zhang_2021_topic_evolution_disruption_and_resilience_in_early_covid_19_research

Year

2021

Author(s)

Zhang, Yi and Cai, Xiaojing and Fry, Caroline V. and Wu, Mengjia and Wagner, Caroline S.

Title

Topic evolution, disruption and resilience in early COVID-19 research

Venue

Scientometrics

Topic labeling

Fully automated

Focus

Secondary

Type of contribution

Established?

Underlying technique

Cosine similarity topic-term extraction

Topic labeling parameters

\

Label generation

We labelled a descendant-topic via the term with the highest similarity to all other terms in this topic. If the term has been used before, we would choose the term with the second highest similarity, et cetera.

This labelling strategy will select high-frequency terms in early time slices but along with time relatively low-frequency terms will be highlighted. This strategy provides a solution of using a set of labels to comprehensively describe a community, described in Fig. 1–imaging some high-frequency terms representing basic knowledge in the root and some relatively low-frequency terms at the end representing their follow-up evolution.

Due to the use of low-frequency terms, this labelling strategy may result in certain unexpected topics, whose labels could not exactly reflect the main content of their involved articles, because a perfect label for this content has been used by other topics but most of those topics might be their predecessors in the same community.

Table 3 Topic extraction for the pre- and COVID-19 periods

ID	Topic label	Topic description
Pre COVID-19 period (2009–2019)		
1	Epidemiology (996)	Host cells (539), United States (463), infected cells (450), spike protein (393), co-infection (357), Central Nervous System (287), influenza-like illness (252), early stage (251), T cells (243), healthcare workers (228), antibody response (225), S protein (217), host response (203), nucleic acid (200), dendritic cells (190), nucleocapsid protein (181), Receptor-Binding Domain (176), cross-sectional study (175), flow cytometry (173), mammalian cells (172)
2	Viral infection (1482)	Viral replication (714), Saudi Arabia (585), public health (569), viral pathogens (371), viral RNA (368), viral proteins (320), World Health Organization (267), viral genome (264), human health (256), infection control (254), viral load (237), viral entry (228), genetic diversity (220), human infection (195), Intensive Care Unit (191), case report (184), interferon (177), viral diseases (173), PEDV infection (171), health care workers (168)
3	Infectious diseases (1392)	Fever (480), severe disease (417), cell culture (416), disease control (416), clinical signs (408), infectious agents (259), young children (248), feline infectious peritonitis (241), clinical trials (231), Dromedary Camels (219), clinical features (203), control group (199), developing countries (197), human disease (196), West Africa (193), clinical characteristics (189), Clinical presentation (179), IFN-gamma (159), prevention (155), common cold (152)
4	Respiratory viruses (1081)	Respiratory syncytial virus (1061), respiratory infections (462), respiratory viral infections (363), respiratory disease (358), respiratory tract infections (354), Middle East (317), acute respiratory infections (300), respiratory syndrome virus (300), respiratory tract (275), respiratory pathogens (249), acute respiratory distress syndrome (210), Feline coronavirus (203), respiratory virus (203), respiratory symptoms (201), coronavirus infection (182), Bovine Coronavirus (180), respiratory illness (164), human coronaviruses (161), human coronavirus (160), Acute respiratory tract infections (147)
5	SARS-CoV (2370)	Immune response (771), HIV (588), gene expression (321), immune system (321), innate immune response (303), H5N1 (283), South Korea (280), Molecular Mechanisms (270), Monoclonal Antibodies (268), mouse model (256), study period (212), electron microscopy (202), inflammatory response (200), inhibitory effect (183), host immune response (175), molecular characterization (172), adaptive immune responses (166), mathematical model (156), endoplasmic reticulum (150), multiplex PCR (150)
6	MERS-CoV (2403)	Phylogenetic analysis (683), antiviral activity (508), human metapneumovirus (408), animal models (352), causative agent (324), Hong Kong (301), real-time PCR (282), vaccine development (263), ages (233), human population (230), clinical samples (202), crystal structure (201), high mortality (200), age groups (198), human bocavirus (197), licensee MDPI (193), virus-host interactions (189), antiviral effects (180), fecal samples (179), etiological agent (177), mortality rate (177)

Table 3 (continued)

