# Getting Insights from a Large Corpus of Scientific Papers on Specialisted Comprehensive Topics - the Case of COVID-19

Bernard Dousset, Josiane Mothe

Available online at www.sciencedirect.com

Procedia Computer Science 00 (2020) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

# Getting Insights from a Large Corpus of Scientific Papers on Specialisted Comprehensive Topics - the Case of COVID-19

Bernard Dousset[a], Josiane Mothe[b,*]

[a] *IRIT, UMR5505, CNRS & Univ. Toulouse, France*
[b] *IRIT, UMR5505, CNRS & INSPEE UT2J, Univ. Toulouse, France*

**Abstract**

COVID-19 is one of the most important topics these days, specifically on search engines and news. While fake news is easily shared, scientific papers are reliable sources where information can be extracted. With about 24,000 scientific publications on COVID-19 and related research on PubMed, automatic computer-assisted analysis is required. In this paper, we develop two methodologies to get insights on specific sub-topics of interest and latest research sub-topics. These rely on natural language processing and graph-based visualizations. We run these methodologies on two cases: the virus origin and the uses of existing drugs.



*Keywords:* Topic analysis; Automatic mining of scientific publication; COVID-19; automatic keyphrase extraction;

## 1. Introduction

COVID-19 is one of the most important topics these days, specifically on search engines and news. It is a worldly shared topic of interest. While news is looped on TV, a very few specialists know deeply on COVID-19. On the Internet, a lot of fake news started to circulate and spread as fast as the virus itself. In such situation, citizens and decision makers need reliable sources of information. Scientific papers are certainly such reliable sources that could be used for helping citizens knowing more about it and being informed on both a reliable and accurate way.

Not only scientific sources can be used to try to answer some specific questions that arise but it is also a way to know the main researchers, groups or institutes that work on a specific sub-topic, what the collaborations are, ... Both in-depth analyses and overviews on a large quantity of research papers can help decision makers or even citizens to be better educated on the state of the art. For people, it can help move aside fake news. It can also help new-comers in the COVID-19 research field providing overviews of sub-topics first (main publication venues, ... main authors).

* Corresponding author. Tel.: +33-6-86724870
*E-mail address:* Josiane.Mothe@irit.fr

Indeed the research in the domain is quite huge specifically if we consider other forms of COVIDs, Severe Acute Respiratory Syndrome, and Middle East Respiratory Syndrome. For example, the recently released collection named COVID-19 Open Research Dataset [1] consists of more than 40,000 articles. Some papers start to provide reviews [6, 10, 14] which are indeed very useful. This paper aims at providing a systematic methodology to mine such a large publication set while giving some specific focuses on topics of interest.

The rest of this paper is organized as follows: Section 2 presents some related works; Section 3 describes the analysis framework including the data description and the methodology for analysis that was followed; Section 4 describes the methodology we developed to get insights on a specific sub-topic or on latest research. Section 5 focuses on a deeper analysis on the origins of COVID-19. Section 6 focuses on the analysis of the terminology related to the latest research. Finally Section 7 concludes this paper.

## 2. Related work

**Publication analysis on COVID-19.** Some related work reports either studies on various topics or focuses on one or two topics. For example, the report from Nature Medecine [6] presents a focus on clinical trials and human studies, Preclinical studies, and Epidemiology. Another distinction on related work is what level as automatic assistance they benefit from. In most of the cases, it is difficult to know that level as it is not necessarily depicted. For example, [6] seems to be a manual analysis of a few papers on the topic since the references are few. Harapan *et al.* [10] presents a literature review on Coronavirus disease 2019 (COVID-19). This study cites 74 references and presents a comprehensive state of the art on different sub-topics such as COVID-19 transmission, risk factors, diagnosis or treatments. Huang *et al.* [14] reviewed 1,281 abstracts from which they identified 322 manuscripts relevant to 5 areas of interest for their study. The five topics they chose are as follows: antibody kinetics, correlates of protection, immunopathogenesis, antigenic diversity and cross-reactivity, and population seroprevalence. Dousset *et al.* [8] presents a short analysis of a larger dataset which size is similar to the one that we are using in this study. However their study does not go in-depth in the publication content but rather analyze the collaborations between researchers and countries. They do not analyze any specific topic.

Different from the previous analysis, this paper combines (a) the analysis of a large data set of about 25,000 references and (b) highly assisted analysis. While the methodology could be applied to various COVID-19 subtopics, we focus here on two main topics: the origin of the virus and the use of other disease treatment.

**Mining scientific publications.** Most of the work in automatic publication analysis is related to scientometrics [25] which has been defined as "quantitative study of science, communication in science, and science policy" [12]. In this paper, publication mining is not strictly a question of scientometrics, rather the objective is to build knowledge from a large set of publications.

Shibata et al. [30] uses citation network analysis in order to detect emerging research. Small et al. also use direct citation and co-citation analysis in order to identify emerging topics [31]. Unfortunately, this information is not necessarily provided in digital database for large quantities of documents. Buscaldi et al. mine scholarly publications to build scientific knowledge graphs [5]. As in our approach, the authors use existing natural language processing and mining tools in their approach; however, they focus on the textual parts of the publications. Ronzano and Saggion developed a platform to automatically extract and enrich structural and semantic aspects of scientific publications. Their approach also focuses on the textual content and their applications are related to rhetorical sentence classification and extractive text summarization [28]. Mothe et al. present a platform to mine scientific publications. In their paper, they focus on detecting the main collaborations, focusing on the geographical structure of a domain [23].

