Scholytics Metrics guide

Ver.1.01, Updated October 2020

Scholytics

Table of Contents

01
Overview

1.1 Approach to research evaluation

Scholytics is a research evaluation service. By providing diverse metrics for institutions and authors, Scholytics shows the relative performance of authors and institutions. In other words, users can evaluate the performance of institutions and authors by comparing various metrics. The unique feature of Scholytics is to provide metrics for research units in institutions and universities as well. Most of other research evaluation platforms are providing metrics only for institution level, not for department level organizations. For example, MIT has lots of divisions and research centers in its organization. If a user want to check only the performance of Plasma Science and Fusion Center in MIT, excluding the performance of other division, Scholytics will be the perfect choice for the user because metrics for research unit level is the unique feature of Scholytics.

One simple keyword to explain the approach of Scholytics is openness. Scholytics is trying to cover as broad scholarly documents as possible, not intentionally selecting limited number of journals. Of course, Scholytics does not currently cover all the academic documents in the world, however, the ultimate goal is to cover 100% of the scholarly documents in the world. The coverage is going higher year by year.

Major metrics in Scholytics will be almost similar with them in the other research evaluation platforms due to the principle of statistics. If users in Scholytics use diverse filter options, however, users will have unlimited choices and can have real diverse perspectives.

1.2 Data

Data in Scholytics are basically come from Naver Academic, and part of the data in Scholytics are from secondary data not from Naver Academic. Naver Academic is one of the biggest scholarly search sites in the world. Similar web services are Google Scholar, Microsoft Academic and Baidu Scholar because all of these services are scholarly search services made by famous search companies in the world. Just like other search company-based services, Naver Academic is trying to cover all of the scholarly documents in the world. Global top publishers and aggregators are providing their metadata based on agreements.

Not all the data in Naver Academic is useful for Scholytics because some data do not have the information of authors and because some data do not have the information of affiliations. Therefore, the total number of scholarly documents in Scholytics is not identical with that in Naver Academic. More than 200 million documents are in Naver Academic, however, only a little less than 100 million documents are identified in Scholytics.

The other aspect of data in Scholytics is openness to local data. Lots of scholarly documents are written in local languages and those documents are also a great asset for the future of human. However, it is hard to find a research evaluation platform paying some attention to local data. Because the philosophy of Scholytics is openness and unlimited choices, Scholytics is trying to acquire and handle local data written in each local language. Therefore, users of Scholytics will have an additional option to view with or without local data.

1.3 Identification

Identification is the core process in research evaluation platform. In metadata of academic articles, authors and affiliations are written in text, and there is no universal ID working for all the authors and affiliations until now. Identification is the organizing process to recognize authors or affiliations in multiple documents and then add IDs as the same authors or affiliations.

To identify authors, Scholytics is using ORCID, email, co-author information, affiliation and other available information. Directly acquired Information from universities and institutions is also a great source for the quality of identification, and manual correction also enhances the quality as well.

There are two indices to check quality of identification, recall and precision. Those indices are widely used in many other quality tests in the process of data science. In the identification for author, recall means the biggest ratio of documents grouped as an author's from identification over all the real documents of the author. There can be several groups collected for the documents of an author. For example, if an author wrote 100 articles and 90 of them are grouped as the document of the author after identifying, recall for the author is 90%. There can be other groups for the author, however recall is defined as the ratio of the biggest group. The other core index, precision means the ratio of correctly grouped documents within actually grouped documents. In an identified group, for example, 99 documents were written by an author A, however, 1 document was written by an author B. Then, the precision of the identified group is 99%. Therefore, recall is mostly related with coverage, and precision is with correctness.

In Scholytics, recall for authors is at least 90%, and precision for authors is at least 99%. However, the ratio will go up because algorithm in Scholytics is gradually enhancing, and manual modification is ongoing to make the ratio higher.

To identify affiliations, Scholytics is using Ringgold, Grid and FundRef as the basement of affiliation database. Rather heavy algorithms are used to analyze full text expression of affiliations. By analyzing text information of affiliations and then matching with basement database, Scholytics identifies high level institutions and low level research units.

In Scholytics, recall for affiliations is also at least 90%, and precision for affiliations is at least 99%.

