Scientometrics (2019) 121:1165–1188
https://doi.org/10.1007/s11192-019-03213-w

Software survey: ScientoPy, a scientometric tool for topics trend analysis in scientific publications

Juan Ruiz-Rosero[1] · Gustavo Ramirez-Gonzalez[1] · Jesus Viveros-Delgado[2]

Received: 11 June 2019 / Published online: 30 August 2019

Abstract

Bibliometric analysis is growing research filed supported in different tools. Some of these tools are based on network representation or thematic analysis. Despite years of tools development, still, there is the need to support merging information from different sources and enhancing longitudinal temporal analysis as part of trending topic evolution. We carried out a new scientometric open-source tool called ScientoPy and demonstrated it in a use case for the Internet of things topic. This tool contributes to merging problems from Scopus and Clarivate Web of Science sources, extracts and represents *h*-index for the analysis topic, and offers a set of possibilities for temporal analysis for authors, institutions, wildcards, and trending topics using four different visualizations options. This tool enables future bibliometric analysis in different emerging fields.

Keywords ScientoPy · Scientometrics · Science mapping · Bibliometrics · Internet of things · Wildcards

Introduction

Scientometric is the study of measuring research quality and impact, understanding the processes of citations, scientific mapping fields, and the use of indicators in research policy and management (Mingers and Leydesdorff 2015). Nowadays, we can find a broad range of scientometrics tools: SciMAT (Cobo et al. 2012), Bibexcel (Persson et al. 2009), CiteSpace (Chen 2006), CoPalRed (Bailón-Moreno et al. 2006), Network Workbench Tool (Boerner et al. 2010), Sci2 (Lewis and Alpi 2017), VOSViewer (van Eck and Waltman 2010), BibliOTools (Grauwin and Jensen 2011), Publish or Perish (Harzing 2014), Bibliometrix (Aria and Cuccurullo 2017), and others. Most of these tools are specialized in science mapping, that aims to build bibliometric maps that describe how research fields are structured and connected through a network representation (Small 1997). Others tools are specialized

✉ Juan Ruiz-Rosero
jpabloruiz@unicauca.edu.co

[1] Departamento de Telemática, Universidad del Cauca, Calle 5, No. 4-70, Popayán, Cauca 190002, Colombia

[2] Departamento de Ingeniería Electrónica, Universidad de Nariño, Calle 18 Cra 50, Pasto, Nariño 520001, Colombia

in temporal analysis, which aims to identify the nature of phenomena represented by a sequence of observations across different periods of times.

The scientometric analysis shows the topics inside a search criterion, for example, the top countries evolution inside the criterion countries, or a list of specific author keywords inside the criterion author keywords. The temporal analysis allows us to find when a new phenom starts, and when it advances to a trending or emerging topic. The scientometrics tools have developed a kind of algorithms to perform the temporal analysis and find the trending topics, such as strategic diagrams (Cobo et al. 2011a, 2012) and Kleinberg's burst detection algorithm (Kleinberg 2003; Chen 2006). These kinds of analysis are performed in datasets that generally are extracted from a single bibliometric database, like Scopus or WoS, because, most of the tools can not merge the information successfully from different databases. Also, there is not a longitudinal graph representation of the trending topics evolution provided by all the actual tools.

In this article, we present a new open-source[1] scientometric tool called ScientoPy. This tool is a Python script based tool specialized in the temporal scientometric analysis. The full source code, instructions manual, and example commands are available in the public repository: https://github.com/jpruiz84/ScientoPy, and https://github.com/jpruiz84/ScientoPyUI for the user interface. Moreover, the tool described here has the following main characteristics:

- Import Clarivate Web of Science (WoS) and Scopus datasets
- Filter publications by document type
- Merge WoS and Scopus datasets based on a field tags correlation table
- Find and remove duplicated documents
- *H*-index extraction for the analyzed topics
- Country and institution extraction from author affiliations
- Top authors, countries, or institutions extraction based on the first document's authors or all document's authors
- Preprocessing brief graph and report table
- Top topics and specific topics analysis
- Wildcard topics search
- Absolute and relative growth indications
- Trending topics using the top average growth rate (AGR)
- Five different visualization graphs: timeline, bar, bar trends, evolution, and word cloud
- Command line and graphical user interfaces.

In this paper, we describe widely the preprocess and analysis steps performed by ScientoPy. Also, we use the Internet of things dataset [an updated dataset from the scientometric review (Ruiz-Rosero et al. 2017)] as a case study to show ScientoPy capabilities. In this way, we organized this paper as follows. First, "Methodology" section describes ScientoPy workflow to perform a scientometric analysis, divided by the dataset extraction, preprocessing steps, data analysis, and visualization. Then, we present a section "Case study: Internet of things" with an updated Internet of things (IoT) dataset, where we show all ScientoPy capabilities. In "Discussion" section, we compare ScientoPy skills against

[1] With MIT license (for more information, see https://opensource.org/licenses/MIT).

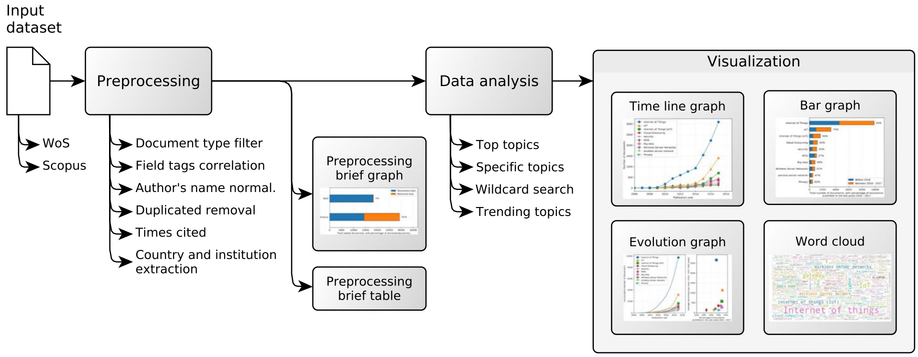


Fig. 1 ScientoPy scientometrics analysis workflow steps

others scientometric tools. Finally, in "Conclusion" section we summarize all the concluding remarks.

Methodology

In this section, we describe all of ScientoPy's capabilities for scientometric analysis. Figure 1 briefs the ScientoPy workflow steps for this kind of analysis. The first step is the input dataset extraction, where we explain which kind of databases and datasets are supported by ScientoPy. Second, the preprocessing that improves the dataset readability and precision, which includes document type filter, field tags correlation, author's name normalization, duplicated removal, times cited, and country/institution extraction. The preprocess results are summarized by the preprocessing brief graph and the preprocessing brief table. Third, in the data analysis, we can perform different operations to extract the top topics, specific topics trends, topic search based on wildcards, or trending topics, inside a selected criterion field (author names, country, author keywords, etc.). Finally, in the visualization step, we can observe the results that we have obtained from the data analysis step, using various graph types such as timeline graph, bar graph, evolution graph, and word cloud.

Dataset extraction

ScientoPy can process datasets from the two main bibliographic databases: Clarivate Web of Science (WoS), and Scopus. For WoS the supported format is *Tab-delimited (Win, UTF-8)*, which can be selected through the export option *Save to Other File Formats*. Also, we recommend saving the *Full Record* option in *Record Content*, to get the full document's information related to authors corresponding address used after for country and institution analysis. On the other hand, to get the dataset from Scopus, on the Scopus's *Export document settings* section (as the method of export) we must use *CSV Excel*. Then, on the information that we want to export, we select *Citation information*, *Bibliographical information*, and *Abstract and keywords* options. Now, we put all dataset files extracted in a single folder so that ScientoPy preprocessing script can handle them. More detailed instructions to get the dataset is available in ScientoPy's user manual from the public repository.

Preprocessing

ScientoPy uses the preprocessing steps techniques to improve the analysis readability and precision. The following subsections describe these techniques.