## 3. Analysis Framework

**COVID-19 scientific publication set.** Recently, publishers have released the COVID-19 Open Research Dataset. It is available at `https://pages.semanticscholar.org/coronavirus-research`. This data set consists of multiple files.

[1] `https://pages.semanticscholar.org/coronavirus-research`

Table 1. A few statistics on the data collection. An author may appear in the collection with different spellings; for example there are 6 spellings for R.A. BARIC and two for L. ENJUANES. Moreover, while the collected papers range from 1952 to 2020, 15 years have more than 500 publications.

| Label | Number of | Occur. at least twice | at least 10 times |
|---|---|---|---|
| Publication | 23,784 | | |
| Abstract | 17,116 | | |
| Author | 75,147 | 9,211 | 1,145 |
| Venues | 3,120 | 1,832 | 431 |
| Year | 60 | 54 | 48 |

Table 2. Authors (full author names) with more than 100 publications in the collection and Most frequent publication venues in the collection.

| **Authors** | **Venues** |
|---|---|
| YUEN, KWOK-YUNG (Univ. of Hong-Kong) | JOURNAL OF VIROLOGY |
| PERLMAN, STANLEY (Univ. of Iowa) | ADVANCES IN EXP. MEDICINE AND BIOLOGY |
| DROSTEN, CHRISTIAN (Berlin Institute of Virology, Germany) | VIROLOGY |
| BARIC, RALPH S. (Univ. of North Carolina, USA) | EMERGING INFECTIOUS DISEASES |
| MEMISH, ZIAD A (Alfaisal University, Saudi Arabia) | BMJ (CLINICAL RESEARCH ED.) |
| JIANG, SHIBO (New York Blood Center, USA) | JOURNAL OF MEDICAL VIROLOGY |
| ENJUANES, LUIS (Campus Univ. Autónoma de Madrid, Spain) | THE JOURNAL OF GENERAL VIROLOGY |

Among them, the Metadata file (60Mb) is a CSV file corresponding to 44,270 research articles with links to PubMed[2], Microsoft Academic[3] and the WHO (World Health Organization)[4] COVID-19 database of publications. Record fields are the following: title, doi, abstract, date of publication, authors, journal, as well as internal document ids. (PMC ID, PUBMED ID, Microsoft Academic Paper ID, WHO ID) and information whether the full text is available or not. While the meta file is a rich source of information, other pieces of information that are missing in that data file can be very useful such as the affiliation of the authors for example. For this reason we also considered a more complete set regarding the attributes that are provided.

We chose to focus on the documents from PubMed only, which is known to be realable source. It does not contain all the 44k scientific papers from the COVID-19 Open Research Dataset (March, 2020) but about 24,000 papers.

The query used to query the PubMed collection (https://www.ncbi.nlm.nih.gov/pubmed) on March 25, 2020 was the same as the one used in the sub-mentioned data set:
*"COVID-19" OR "covid19" OR "Coronavirus" OR "Corona virus" OR "2019-nCoV" OR "SARS-CoV" OR "MERS-CoV" OR "Severe Acute Respiratory Syndrome" OR "Middle East Respiratory Syndrome"*

**A few elements of the document collection**

A few statistics are presented in Table 1 where the top 7 authors (the ones that authored the larger number of publications in the analyzed collection) are listed in Table 2. On average, there is 5.6 authors per publications. Not all the publications come with an abstract (72% have an abstract for a total of 17,116). Table **??** lists the most frequent venues where the scientific papers have been published.

**Topics of interest**. While researchers have their topics of interest driven by their funding, project and research, topics of interest also come from the civil society on COVID-19.

It is worth to mention that NIST/TREC has formed a joint effort called TREC-COVID[5]. Like the other TREC tracks[6], TREC-COVID aims at gathering research teams in information retrieval to evaluate search engines on specific

[2] https://www.ncbi.nlm.nih.gov/pubmed
[3] https://academic.microsoft.com/
[4] https://www.who.int/emergencies/diseases/novel-coronavirus-2019
[5] https://ir.nist.gov/covidSubmit/
[6] TREC: Text Retrieval Conference is co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense supports research for large-scale evaluation of text retrieval methodologies trec.nist.gov

tasks. Mid April 2020, TREC-COVID has release a set of 30 topics of interest. A topic of interest is for example the "Coronavirus origin" or "early symptoms".

These topics also echo the most popular questions web searchers are interested on. Google trends mentions "Where did coronavirius start?" and "How to know if you have coronarius" among the most asked questions.

Finally, these topics also echo the ones that have been considered in some related work papers [10, 6] that generally target the COVID-19 origins, its symptoms, spreading, risk factors and treatments.

## 4. Methodology

**Overview.** The information we use is the data as collected. While automatic analysis helps in handling large quantities of publications, the conclusions drawn have to be handled with caution because there is no manual analysis of the content and no checking. Moreover, we did not solved content anomalies such as variants of entities spelling (e.g. author names). There are also missing values that we did not consider either and not resolved (For example while there are 23,784 publications, there are 17,116 non empty abstract only, See Table 1).

