BIP4COVID19: Releasing impact measures for articles relevant to COVID-19

Thanasis Vergoulis[1,*], Ilias Kanellos[1], Serafeim Chatzopoulos[1,2], Danae Pla Karidi[1,3], Theodore Dalamagas[1],

1 "Athena" Research Center, Greece
2 Univ. of the Peloponnese, Greece
3 NTU Athens, Greece

* vergoulis@athenarc.gr

Abstract

Since the beginning of the 2019-20 coronavirus pandemic, a large number of relevant articles has been published or become available in preprint servers. These articles, along with earlier related literature, compose a valuable knowledge base affecting contemporary research studies, or even government actions to limit the spread of the disease and treatment decisions taken by physicians. However, the number of such articles is increasing at an intense rate making the exploration of the relevant literature and the identification of useful knowledge in it challenging. In this work, we describe BIP4COVID19, an open dataset compiled to facilitate the coronavirus-related literature exploration, by providing various indicators of scientific impact for the relevant articles. Finally, we provide a publicly accessible Web interface on top of our data, allowing the exploration of the publications based on the computed indicators.

Introduction

COVID-19 is an infectious disease caused by the coronavirus SARS-CoV-2, which may result, for some cases, in progressing viral pneumonia and multi-organ failure. After its first outbreak in Hubei, a province in China, it subsequently spread to other Chinese provinces and many other countries. On March 11th 2020, the World Health Organisation (WHO) declared the 2019–20 coronavirus outbreak a pandemic. Until the end of May more than $4,000,000$ cases had been recorded in more than 200 countries, counting more than $320,000$ fatalities.

At the time of writing, an extensive amount of coronavirus related articles have been published since the virus' outbreak (indicatively, our collected data contain about $14,954$ articles published in 2020). Taking additionally into account previous literature on coronaviruses and related diseases, it is evident that there is a vast literature on the subject. However, it is critical for researchers or other interested parties (e.g., government officers, physicians) to be able to identify high-impact articles. A variety of impact measures have been proposed in the fields of bibliometrics and scientometrics [7,8]. Some of them rely on the analysis of the underlying citation network. Other approaches utilise measures commonly known as "altmetrics", which analyse data from social media and/or usage analytics in online platforms (e.g., in publishers' websites). Both approaches have their benefits and shortcomings, each capturing different aspects of an article's impact. Thus, by considering a wide range of different measures we can better uncover a comprehensive view of each article's impact.

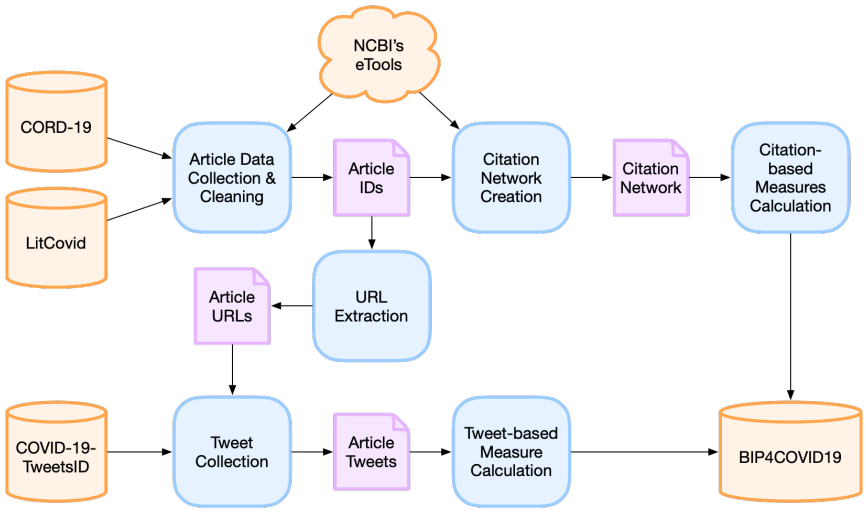

Fig 1. The data update workflow of BIP4COVID19

In this context, the objective of this work is to produce *BIP4COVID19*, an openly available dataset, which contains a variety of different impact measures calculated for COVID-19-related literature. Two citation-based impact measures (PageRank [5] and RAM [6]) were chosen to be calculated, as well as an altmetric indicator (tweet count). The selected measures were chosen so as to cover different impact aspects of the articles. Furthermore, to select a representative set of publications, we rely on two open datasets of COVID-19-related articles: the CORD-19 [3] and the LitCovid [2] datasets. BIP4COVID19 data are updated on a regular basis and are openly available on Zenodo [1].