Document type filtering

By default, ScientoPy filters publications which are classified in one or more of the following document types:

- Conference Paper
- Article
- Review
- Proceedings Paper
- Article in Press

Because this kind of documents represent research works with a higher SJR (SCImago Journal Rank) and JCR (Journal Citation Reports) indicators. Others documents, such as book chapters, short surveys, letters, notes, books, editorials, erratum, reports, retracted documents, meeting abstracts, corrections, software reviews, and hardware reviews are excluded. Nevertheless, by modifying the ScientoPy global settings file, we can customize this document type filter.

Field tags correlation

WoS and Scopus use different field tags for their exported datasets. WoS uses two-character field tags, and Scopus uses a full sentence to describe each field. Table 1 describes the correlation that ScientoPy uses to convert the WoS to Scopus style fields. Moreover, ScientoPy uses Scopus style to handle and save the preprocessed dataset.

Author's name normalization

To have coherence in the document's author names, these fields must have consistency between the two input databases, for that reason a author names normalization is need here. For this case, WoS and Scopus have the following inconsistencies in the author's names:

- Scopus uses one comma to divide the authors, and WoS uses a semicolon.
- Scopus uses one dot after the first name initial, and WoS does not.
- WoS uses one comma to separate the author's last name and author's first name initial.
- Some journals use the author name with accents marks, special characters, and others do not.

Table 1 WoS and Scopus correlation field tags

WoS field tags	Scopus field tags
AU	Authors
BE	Editors
TI	Title
SO	Source title
LA	Language of original document
DT	Document type
DE	Author keywords
ID	Index keywords
AB	Abstract
C1	Affiliations
EM	Correspondence address
OI	ORCID
CR	Cited references
Z9	Cited by
PU	Publisher
SN	ISSN
BN	ISBN
J9	Abbreviated source title
PY	Year
VL	Volume
IS	Issue
BP	Page start
EP	Page end
AR	Article number
DI	DOI
PG	Page count
SC	Subject
UT	EID
PM	PubMed ID

Table 2 Two documents author's names fields example from WoS and Scopus (Munoz-Organero et al. 2011; Ciftler et al. 2017)

Database	Original author's names	Normalized author's names
WoS	Munoz-Organero, M; Ramirez, GA; ...	Munoz-Organero M; Ramirez GA; ...
Scopus	Muñoz-Organero M., Ramírez G.A., ...	Munoz-Organero M; Ramirez GA; ...
WoS	Ciftler, BS; Kadri, A; Guvenc, I	Ciftler BS; Kadri A; Guvenc I
Scopus	Çiftler B.S., Kadri A., Güvenç I.	Ciftler BS; Kadri A; Guvenc I

For instance, Table 2 shows two documents original author's names extracted from WoS and Scopus. Here, we observe the differences described previously.

Table 3 Document's titles examples from WoS and Scopus (Gezer and Taskin 2016; Moulin and Simon 2016) with the original language name in square brackets

Database	Language	Title
WoS	Turkey	An Overview of oneM2M Standard
Scopus	Turkey	An overview of oneM2M standard [oneM2M Standardina Genel Bir Bakis]
WoS	France	e-Health - The internet of things and telemedicine
Scopus	France	E-Health - The internet of things and telemedicine [E-santé - Objets ...]

These author's name inconsistencies generate problems to find similar author's names. For that reason, ScientoPy preprocessing script applies, in this order, the following steps to normalize author's name fields:

1. For Scopus replace the dot and coma (.,) with a semicolon (;)
2. For WoS and Scopus remove dots and comma
3. For WoS and Scopus remove accents marks

Table 2 in the third column shows the author's names normalized. Here, we find that the author's names of the same document in the two databases match.

Duplicated removal

Duplication removal is a critical step during the preprocessing procedure. If a dataset has duplicated items, the analysis scripts give us results that are not consistent and reliable. This duplication removal filter is based in the DOI and on the document's normalized title and first author last name. For the document title, unfortunately, for some documents that the original language is not English, Scopus adds the original title in square brackets, after the English title, and WoS does not. For instance, Table 3 shows two documents that are duplicated on WoS and Scopus with a different title. To solve this problem, for duplication removal, the ScientoPy's preprocessing script normalizes the title by removing the square brackets and the text inside them from the tile and converts it to upper case. In the case of authors name, the first author last name is normalized by removing the accents marks, special characters, and converting it to upper case.

Once ScientoPy normalizes the document's titles and first author last name, it runs the steps described on Algorithm 1 for duplication removal. Here, the documents are sorted by the database with WoS first than Scopus. Then, this script sorts the documents by the normalized title. In that way, the documents (before the for loop) are sorted first by the normalized title, and then by the database. In the for loop, ScientoPy processes each document of the documents list. Here, the document DOI, the normalized title, and the normalized first author last name are extracted from the actual processed document by the for loop ($D_{doi[i]}$, $D_{t[i]}$, and $D_{fa[i]}$), and for the document of the next for loop iteration ($D_{doi[i+1]}$, $D_{t[i+1]}$ and $D_{fa[i+1]}$). Because, the documents are sorted by title if there is a duplicated document, the DIOs match, or the normalized title and the normalized first author last name match between two consecutive documents. If two documents match in the for loop iteration, the document of the next iteration is removed from the document list. For the case, if there were two documents from different databases, the Scopus document is removed, because, they were sorted secondly by the database, with WoS document first.

Algorithm 1 Duplicate removal algorithm

```
 1: procedure RemoveDuplicates(documentList)
 2:     Sort documentList by database, first WoS than Scopus
 3:     Sort documentList by normalized title
 4:     for each item $D$ in documentList do
 5:         $D_{dio[i]} \leftarrow D$ document DIO
 6:         $D_{t[i]} \leftarrow D$ title normalized
 7:         $D_{fa[i]} \leftarrow D$ first author last name normalized
 8:
 9:         $D_{dio[i+1]} \leftarrow$ Next $D$ document DIO
10:         $D_{t[i+1]} \leftarrow$ Next $D$ title normalized
11:         $D_{fa[i+1]} \leftarrow$ Next $D$ first author last name normalized
12:
13:         if $D_{doi[i]} == D_{doi[i+1]}$ OR ($D_{t[i]} == D_{t[i+1]}$ AND $D_{fa[i]} == D_{fa[i+1]}$) then
14:             Get the average times cited between $D_{[i]}$ and $D_{[i+1]}$
15:             Set the average times cited to $D_{[i]}$
16:             Remove $D_{[i+i]}$ from documentList
17:     return documentList
```

Times cited

Scopus and WoS databases report their *time cited* count or cited by number for each document. When a document is duplicated, most of the time, the *times cited* field does not match. For these cases, ScientoPy gets the average **times cited** between the two duplicated documents and sets it in the document that is going to keep (see Algorithm 1, lines 14 and 15). Using this *times cited* field, ScientoPy calculates the h-index of each topic for the different categories, such as authors, countries, institutions, and others.

Document's country

To get the document's countries, ScientoPy extracts it from all author's affiliations (last section after the comma of this field). Thus, each document could have one country or many countries associated with it. Nevertheless, if two or more authors have the same country on the affiliation, ScientoPy only associates that country once to the document, to avoid countries duplication per document, and in that way, a document can not add more than once the same country to ScientoPy analysis calculations. For normalization, ScientoPy removes the dot or dots in the country field. Furthermore, some authors use different naming to refer to the same country (such as USA and United States). For that reason, some country names were replaced based on Table 4.

For the match criterion *Equal to country*, ScientoPy replaces the original country field only if it is equal to the comparison string. On the other hand, for the match criterion *In country*, the original country is replaced if the comparison string is in the original country field like an asterisk surrounds the comparison string in a wildcard based search. For instance, "TX 77843 USA" is replaced by United States, and "Peoples R China" or "People's Republic of China" are replaced to China.