The methods we use are usual data mining tools like frequency, graph-based visualization, factorial analysis. We consider crossing meta-data with content-based information from free texts as detailed later on.

The analysis considers both meta-data, such as the source, publication date and author fields and free text data such as the title and abstract fields. With regard to the titles and abstracts, we consider both single words as well as phrases that we automatically extracted.

**Keyphrase extraction**. Different methods have been developed to extract key-phrases from free texts [11, 2]; among the most popular are graph-based extraction[3, 4, 24], co-occurrence-based methods [15] and more recently embeddings [29].

In our approach, we use a n-gram word extraction where we skip stop words. More precisely, we extract the most frequent n-grams after stop word removal, but without stemming to keep more precise semantics. We also consider an initial lexicon from composed terms (as written by the authors e.g. "anti-malagia" or "animal-origin") as initial phrases that are enriched by the n-gram extracted ones. Undeniably, the method used for keyphrase extraction has an influence on the results, we keep this for future work.

**Graph-based visualization.** Graphs are among visualization tools the most used in the literature, as linking concepts or objects is the most common mining technique [23]. Graph-based visualizations are widely used to visualize bibliometric networks [9, 33].

In this paper, we mainly use bipartite graphs. A bipartite graph is a graph whose vertices (nodes) can be divided into two disjoint and independent sets and where edges connect a node of each type (Wikipedia). A bipartite graph does not contain any odd-length cycle. This type of representation is also widely used for document analysis and visualizations [7, 1]. In this paper, bipartite graphs are used to visualize the results of crossing meta-data and keyphrases extracted from publications.

**Process.** In this paper, we developed two different processes: the first one can be applied to focus on any specific sub-topic of interest, this is the process we applied to the "Origin" of the virus sub-topic (See Section 5). The second process is related to the latest research and we apply it to detect some topic clusters (See Section 6).

*Getting insights on a specific topic*

This process consists into three steps:

- Select the keyphrases related to the topic of interest. It can be computer-assisted by considering strings of characters (e.g. "ORIGIN" is a relevant character string to extract many relevant key-phrases such as "BAT-ORIGIN", "HUMAN-ORIGIN",...) ; stems of the topic word(s) is a good start. The automatically obtained list should then be manually checked in order to remove non-relevant terms (e.g. "ORIGINALITY")
- Build a bipartite graph where vertices are key-phrases in the one hand and publication identifiers in the other hand. This first representation provides a quick overview of the use of the terms: are the terms shared among publications or are they rather partitioned (each publication is more focusing on one of the aspects). It is also a means to directly go to the associated publications;

- For one specific term, build the bipartite graph where the other vertices are authors names. This graph can be weighted by the number of publications of each author that mention the terms. This graph shows the most "important" authors related to that term (who are likely to be specialists). Such a graph can also be built considering several related terms at the same time when terms are non independents in publications (cf. previous step). The latter graph makes it possible to highlight the authors that tackles several aspects of the topic.

Other meta data could be also crossed to include additional steps in that process (e.g. considering the authors' affiliation or affiliation countries). This step was not included in this paper.

*Getting the latest research topic of interest*

This process also consists into three steps:

- Extract the latest terminology: In this case these are n-grams (keyphrases) that occur in the latest year(s) but not before. They are obtained by crossing the keyphrases and the year of the publications the keyphrases occur in. We consider also the occurrence frequency when selecting them;
- Extract highly connected graphs ("communities" of keyphrases). This step results in sub-topics of interest consisting of terms of various nature but often used together in recent publications;
- Build bipartite graphs where vertices of the first type are the terms from one community and vertices of the second type are publication identifiers to identify the relevant publications with regard to the group of terms.

This process is used to get the latest research topics and associated publications.

## 5. Origins

The topic of interest we tackle in this section is related to the COVID-19 origins and uses the first process as described in Section 4.

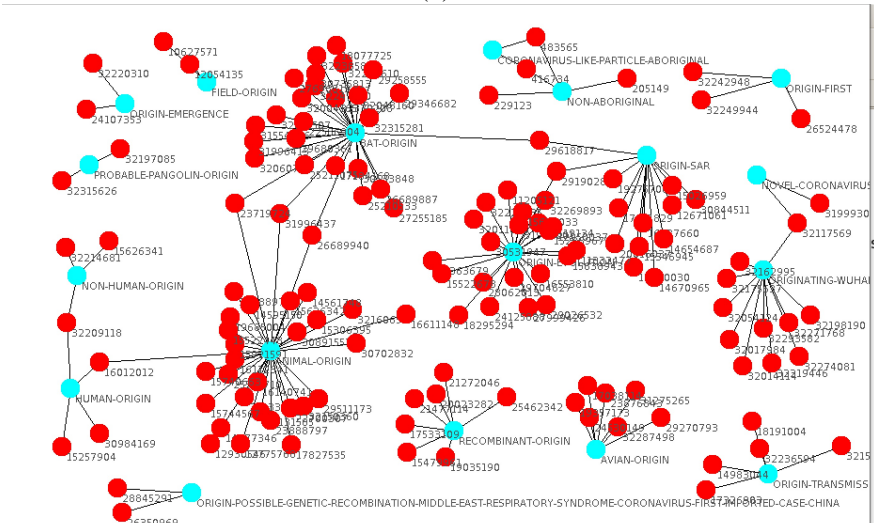
Figure 1(a) displays the terms related to the ORIGINs of the virus (extracted from the publications titles and abstracts when available) as well as their frequency (See Section 4.2 for keyphrase extraction). For example, 33 publications study BAT-ORIGIN and 7 AVIAN-ORIGIN. We also used these terms to extract the associated publications. For example, Figure 1(b) displays the PMID of the 33 publications that mention the "BAT-ORIGIN" (here just an extract). The terms from Figure 1(a) are the one used in the graph from Figure 1(c) as blue nodes. Indeed, Figure 1(c) is a bipartite graph where blue nodes are the ORIGIN terms while the red nodes are the publication identifiers (PMID). This graph is not fully linked since some publications mention one of these "ORIGIN" terms only.

