Fifteen years of YouTube scholarly research: knowledge structure, collaborative networks, and trending topics

Mohamed M. Mostafa 1 · Ali Feizollah 2 · Nor Badrul Anuar 2

Received: 13 December 2021 / Revised: 21 February 2022 / Accepted: 12 September 2022 / Published online: 19 September 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
Since its inception, YouTube has been a source of entertainment and education. Everyday millions of videos are uploaded to this platform. Researchers have been using YouTube as a source of information in their research. However, there is a lack of bibliometric reports on research carried out on this platform and the pattern in the published works. This study aims at providing a bibliometric analysis on YouTube as a source of information to fill this gap. Specifically, this paper analyzes 1781 articles collected from the Scopus database spanning fifteen years. The analysis revealed that 2006-2007 were initial stage in YouTube research followed by 2008 -2017 which is the decade of rapid growth in YouTube research. The 2017 -2021 is considered the stage of consolidation and stabilization of this research topic. We also discovered that most relevant papers were published in small number of journals such as New Media and Society, Convergence, Journal of Medical Internet Research, Computers in Human Behaviour and the Physics Teacher, which proves the Bradford’s law. USA, Turkey, and UK are the countries with the highest number of publications. We also present network analysis between countries, sources, and authors. Analyzing the keywords resulted in finding the trend in research such as “video sharing” (2010-2018), “web -based learning” (2012-2014), and “COVID -19” (2020 onward). Finally, we used Multiple Correspondence Analysis (MCA) to find the conceptual clusters of research on YouTube. The first cluster is related to user - generated content. The second cluster is about health and medical issues, and the final cluster is on the topic of information quality.

Keywords YouTube · Bibliometric analysis · Co -citation networks · Keyword co -occurrence networks
1 Introduction

Since its acquisition by Google in 2005, YouTube has been a video-sharing social media and a search engine with over 2 billion views per month [41]. It allows users to upload videos and share their content. It is a preferred search engine for contents like cooking recipes because of its audio and visual medium of communication. In addition to watching the videos, users can leave their comments and feedbacks for each video. The combination of audio, video, and comments make YouTube a valuable source of data. Researchers have been using this source of data to analyze various topics across wide range of research domains. One of the research domains that utilizes YouTube is health and healthcare. Educational videos and users’ feedback towards them have been a common research topic. For example, Li et al. [41] examined YouTube as a source of information on COVID-19 pandemic. Khatri et al. [34] also researched YouTube as a source of information on COVID-19 on English and Mandarin content. Hussein et al. [30] evaluated YouTube as a source of information by measuring the information on this platform and by auditing misinformation in videos. Indirectly, some research works developed methods that can be used in YouTube video analysis [42–45].

Analyzing research trends in YouTube papers requires the use of the bibliometric method. Bibliometric is a quantitative analysis of papers published in a specific research domain [46]. The bibliometric study analyzes the authors’ activities, publication trends, and collaborations among institutions and countries. The bibliometric analysis evaluates impact of published papers and reveals the potential gaps and future directions in a research area, which increases interest and attention of researchers and funding bodies. The bibliometric study has been used in many research areas like COVID-19 pandemic [26], agricultural [47], accounting [50], and economic [8]. The advantages of using a bibliometric study are: 1) reveals important research works in a research domain; 2) helps to discover the gaps need to be addressed by researchers; 3) gives young researchers a holistic view of a research area.

To scrutinize research trends and direction on YouTube, this study aims at performing a bibliometric analysis on research works focused on YouTube published between 2006 and 2021. We propose the following research questions to fulfill the aim of this study: 1) what are the trends and directions in YouTube research? And 2) what information can be discovered related to YouTube research? The contributions of this study are as following:

1) we found only one published paper on YouTube bibliometric study, which presents number of papers, citations, and countries that published research works related to YouTube [56]. However, our work presents a more comprehensive analysis of the YouTube papers by providing network analysis, research structure, and thematic mapping.
2) we present a comprehensive network analysis like co-citation network, co-cited sources network, authors’ collaboration network, institutions’ collaboration network, nations’ network, as well as keywords and co-occurrence network.
3) We analyze the trending research and provide a structured research trends as well as thematic and historiographic mapping.
4) we adopt dominance factor, Bradford’s law, and Lotka’s law to analyze the published works using scientific methods.
This article is organized as follows. Section 2 describes the methodology used to carry out the analysis. Section 3 deals with research findings. Section 4 discusses the research findings. The last section deals with research limitations and explores potential avenues for future research.

2 Method

This study is guided by the following four steps:

1. Selecting the database and defining the search terms.
2. Conducting the preliminary statistical analysis.
3. Performing the bibliometric network analysis.
4. Performing the conceptual structure, thematic and historiographic mapping.

To conduct the analysis, the R version 4.1 software [58] was used along with several libraries such as the bibliometrix, wordcloud and ggplot2. For network visualization, we used the VOSviewer software [61]. We discuss here the steps outlined above in some detail.