Table 4 Document's countries names replacing the table

Comparison string	Criterion	Replaced to
"Bosnia and Herceg"	Equal to country	Bosnia and Herzegovina
"China"	In country	China
"England" OR "Scotland" OR "Wales"	In country	United Kingdom
"Kingdom of Saudi Arabia"	Equal to country	Saudi Arabia
"Russia"	In country	Russian Federation
"Trinid & Tobago"	Equal to country	Trinidad and Tobago
"U Arab Emirates"	Equal to country	United Arab Emirates
"UK"	Equal to country	United Kingdom
"USA"	In country	United States
"Viet Nam"	Equal to country	Vietnam

If the original country name meets the criteria, it is replaced. For *equal to country* criterion, the original country field is replaced only if it is equal to the comparison string. For *in country* criterion, the original country is replaced if the comparison string is in the original country field

Table 5 Document's affiliation examples from WoS and Scopus (Kim et al. 2016; Paethong et al. 2016; Ruiz-Rosero et al. 2017; Savaglio and Fortino 2015)

Source	Affiliations
WoS	*Sejong Univ*, Dept Informat & Commun Engn, Seoul 05006, South Korea
Scopus	Department of Information and Communication Engineering, *Sejong University*, Seoul, South Korea
WoS	*Tokyo Univ Agr & Technol*, Tokyo, Japan
Scopus	*Tokyo University of Agriculture and Technology*, Tokyo, Japan
WoS	*Univ Cauca*, Dept Telemat, Popayan 190002, Cauca, Colombia
Scopus	Departamento de Telemática, *Universidad del Cauca*, Popayán, Cauca, Colombia
WoS	*Univ Calabria*, DIMES, I-87036 Arcavacata Di Rende, CS, Italy
Scopus	DIMES, *Università della Calabria*, Via P. Bucci, cubo 41C, Rende (CS), Italy

Institutions italicized to show that Scopus affiliation not always put it on the same position, and in English

Document's institutions

The document's institutions, are extracted only from all WoS documents author's affiliation fields (first section before comma of this field), because two problems where found in Scopus documents affiliations related to institutions: this field is not always in the same position, and sometimes it is written in the author country's language in Scopus dataset (see Table 5). As country, each document could have one institution or many institutions associated with it. Nevertheless, if two or more authors have the same institution on the affiliation, ScientoPy only associates that institution once in the document, to avoid institutions duplication per document, and in that way, an institution can not add more than once the same institution to ScientoPy analysis calculations.

Table 6 ScientoPy criteria description

Criterion	Description
author	Authors last name and first name initial
sourceTitle	Journal name
subject	Research area (only from WoS documents)
abstract	Document's abstract
authorKeywords	Author's keywords
indexKeywords	Keywords generated by the index. From WoS {Keyword Plus}, and from Scopus {Indexed keywords}
bothKeywords	AuthorKeywords and indexKeywords are used for this search
documentType	Type of document
dataBase	Database where the document was extracted (WoS or Scopus)
country	Country extracted from authors affiliations
institution	Institution extracted from authors affiliations (only from WoS documents)

Data analysis

ScientoPy can perform different types of data analysis, including top topics finding and evolution, specific topic evolution, wildcard search, and trending topics. We describe these analysis types below.

Top and specific topics

One of the main ScientoPy capabilities is to extract the top topics of a selected criterion. Top topics are the ones that have more documents count in the processed dataset. Table 6 shows the criterion options available on ScientoPy. By default, the top topic analysis extracts the 10 top topics on the default criterion (author keywords), and graphs the documents count per topic in horizontal bars. Similarly, we can perform this analysis with specific topics inside a selected criterion. For example, we can compare the documents growth of two specific countries, or the evolution of two different technologies by the author keywords.

Moreover, some documents criterion have multiple items. Like a document with multiple authors has multiple author's names and multiple author's affiliations. When we use ScientoPy to extract the top topics inside a criterion, it uses by default the multiple document's fields to calculate the topic count. For example, to extract the top authors, ScientoPy uses all authors of each paper to extract the total top author's list. Similarly, it uses all document's fields inside the topics that could have more than one item, such as authors names, authorKeywords, indexKeywords, bothKeywords, countries, and institutions. Nevertheless, if we want that ScientoPy only uses the first item in the selected criterion to extract the top topics, we can do this by a command option described in the user manual (–onlyFirst command option). In this case, for instance, ScientoPy uses only the first author to extract the top author's list of the dataset.

Wildcard search

Wildcards are very useful to find topics that come in plural and singular, such as network and networks. Also, they are effective to find topics inside some defined categories that starts or ends with certain words or phrases, such as:

- ** latency* operational latency, mechanical latency, WAN latency.
- *blood* * blood pressure, blood glucose, blood platelets.

For that reason, in the topic analysis, we can use the asterisk (*) wildcard to find, for example, the distribution of the keywords that starts or ends with a particular word or phrase. Also, with this characteristic, we can use ScientoPy to find the documents that in the abstract contain that particular word or phrase.

Topics growth indicators

ScientoPy uses three different topic growth indicators to find trending topics and its relative/absolute growth.

Average growth rate (AGR)

ScientoPy finds the top trending topics based on the higher average growth rate (AGR). The AGR is the average difference between the number of documents published in one year with the number of documents published in the previous year. It indicates how the number of documents published for a topic has growth (positive number) or decline (negative number) on average inside a time frame. This AGR is calculated using the Eq. 1:

$$\mathrm{AGR} = \frac{\sum_{i=Y_\mathrm{s}}^{Y_\mathrm{e}} P_i - P_{i-1}}{(Y_\mathrm{e} - Y_\mathrm{s}) + 1} \tag{1}$$

where AGR = average growth rate; Y_e = end year; Y_s = start year; P_i = number of publications on year i.

For the end year Y_e, ScientoPy uses the default global end year configured in the global options or/in ScientoPy command parameters. The start year Y_s is calculated from the end year Y_e, as indicated in the Eq. 2

$$Y_\mathrm{s} = Y_\mathrm{e} - (\mathrm{WindowWidth} + 1) \tag{2}$$

The default WindowWidth is 2 years. Thus, if the end year is 2018, the AGR is the average growth rate between 2017 and 2018.

Average documents per year (ADY)

The average documents per year (ADY) is an absolute indicator that represents the average number of documents published inside a time frame for a specific topic. The ADY is calculated using the Equation 3:

$$\mathrm{ADY} = \frac{\sum_{i=Y_\mathrm{s}}^{Y_\mathrm{e}} P_i}{(Y_\mathrm{e} - Y_\mathrm{s}) + 1} \tag{3}$$

where ADY = average documents per year; Y_e = end year; Y_s = start year, calculated as described in Eq. 2; P_i = number of publications on year *i*.

Percentage of documents in last years (PDLY)

Percentage of documents in last years (PDLY) is a relative indicator that represents the percentage of the ADY relative to the total number of documents for a specific topic. In this way, the PDLY is calculated using the Eq. 4:

$$\mathrm{PDLY} = \frac{\sum_{i=Y_\mathrm{s}}^{Y_\mathrm{e}} P_i}{(Y_\mathrm{e} - Y_\mathrm{s} + 1) * \mathrm{TND}} * 100\% \tag{4}$$

where PDLY = percentage of documents in last years; Y_e = end year; Y_s = start year, calculated as described in Eq. 2; P_i = number of publications on year *i*; TND = total number of documents.

User graphic interface (ScientoPyUI)

We developed a simple graphic user interface (GUI) for the use of ScientoPy called ScientoPyUI (available in https://github.com/jpruiz84/ScientoPyUI). This interface allows us to select the preprocess folder and perform the different kind of temporal analysis (see Fig. 2).