ID	Topic label	Topic description
7	H1N1 (558)	RT-PCR (432), innate immunity (302), Polymerase chain reaction (273), sequence analysis (228), community-acquired pneumonia (197), rapid detection (193), Escherichia coli (187), enzyme-linked immunosorbent assay (182), H7N9 (180), cross-reactivity (173), real-time RT-PCR (171), complete genome sequence (168), complete genome (167), porcine epidemic diarrhea (160), Multiple sclerosis (150), nasopharyngeal aspirates (150), host factors (147), control measures (141), protective immunity (140)
8	Porcine epidemic diarrhea virus (719)	Infectious bronchitis virus (654), influenza virus (639), virus infection (552), RNA viruses (437), virus replication (421), influenza viruses (379), pandemic influenza (309), Ebola virus (271), transmissible gastroenteritis virus (268), mouse hepatitis virus (239), hepatitis C virus (229), avian influenza (211), virus entry (211), Dengue virus (186), Zika virus (186), enveloped viruses (176), influenza virus infections (161), Ebola virus disease (155), virus detection (152), influenza vaccination (149)
COVID-19 Period (2020)		
1	COVID-19 (2235)	COVID-19 outbreak (230), COVID-19 epidemic (127), clinical characteristics (116), United States (75), clinical features (74), mainland China (52), retrospective study (33), clinical manifestations (32), COVID-19 transmission (23), clinical outcomes (22), severe COVID-19 (22), clinical symptoms (21), Hong Kong (20), COVID-19 spread (18), traditional Chinese medicine (16), travel restrictions (16), Chinese government (15), retrospective cohort study (14), modeling studies (13), Case Study (12)
2	SARS-CoV-2 (751)	Disease control (41), healthcare workers (34), common symptoms (27), chest CT (26), Saudi Arabia (24), viral pneumonia (24), Intensive Care Unit (23), CT images (21), Informa UK (21), global spread (20), clinical course (19), clinical practice (18), etiological agent (17), Molecular Mechanisms (17), SARS-CoV-2 outbreak (17), intensive care (16), SARS-CoV-2 pandemic (16), C-reactive protein (14), CT findings (14), viral genome (14)
3	Wuhan (635)	Hubei province (131), fever (99), coronavirus disease (62), confirmed cases (54), mathematical model (50), severe disease (49), coronavirus (41), epidemiological characteristics (39), spike protein (39), phylogenetic analysis (38), immune response (30), personal protective equipment (29), angiotensin-converting enzyme 2 (27), rapid spread (26), porcine epidemic diarrhea virus (21), retrospective analysis (21), severe pneumonia (21), suspected cases (21), severe cases (20), transmission dynamics (20)

Table 3 (continued)

ID	Topic label	Topic description
4	SARS-CoV (254)	South Korea (40), incubation period (32), respiratory infections (31), early detection (24), cardiovascular diseases (21), preventive measures (15), Open Access article (14), co-infection (13), online version (13), viral load (13), high morbidity (12), exponential growth (11), cross-infection (10), Pleural effusion (10), acute respiratory infections (8), bacterial infections (8), Chinese General Practice (8), early identification (8), Feline coronavirus (8), medical countermeasures (8)
5	COVID-19 pandemic (237)	Global pandemic (53), International Concern (51), ongoing outbreak (41), close contact (35), medical staff (33), causative agent (32), median age (30), imported case (24), Coronavirus Pandemic (23), coronavirus Outbreak (22), machine learning (20), healthcare systems (18), mechanical ventilation (17), global concern (14), case definition (13), Monoclonal Antibodies (13), real time (13), age groups (12), illness onset (12), diagnostic tests (11)
6	MERS-CoV (183)	Early stage (43), case fatality rate (30), respiratory syncytial virus (28), early phase (26), host cells (23), Receptor-Binding Domain (22), mortality rate (21), respiratory illness (20), Cytokine Storm (19), infectious bronchitis virus (17), Shanghai Shangyixun Cultural Communication Co. Ltd (17), genome sequence (14), convalescent plasma (13), decision-making (12), intermediate host (12), adverse effects (11), family Coronaviridae (11), family members (11), John Wiley (11), serial interval (11)
7	Epidemiology (112)	Infectious diseases (91), ill patients (36), case report (35), urgent need (35), infected patients (31), clinical trials (30), general population (25), influenza virus (25), Clinical presentation (18), immune system (18), cancer patients (15), infected individuals (14), Clinical management (13), influenza viruses (12), Lopinavir/ritonavir (12), severe illness (12), antibody response (11), HIV (11), Northern Italy (11), pediatric patients (11)
8	World Health Organization (90)	Public health (73), public health emergency (64), viral infection (56), control measures (40), acute respiratory distress syndrome (38), clinical data (33), infection control (30), pregnant women (30), respiratory viruses (30), coronavirus infection (29), global health (28), human-to-human transmission (28), respiratory disease (28), RT-PCR (27), viral replication (27), antiviral activity (26), vaccine development (25), licensee MDPI (24), symptom onset (23), infection prevention (22)

The number following each term indicates the frequency of the term in the given dataset

Motivation

Topic modeling

K-means approach

Topic modeling parameters

Based on the phrase vectors, topic extraction was employed

1. Determine the number of topics *k* and the maximum times of iteration.
2. Randomly initialize *k* phrase vectors as the starting centroids *C* of *k* topics.
3. Assign each phrase vector *v* to its nearest centroid using cosine similarity maximization

$$\text{Cosine Similarity}(v, C) = \frac{v \cdot C}{\sqrt{v \cdot v} \cdot \sqrt{C \cdot C}}$$

4. Recalculate every centroid by averaging all allocated phrase vectors

$$C_i = \frac{1}{Num_i} \sum_{j=1}^{Num_i} v_{i,j}$$

   where *C_i* and *v_i,j* respectively represent the centroid of Topic *i* and the *j* th phrase vector in Topic *i*, and *Num_i* is the total number of phrase vectors in Topic *i*;
5. Iterate Steps 3 and 4 until all the *k* centroids stop moving or the maximum iteration is reached.