2.1 Database and documents’ extraction

Following Sigala et al. (2021), the Scopus database was selected to conduct the analysis. As the largest database for peer-reviewed journals (Norris & Oppenheim, 2007), Scopus is frequently used by researchers to conduct bibliometric analysis (Cunill et al., 2019; Hassan et al., 2021). Having selected the database, we extracted bibliographic records related to the selected documents, including relevant information about documents’ titles, authors, and keywords. Retrieved documents were then transformed to a plain text format for further filtering and analysis. Choosing a particular type of document for bibliometric analysis has long been the subject of debate [51, 52]. For instance, journal articles only have been selected in prior studies (e.g., [20]), whereas some authors have focused on both books and journal articles (e.g., [4]), yet others excluded only meeting abstracts, corrections, and editorial material, (e.g., [2]). Here, we opted for peer-reviewed articles only because such articles “usually undergo a meticulous peer-review process and are generally of high quality” ([16], p. 206). To avoid false-positive results, only article titles, abstracts and keywords were searched using the terms “YouTube.” Figure 1 plots the search procedure followed to extract the articles used in this analysis. We limited the selection to documents written in English and we chose 2006 as the date of reference because YouTube was launched in 2006.

Having selected the database, we extracted bibliographic records related to the selected documents, including relevant information about documents’ titles, authors, and keywords. Retrieved documents were then transformed to a plain text format for further filtering and analysis. Choosing a particular type of document for analysis has long been the subject of debate [51, 52]. For instance, journal articles only have been selected in prior studies (e.g., [20]), whereas some authors have focused on both books and journal articles (e.g., [4]), yet others excluded only meeting abstracts, corrections, and editorial material, (e.g., [2]). Here, we opted for peer-reviewed articles only because such articles “usually undergo a meticulous peer-review process and are generally of high quality” ([16], p. 206).

Table 1 shows the main information about the YouTube research data.
The table reveals that 1781 research articles were extracted. The articles were written by 4699 authors, and they include 65,677 references. 417 articles were written by single authors, whereas 567 were written by multi-authors, with a collaboration index of 3.26. This index is calculated by dividing the total authors of multi-authored articles by total multi-authored articles \([23, 36]\). Our result indicates that the average YouTube research team falls between 3 and 4.

![Fig. 1 Schematic flowchart of data acquisition and methodology (Adapted from [15])](image)

The table reveals that 1781 research articles were extracted. The articles were written by 4699 authors, and they include 65,677 references. 417 articles were written by single authors, whereas 567 were written by multi-authors, with a collaboration index of 3.26. This index is calculated by dividing the total authors of multi-authored articles by total multi-authored articles [23, 36]. Our result indicates that the average YouTube research team falls between 3 and 4.

| Description                        | Results         |
|------------------------------------|-----------------|
| Timespan                           | 2006–2021       |
| Documents                          | 1781            |
| Average citations per documents    | 14.89           |
| Average citations per year per doc | 2.494           |
| References                         | 65,677          |
| Document types                     |                 |
| article                            | 1781            |
| Document contents                  |                 |
| Keywords Plus (ID)                 | 4088            |
| Author’s Keywords (DE)             | 4252            |
| Authors                            |                 |
| Authors                            | 4699            |
| Authors of single-authored documents | 417          |
| Authors of multi-authored documents| 4282            |
| Authors collaboration              |                 |
| Documents per Author               | 0.379           |
| Authors per Document               | 2.64            |
| Co-Authors per Documents           | 3.12            |
| Collaboration Index                | 3.26            |
2.2 Bibliometric network analysis

A network can be regarded as “a structure composed of a set of actors, some of whose members are connected by a set of one or more relationships” ([35], p. 8). In social network analysis (SNA), an edge connecting two nodes represents a relationship. Khan and Wood [32] noted that “when used to synthesize the existing literature from a network perspective, the SNA technique can reveal valuable invisible patterns that can certainly facilitate theory development and uncover areas for future research.” There has been extensive prior research using network analysis in areas as diverse as exploring individual scientific collaboration networks [11, 27, 66], collaboration among research institutions [21] and keywords co-occurrence networks [7].

2.3 Thematic and conceptual structure maps

Thematic maps or strategic diagrams were suggested by Law et al. [39]. The map is usually employed to reveal the clusters’ dynamics based on analyzing the keywords or co-word occurrences [29]. The Callon et al. [10] density and centrality metrics are generally used to construct the map. The map also draws heavily on the financial portfolio analysis and concepts based on co-word networks [5]. Due to its usefulness, the map has been used in a plethora of research articles [33, 40, 65]. On the other hand, conceptual structure maps can be employed to investigate the conceptual structure of a research area by breaking down a research domain into clear “knowledge clusters” [63].

3 Results

3.1 Scientific output, core journals and impactful authors

We extracted 1781 Scopus documents related to YouTube. The documents were written by 4699 authors representing 70 nations. Timewise, the documents covered almost fifteen years (2006-2021). Figure 2 plots the scientific output trends in the field. Although the figure reveals an exponential annual growth rate, this rate is not evenly distributed. For instance, in the first two years there was a paucity in YouTube research with only a handful of papers per year. These two years might be referred to as “the initial stage in the YouTube research.” However, the next decade (2008-2017) appears to witness a tremendous increase in research dealing with YouTube. This decade might be called “the rapid growth stage.” Indeed, this period represents the highest growth rate. The final stage (2018-2021) might be called the “consolidation and stabilization stage” because the YouTube research reached the “saturation/maturity” stage. This result is in line with several bibliometric studies conducted in several research areas [53, 66].