02
Metrics

2.1 Metrics for publications

2.1.1 Publication

Publication is the total number of documents published within selected years. Because Scholytics displays metrics for an entity, publication is the number of documents published by authors in a selected entity. In addition, diverse filters can be applied to this metrics. Total number of publications is a great index to show the performance of an entity, however filters can make this metric more valuable.
One of the filters is document type. In Scholytics, there are five document types: articles, conferences, trends/reports, books and others. These document types are the same types in Naver Academic. Among these types, articles, conferences and books have categories. Categories in Scholytics are identical with those in Naver Academic. In Naver Academic, there are ten top level categories and 175 second level categories. Therefore, 175 categories are also used as the second level categories in Scholytics. Those second level categories are applied to calculate Field-Weighed Citation Impact.

Table 1. Categories in Scholytics and Naver Academic

Top level categories	Second level categories
Humanities	Lexicography, History, Philosophy, Ethics …
Arts and Sports	Music, Fine Art, Cinema Studies …
Social Sciences	Political Science, Economics, Business …
Natural Sciences	Mathematics, Biology, Chemistry …
Engineering	Mechanical Engineering, Electronic Engineering ...
Medicine	Anatomy, Biotechnology, Immunology ...
Marine Studies	Agriculture, Forestry, Landscape Architecture ...
Interdisciplinary	Technology Policy ...
Education	Science Education, Ethics Education …
General Works	—

Currently, articles are categorized by journal-driven classification, however, conferences and books are categorized by publication-driven classification. Other category systems such as FOS, THE, QS are also scheduled to be used in Scholytics in a near future.
There are traditionally well-known journal indices in the world: SCIE, SSCI, A&HCI and Scopus. In Scholytics, users can use filters based on these indices. In addition, local indices mostly used in specific local counties can also be used. One of the examples is KCI. KCI is famous journal index in South Korea. Because there are many other journal indices mostly used in Asian countries, Scholytics will apply as many indices as possible in a near future. Because Scholytics is using journal index at the last month of each year, there can be a little difference from actual index. However, the difference is negligible, and feedback from clients will be accepted to correct metrics from the index.
Years in Scholytics are publication years. Therefore, if you check the graph of citations, you will find downside graphs in most cases because years in Scholytics are publication years, not cited years. In addition, years in Scholytics starts from 2010. Metadata in our academic world are enhancing year by year, and data from 2010 have relatively much better quality than ever. Although Naver Academic has documents covering all years even before 19 th century, Scholytics displays only publications published from 2010.

2.1.2 Publication in top journal percentiles

Top journals in Scholytics are selected based on FWCI calculated in a limited condition to consider different environment of each categories. For example, most categories in chemical or medical area earn normally higher citations rather than categories in humanities. If Scholytics calculated top journals only based on the number of citations, Scholytics would lose balance between categories. Therefore, FWCI in a limited condition is used in Scholytics to define top journals.
With articles published in the previous two years on a specific journal, Scholytics calculate FWCI of this year based on citations in this year. Therefore, FWCI of journal A in 2019 can be calculated as the numerical mean of FWCIs from articles published in 2017 and in 2018, and the FWCIs are calculated from citations made in 2019.

Scholytics is currently using only top 10 percentile when calculating top journal percentile, however more diverse percentiles will be added in a near future.
Publication in top journal percentiles is the number of articles published in journals selected as top journal percentiles.

2.1.3 Publication in top citation percentiles

Publication in top citation percentiles is the number of publications having citation numbers more than the threshold of top citation percentiles. Just like publication in top journal percentiles, publication in top citation percentiles is showing the performance of an entity with focusing on quality.
To make this metric available, threshold should be first calculated for each category and each publication year. Therefore, the competitors of each document are in the same category and published in the same year. Threshold for each category and year are calculated in advance, and then publications can be divided with the standard of the threshold.
The threshold is integer. If the threshold is zero for a category in a year, there is no publication over the threshold for the category in the year because zero means that top citation percentile is not defined at that time.
Scholytics is currently using only top 10 percentile when calculating top citation percentile, however more diverse percentiles will be added in a near future.

2.2 Metrics for citations

2.2.1 Citation

Citation in Scholytics is the number of lifetime citation received by publications of an entity. Because years in chart of Scholytics mean publication years, citation charts in Scholytics display downside graphs in most cases. It is natural that recently published documents have low chance to be cited by other publications.

2.2.2 Self-citation

Self-citation in Scholytics is citation received from publications of the same authors in an entity. Other research evaluation platforms may use different meaning concerning self-citation, however Scholytics use self-citation only in author-level. Although a selected entity can be an institution or a country, self-citation in Scholytics always means self-citation in author-level. Therefore, if users select to exclude self-citations in Scholytics, the result means author-level exclusion.