Materials and methods

BIP4COVID19 is a regularly updated dataset. Data production and update is based on the semi-automatic workflow presented in Figure 1. In the following subsections the major processes involved are elaborated.

Article Data Collection and Cleaning

The list of COVID-19-related articles is created based on two main data sources: the *CORD-19*[1] Open Research Dataset [3], provided by the Allen Institute for AI, and the *LitCovid*[2] collection [2] provided by the NLM/NCBI BioNLP Research Group. CORD-19 offers a full-text corpus of more than 63,000 articles on coronavirus and COVID-19, collected based on articles that contain a set of COVID-19 related keywords from PMC, arXiv, biorXiv, and medRxiv and the further addition of a set of publications on the novel coronavirus, maintained by the WHO. LitCovid, is a curated dataset which currently contains more than 13,000 papers on the novel coronavirus.

The contents of the previous datasets are integrated and cleaned. During this process, the eSummary tool[3] from NCBI's eTool suite is utilised to collect extra metadata for each publication using the corresponding PubMed or PubMed Central identifiers (pmid and pmcid, respectively), where available. The collected metadata are

[1] https://pages.semanticscholar.org/coronavirus-research
[2] https://www.ncbi.nlm.nih.gov/research/coronavirus/
[3] https://www.ncbi.nlm.nih.gov/books/NBK25500/

semi-automatically processed to remove duplicate records. The resulting dataset contains one entry for each distinct article. Each entry contains the pmid, the DOI, the pmcid, and the publication year of the corresponding article. This information is the minimum required for the calculation of the selected impact metrics.

Calculation of Citation-based Measures

A prerequisite for calculating the citation-based impact measures of the collected articles, is the compilation of their citation network, i.e., the network which has articles as nodes and citations between them as directed edges. The citations of the articles required to construct this network are gathered using NCBI's eLink tool. The tool returns for a given article the identifiers (pmids/pmcids) of all articles that cite, or are cited by it. Two citation-based impact measures are calculated on the constructed network: the PageRank [5] and the RAM scores [6]. These two measures were selected based on the results of a recent experimental study [7], which found them to perform best in capturing the overall and the current impact of an article (i.e., its "influence" and its "popularity"), respectively. Both measures are calculated by performing citation analysis. PageRank evaluates the overall impact of articles by differentiating their citations based on the importance of the articles making them. However, it is biased against recent articles that haven't accumulated many citations yet, but may be the current focus of the research community. RAM alleviates this issue by considering recent citations as being more important.

Calculation of Tweet-based measure

In addition to the citation-based measures, for each article, the number of tweet posts mentioning it is calculated as well. This is considered a measure of its social media attention. The *COVID-19-TweetIDs*[4] dataset [4] is used for the collection of COVID-19-relevant tweets. This dataset contains a collection of tweet IDs, each of them published by one of 9 predetermined Twitter accounts (e.g., @WHO) and containing at least one out of 71 predefined coronavirus-related keywords (e.g., "Coronavirus", "covid19", etc). At the time of writing, a subset of this dataset containing tweets posted from January 21^{st} to March 31^{st}($83,998,659$ unique tweet IDs) has been integrated in BIP4COVID19. The corresponding Tweet objects were collected using the Twitter API. The result was a collection of $76,046,064$ tweet objects (66, 1 GB in zipped format). The difference between the number of IDs and hydrated objects is due to the fact that $7,952,595$ tweets have been deleted in the meantime (9%) and are, therefore, impossible to retrieve.