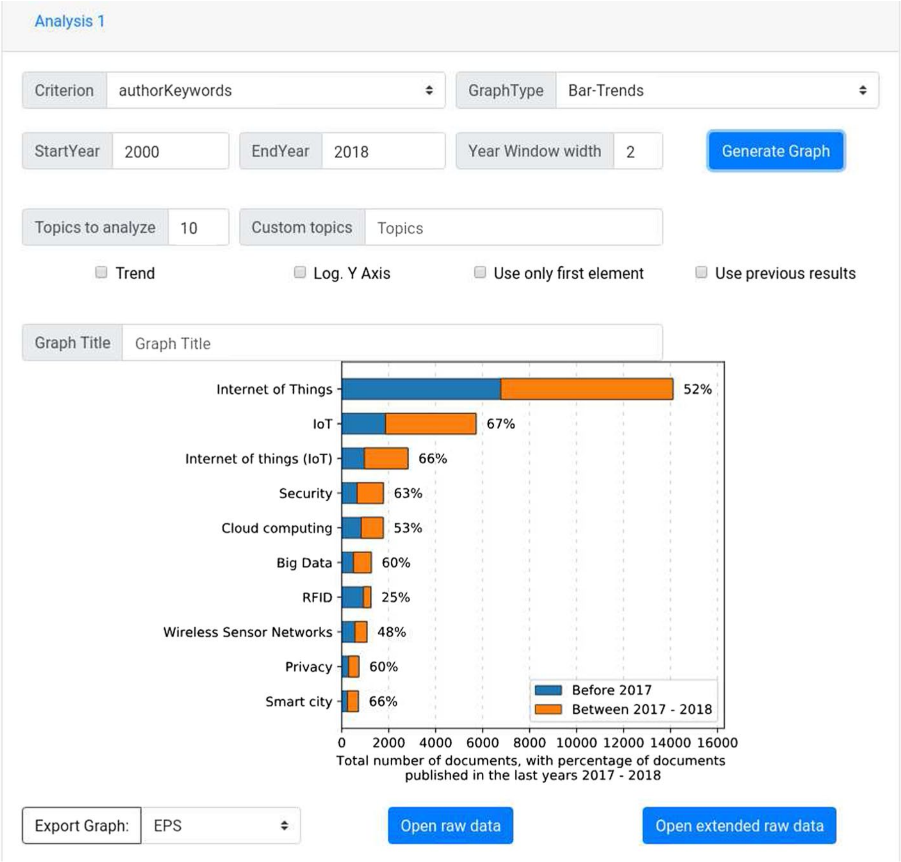
The main features and capabilities of this GUI are:

- Select the input dataset folder
- Perform the dataset preprocessing with and without the remove duplicate filter
- Select the analysis criterion and graph type
- Define the year range analysis time frame, and the year window width
- Set the topic list length (topics to analyze)
- Set the custom topics to analyze
- Define customs analyzes options (trend analysis, *Y* axis in log scale, only first element to analyze, and use previous results)
- Set the graph title
- Export the graph in three different formats (EPS, SVG, and PNG)
- Open the raw and extended raw output data

You can find more information about the different capabilities and the installation instructions of ScientoPyUI by the user manual available in the following link: manual.

Case study: Internet of things

We selected Internet of things (an updated dataset from the scientometric review (Ruiz-Rosero et al. 2017) available to download from this link) as a case study to show ScientoPy capabilities. The dataset was built using two bibliographic databases: Clarivate Web of Science (WoS), and Scopus. The search string for this analysis was “Internet of Things”. We used this string as the topic search in WoS and Scopus, which includes



Analysis 1

Criterion: authorKeywords

GraphType: Bar-Trends

StartYear: 2000

EndYear: 2018

Year Window width: 2

Topics to analyze: 10

Custom topics: Topics

□ Trend □ Log. Y Axis □ Use only first element □ Use previous results

Graph Title: Graph Title

Export Graph: EPS

Fig. 2 ScientoPyUI, graphical user interface analysis window

title, abstract, author's keywords, and KeyWords Plus® (for WoS). With this search criteria, we downloaded the dataset within a day on May 27th, 2019. Table 7 shows the preprocess brief table information generated by ScientoPy preprocess script. This table describes the dataset information including in the second column (*Number*) the number of publications after and before the duplication removal filter per database, and the third column (*Percentage*) the relative percentages (see table description for detailed information about these percentages). *Loaded papers* represents the total documents loaded. *Omitted papers by document type* are the documents outside the default documents type filter (see "Document type filtering" section). *Total papers after omitted papers removed* are the documents inside the default documents type filter. *Loaded papers from WoS*, and *Loaded papers from Scopus* are the number of documents from each database after the omitted papers removed. *Duplicated papers found* are the duplicated documents total number found and removed. *Removed duplicated papers from WoS*, and *Removed duplicated papers from Scopus* are the number of documents removed from each database after the duplication removal filter. *Duplicated documents with different cited by* are the duplicated documents found with a different "cited by" or time cited number between the document removed and the documents kept. *Total papers after rem.*

Table 7 Internet of things preprocess brief table

Information	Number	Percentage
Loaded papers	80,818	
Omitted papers by document type	3571	4.4
Total papers after omitted papers removed	77,240	
Loaded papers from WoS	28,323	36.7
Loaded papers from Scopus	48,921	63.3
Duplicated removal results		
Duplicated papers found	25,046	32.4
Removed duplicated papers from WoS	236	0.8
Removed duplicated papers from Scopus	24,808	50.7
Duplicated documents with different cited by	14,420	57.6
Total papers after rem. dupl.	52,194	
Papers from WoS	28,087	53.8
Papers from Scopus	24,113	46.2

Omitted papers by document type percentage relative to *Total loaded papers*. *Papers from WoS*, *Papers from Scopus*, and *Total duplicated papers found* percentages relative to *Total loaded papers*. *Removed duplicated papers from WoS* percentage relative to *Papers from WoS*. *Removed duplicated papers from Scopus* percentage relative to *Papers from Scopus*. *Duplicated documents with different cited by* percentage relative to *Total duplicated papers found*. *Papers from WoS/Scopus* percentage relative to *Total papers after rem. dupl.*

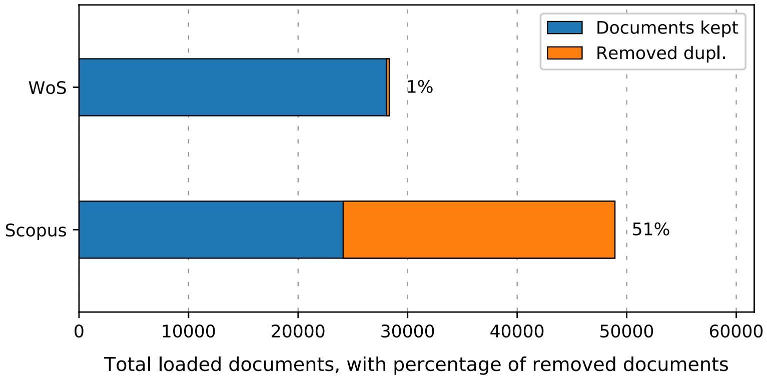


Fig. 3 Internet of things preprocess brief graph, with total loaded documents from Clarivate Web of Science (WoS), and Scopus databases, and the percentage of removed documents after duplication removal filter

dupl. are the output documents number after duplication removal from the preprocessing script. Finally, *Papers from WoS/Scopus* are entire documents from WoS and Scopus, respectively after the duplication removal filter.

Figure 3 shows the preprocess brief graph that presents the loaded documents for each database and the removed duplicated documents, respectively. As since ScientoPy preprocess script keeps WoS documents over Scopus documents, after duplication removal filter we see more documents from WoS than Scopus database.