Table 2 shows the most important Scopus-indexed journals publishing YouTube research. The table reveals that the most relevant sources publishing YouTube research include journals such as New Media and Society, Convergence, Journal of Medical Internet Research, Computers in Human Behavior and the Physics Teacher. Another way to examine the journals’ influence is known as the Bradford’s law [37]. This law was first proposed by Bradford [9], who noted that “if scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several groups or zones containing the same number of articles as the
Fig. 2 YouTube research annual scientific production (2006–2021)

nucleus.” Fig. 3 plots the Bradford’s law in YouTube research. From the graph, we see that the “core zone” is dominated by just few journals, including New Media and Society, Convergence, Journal of Medical Internet Research, etc. Such journals are considered the outlets publishing the “core” YouTube research.

The YouTube research growth is also evident from the corresponding author’s country involved (Fig. 4).

Table 2 Most relevant sources

| Sources                                      | Articles |
|----------------------------------------------|----------|
| New Media and Society                        | 22       |
| Convergence                                  | 20       |
| Journal of Medical Internet Research         | 19       |
| Computers in Human Behavior                  | 16       |
| Physics Teacher                              | 16       |
| Social Media and Society                     | 16       |
| Plos One                                     | 15       |
| Cyberpsychology Behavior and Social Networking| 11       |
| Journal of Information Technology and Politics| 11       |
| Information Communication and Society        | 10       |
| Health Communication                         | 9        |
| International Journal of Communication      | 9        |
| JMIR Public Health and Surveillance          | 9        |
| Television and New Media                     | 9        |
| First Monday                                 | 8        |
| International Journal of Environmental Research and Public Health | 8 |
| Journal of Cancer Education                  | 8        |
| Journal of Pragmatics                        | 8        |
| Media and Communication                      | 8        |
| Multimedia Tools and Applications            | 8        |
Table 3 shows the most cited articles in YouTube research. The table shows that Smith et al. (2012) paper in the *Journal of Interactive Marketing* is the most cited paper as it was cited 457 times. In this article, the authors compared brand-related user-generated content between three social media platforms, namely Twitter, Facebook, and YouTube. Results provide a general theoretical framework demonstrating how consumer-generated brand communications are influenced by a particular social media channel. The second most cited paper (443 citations) is Lang (2007) paper published in the *Journal of Computer-Mediated Communication*. In this paper, the author employed ethnographic methodology to analyze how
YouTube participants develop and maintain social networks related to video sharing activities. With 333 citations, Susarla et al. (2011) article is the third most cited paper. In this article published in *Information Systems Research*, the authors analyzed the networked structure of interactions on YouTube. Results revealed that “social interactions are influential not only in determining which videos become successful but also on the magnitude of the impact.” (p. 23).

Halpern and Gibbs (2013) paper published in *Computers in Human Behavior* was cited 298 times. In this paper the authors used two social media platforms, namely YouTube and Facebook to examine how social media can be used to foster democratic deliberations. Results showed that the “Facebook expands the flow of information to other networks and enables more symmetrical conversations among users, whereas politeness is lower in the more anonymous and deindividuated YouTube” (p. 1159). Khan’s (2017) paper published in *Computers in Human Behavior* was cited 291 times. In this paper the author investigated motives behind YouTube users’ engagement. Results revealed that YouTube participation is driven mainly by the relaxing/entertainment motive. However, passive content viewing was mainly driven by reading comments posted on the platform. Table 4.

The dominance factor is a bibliometric measure that calculates authors dominance by dividing the number of multi-authored articles in which the author is the first author by the total number of multi-authored articles [38]. This metric has been used widely in the literature [23, 25]. Figure 5 shows the dominating authors over time. From the figure, we see that the most dominating authors were C Basch from 2015 till 2021, Riendeau from 2009 till 2012 and S Azer from 2012 to 2021. Newcomers to the field have also achieved some dominance. Examples include J Yin (2017-2019) and J Park (2016-2021).

In bibliometric studies, “Evenness/concentration of authors’ contribution” is a widely used metric [49]. This metric can be quantified using Lotka’s law (Lotka, 1926). Based on the well-known Zipf’s law, Lotka’s law implies that “the number of authors producing a certain number