To find those tweets which are related to the articles in our database, we rely on the URLs of the articles in doi.org, PubMed, and PMC. These URLs are easily produced based on the corresponding identifiers. In addition, when possible, the corresponding page in the publisher's website is also retrieved based on the doi.org redirection. After the collection of the URLs of all articles, the number of appearances of the URLs related to each one are produced. However, since the Twitter API returns either shortened or not fully expanded URLs, the fully expanded URLs are collected using the unshrtn[5] library.

[4] https://github.com/echen102/COVID-19-TweetIDs
[5] https://github.com/docnow/unshrtn

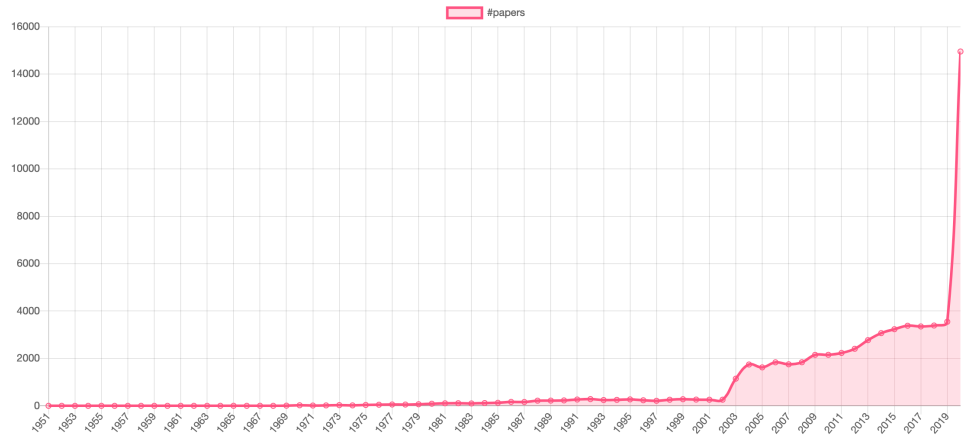

Fig 2. COVID-19-related articles per year.

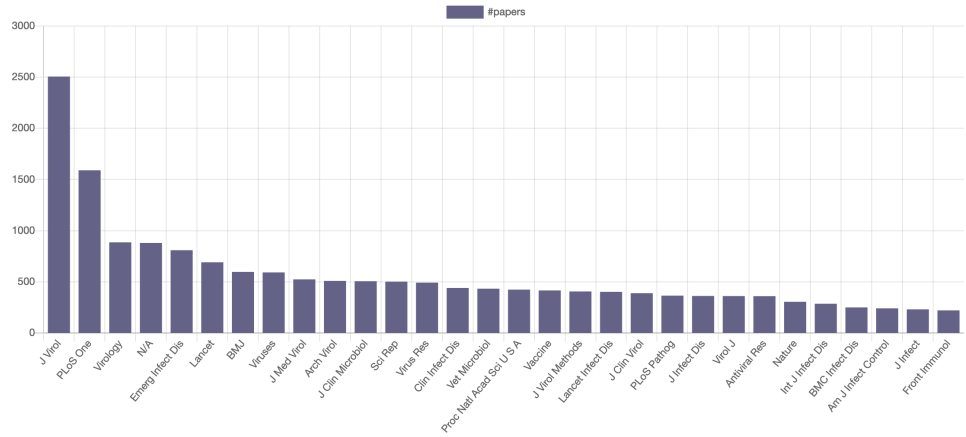

Fig 3. Top 30 venues in terms of published COVID-19-related articles.

Results

Data Set Details

The BIP4COVID19 dataset, produced by the previously described workflow, is openly available on Zenodo [1], under the Creative Commons Attribution 4.0 International license. At the time of publication, the ninth release of this dataset (v4) is available[6], counting 61,746 records in total. Of these, 54,868 correspond to entries in PubMed, 54,891 to entries in PMC, while 59,172 have an associated DOI. All publications included were published from 1951 to 2020. The distribution of publication years of the articles recorded in the dataset is illustrated in Figure 2. 14,954 of these articles were published in 2020, i.e., after the coronavirus pandemic outbreak, while 46,792 were published from 1951 to 2019. Moreover, the number of articles per venue for the top 30 venues (in terms of relevant articles published) are presented in Figure 3.