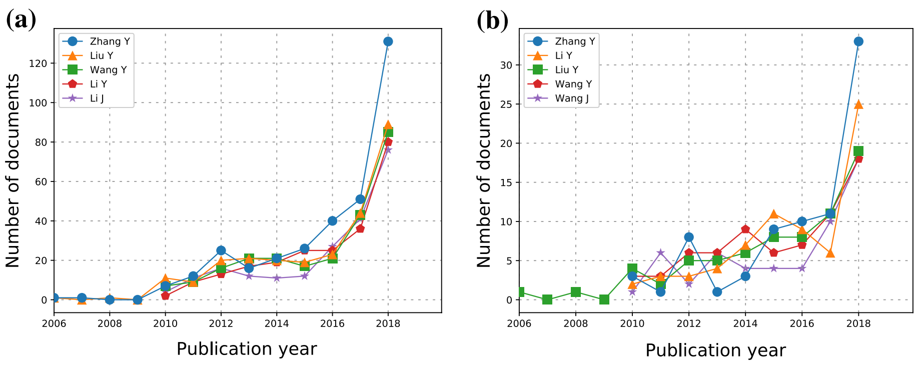


Fig. 4 Internet of things top 5 authors in documents per year. **a** based on all document's authors; **b** based on first document's author

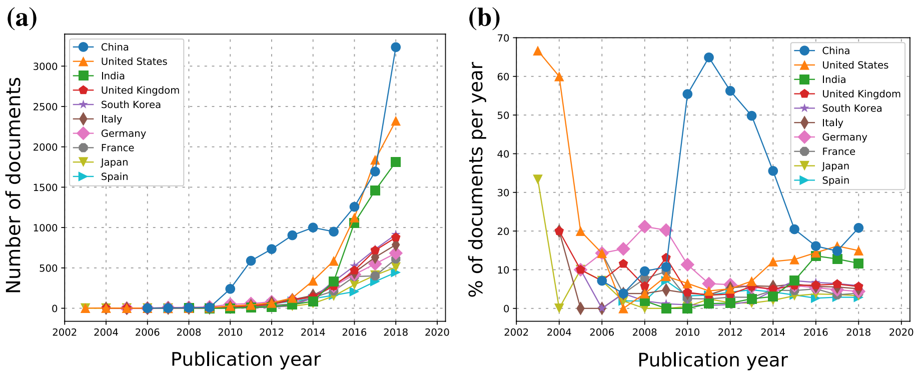


Fig. 5 Internet of things top 10 countries in documents per year. **a** number of documents per year on *Y*-axis; **b** percentage of documents per year on *Y*-axis

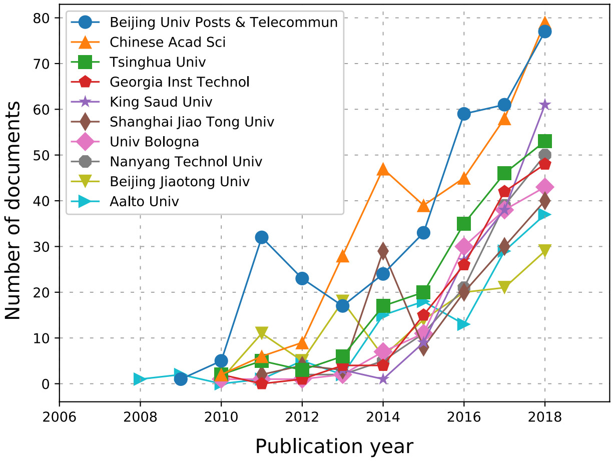
Top authors analysis

The author criterion in ScientoPy let us analyze the top authors. By default, ScientoPy takes all authors from each paper to extract the top author's list (see Fig. 4a). Nevertheless, an analysis based on the first document's author can be done (see Fig. 4b). Figure 4 shows the top 5 author's number of documents versus the publication year for both described cases. This graph lets us understand when a top author has started to publish in the field that we are studying. Also, we can analyze here the publication's growth of each one.

Top countries analysis

As described in "Document's country" section, ScientoPy can extract the document country based on the author's affiliations. Figure 5a shows the top 10 document's countries, in a graph that describes the number of documents per country versus the publication year. This

Fig. 6 Internet of things top 10 institutions in documents per year



graph is useful to show which country is leading the publications count, especially inside the 2010–2018 time frame. However, it is not effective to describe the leading countries inside the 2002–2008 time frame for this case of study. Hence, ScientoPy can change the graph's *Y*-axis from the number of documents to the percentage of documents per year (see Fig. 5b). Using this graph, we can analyze, in any period, the leading country and its participation percentage for each year.

Institutions time analysis

Another ScientoPy characteristic is to extract the document's institutions from the author's affiliation, as described in "Document's institutions" section. Here, Fig. 6 shows the top 10 document's institutions with a graph that describes the number of documents per institution versus the publication year.

Institutions time analysis from a specific country

Additionally, with ScientoPy, we can extract the top institutions from a specific country. Figure 7 shows the China, United States, Spain, and Colombia top 20 institutions with the total number of documents related to the Internet of things.

Top author keywords

To know the top research fields and them growth inside a research area, ScientoPy let us extract these fields based on the top author keywords topics. Figure 8a shows the 10 top author keywords bar trends graph related to the Internet of things. Nevertheless, in this graph, the *Internet of Things*, *IoT*, and the *Internet of Things (IoT)* keywords overshadow the other topics because they are the main keywords for this research field. Here, ScientoPy can filter the first elements in the topic list to generate a new graph (see Fig. 8b).

Figure 8b also shows the percentage of documents published in the last years (2017–2018) as a relative growth indicator. With this indicator, we notice that despite *fog*

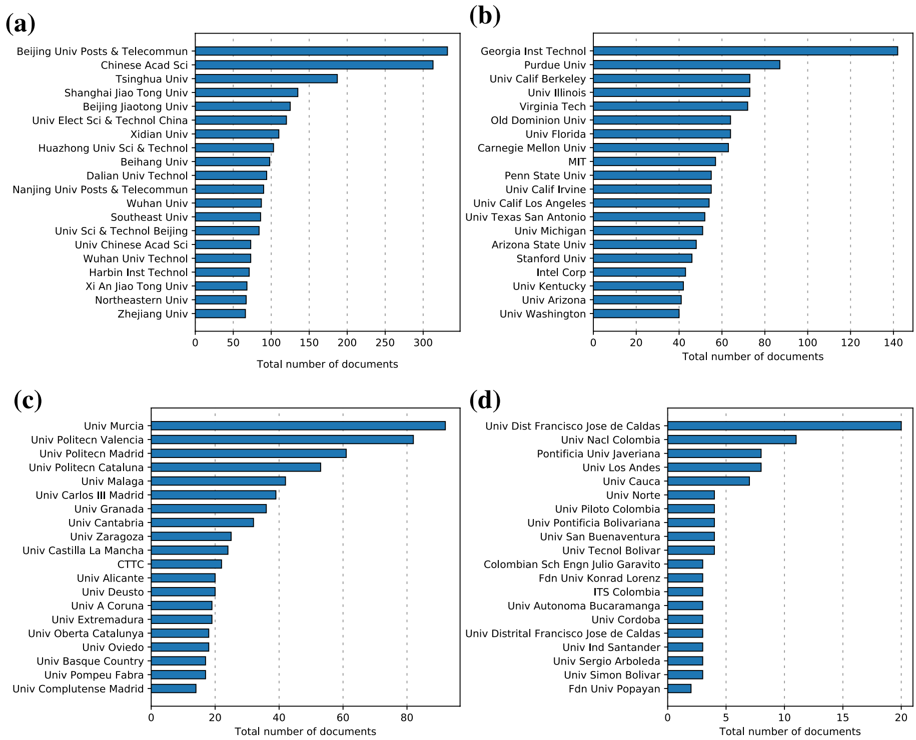


Fig. 7 Internet of things top 20 institutions per country, **a** China; **b** United States; **c** Spain; **d** Colombia

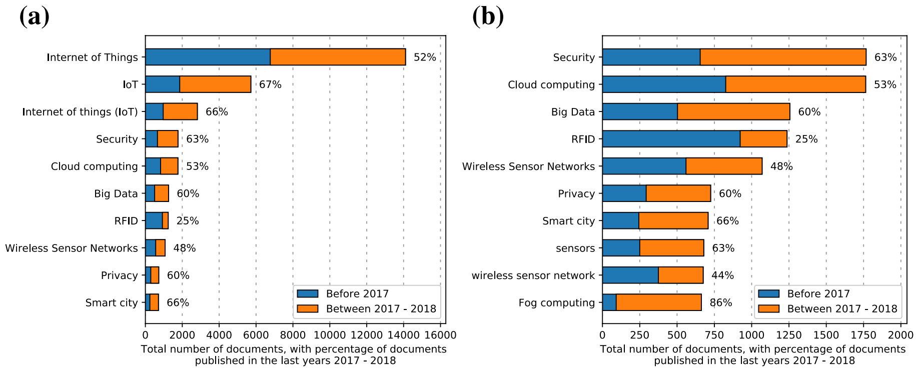


Fig. 8 Internet of things top 10 author's keywords bar trends graph. **a** including all top 10 author's keywords; **b** filtering the first top three author's keywords

computing is the 10th topic in this list, it has the highest PDLY, which indicates that this is the topic that has grown more in this list during the last 2 years.