From this graph, we can also identify the publications that mention several of the ORIGIN terms as PMID-26689940 [13], 31996437 [34] and 23719724 [21] which mention both "ANIMAL-ORIGIN" and "BAT-ORIGIN". Another example is PMID- 16012012 [17] which mention both "HUMAN-ORIGIN" and "ANIMAL-ORIGIN" (See Figure 1(c)).

With regard to the publications associated with BAT-ORIGIN terms in Figure 1(b), we can mention Ren et al. [27] (PMID- 32004165) who report a novel bat-origin CoV causing severe and fatal pneumonia in humans. The identified virus is phylogenetically closest to a bat SARS-like CoV. PMID-31996413 [19] employs a capture-based NGS approach for virus discovery. Since (SARS)-COV and (MERS)-CoV both originated from bat, active surveillance is recommended in this paper. Yang et al. [36] (PMID- 31554686) study potential cross-species transmissibility of SADS-CoV. PMID- 30533848 [20] studies the full-length genome sequence of a novel swine acute diarrhea syndrome coronavirus (SADS-CoV), CH/FJWT/2018 which is closely related to CN/GDWT/2017. Also, we can mention PMID-29618817 (see also Figure 1(c)) that makes the link between the term BAT-ORIGIN and ORIGIN-SARS. In this paper from 2018, the authors focus on cross-species transmission and more specifically on bats being an important reservoirs for emerging viruses and the transmission of a coronavirus to humans. They provide evidence that HKU2-related bat coronavirus (SADS-CoV) is the aetiological agent that was responsible for fatal disease in pigs in China; they also highlight striking similarities between the SADS and SARS [37]. A link is also made with bat-origin SADS-related coronaviruses. Wang et al. (PMID- 29680361 [35]) report cross-species transmission due to a large number of mutations on the receptor-binding; here a novel bat-origin coronaviruses found in pigs is considered.

| | |
|---|---|
| 34 ANIMAL-ORIGIN | |
| 33 BAT-ORIGIN | |
| 24 ORIGIN-EVOLUTION | PMID- 32330208 |
| 15 ORIGIN-SAR | PMID- 32315281 |
| 11 ORIGINATING-WUHAN | PMID- 32266610 |
| 7 RECOMBINANT-ORIGIN | PMID- 32247050 |
| 7 AVIAN-ORIGIN | PMID- 32238584 |
| 6 NON-ABORIGINAL | PMID- 32060789 |
| 5 ORIGIN-TRANSMISSION | PMID- 32048160 |
| 4 NON-HUMAN-ORIGIN | PMID- 32015507 |
| 4 HUMAN-ORIGIN | PMID- 32004165 |
| 3 ORIGIN-FIRST | PMID- 31996437 |
| 2 PROBABLE-PANGOLIN-ORIGIN | PMID- 31996413 |
| 2 ORIGIN-POSSIBLE-GENETIC-RECOMBINATION-I | PMID- 31554686 |
| 2 ORIGIN-EMERGENCE | PMID- 30735813 |
| 2 NOVEL-CORONAVIRUS-ORIGINATING-WUHAN | PMID- 30533848 |
| 2 FIELD-ORIGIN | PMID- 29680361 |
| 2 CORONAVIRUS-LIKE-PARTICLE-ABORIGINAL | PMID- 29618817 |
| 1 PROTEIN-OF-ORIGIN | |
| 1 OSTRICH-ORIGINATING | |
| 1 ORIGIN-UNKNOWN | |
| 1 ORIGINATED-HUBEI | |

(a) Terms related to the ORIGINs of the virus (b) Publication PMID that mention BAT-ORIGIN




(c) Graph of the ORIGIN terms and associated PUB-ID

Fig. 1. Focus on ORIGIN

We also had a look to the associated authors (See Figure 2(a)). In this figure, the only blue node is BAT-ORIGIN term, while the red nodes are authors of publications that mention this term. The value on the link indicates the number of publications a researcher authored that mention BAT-ORIGIN. Figure 2(b) shows, for those authors that mention BAT-ORIGIN, the other ORIGIN terms also mentioned by them. The thickness of the link is an indication of the number of publications as well as the size of the nodes.