| Table 3  | Most cited papers                                                        |
|---------|---------------------------------------------------------------------------|
| Paper   | Total Citations | TC per Year | Normalized TC |
| Smith AN, 2012, J Interact Mark     | 457          | 45.7        | 11.3518       |
| Lange PG, 2007, J Computer-Mediated Commun | 443          | 29.5333     | 6.6976        |
| Susarla A, 2012, Inf Syst Res          | 333          | 33.3        | 8.2716        |
| Halpem D, 2013, Comput Hum Behav       | 298          | 33.1111     | 7.807         |
| Khan ML, 2017, Comput Hum Behav        | 291          | 58.2        | 18.32         |
| Zink M, 2009, Comput Networks          | 270          | 20.7692     | 8.6269        |
| Lee DY, 2013, Comput Educ              | 249          | 27.6667     | 6.5233        |
| Haridakis P, 2009, J Broadcast Electron Media | 221          | 17          | 7.0613        |
| Shifman L, 2012, New Media And Society | 207          | 20.7        | 5.1418        |
| Gueorguieva V, 2008, Soc Sci Comput Rev | 192          | 13.7143     | 5.992         |
| Lee JE, 2016, J Bus Res              | 183          | 30.5        | 8.4662        |
| Syed-Abdul S, 2013, J Med Internet Res | 179          | 19.8889     | 4.6895        |
| Singh AG, 2012, J Rheumatal          | 177          | 17.7        | 4.3966        |
| Pandey A, 2010, Am J Prev Med        | 171          | 14.25       | 4.2993        |
| Briones R, 2012, Health Commun       | 164          | 16.4        | 4.0737        |
| Walther JB, 2010, Hum Commun Res      | 161          | 13.4167     | 4.0479        |
| Jaffar AA, 2012, Anat Sci Educ        | 153          | 15.3        | 3.8005        |
| Steinberg PL, 2010, Urology           | 151          | 12.5833     | 3.7965        |
| Cheng X, 2013, IEEE Trans Multimedia  | 139          | 15.4444     | 3.6415        |
| Poell T, 2012, Journalism            | 133          | 13.3        | 3.3037        |
of articles is a fixed ratio, 2, to single-article authors.” Results suggest that the Lotka’s law seems to hold in YouTube research ($K$-S two sample test $p > 0.05$).

### 3.2 Network analysis

#### 3.2.1 Co-citation networks

A co-citation network is formed when two authors are cited together in a third reference. Figure 6 displays the YouTube research co-cited authors’ network. Based on the color used, the graph

![Figure 5: YouTube authors dominance over the time](image-url)
reveals four distinct clusters. The red cluster includes authors such as J Burgess, M Thelwall and J Green. The size of the node indicates which author occupies a central position in the cluster. Such author(s) might be regarded as influential as they have disproportionate impact on the information diffusion on the network [6]. From the graph, we also see that some nodes are quite close to each other, whereas others drift further away. McPherson et al. [48] argued that closeness signifies a strong “homophily effect,” which occurs when authors in a virtual-room-like environment discuss common topics [24]. In bibliometrics, homophily is an indicator of “disciplinary or thematic similarity” [31]. For example, the nodes representing both R Schatz and A Finamore are very close to each other, indicating possible “homophily effect.”

The green cluster includes authors such as C Basch, J Keelan, A Pandy and S Sarangi. The blue cluster includes sixty-two authors such as J Baker, D Charnock, A Rapp and J Lee. The yellow cluster is the smallest and it includes ten authors such as A Finamore, R Schatz and J Wang. The centrally located authors in each cluster might be regarded as influential authors as they “tend to anchor each community and they have a large impact on other communities as they control and stimulate information diffusion in the network through research activities” ([53], p. 664).

Figure 7 displays the YouTube research co-cited sources’ network. The graph reveals five distinct clusters. For example, the Journal of Clinical Rheumatology, Epilepsy Behavior and the Journal of Cancer Education are co-cited together as they belong to the same cluster. The American Sociological Review is co-cited with Discourse and Society, and Feminist Media Studies. The Journal of Advertising is co-cited with the Journal of Business Research and the Journal of Consumer Research, whereas Body Image is co-cited with the Journal of Pragmatics and Sex Roles. Interestingly, “core” journals occupy central position in the network with a minimal interaction among the distinct clusters, confirming what Glotzl and Aigner [28] term “the orthodox core-heterodox periphery” phenomenon within the field of
YouTube research. Dobusch and Kapeller [22] found that “orthodox journals” tend to be heavily cited, whereas “heterodox journals” tend to be drifted towards the periphery.

### 3.2.2 Collaboration networks

The collaboration network among authors is depicted in Fig. 8. The thickness of the link in this graph is proportionate to articles coauthored, whereas the node size is formed based on the author’s publications. A glance at the graph reveals that the sparse network is formed by seven distinct communities, signifying a limited cooperation among authors. The sparse network implies that impactful researchers in the field work in isolated “silos” [62].

Figure 9 depicts the collaboration network at the institutional level. The thickness of the link is proportional to the institution’s collaboration, whereas the node size is formed based on each institution’s publications. From the graph, we see that there are seven distinct clusters. For example, there is a strong collaboration between Columbia University, the New York University and the William Paterson University in the US. Zou et al. [66] argued that this type of sparse collaboration reflects a “locally -centralized -globally -discrete” cooperation. It also reflects a “North -South” divide, with a clear lack of cooperation between developed/developing world institutions.