Another useful way to show the top topics based on author keywords is using a word cloud visual representation. Figure 9 shows the top 1000 author keywords represented in a word cloud, where the size of the word or phrase is proportional to the number of

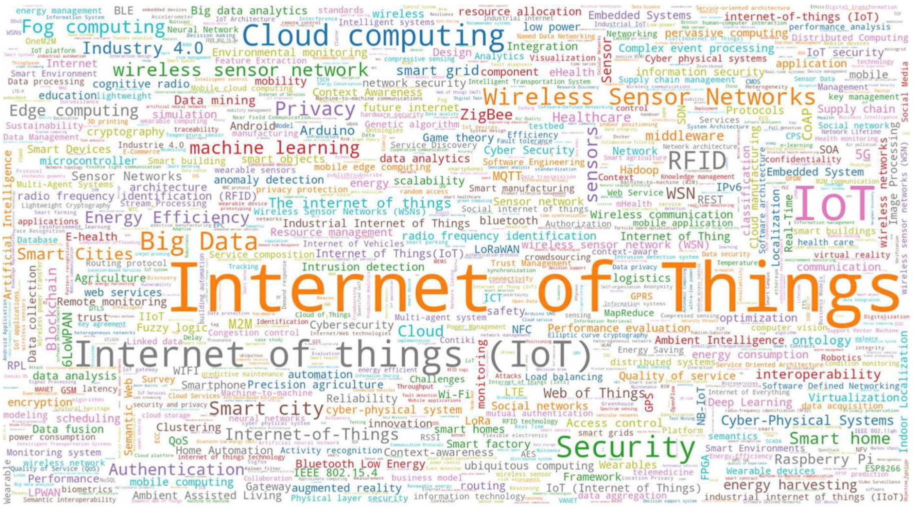


Fig. 9 Internet of things top 1000 author keywords word cloud. Word or phrase size is proportional to the number of documents

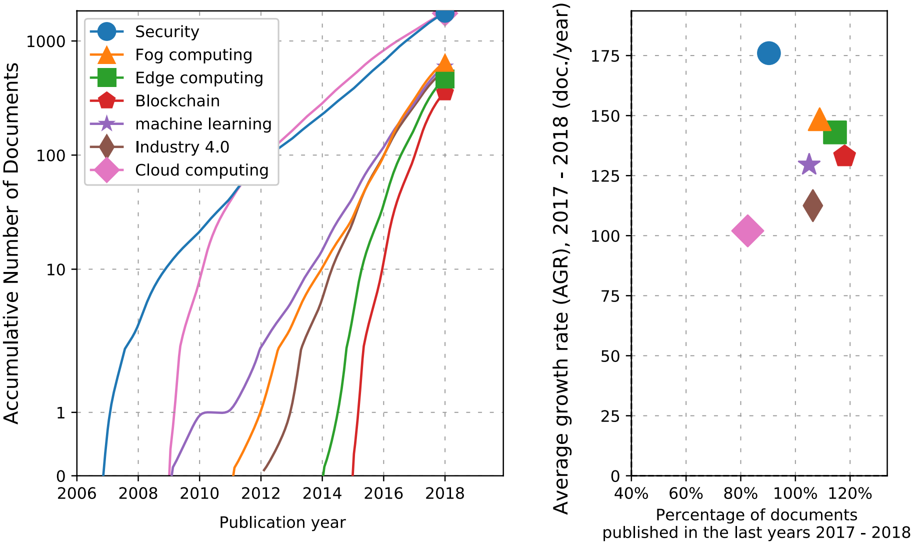


Fig. 10 Internet of things top 10 trending topics based on author's keywords evolution plot

documents related to it. This representation is useful for using in slides, to show how big it could be a research area.

Trending topics

ScientoPy can find trending topics by looking at the top author keywords based on the highest AGR (described in "Topics growth indicators" section). Figure 10 shows the

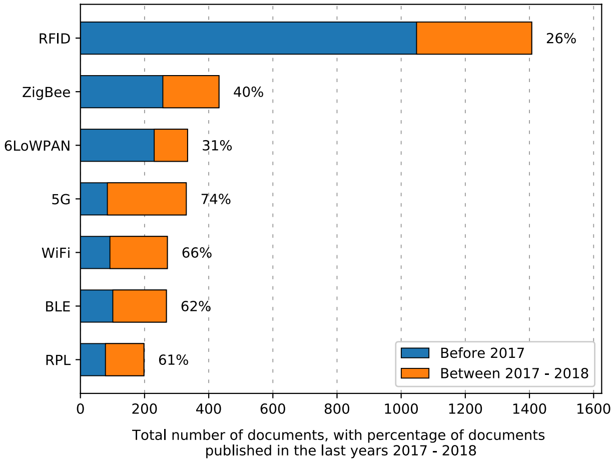


Fig. 11 Internet of things main media layer communication protocols bar trends plot

Internet of things top trending topics evolution plot. This evolution plot shows on the left side the accumulative number of documents (in logarithmic scale) versus the publication year. In that way, the beginning of the line in the *X*-axis represents the year when the topic research started, and the final of each line in the *Y*-axis represents the total number of documents published of each topic. On the right side, the *Y*-axis represents the AGR of each topic for the 2017–2018 period, and the *X*-axis represents PDLY. With this graph, we can analyze which are the topics with the higher AGR, and the higher PDLY. In this way, we see that the trending topic with the highest absolute growth is *security*, and the trending topic with the highest relative growth is *blockchain*.

Specific topic analysis

Alongside top topic analysis, ScientoPy can analyze specific topics inside any of the predefined criterion. For example, from author's keywords, we can define and extract the main media layer communications protocols that have been used in the Internet of things, to show the results in a bar trends graph (see Fig. 11). From this plot, we find that *RFID* is the most popular media layer protocol for IoT research documents. Also, we find here that *5G* is the topic with the highest relative growth, with 74% of its documents published in the last two years.

Furthermore, we can use specific topic analysis with other criterion, such as institutions. Figure 12 shows the top 3 Colombian institutions contrasted with our local Colombian institution Universidad del Cauca (Univ Cauca). Here, we found that Univ Cauca is the one with fewer publications on the Internet of Things, but one of the two with the higher PDLY.

Moreover, in specific topics, we can use the asterisk wildcard to find all the author keywords which start or end with a particular word or phrase. Figure 13 shows the most used author keywords that start with "Ubiquitous" (Fig. 13a), and the ones that start with "Smart" (Fig. 13b). These results show the top topics inside the category ubiquitous and the category smart applications for the Internet of things.

