5	18	1	1	75711	138
6	35	1	1	117956	134
7	5	1	1	56039	131
8	5	1	1	84371	131
9	92	1	1	63627	127
10	16	1	1	75645	127
11	109	1	1	30050	126

TABLE VIII
QUARTILES OF H-INDEX OF GOOGLE SCHOLAR AND SCOPUS

Database services	Quartiles	1 (25%)	2 (50%)	3 (75%)	4 (100 %)
Scopus	H-index range	0-61	62-74	75-88	89-195
	Number of H-index of researchers	290	329	316	292
Google Scholar	H-index range	100-106	107-114	115-128	129-255
	Number of H-index of researchers	332	289	311	295

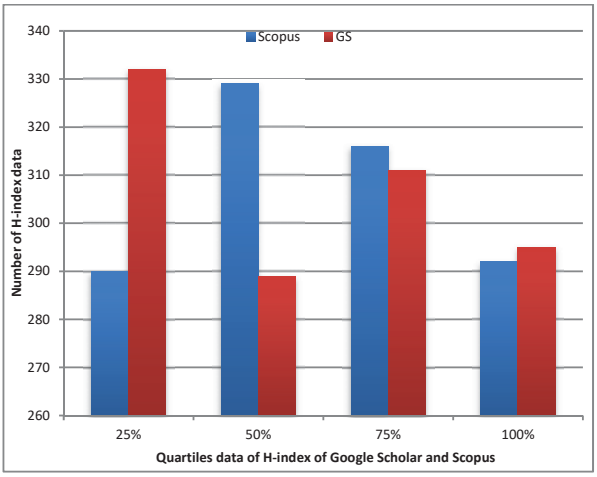

Fig. 5 Distribution H-index data by quartiles representation

```
> cor.test(gsotemp,scotemp)

        Pearson's product-moment correlation

data:  gsotemp and scotemp
t = 18.15, df = 1224, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4152118 0.5034986
sample estimates:
      cor
0.4604933
```

Fig. 6 Correlation test of the Google Scholar and Scopus by Pearson's Method

ACKNOWLEDGMENT

This research is funded by the grant from the Ministry of Research and Higher Education of the Republic of Indonesia number 3202/UN2.R3.1/PPM.00.01/2017.

REFERENCES

[1] J. R. Lacasse, D. R. Hodge, and K. F. Bean, "Evaluating the Productivity of Social Work Scholars Using the h-Index," Res. Soc. Work Pract., vol. 21, no. 5, pp. 599–607, 2011.
[2] G. Mester, "Rankings Scientists, Journals and Countries using h-Index," Interdiscip. Descr. Complex Syst., vol. 14, no. 1, pp. 1–9, 2016.
[3] I. Díaz, M. Cortey, À. Olvera, and J. Segalés, "Use of H-Index and other bibliometric indicators to evaluate research productivity outcome on swine diseases," PLoS One, vol. 11, no. 3, pp. 1–21, 2016.
[4] Ministry of Research, Technology and Higher Education of the Republic of Indonesia "Science and Technology Index," website, 2017. (Online). Available: http://sinta2.ristekdikti.go.id/about.
[5] I. F. Aguillo, "Highly Cited Researchers (h>100) according to their Google Scholar Citations public profiles," CSIC, Madrid, Spain, 2017. (Online). Available: http://www.webometrics.info/en/node/58. (Accessed: 01-Sep-2017).
[6] Scopus, "Content - Scopus - Solutions | Elsevier," Scopus, 2017. (Online). Available: https://www.elsevier.com/solutions/scopus/content. (Accessed: 14-May-2017).
[7] E. Orduña-Malea and E. Delgado López-Cózar, "Google Scholar Metrics evolution: an analysis according to languages," Scientometrics, vol. 98, no. 3, pp. 2353–2367, 2014.
[8] J. Mingers and L. Leydesdorff, "A review of theory and practice in scientometrics," Eur. J. Oper. Res., vol. 246, no. 1, pp. 1–19, 2015.
[9] A. Harzing, "A Google Scholar H-Index for Journals : An Alternative Metric to Measure Journal Impact in Economics & Business," J. Am. Soc. Inf. Sci. Technol., vol. 60, pp. 41–46, 2009.
[10] B. Minasny, A. E. Hartemink, A. McBratney, and H.-J. Jang, "Citations and the h index of soil researchers and journals in the Web of Science, Scopus, and Google Scholar.," PeerJ, vol. 1, no. 1955, p. e183, 2013.
[11] J. Bar-Ilan, "Citations to the 'Introduction to informetrics' indexed by WOS, Scopus and Google Scholar," Scientometrics, vol. 82, no. 3, pp. 495–506, 2010.
[12] "What is webometrics? | Radek Malinský." [Online]. Available: http://malinsky.eu/blog/what-is-webometrics/. [Accessed: 14-Sep-2015].
[13] S. M. Ross, "Peirce's criterion for the elimination of suspect experimental data," J. Eng. Technol., vol. 20, no. 2, pp. 1–12, 2003.
[14] L. Lin and P. D. Sherman, "Cleaning Data the Chauvenet Way," SESUG 2007 Proc. SouthEast SAS Users Gr., no. c, pp. 1–11, 2007.