The BIP4COVID19 dataset is comprised of three files in tab separated (TSV) format. The files contain identical information, however, in each of them, the records are ordered based on a different impact measure (popularity, influence, social media attention). The data attributes included in each file are summarised in Table 1.

[6]The dataset is updated on a regular basis.

Attribute	Interpretation
PubMed identifier	Unique identifier of the article in PubMed, as collected from the source data files. Articles missing this identifier are indicated with the value "N/A".
DOI	The Digital Object Identifier of the article, as collected from PubMed. Articles missing a DOI are indicated with the value "N/A".
PCM identifier	Unique identifier of the article in Pubmed Central (PMC), as collected from the source data files. Articles missing an identifier in PMC have the value "N/A".
Popularity score	The value of the corresponding citation-based measure (RAM [6]) for the respective article.
Influence score	The value of the corresponding citation-based measure (PageRank [5]) for the respective article.
Social media attention	The calculated tweet count for the article corresponding to the record.

Table 1. Data attributes inside the TSV files.

Web Interface

A Web interface has been developed on top of the BIP4COVID19 data.[7] Its aim is to facilitate the exploration of COVID-19-related literature. The option to order articles according to different impact measures is provided. This is expected to be useful since users can better prioritise their reading based on their needs. For example, a user that wants to delve into the background knowledge about a particular COVID-19-related sub-topic could select to order the articles based on their influence. On the other hand, another user that needs to get an overview of the latest trends in the same topic, could select to order the articles based on their popularity.

The information shown to users, per publication, includes its title, venue, year, and the source dataset where it was found. Moreover, each result is accompanied by color coded icons that denote the publication's importance based on each calculated impact measure. In this way, the users can easily get a quick insight about the different impact aspects of each article. The tooltips of these icons provide the exact scores for each measure. Each publication title functions as a link to the corresponding article's entry in its publisher's website, or to Pubmed. Finally, a page containing interesting statistics is provided. This page contains various charts that visualise, for example, the number of articles per year, or the number of articles that have substantial impact based on each of the provided impact measures, per year.

Discussion

To ensure the proper integration and cleaning of the CORD-19 and LitCovid datasets, we rely on NCBI's eTool suite. In particular, we collect pmids and pmcids from both datasets and use them as queries to gather each article's metadata. After cleaning the article title (e.g., removing special characters) we automatically identify duplicates by comparing each record's complete content and eliminate them. Finally, manual inspection is performed to produce the correct metadata for a limited number of duplicates that remain (e.g., duplicate records containing the title of the same publication in two different languages).

[7] `https://bip.covid19.athenarc.gr/`

Further, to guarantee the correctness of the compiled citation graph we apply the following procedures. After gathering all citing - cited records using NCBI's eTools, those that include identifiers not found in the source data are removed. Since many citing - cited pairs may have been found both with pmids and pmcids, the resulting data may still contain duplicate records. These records are removed, after mapping all pmids/pmcids to custom identifiers, with pmid-pmcid pairs that refer to the same article being mapped to the same identifier. The final resulting citation graph is based on these mapped identifiers. As an extra cleaning step, any links in the graph that denote citations to articles published at a later time than the citing article are removed.[8]

To ensure that we retrieve a set of tweets about each article that is as comprehensive as possible, we collect not only the URLs in doi.org, Pubmed, and PMC, but also the URL to the article in its publisher's website, where possible. These latter URLs are very important, since they are widely used in tweets. To collect them we utilize doi.org redirections. To avoid incorrect tweet counts due to duplicate tweets, we used a simple deduplication process after the Tweet object retrieval. Moreover, the use of the unshrtn library to expand the short URLs from tweet texts ensures that our measurements derive from all available URL instances of each publication record, no matter how they were shortened by users or Twitter.