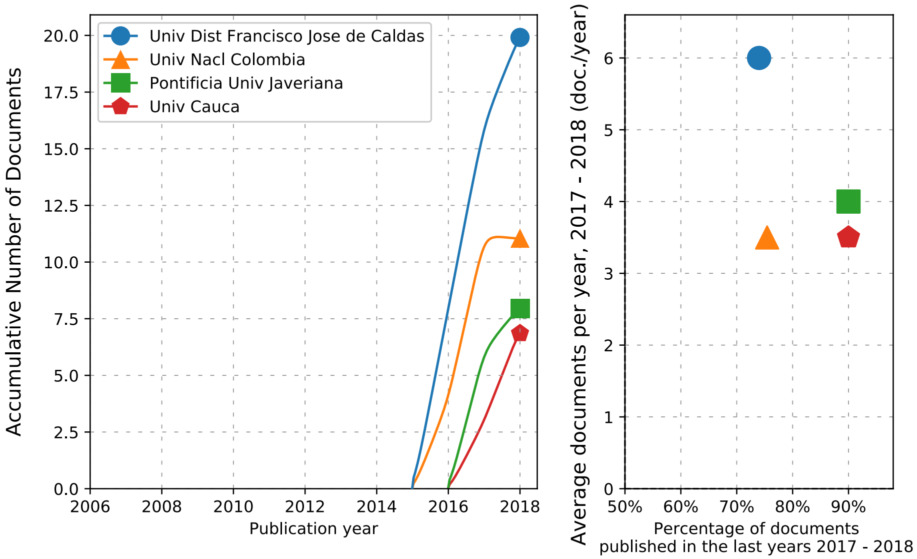


Fig. 12 Internet of things top 3 Colombian institutions contrasted with the our local Colombian institution Universidad del Cauca (Univ Cauca) evolution plot

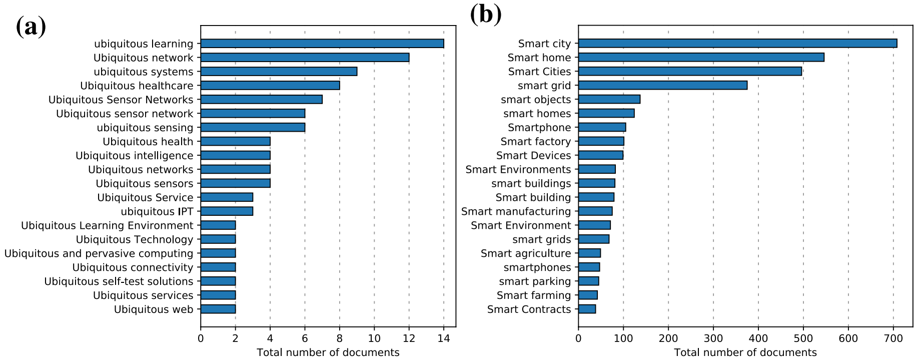


Fig. 13 Internet of things most used author keywords that starts with: **a** ubiquitous; **b** smart

Discussion

In this section, we discuss 8 different tools (including ScientoPy) designed to make a scientometric quantitative analysis which helps scholars to track innovation progress and to find trending topics. These tools were tested and analyzed from July 1st, 2018 to December 17th, 2018, then, any updates of these tools after the last date are not included. Table 8 provides an overview of these tools characteristics. All the tools mentioned here support the datasets from the two main bibliographic databases (WoS and Scopus), nevertheless, some tools like CiteSpace, Bibliometrix, Networkbench, and Sci2 support more than these two main databases. Notwithstanding, not all tools that support more than one database can merge the information successfully from two or more sources because the source databases

Table 8 Scientometrics tools

Tool	Supported databases	Duplication removal	Search with wildcard	Trend analysis	Output	User interface	References
ScientoPy	WoS and Scopus	Yes	Yes	AGR, evolution plot	Graphical images, and TSV files	Command line/windowed JavaScript application	
CiteSpace	WoS, Scopus, CSCD, CSSCI, CNKI, or PubMed	Yes	No	Burst-detection	GraphML, Pajek, HTML, CSV report	Windowed Java application	Chen (2006)
SciMAT	WoS or Scopus	Yes	No	Strategic diagram	Graphical images, and HTML or Latex report.	Windowed Java application	Cobo et al. (2012)
Bibexcel	WoS or Scopus	Yes	No	No	Tabbed data records	Windowed application	Persson et al. (2009)
Bibliometrix	(WoS and Scopus) or Cochrane Library or PubMed	Yes	No	KeywordGrowth, sourceGrowth, thematicMap, thematicEvolution, and histNetwork‘	R data frames, and graphical images	biblioshiny web-interface	Aria and Cuccurullo (2017)
Network Workbench	Scholarly Database (SDB), Bibtex, WoS, Scopus or Google Scholar	Yes		Bursting words	Graphical images, GraphML, XGMML and Pajek files	Windowed application	Boerner et al. (2010)
BiblioTools	WoS or Scopus	No	No	No	Web based interface and Latex report.	Command line/web based	Grauwin and Jensen (2011)
Sci2	WoS, Scopus, Endnote, or BibteX	Yes	No	Burst detection	GraphML, NWB, Pajek, XGMML, and CSV	Windowed application	Lewis and Alpi (2017)

do not always use the same fields labeling or fields representation (like the author's names represented different in Scopus and WoS). From these tools, only ScientoPy and Bibliometrix perform the preprocess and merge steps that allow working with the combined information of the databases that they support.

The preprocessing is one of the most critical steps in scientometrics analysis because the goodness of the result will depend on the quality of the data (Cobo et al. 2011b). Duplication removal is one of the main preprocessing steps which allows the quality of the data, primarily if we work with datasets from two different databases. In this analysis, all tools have an implementation for duplication removal except BiblioTools.

Besides, when we use a bibliometric database, we can do search wildcards, like the asterisk (*), in the database's search engine. These wildcards allow us to find not only plural and singular equivalence, even more, help us to find concepts, like "smart*" (smart homes, smart cities, among others). ScientoPy is the only tool in this list that allows us to find topics using the asterisk wildcard. This capability lets us find a singular, plurals, and word concepts inside the downloaded dataset.

Inside the temporal analysis, trend analysis identifies trending topics and their evolution over time. ScientoPy calculates the AGR, finds the topics with the top ones, makes an evolution plot to show the evolution over the time of each trending topic, and compare them according to the absolute growth with the AGR, and the relative growth with the PDLY. CiteSpace uses the burst-detection Kleinberg's algorithm for detecting sharp increases of interest in a specialty. It finds burst terms extracted from titles, abstracts, descriptors, and identifiers of bibliographic records (Chen 2006). Others tools, like SciMAT, uses the strategic diagram to plot the topics in a Cartesian plane according to their centrality and density over different periods (Cobo et al. 2011a, 2012). As well, Bibliometrix uses the functions KeywordGrotwh and sourceGrowth to calculate yearly published documents for top keywords and source respectively, the thematicMap and thematicEvolution to create a thematic map and evolution analysis based on co-word network analysis and clustering, and the histNetwork to create a historical citation network from a bibliographic data frame (Aria and Cuccurullo 2017, 2018). Finally, Network Workbench and Sci2 use the Kleinberg's burst detection algorithm (Kleinberg 2003) to identify sudden increases in the usage frequency of words over time (Boerner et al. 2010).

The output data could be represented differently by the tools here discussed. The tools specialized in network analysis (CiteSpace, Network Workbench, and Sci2) export the result data into GraphML files. SciMAT and BiblioTools generate Latex reports with graphical and numerical information. Meanwhile, Bibexcel exports tabbed data records to be imported directly to Excel or any spreadsheet tool. Bibliometrix generates R data frames structures with the data obtained after the preprocess or analysis routines. Also, it can generate graphical images. ScientoPy can generate different kind of graphical images that summarize the preprocess or analysis results. Also, it generates CSV (Comma Separated Values) files that we can easily import into any spreadsheet tool.

The user interface allows us to interact with the different characteristics of each tool. In this field, CiteSpace, SciMAT, Bibexcel Network Workbench, and Sci2 offer a windowed user interface that allows to load the dataset, run the preprocessing steps, and execute the data analysis. In this classification, SciMAT highlights with a wizard menu to perform the analysis steps. In BiblioTools, the user runs some command lines scripts to execute the preprocessing and data analysis. Then the user can navigate through the result plots using a web-based application.