## 6. Latest research

While in the previous section we did not consider the year of publication, in this section, we focus on the latest research and the associated terminology. Before building any graphs, we consider the publications that are published in the 30 last years only (1991 to 2020). We then keep the only phrases or automatically extracted keywords from the titles and abstracts (see Section 3) that mainly occurs in 2020. More precisely, we selected the keyphrases whose occurrences are 80% in 2020. These keyphrases (there are about 1,500) are thus the keyphrases of current interest. We built the co-occurrence network where the weights between two nodes (keyphrases) are calculated as follows: $W(t_i, t_j) = \frac{df_{ij}}{df_i.df_j}$ where $t_i$ and $t_j$ are keyphrases that mainly occurs in 2020, $df_{ij}$ is the number of documents that contain both $t_i$ and $t_j$ and $d_i$ (resp. $d_j$ is the number of documents that contains $t_i$ (resp.$t_j$). This ensure a normalization of the weights through columns and raws. We then extracted clusters of terms that are closely related as communities

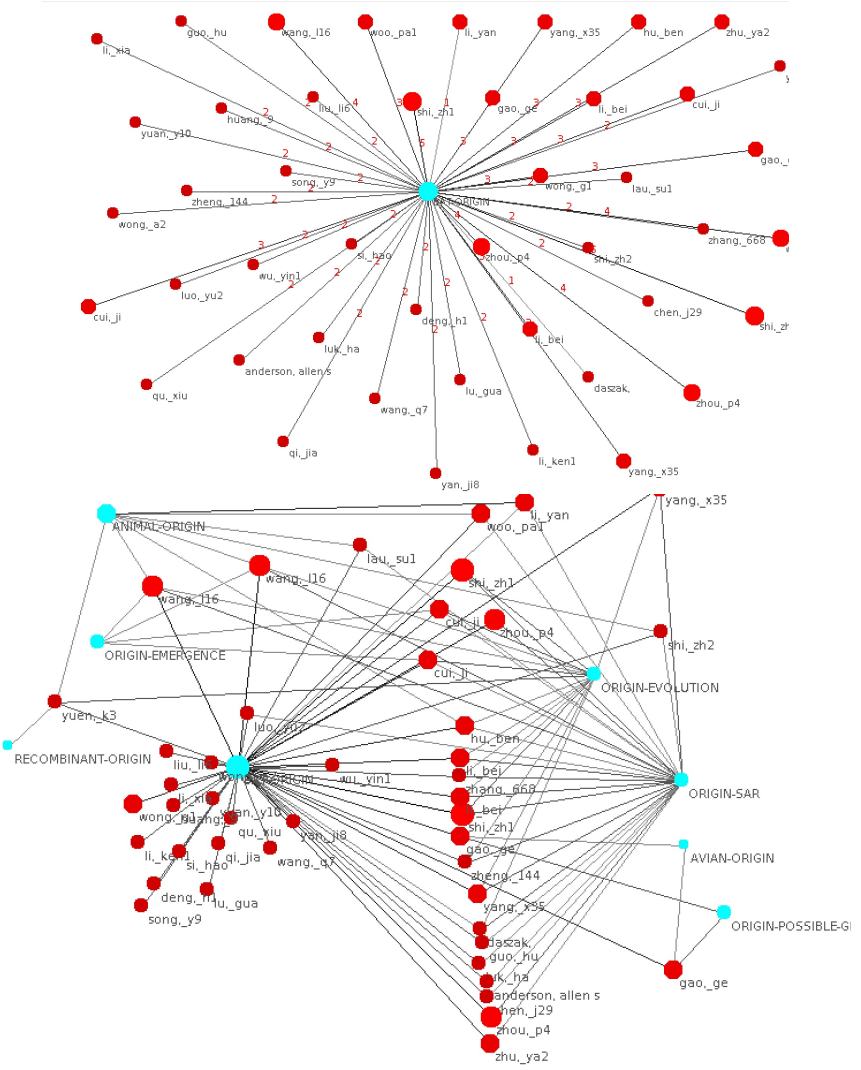


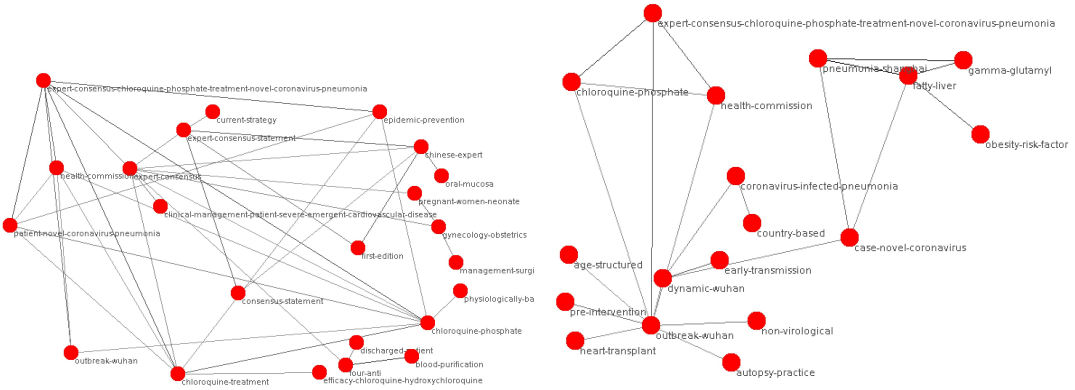

Fig. 2. Authors associated to (a) BAT-ORIGIN term (b) ORIGIN terms from the one listed in Figure 1

(nodes that are highly inter-connected together and weakly connected with other nodes). They are the connected sub-graphs after removing the weaker connections (above 0.15). That provides clusters from different focuses.