Figure 10 shows the collaboration at the nations’ level, with a total of 62 nations collaborating in the scientific production of YouTube research. The figure shows that US tops the world in terms of the total collaboration links, followed by the UK and Australia. A closer look at the graph reveals that some clusters are formed based on geographic distance or linguistic
similarity. For example, Spain cooperates with Colombia, Ecuador and Mexico. The cluster that includes Egypt also includes Kuwait and Saudi Arabia. Figure 11 plots the “geographic atlas” of the countries producing the YouTube research.

3.2.3 Keywords and co-occurrence network analysis

Due to their abstract nature [12], keywords can be used to reveal the content of a paper. Figure 12 shows a simple wordcloud constructed based on the author-provided keywords. A wordcloud plot is an appealing visual tool that can be used to summarize textual data. The size of each word and its closeness to the cloud center determine its significance [42, 43]. From the

**Fig. 10** Collaboration network among nations producing YouTube scholarly research (documents ≥ 2 articles)

**Fig. 11** Geographic atlas of collaboration among nations producing YouTube scholarly research

@Springer
figure we see that the most relevant/frequent keywords used are “Youtbe”, “social media” and “Internet.”

To further scrutinize how frequently keywords co-occur in the same document, we also used the author-provided keywords to construct the YouTube keyword co-occurrence network because “authors of a paper should be the ones that have the best feel as to what areas are spoken to by the paper” [19]. Figure 13 displays the resulting co-occurrence

![Co-occurrence network for author-provided YouTube keywords](image-url)

**Fig. 13** Co-occurrence network for author-provided YouTube keywords
network. The graph reveals eight main clusters. For example, the first cluster in blue deals with medical/health use of the YouTube and includes words such as “health communication”, “health education” and “health information”. The second cluster (green-colored) deals with consumer comments and includes words such as “user-generated content”, “social network” and “Web 2.0”. The third cluster (yellow-colored) deals mainly with the educational use of the YouTube and includes words such as “e-learning”, “medical education” and “online videos”.

A three-field plot, also known as a Sankey diagram, was also used to contextualize the flow trend linking keywords (left), authors (middle) and sources (right). In this diagram the size of the boxes is proportional to the related quantity (keyword, author, or source). Figure 14 displays the YouTube research Sankey diagram. Not surprisingly, edge widths flowing from keywords as “YouTube”, “social media”, and “Internet” are the largest, signifying that such keywords were used by several authors in their publications. We see also see that while some authors have used an extensive list of the keywords reflecting the diversity of their research (C Basch), others used a unique keyword (J Kim).

### 3.2.4 Trending topics and thematic evolution

Figure 15 plots the major YouTube research trending topics. From the graph we see that there is a move from established YouTube topics such as “video sharing” (2010-2018) and “web-based learning” (2012-2014) to new topics such as “COVID-19” (2020 onwards) and “misinformation” (2020 onwards). Such topics might be regarded as “trending topics/hotspots” in the scholarly publications dealing with YouTube because it has been argued that trending topics usually represent hotspots or evolving themes in a specific research domain [13, 14, 54, 60]. Abrupt burst or surge in keywords might be also an indicator of “potential fronts” [57] as “the body of knowledge in a certain discipline can be seen as a sequence of topics that appear, grow in importance for a particular period and then disappear” [18].
3.3 Conceptual structure and thematic maps

We applied the Multiple Correspondence Analysis (MCA) method on the author-provided keywords. The MCA is an extension of correspondence analysis, akin to the Principal Component Analysis (PCA), that helps to analyze the pattern of relationships of categorical data [1]. It was selected since the results of this method is proved to be better on categorical data compared to other methods [1]. Figure 16 depicts the resulting YouTube research conceptual structure over four decades. From the graph, we see that the best dimension reduction achieved for the first two dimensions of the MCA account for roughly 72% of the total variability. In this graph, the closer the dots, the similar the profile they represent, whereas each cluster of dots represents discriminating profiles [64].

An inspection of the graph reveals the depth and breadth of the domain. For instance, the largest red cluster comprises keywords emphasizing the consumer-generated content such as “user-generated content”, “web 2.0” and “online video.” The second cluster (in green) appears to deal with health and medical issues and includes keywords such as “health communication”, “health information” and “misinformation.” The third cluster in blue appears to deal with YouTube research within the context of information quality and includes keywords such as “internet”, “information” and “quality.”

A thematic/strategic map is also shown in Fig. 17. In this graph, average values of both axes are represented by a dotted line dividing the map into four quadrants. Each quadrant in this graph represents a different theme, whereas the bubble size is drawn in proportion to the frequency of documents in which the keywords is used. The first quadrant represents “motor themes” that are well-developed both internally and externally as it is characterized by high density and centrality. [17]. Within the YouTube research, such themes include “user-generated content”, “new media”, “influencers”, and “gender.” The second one is usually
labeled the “highly-developed-and-isolated themes” quadrant as it deals with niche themes. With high-density-low-centrality structure, this quadrant highlights the fact that while the themes it comprises are well-developed internally, they are marginally important externally. Within the YouTube research, such themes include “education,” “medical education,” and “technology.” The low-density-low-centrality third quadrant is termed the “emerging-or-
declining themes” quadrant. This implies that the themes in this quadrant are characterized by weak ties at the internal and external levels. Such themes might indicate potential hotspots in YouTube research. Examples include “COVID-19”, “health communication” and “Twitter.” Finally, the “basic -and -transversal themes” quadrant (low density -high -centrality) comprises themes that are weakly developed in terms of internal ties. Nevertheless, they are characterized by important external ties. Within the YouTube research, such themes include “Social media” and “internet.”