2.2.3 H5-index, H10-index

H-index is defined as the maximum value of h while an entity has published h papers that have each been cited at least h times. H-index is mostly used to show a balanced view over productivity and citation impact. To display stable and qualified values, Scholytics is currently providing h5-index and h10-index with limitation of publications years within 5 or 10 years.
If a user selects publication years from 2016 to 2018, Scholytics will display h5-index with publications from 2014 to 2018 with citation numbers from 2014 to 2018. Therefore, the last selected year is the control point in Scholytics.

2.2.4 Citation per publication

Citation per publication is the number of citations divided by the number of publications. Citation per publication was mostly used in the past if there is no access to research evaluation system. Although this metric is traditionally a great indicator in checking the performance of quality side, balance between categories is the clear weakness of this metric. In some categories such as medical and chemical categories, high citation per publication is usual. However, there are other categories with low citation per publication such as humanities. Therefore, we need to be careful when comparing entities with citation per publication if the compositions of categories in the entities are different.

2.2.5 Field-Weighted Citation Impact

FWCI, Field-Weighted Citation Impact is the core metric in Scholytics. Citation itself is a good performance indicator, however publications in a different categories cannot be compared fairly if we do not consider contextual information. For example, publications in engineering or medical category earn normally higher citations rather than publications categorized in humanities. Therefore, FWCI can be more balanced metric in comparing publications in different categories.

$$FWCI_i = \frac{C_i}{B_i}$$

C_i : Citations received by publication i received from the publication year.

B_i : Base number of citations per publication received by similar publications published in the same year. Similar publications are in the same second level category and in the same document type of the publication i.

To calculate FWCI of an entity, following steps are needed.
First, denominator is citations per publications by similar publications in the same category, in the same document type and in the same published year.
Second, numerator is citations of the publication. In Scholytics, citations are lifetime citations in calculating FWCI.
Third, calculate FWCI of a publication by dividing numerator by denominator.
Fourth, if an entity has many publications, calculate FWCI of the other publications as well in the same way and then calculate numerical mean of all the FWCIs.
Categories in calculating FWCI are the second level categories. If a publication has multiple categories, harmonic mean will be used when calculating denominator.
The meaning of FWCI is rather simple and intuitive. If FWCI is 1, it means that the publication received the same number of citations with similar publications in the same category, publication year and document type. If FWCI is 2, it means that the publication received twice more citations than similar publications. If FWCI is 0.5, it means that the publication received only half of the citations received by similar publications.

2.3 Metrics for collaborations

2.3.1 Collaboration

Collaboration in Scholytics means the degree of each type of co-authorship: International, national, institutional co-authorship, and single authorship.
All publications in Scholytics can be divided as one of these four authorship types. To define authorship types, there are three steps described below.
First, if there are multiple authors, it means the publication is not a single authorship. If it is not a single authorship, go to the next step. If not, the collaboration type is single authorship.
Second, if affiliations of authors have multiple countries, it means international collaboration. If not, go to the next step.
Third, if there are multiple top level institutions, it means the collaboration is a national collaboration. If not, the collaboration is an institutional collaboration.
Sometimes, top level institution can be a government of each country. In that case, Scholytics only checks lower level institutions, excluding the top level government, while defining authorship types.
In many cases, higher ratio of international collaboration type is related with better quality of research projects. Therefore, ratio of collaboration types is also a good indicator in checking performance.

2.3.2 Academic-Corporate Collaboration

Academic-Corporate Collaboration in Scholytics is the degree of collaboration between academic sector and corporate sector.
Affiliations in Scholytics can be defined diverse sectors: academic, corporate, government, hospital, public sector, and others. If some authors in a publication are from academic sector and the others are from corporate sector, it can be defined as academic-corporate collaboration.
In many cases, higher ratio of academic-corporate collaboration is related with higher industrial usage. Therefore, ratio of academic-corporate collaboration is also a good indicator in checking

performance of an entity. Industrial usage is important especially in some categories such as engineering categories.

2.3.3 Field-Weighted Collaboration

FWC, Field-Weighted Collaboration in Scholytics is a metrics using similar algorithm with FWCI. Just as the magnitude of citation can be different in each category, the magnitude of collaboration can also be diverse in each category. To earn more balanced view over collaboration, Scholytics set publications in the same category, document type and published year as similar publications and then calculate relative index to the competitors.
The meaning of FWC is rather simple and intuitive. If FWC is 1, it means that publications of an entity have the same ratio of a collaboration type with similar publications in the same category, publication year and document type. If FWC is 2, it means that publications of an entity have twice higher ratio of a collaboration type than similar publications. If FWC is 0.5, it means that publications of an entity have only half of the ratio of a collaboration type comparing with the ratio of similar publications.