The example in Figure 3(a) focuses on the Chloroquine-treatment while Figure 3(b) is related to Obesity risk factor. From these figures, we can also identify the related terms.

COVID-19 is an infectious disease caused by SARS-CoV-2 and several papers investigate the use of existing anti-viral treatments. For examples, some of the publications mention ”anti-” (e.g. anti-flu, anti-malaria, ..). We thus had a closer look to the network related to treatment used for other diseases considering various ”anti-” terms as shown in Figure 4. As we did in Section 5 and illustrated in Figure 1(c), we look at the publications associated with these terms. We found 33 publications directly related to these terms. The PMID of these publications as well as related terms are presented in Figure 5.

From this Figure 5, we can see that some of the papers are more connected than others; suggesting they potentially mention more topics of interest. For example, PMID 32277367 [22], which is marked up with a black square on the right-bottom part of Figure 5, was published on the 10th of April, in J Neurooncol. and results from a collaborative work of India, Russia, and UK and concludes that “it may be cautiously recommended to continue glucocorticoids and other disease-modifying antirheumatic drugs (DMARDs) in patients receiving these therapies, with discontinuation




(a) Chloroquine-treatment term network (b) Obesity Risk Factor network

Fig. 3. Examples of most recent terms and semantic network. Figure 5 shows two interconnected sets of keyphrases as just defined. The left-side one is related to cancer while the right-side one is more related to patient rehabilitation. These are two examples of recent research topic of interest in the domain.

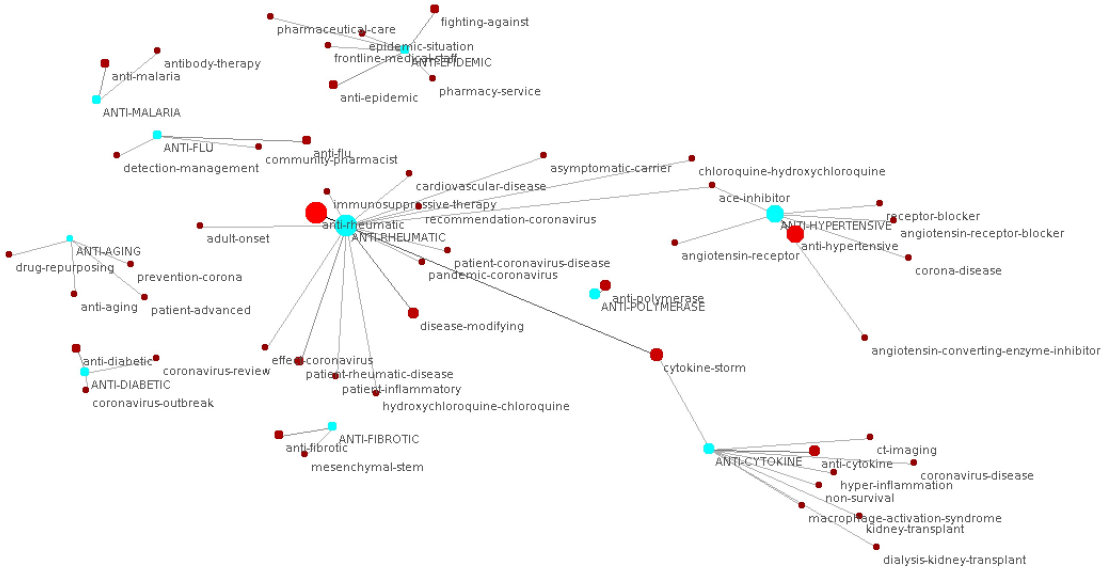



Fig. 4. Various "ANTI-" uses.

of DMARDs during infections as per standard practice". This paper also mentions various evidence that suggest potential benefits of various drugs. Reversely, we can see in the same figure that some keywords are also more shared than others. As an example, anti-rheumatic is mentioned in 9 of the displayed publications.

Among these publications, that mainly report observations and not clinical tests, we can quote Perricone [26] (PMID- 32317220) discusses the anti-viral aspect of immunosuppressants for searching for a potential treatment for SARS-CoV-2 infection. Lehrer et al. [18] (PMID- 32313883) discusses the effects of biguanides on influenza and coronavirus. Kumar et al. (PMID- 32313660) study the antibody therapy as an immediate strategy for emergency prophylaxis and SARS-CoV-2 therapy [16]. Song et al. (PMID- 32314010) reports a case of COVID-19 pneumonia on a 61-year-old female rheumatoid arthritis; she was treated with antiviral agents (lopinavir/ritonavir), and treatment with cDMARDs was discontinued except hydroxychloroquine. Her symptoms gradually improved and three weeks later, real-time PCR for COVID-19 showed negative conversion [32].