4 Discussion

This study examined published research works related to YouTube between 2006 and 2021. At this point, we can answer the research questions. To answer the first research question about trends and directions, we found that between 2006 and 2008, there was a slow growth in publications since the YouTube platform was new. Then from 2008 to 2017, there was a rapid growth in research on YouTube. Afterwards, the trend is still upward with a slower pace. We also found that the trending topic changes over time. While “gaming” and “video sharing” were trending topics in some time period, the trend shifted towards topics like “COVID-19” and “misinformation”.

The second research question is related to the information discovered from YouTube research. We discovered the most cited papers, authors, and countries with highest number of publications. We also discovered the network between the published works. Specifically, the authors’ collaboration network, collaboration between institutions, and collaboration between countries. We also analyzed the collected works regarding the Bradford’s law and Lotka’s law. It was proved that large number of papers were published in a small group of journals, which followed the Bradford’s law. Also, it was proved that the frequency indexes of author productivity distribution followed Lotka’s law. Additionally, the MCA algorithm was used to find the conceptual structure map related to YouTube papers. The output shows three clusters, consumer-generated content, health and medical issues, and information quality.

Based on this paper’s results, large number of works are related to health and medical issues. Among the institutions, department of public health appeared more than other institutions. Additionally, the journal of medical internet research is in the third spot of the most relevant sources. The MCA algorithm dedicated one cluster for health and medical issues. Furthermore, “medical education” topic started trending in 2014 and is still trending, based on Fig. 15, which is one of the longest trending topics. It is clear that researchers are interested in analyzing YouTube about health-related issues. These points coincide with studies on effectiveness of YouTube videos as a health educational platform. Allgaier mentioned that many people use YouTube as a source of information on science, technology, and health [3]. It is also assumed that because of the sensitivity of health and medical related issues, researchers focused more on the health aspect of the YouTube to find information and misinformation in videos. They analyzed videos and comments to understand users’ feedback on the health-related videos [59].

5 Limitations and future research

Despite the major contributions of this study, it suffers from some limitations. First, we relied only on the Scopus database to conduct our bibliometric analysis. Thus, we unavoidably
commit a selection bias. Subsequently, we believe that future research should test the robustness of our finding by merging several databases such as WoS and Google Scholar. However, it has been argued that the Google Scholar database is less stringent as it comprises citations from unpublished manuscripts, blogs, etc. (Gavel & Iselid, 2008; [55]). Second, we limited the selection of documents to articles published in English. Thus, our results might be limited in terms of coverage [57]. Future research might add other languages to test the generalizability of our findings. Finally, although we conducted a comprehensive study on the whole domain of YouTube research, future research might focus on specific journals publishing YouTube research such as New Media and Society, Convergence, Journal of Medical Internet Research, Computers in Human Behaviour, and the Physics Teacher, among others.

6 Conclusion

This work conducted a bibliometric study on YouTube, as a research topic, in the literature between 2006 and 2021. The search in Scopus database resulted in 1781 research works, which were collected along their meta data such as authors name, keyword, etc. The collected data were analyzed, and the results were presented in the form of network of collaborations between authors, institutions, and countries. We also show the results of networks of keywords. We then created a thematic map based on the keywords to find the trending topic in research related to YouTube. The analysis revealed that 2006 -2007 were initial stage in YouTube research followed by 2008 -2017 which is the decade of rapid growth in YouTube research. The 2017 -2021 is considered the stage of consolidation and stabilization of this research topic. We also found that the trending topic changes over time. While “gaming” and “video sharing” were initially trending, the trend shifted towards topics like “COVID -19” and “misinformation”.