Similarly, Bibliometrix from the version 2.0.0 includes a web-interface, namely biblioshiny, which implements the main features of Bibliometrix including dataset loading, filtering, descriptive analysis (longitudinal analysis, keywords and sources trending), conceptual, intellectual and social structure analysis. Alternatively, for ScientoPy, the users can perform all operations through the command line scripts (preprocessing, and data analysis), or from the ScientoPyUI graphical interface. The command line option allows the user to generate a single batch script that performs all the operations needed for a particular analysis.

Conclusions

This paper describes a new scientometric open-source tool called ScientoPy. This tool supports datasets from WoS and Scopus. It can load the datasets simultaneously from both mentioned databases, and merge them successfully thanks to its field tags correlation table. Also, in the preprocess, it performs an author's name normalization to have consistency information between the dataset imported from the two databases. The ScientoPy's duplication removal filter finds documents with the same title and same first author last name. Then it removes the duplicated instances found, keeping WoS documents over Scopus ones. Besides, the preprocess extracts the highest times cited field from the duplicated documents, to be used later for the h-index calculation. Besides, in this step, the country and institution are extracted from all document author's affiliation. Finally, the preprocess summarizes the results with a prepossessing brief graph and table, getting the dataset ready for the data analysis phase.

ScientoPy can perform data analysis by extracting the top topics of a selected criterion, such as top authors, top countries, top institutions, top author's keywords, and others. Besides, it can plot the data analysis results in four different ways: timeline graph, bar graph, evolution graph, and word cloud. Additionally, we can perform in ScientoPy another type of data analysis such as trending topics, specific topic search, and wildcard search. With these capabilities, we can find several results from a bibliographic database, like first top author's evolution, countries or institutions evolution and participation over the time, institutions participation of a selected country, and specific topic or trending topics evolution based on author's keywords.

The trending topics analysis finds the topics with the highest growth rate during the last years. Also, this analysis shows the longitudinal evolution of these topics by a evolution plot that shows the accumulative number of documents versus the publication year to show when the topic has started, how it has evolved over the time, and what is the total number of documents published for this one. Also, this graph shows the AGR versus the PDLY to know which topic has higher relative and absolute growth.

Wildcard search allows us to find the top topics inside a specific research category. For the Internet of things case study, this capability let us find the top topics inside the category ubiquitous and inside the category smart applications, by searching the top topics that start with "Ubiquitous" and the ones that start with "Smart" respectively.

The main difference of ScientoPy with respect to the other tools are: (a) the capability to support seamlessly the both databases (WoS and Scopus); (b) the ability to find topics using search wildcards; (c) trend analysis based on AGR; (d) command line user interface that allows us to generate a single batch script that performs all the operations needed for our analysis, with only a single order or command execution.

Finally, ScientoPy were tested and validated for Ph.D. and master students of the University of Cauca (Universidad del Cauca), using a different kind of datasets. Also, we improved many characteristics of this tool according to these students suggestions.

Acknowledgements This research is funded by Colciencias Doctoral scholarship, from the Departamento Administrativo de Ciencia, Tecnología e Innovación (647-2014) for the Ph.D. in Telematic Engineering at the Universidad del Cauca, Popayán, Colombia. Also, this work was supported by the Universidad del Cauca (501100005682).

References

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975.

Aria, M., & Cuccurullo, C. (2018) bibliometrix v 2.0.2, reference manual. Accessed December 17, 2018.

Bailón-Moreno, R., Jurado-Alameda, E., & Ruiz-Baños, R. (2006). The scientific network of surfactants: Structural analysis. *Journal of the American Society for Information Science and Technology*, *57*(7), 949–960.

Boerner, K., Huang, W., Linnemeier, M., Duhon, R. J., Phillips, P., Ma, N., et al. (2010). Rete-netzwerk-red: Analyzing and visualizing scholarly networks using the Network Workbench Tool. *Scientometrics*, *83*(3), 863–876.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377.

Ciftler, B. S., Kadri, A., & Guevenc, I. (2017). IoT localization for bistatic passive UHF RFID systems with 3-D radiation pattern. *IEEE Internet of Things Journal*, *4*(4, SI), 905–916.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011a). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. *Journal of Informetrics*, *5*(1), 146–166.

Cobo, M. J., Lopez-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011b). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, *62*(7), 1382–1402.

Cobo, M. J., Lopez-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, *63*(8), 1609–1630.

Gezer, C., & Taskin, E. (2016). An Overview of oneM2M standard. In *2016 24th signal processing and communication application conference (SIU)* (pp. 1705–1708). IEEE; Bulent Ecevit University, Department of Electrical and Electronic Engineering; Bulent Ecevit University, Department of Biomedical Engineering; Bulent Ecevit University, Department of Computer Engineering, Zonguldak, Turkey, May 16–19, 2016.

Grauwin, S., & Jensen, P. (2011). Mapping scientific institutions. *Scientometrics*, *89*(3), 943–954.

Harzing, A.-W. (2014). A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics*, *98*(1), 565–575.

Kim, J., Lim, H., Han, S., Jung, Y., & Lee, S. (2016). Compensation algorithm for misrecognition caused by hard pressure touch in plastic cover capacitive touch screen panels. *Journal of Display Technology*, *12*(12), 1623–1628.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, *7*(4), 373–397.

Lewis, D. M., & Alpi, K. M. (2017). Bibliometric network analysis and visualization for serials librarians: An introduction to Sci2. *Serials Review*, *43*(3–4, SI), 239–245.

Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, *246*(1), 1–19.

Moulin, T., & Simon, P. (2016). e-Health—The internet of things and telemedicine. *Correspondances en Metabolismes Hormones Diabetes et Nutrition*, *20*(3), 58–64.

Munoz-Organero, M., Ramirez, G. A., Munoz-Merino, P. J., & Kloos, C. D. (2011). Framework for contextualized learning ecosystems. In C. D. Kloos, D. Gillet, R. M. G. Garcia, F. Wild, & M. Wolpers (Eds.), *Towards ubiquitous learning, EC-TEL 2011, volume 6964 of Lecture Notes in Computer Science. 6th European conference on technology-enhanced learning (EC-TEL)*, Palermo, Italy, September 20–23, 2011.

Paethong, P., Sato, M., & Namiki, M. (2016). Low-power distributed NoSQL database for IoT middleware. In J. L. Mitrpanont (Ed.), *2016 Fifth ICT international student project conference (ICT-ISPC)* (pp. 158–161). ICT; Mahidol University, Faculty of Information and Communication Technology; TAT; Universiti Teknologi Malaysia. *5th ICT international student project conference (ICT-ISPC)*, Nakhon Pathom, Thailand, May 27–28, 2016.

Persson, O., Danell, R., & Schneider, J. W. (2009). How to use Bibexcel for various types of bibliometric analysis. In F. Åström, R. Danell, B. Larsen, & J. W. Schneider (Eds.), *Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th Birthday* (Vol. 5, pp. 9–24). Berlin: International Society for Scientometrics and Informetrics.

Ruiz-Rosero, J., Ramirez-Gonzalez, G., Williams, J. M., Liu, H., Khanna, R., & Pisharody, G. (2017). Internet of things: A scientometric review. *Symmetry-Basel*, *9*(12), 301.

Savaglio, C., & Fortino, G. (2015). Autonomic and cognitive architectures for the Internet of things. In G. DiFatta, G. Fortino, W. Li, M. Pathan, F. Stahl, & A. Guerrieri (Eds.), *Internet and distributed computing systems, IDCS 2015, volume 9258 of Lecture Notes in Computer Science* (pp. 39–47). *8th annual international conference on internet and distributed computing systems (IDCS)*, Windsor, England, September 02–04, 2015.

Small, H. (1997). Update on science mapping: Creating large document spaces. *Scientometrics*, *38*(2), 275–293.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523–538.