Nr. of topics

Label

Term extracted from the topic

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper: Health (COVID-19)

Dataset: Health (COVID-19)

Problem statement

Drawing expectations from resilence theory, this paper explores how the trajectory of and research community around the coronavirus research was affected by the COVID-19 pandemic. Characterizing epistemic clusters and pathways of knowledge through extracting terms featured in articles in early COVID-19 research, combined with evolutionary pathways and statistical analysis.

Corpus

Origin: Clarivate Web of Science (WoS), Elsevier Scopus, PubMed Central, and Dimensions

Nr. of documents: 27,424

Details:

- articles about coronavirus in the 10 years leading up to the COVID- 19 crisis (between January 1, 2009 and December 31, 2019).

Origin: Clarivate Web of Science (WoS), Elsevier Scopus, PubMed Central, and Dimensions

Nr. of documents: 3128

Details:

- articles, notes, letters, and preprints about coronavirus research during the COVID-19 crisis period (between January 1, 2020 and April 23, 2020).

Document

Title and abstract of a scientific article

Pre-processing

- Duplicate removal

- retrieve terms (i.e., multi-word phrases) from the combined field (titles and abstracts), and then performing a term-clumping process to identify core terms by removing noise and consolidating synonyms.

Table 2 Stepwise term clumping process for identifying core terms on coronavirus-related research

Step	Description	#Terms
1	Raw terms retrieved by an NLP function integrated in VantagePoint	601,103
2	Remove meaningless terms, e.g., pronouns, prepositions, and conjunctions	594,116
3	Remove common terms in scientific articles, e.g., "methods"	584,465
4	Remove terms starting with non-alphabetic characters, e.g., "step 1" or "1.5 m/s"	517,502
5	Consolidate terms with specific rules, e.g., abbreviations and related full names	506,283
6	Remove terms appearing in only one record	89,497
7	Consolidate terms with the same stem, e.g., "infectious disease" and "infectious diseases"	81,871
8	Remove single-word terms, e.g., "virus"	68,055
9	Consolidate terms based on given topics, e.g., "MERS" and "MERS-COV"	64,776

- applied the Word2Vec model to the raw text of the combined field and generated phrase vectors by matching core terms and word vectors (each word is represented by a vector, which is the raw output of the Word2Vec model). This set becomes an input in the topic extraction phase.

```
@article{zhang_2021_topic_evolution_disruption_and_resilience_in_early_covid_19_research,
    abstract = {The COVID-19 pandemic presented a challenge to the global research community as scientists rushed to find solutions to the devastating crisis. Drawing expectations from resilience theory, this paper explores how the trajectory of and research community around the coronavirus research was affected by the COVID-19 pandemic. Characterizing epistemic clusters and pathways of knowledge through extracting terms featured in articles in early COVID-19 research, combined with evolutionary pathways and statistical analysis, the results reveal that the pandemic disrupted existing lines of coronavirus research to a large degree. While some communities of coronavirus research are similar pre- and during COVID-19, topics themselves change significantly and there is less cohesion amongst early COVID-19 research compared to that before the pandemic. We find that some lines of research revert to basic research pursued almost a decade earlier, whilst others pursue brand new trajectories. The epidemiology topic is the most resilient among the many subjects related to COVID-19 research. Chinese researchers in particular appear to be driving more novel research approaches in the early months of the pandemic. The findings raise questions about whether shifts are advantageous for global scientific
```

```
progress, and whether the research community will return to the original
equilibrium or reorganize into a different knowledge configuration.},
    author = {Zhang, Yi and Cai, Xiaojing and Fry, Caroline V. and Wu, Mengjia
and Wagner, Caroline S.},
    date-added = {2023-04-13 22:47:24 +0200},
    date-modified = {2023-04-13 22:47:24 +0200},
    day = {01},
    doi = {10.1007/s11192-021-03946-7},
    issn = {1588-2861},
    journal = {Scientometrics},
    month = {May},
    number = {5},
    pages = {4225--4253},
    title = {Topic evolution, disruption and resilience in early COVID-19
research},
    url = {https://link.springer.com/content/pdf/10.1007/
s11192-021-03946-7.pdf},
    volume = {126},
    year = {2021},
    bdsk-url-1 = {https://link.springer.com/content/pdf/10.1007/
s11192-021-03946-7.pdf},
    bdsk-url-2 = {https://doi.org/10.1007/s11192-021-03946-7}}
```

#Thesis/Papers/FS#