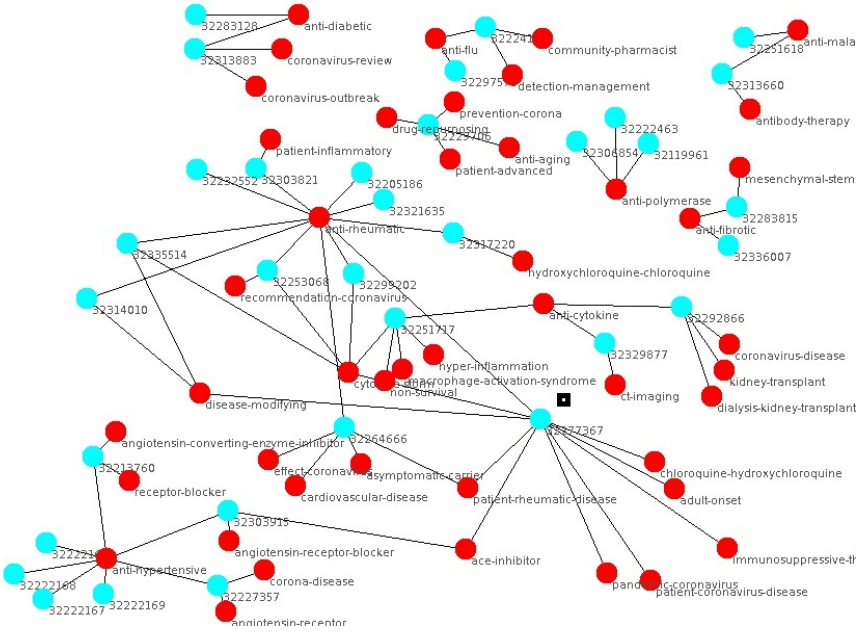



Fig. 5. PMID of the publications that mention an ”ANTI” terms and the other related terms.

## 7. Conclusion

This paper presents two analytical processes in order to mine scientific papers that are illustrated on COVID-19 scientific publications. The results are knowledge graphs of various natures that helps getting insights on specific subtopics or recent research topics.

Typically, analytical work for medical applications relies on data sources and links to physicians. Here, however, the large amount of published material provides a very important data source for themes and issues that are highly relevant to disease and medical practice.

While scientific papers are reliable sources of knowledge, on COVID-19, other sources of information such as the World Health Organization reports would also worth being analysed using the same type of methodology.

## References

[1] Alsallakh, B., Micallef, L., Aigner, W., Hauser, H., Miksch, S., Rodgers, P., 2014. Visualizing sets and set-typed data: State-of-the-art and future challenges, in: Eurographics Conference on Visualization (EuroVis), pp. 1–22.

[2] Beliga, S., 2014. Keyword extraction: a review of methods and approaches. University of Rijeka, doi 10.1.1.704.9230 , 1–9.

[3] Boudin, F., 2013. A comparison of centrality measures for graph-based keyphrase extraction, in: Proceedings of the sixth international joint conference on natural language processing, pp. 834–838.

[4] Boudin, F., 2018. Unsupervised keyphrase extraction with multipartite graphs. arXiv preprint arXiv:1803.08721 .

[5] Buscaldi, D., Dessì, D., Motta, E., Osborne, F., Recupero, D.R., 2019. Mining scholarly publications for scientific knowledge graph construction, in: European Semantic Web Conference, Springer. pp. 8–12.

[6] Carvalho, T., 2020. Covid-19 research in brief: 18 april to 24 april, 2020. NatureMedecine, News, 24 April .

[7] Crossno, P.J., Wilson, A.T., Shead, T.M., Dunlavy, D.M., 2011. Topicview: Visually comparing topic models of text collections, in: 2011 ieee 23rd international conference on tools with artificial intelligence, IEEE. pp. 936–943.

[8] Dousset, B., Mothe, J., 2020. A short study on covid-19 open research scientific papers. Internal Report, March URL: http://www.irit.fr/publis/SIG/2020_COVID_BD_JM.pdf.

[9] van Eck, N.J., Waltman, L., 2014. Visualizing Bibliometric Networks. Springer International Publishing, Cham. chapter 13. pp. 285–320. URL: https://doi.org/10.1007/978-3-319-10377-8_13, doi:10.1007/978-3-319-10377-8_13.

[10] Harapan, H., Itoh, N., Yufika, A., Winardi, W., Keam, S., Te, H., Megawati, D., Hayati, Z., Wagner, A.L., Mudatsir, M., 2020. Coronavirus disease 2019 (covid-19): A literature review. Journal of Infection and Public Health .