The following limitations should be taken into consideration with respect to the data: while we take effort to include as many articles as possible, there are many cases where our source data do not provide any pmids or pmcids. As a consequence, no data for these articles are collected and they are not included in the BIP4COVID19 dataset. Furthermore, with respect to the calculated impact scores, it should be noted that the citation analysis we conduct is applied on the citation graph formed by citations *from* and *to* collected publications only, i.e., our analyses are not based on pubmed's complete citation graph, but on a COVID-19-related subgraph. Consequently, the relative scores of publications may differ from those calculated on the complete PubMed data. Finally, regarding the tweet-based analysis, since our data come from the COVID-19-TweetIDs dataset which only tracks tweets from a predefined set of accounts and which is based on a particular set of COVD-19-related keywords, the measured number of tweets is only based on a subset of the complete COVID-19-related tweets.

Our data are available in files following TSV format, allowing easy import to various database management systems and can be conveniently opened and edited by any text editor, or spreadsheet software. We plan to update the data regularly, incorporating any additions and changes from our source datasets, as well as to expand the tweet counts based on all available data for 2020. Additionally, we plan to incorporate any further sources on coronavirus related literature that may be released and which will index the literature based on pmids and/or pmcids.

The contents of the BIP4COVID19 dataset may be used to support multiple interesting applications. For instance, the calculated scores for each impact measure could be used to rank articles based on their impact to help researchers prioritise their reading. In fact, we used our data to implement such a demo as previously described. Additionally the rank scores may be useful for monitoring the research output impact of particular sub-topics or as features in machine learning applications that apply data mining on publications related to coronavirus.

Conclusion

We presented BIP4COVID19, an openly available dataset, providing impact scores for coronavirus related scientific publications. Our dataset can be potentially useful both

[8] Such references to future articles are often observed in citation data due to various reasons. Hence, a common practice is to remove them [6].

for researchers in need to prioritize their reading, as well as for in various applications (e.g., applications using impact scores as machine learning features). We have additionally built on our dataset, providing a web interface that allows for the ordering of coronavirus literature based on the various impact measures. Finally, our dataset is regularly updated.

Acknowledgements

We acknowledge support of this work by the project "Moving from Big Data Management to Data Science" (MIS 5002437/3) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

References

1. Vergoulis T, Kanellos I, Chatzopoulos S, Pla Karidi D, Dalamagas T. BIP4COVID19: Impact metrics and indicators for coronavirus related publications. *Zenodo* 10.5281/zenodo.3747809 (2020).

2. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. Natur. 2020 Mar;579(7798):193-.

3. COVID-19 Open Research Dataset (CORD-19). 2020. Version 2020-05-12. Semantic Scholar: https://pages.semanticscholar.org/coronavirus-research. Accessed 2020-05-17. doi:10.5281/zenodo.3715505

4. Chen E, Lerman K, Ferrara E. Covid-19: The first public coronavirus twitter dataset. arXiv preprint arXiv:2003.07372. 2020 Mar 16.

5. Page L, Brin S, Motwani R, Winograd T. The pagerank citation ranking: Bringing order to the web. Stanford InfoLab; 1999 Nov 11.

6. Ghosh R, Kuo TT, Hsu CN, Lin SD, Lerman K. Time-aware ranking in dynamic citation networks. In 2011 IEEE 11th International Conference on Data Mining Workshops 2011 Dec 11 (pp. 373-380). IEEE.

7. Kanellos I, Vergoulis T, Sacharidis D, Dalamagas T, Vassiliou Y. Impact-Based Ranking of Scientific Publications: A Survey and Experimental Evaluation. IEEE Transactions on Knowledge and Data Engineering. 2019 Sep 13.

8. Piwowar H. Introduction altmetrics: what, why and where?. Bulletin of the American Society for Information Science and Technology. 2013 Apr;39(4):8-9.