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdi H, Valentin D (2007) Multiple correspondence analysis. Encycl Meas Stat 2(4):651–657
2. Al-Khalifa H (2014) Scientometric assessment of Saudi publication productivity in computer science in the period of 1978-2012. Int J Web Inf Syst 10:194–208
3. Allgaier J (2019) Science and environmental communication on YouTube: strategically distorted Communications in Online Videos on climate change and climate engineering. Front Commun 4(36):1–14. https://doi.org/10.3389/fcomm.2019.00036
4. Aryadoust V, Ang B (Forthcoming) Exploring the frontiers of eye tracking research in language studies: A novel co-citation scientometric review. Comput Assist Lang Learn
5. Ávila-Robinson A, Wakabayashi N (2018) Changes in the structures and directions of destination management and marketing research: A bibliometric mapping study, 2005-2016. J Destin Mark Manag 10:101–111
6. Bakshy E, Hofman J, Mason W, Watts D (2011) Everyone’s an influencer. In: King I, Nejdl W, Li H (eds) Proceedings of the 4th ACM International conference on web search and data mining – WSDM’11. ACM Press, New York, p 65
7. Banckendorff P (2009) Themes and trends in Australian and New Zealand tourism research: A social network analysis of citations in two leading journals (1994-2007). J Hosp Tour Manag 16:1–15
8. Bonilla CA, Merigó JM, Torres-Abad C (2015) Economics in Latin America: a bibliometric analysis. Scientometrics 105(2):1239–1252. https://doi.org/10.1007/s11192-015-1747-7
9. Bradford S (1934) Sources of information on specific subjects. Eng Illus Wkly J 137:85–86
10. Callon M, Courtial J, Laville F (1991) Co-word analysis as a tool for describing the network of interactions between basic and technological research: the case of polymer chemistry. Scientometrics 22:155–205
11. Chen X, Liu Y (2020) Visualization analysis of high-speed railway research based on CiteSpace. Transp Policy 85:1–17
12. Chen C, Song J, Yuan X, Zhang J (2008) The thematic and citation landscape of data and knowledge engineering (1985-2007). Data Knowl Eng 67:234–250
13. Chen C, Hu Z, Liu S, Tseng H (2012) Emerging trends in regenerative medicine: A scientometric analysis in CiteSpace. Expert Opin Biol Ther 12:593–608
14. Chen C, Dublin R, Kim M (2014) Orphan drugs and rare diseases: A scientometric review (2000-2014). Expert Opin Orphan Drugs 2:709–724
15. Chen X, Zou D, Cheng G, Xie H (2020) Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & education. Comput Educ 151:103855
16. Chen X, Zou D, Xie H, Cheng G (2021) Twenty years of personalized language learning: topic modeling and knowledge mapping. Educ Technol Soc 24:205–222
17. Cobo M, Lopez-Herrera A, Herrera-Viedma E, Herrera F (2011) An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. J Inflametrics 5:146–166
18. Colicchia C, Creazza A, Noe C, Strozzi F (2019) Information sharing in supply chains: A review of risks and opportunities with the systematic literature network analysis (SLNA). Supply Chain Manag 24:5–21
19. Corbet S, Dowling M, Gao X, Huang S, Lucey B, Vigne S (2019) An analysis of the intellectual structure of research on financial economics of precious metals. Res Policy 63(101416):101416
20. Corté V, Gaudio G, Sepe F (2018) Ethical food and the kosher certification: A literature review. Br Food J 120:2270–2288
21. Ding Y (2011) Scientific collaboration and endorsement: network analysis of co-authorship and citation networks. J Inflam 5:187–203
22. Dobusch L, Kapeller J (2012) A guide to paradigmatic self-marginalization: lessons for post-Keynesian economists. Rev Polit Econ 24:469–487
23. Elango B, Rajendran P (2012) Authorship trends and collaboration pattern in the marine sciences literature: A scientometric study. Int J Inf Dissem Technol 2:166–169
24. Findlay K, van Rensburg O (2018) Using interaction networks to map communities on twitter. Int J Mark Res 60:169–189
25. Firdaus A, Ab Razak M, Feizollah A, Hashem I, Hazim M, Anuar M (2019) The rise of “blockchain”: bibliometric analysis of blockchain study. Scientometrics 120:1289–1331
26. Gautam P, Maheshwari S, Kaushal-Deep SM, Bhat AR, Jaggi CK (2020) COVID-19: A bibliometric analysis and insights. Int J Math Eng Manag Sci 5(6):1156–1169
27. Glänzel W, Schubert A (2005) Analyzing scientific networks through co-authorship. In: Moed H, Glanzel W, Schmoch U (eds) Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems. Springer, Dordrecht
28. Glotzl F, Aigner E (2018) Orthodox core-heterodox periphery? Contrasting citation networks of economics departments in Vienna. Rev Polit Econ 30:210–240
29. González-Valiente C (2019) Redes de citación de revistas iberoamericanas de bibliotecología y ciencia de la información en Scopus. Bibliotecas Anales de Investigación 15:83–98
30. Hussein E, Juneja P, Mitra T (2020) Measuring misinformation in video search platforms: An audit study on YouTube. Proc ACM Human-Computer Interact 4(CSCW1):Article 048. https://doi.org/10.1145/3392854
31. Jiang Y, Ritchie B, Benckendorff P (2019) Bibliometric visualization: An application to tourism crisis and disaster research. Curr Issue Tour 22:1925–1957
32. Khatri P, Singh SR, Belani NK, Yeong YL, Lohan R, Lim YW, Teo WZY (2020) YouTube as source of information on 2019 novel coronavirus outbreak: a cross sectional study of English and mandarin content. Travel Med Infect Dis 35:101636. https://doi.org/10.1016/j.tmaid.2020.101636