[11] Hasan, K.S., Ng, V., 2014. Automatic keyphrase extraction: A survey of the state of the art, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1262–1273.
[12] Hess, D.J., 1997. Science studies: An advanced introduction. NYU press, ISBN: 0814735649.
[13] Hu, B., Ge, X., Wang, L.F., Shi, Z., 2015. Bat origin of human coronaviruses. Virology journal 12, 221.
[14] Huang, A.T., Garcia-Carreras, B., Hitchings, M.D., Yang, B., Katzelnick, L., Rattigan, S.M., Borgert, B., Moreno, C., Solomon, B.D., Rodriguez-Barraquer, I., Lessler, J., Salje, H., Burke, D.S., Wesolowski, A., Cummings, D.A., 2020. A systematic review of antibody mediated immunity to coronaviruses: antibody kinetics, correlates of protection, and association of antibody responses with severity of disease. medRxiv doi:10.1101/2020.04.14.20065771.
[15] Kaur, J., Gupta, V., 2010. Effective approaches for extraction of keywords. International Journal of Computer Science Issues (IJCSI) 7, 144.
[16] Kumar, G., Jeyanthi, V., Ramakrishnan, S., 2020. A short review on antibody therapy for covid-19. New Microbes New Infect doi:10.1016/j.nmni.2020.100682.
[17] Lai, T.S.T., Keung Ng, T., Seto, W.H., Yam, L., Law, K.I., Chan, J., 2005. Low prevalence of subclinical severe acute respiratory syndrome-associated coronavirus infection among hospital healthcare workers in hong kong. Scandinavian journal of infectious diseases 37, 500–503.
[18] Lehrer, S., 2020. Inhaled biguanides and mtor inhibition for influenza and coronavirus. World Academy of Sciences Journal .
[19] Li, B., Si, H.R., Zhu, Y., Yang, X.L., Anderson, D.E., Shi, Z.L., Wang, L.F., Zhou, P., 2020. Discovery of bat coronaviruses through surveillance and probe capture-based next-generation sequencing. Msphere 5.
[20] Li, K., Li, H., Bi, Z., Gu, J., Gong, W., Luo, S., Zhang, F., Song, D., Ye, Y., Tang, Y., 2018. Complete genome sequence of a novel swine acute diarrhea syndrome coronavirus, ch/fjwt/2018, isolated in fujian, china, in 2018. Microbiol Resour Announc 7, e01259–18.
[21] Lorusso, E., Mari, V., Losurdo, M., Lanave, G., Trotta, A., Dowgier, G., Colaianni, M.L., Zatelli, A., Elia, G., Buonavoglia, D., et al., 2019. Discrepancies between feline coronavirus antibody and nucleic acid detection in effusions of cats with suspected feline infectious peritonitis. Research in veterinary science 125, 421–424.
[22] Misra, D.P., Agarwal, V., Gasparyan, A.Y., Zimba, O., 2020. Rheumatologists' perspective on coronavirus disease 19 (covid-19) and potential therapeutic targets. Clinical Rheumatology , 1–8.
[23] Mothe, J., Chrisment, C., Dkaki, T., Dousset, B., Karouach, S., 2006. Combining mining and visualization tools to discover the geographic structure of a domain. Computers, environment and urban systems 30, 460–484.
[24] Mothe, J., Ramiandrisoa, F., Rasolomanana, M., 2018. Automatic keyphrase extraction using graph-based methods, in: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 728–730.
[25] Osabe, Y., Jibu, M., 2018. Introductory chapter: Scientometrics. Scientometrics , 1.
[26] Perricone, C., Triggianese, P., Bartoloni, E., Cafaro, G., Bonifacio, A.F., Bursi, R., Perricone, R., Gerli, R., 2020. The anti-viral facet of anti-rheumatic drugs: Lessons from covid-19. Journal of Autoimmunity , 102468.
[27] Ren, L.L., Wang, Y.M., Wu, Z.Q., Xiang, Z.C., Guo, L., Xu, T., Jiang, Y.Z., Xiong, Y., Li, Y.J., Li, X.W., et al., 2020. Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. Chinese medical journal .
[28] Ronzano, F., Saggion, H., 2016. Knowledge extraction and modeling from scientific publications, in: International workshop on semantic, analytics, visualization, Springer. pp. 11–25.
[29] Sahrawat, D., Mahata, D., Zhang, H., Kulkarni, M., Sharma, A., Gosangi, R., Stent, A., Kumar, Y., Shah, R.R., Zimmermann, R., 2020. Keyphrase extraction as sequence labeling using contextualized embeddings, in: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham. pp. 328–335.
[30] Shibata, N., Kajikawa, Y., Takeda, Y., Sakata, I., Matsushima, K., 2009. Detecting emerging research fronts in regenerative medicine by citation network analysis of scientific publications, in: PICMET'09-2009 Portland International Conference on Management of Engineering & Technology, IEEE. pp. 2964–2976.
[31] Small, H., Boyack, K.W., Klavans, R., 2014. Identifying emerging topics in science and technology. Research policy 43, 1450–1467.
[32] Song, J., Kang, S., Choi, S.W., Seo, K.W., Lee, S., So, M.W., Lim, D.H., 2020. Coronavirus disease 19 (covid-19) complicated with pneumonia in a patient with rheumatoid arthritis receiving conventional disease-modifying antirheumatic drugs. Rheumatology International , 1.
[33] Van Eck, N.J., Waltman, L., 2017. Citation-based clustering of publications using citnetexplorer and vosviewer. Scientometrics 111, 1053–1070.
[34] Wan, Y., Shang, J., Graham, R., Baric, R.S., Li, F., 2020. Receptor recognition by the novel coronavirus from wuhan: an analysis based on decade-long structural studies of sars coronavirus. Journal of virology 94.
[35] Wang, L., Su, S., Bi, Y., Wong, G., Gao, G.F., 2018. Bat-origin coronaviruses expand their host range to pigs. Trends in microbiology 26, 466–470.
[36] Yang, Y.L., Qin, P., Wang, B., Liu, Y., Xu, G.H., Peng, L., Zhou, J., Zhu, S.J., Huang, Y.W., 2019. Broad cross-species infection of cultured cells by bat hku2-related swine acute diarrhea syndrome coronavirus and identification of its replication in murine dendritic cells in vivo highlight its potential for diverse interspecies transmission. Journal of virology 93.
[37] Zhou, P., Fan, H., Lan, T., Yang, X.L., Shi, W.F., Zhang, W., Zhu, Y., Zhang, Y.W., Xie, Q.M., Mani, S., et al., 2018. Fatal swine acute diarrhoea syndrome caused by an hku2-related coronavirus of bat origin. Nature 556, 255–258.