33. Khasseh A, Soheili F, Moghaddam N, Chelak A (2017) Intellectual structure of knowledge in iMetrivs: A co-word analysis. Int J Inf Syst 39:367–397
34. Khan G, Wood J (2016) Knowledge networks of the information technology management domain: A social network analysis approach. Commun Assoc Inf Syst 39:367–397
35. Knoke D, Yang S (2010) Social network analysis. SAGE, Los Angeles
36. Koseoglu M (2016) Mapping the institutional collaboration network of strategic management research: 1980-2014. Scientometrics 109:203–226
37. Kumar H, Dora M (2011) Citation analysis of doctoral dissertations at IIMA: A review of the local use of journals. Libr Collect Acquis Tech Serv 35:32–39
38. Kumar S, Kumar S (2008) Collaboration in research productivity in oil seed research institutes of India. Fourth International conference on webometrics, informatics and Scientometrics. Universitat zu Berlin, Institute for Library and Information Science, 1–18
39. Law J, Baurin S, Courtial J, Wittaker J (1988) Policy and the mapping of scientific change: A co-word analysis of research into environmental acidification. Scientometrics 14:251–264
40. Lee M, Chen T (2012) Revealing research themes and trends in knowledge management: from 1995 to 2010. Knowl-Based Syst 28:47–58
41. Li HO-Y, Bailey A, Huynh D, Chan J (2020) YouTube as a source of information on COVID-19: a pandemic of misinformation? BMJ Glob Health 5(5):e002604. https://doi.org/10.1136/bmjgh-2020-002604
42. Liao H, Tang M, Li Z, Lev B (2019a) Bibliometric analysis for highly cited papers in operations research and management science from 2008 to 2017 based on essential science indicators. Omega 88:228–236
43. Liao X, Yu Y, Li B, Li Z, Qin Z (2019b) A new payload partition strategy in color image steganography. IEEE Trans Circ Sys Video Technol 30(3):685–696
44. Liao X, Li K, Zhu X, Liu KR (2020a) Robust detection of image operator chain with two-stream convolutional neural network. IEEE J Sel Top Signal Process 14(5):955–968
45. Liao X, Yin J, Chen M, Qin Z (2020b) Adaptive payload distribution in multiple images steganography based on image texture features. IEEE Trans on Dependable Secure Comput
46. Liu J, Tian J, Kong X, Lee I, Xia F (2019) Two decades of information systems: a bibliometric review. Scientometrics 118(2):617–643
47. Luo J, Han H, Jia F, Dong H (2020) Agricultural Co-operatives in the western world: A bibliometric analysis. J Clean Prod 273:122945. https://doi.org/10.1016/j.jclepro.2020.122945
48. McPherson M, Smith-Lovin L, Cook J (2001) Birds of feather: Homophily in social networks. Annu Rev Sociol 27:415–444
49. Merediz-Sola I, Bariviera A (2019) A bibliometric analysis of bitcoin scientific production. Res Int Bus Financ 50:294–305
50. Merigo JM, Yang J-B (2017) Accounting research: A bibliometric analysis. Aust Account Rev 27(1):71–100. https://doi.org/10.1111/auar.12109
51. Mostafa M (2015) Do products’ warning labels affect consumer safe behavior? A meta-analysis of the empirical evidence. J Bus Econ Stud 22:24–39
52. Mostafa M (2016) Do consumers recall products’ warning labels? A meta-analysis. Int J Manag Mark Res 9:81–96
53. Mostafa M (2020) A knowledge domain visualization review of thirty years of halal food research: themes, trends, and knowledge structure. Trends Food Sci Technol 99:660–677
54. Neff M, Corley E (2009) 35 years and 160,000 articles: A bibliometric exploration of the evolution of ecology. Scientometrics 80:657–682
55. Neuhaus C, Neuhaus E, Asher A, Wrede C (2006) The depth and breadth of Google scholar: an empirical study. Libr Acad 6:127–141
56. Noruzi A (2017) YouTube in scientific research: A bibliometric analysis. Webology 14(1):1–7
57. Qian J, Law R, Wei J (2019) Knowledge mapping in travel website studies: A scientometric review. Scand J Hosp Tour 19:192–209
58. R Development Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna https://www.R-project.org
59. Teng S, Khong KW, Pahlevan Sharif S, Ahmed A (2020) YouTube video comments on healthy eating: descriptive and predictive analysis. JMIR Public Health Surveill 6(4):e19618–e19618. https://doi.org/10.2196/19618
60. van Eck N, Waltman L (2014) CitNetExplorer: A new software tool for analyzing and visualizing citation networks. J Inf Secur 8:802–823
61. van Eck N, Waltman L (2019) VOSviewer, version 1.6.13
62. Vidgen R, Henneberg S, Naude P (2007) What sort of community is the European conference on information systems? A social network analysis 1993-2005. Eur J Inf Syst 22:317–335
63. Wetstein A, Feisel E, Hartmann E, Benton W (Forthcoming) Uncovering the supplier selection knowledge structure: A systematic citation network analysis from 1991 to 2017. J Purch Supply Manag
64. Wong W, Mittas N, Arvanitou E, Li Y (2021) A bibliometric assessment of software engineering themes. Schools and institutions (2013-2020). J Syst Softw 180:11029
65. Zong Q, Shen H, Yuan Q, Hu X, Hou Z, Deng S (2013) Doctoral dissertations of library and information science in China: A co-word analysis. Scientometrics 94:781–799
66. Zou X, Yue W, Vu H (2018) Visualization and analysis of mapping knowledge domain of road safety. Accid Anal Prev 118:131–145

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.