2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Multiagent-based Youtube Content Discovery Bot

Ishmam Ahmed Solaiman, Nitin Agarwal
COSMOS Research Center, University of Arkansas - Little Rock
Little Rock, Arkansas, USA
{iasolaiman, nxagarwal}@ualr.edu

***Abstract*— YouTube Content Discovery Bot (YTCDB) is a cutting-edge multi-agent system designed to revolutionize the video discovery process. Traditionally, researchers have faced the arduous task of manually sorting through YouTube videos to find relevant content. YTCDB leverages an analytics-driven approach to autonomously locate videos given a seed video. Each task or process within YTCDB, such as comment scraping, gathering statistics, and collecting channel data, can be efficiently handled by one or multiple agents working in tandem. This distributed approach allows for seamless coordination and delegation of tasks, ensuring optimal performance and scalability. Insights gathered from video barcoding and content analysis of the video metadata and transcripts from the initial round of video discovery are used to propagate further exploration and data collection. This feedback loop refines the search criteria to provide a focused search through the endless content on YouTube. This multi-agent system represents a significant advancement in analytics-driven video discovery and in facilitating efficient data collection, analysis, and knowledge extraction in a time-efficient manner.**

Keywords—YouTube data collection, video discovery, multi-agent system, recommendation API, analytics-driven, insights, video content analysis.

I. Introduction

As of 21 August 2023, YouTube has approximately 500 hours of video content uploaded every minute, according to multiple sources.[1] YouTube is the second-most popular search engine after Google, with over 2 billion monthly active users. YouTube viewers watch over 1 billion hours of videos on its platform every day, generating hundreds of millions of videos every day.[5] This makes YouTube a rich depository of video data and a treasure trove of user insights. Data collection from YouTube for research purposes can be a long and arduous task as the content and subject of the video must be analyzed which usually involves spending a considerable amount of time watching videos to verify relevancy. In order to significantly reduce the time overhead of gathering data from YouTube and to provide an analytics-driven approach to collecting YouTube data, this paper proposes a novel software architecture and framework that can ensure rapid collection of data while identifying and exploring relevant search spaces.

ASONAM '23, November 6-9, 2023, Kusadasi, Turkey

ACM ISBN 979-8-4007-0409-3/23/11. . . $15.00
https://doi.org/10.1145/3625007.3627501

This paper highlights a case study to showcase the performance and results of the data collection given a singular seed video. The rest of the paper is organized as follows: Section 2 reviews the literature summarizing related works on YouTube data collection. In Section 3, we follow the journey of the software from the seed video to collecting metadata comments and the analytics that drive the collection. Section 4 discusses the performance and results of the software and showcases the size of data collected and time taken. In Section 5 we conclude the paper and discuss future works and potential improvements of this software.

II. Literature Review

YouTubers need to create fresh video content all the time, and the most consistent channels add new videos at least twice per week. The number of video content hours uploaded every 60 seconds grew by around 40% between 2014 and 2020. YouTube is estimated to reach approximately 900 million users worldwide in 2023[7]. In order to initiate the process of gathering data from YouTube, it is necessary to generate a YouTube Data API key using the Google Developers Console [8]. Fresh API keys for Data access are allocated a quota threshold of 10,000 daily requests, enabling retrieval of public YouTube data. Each interaction with the YouTube API consumes a portion of your daily quota [9]. Kready et. al [3] explore video collection using parallel processing but do not focus on discovering videos, rather focusing on a fixed set of videos and collection time. Other works explore topic-specific crawling [11] using the shark search algorithm but do not use analytics to drive video discovery. This paper proposes a unique framework that uses content from video data such as barcodes and transcripts to expand the discovery search space providing rapid and autonomous video collection and discovery based on an initial seed video or reference keywords. We perform a case study by providing a seed video of the recent Russia–Ukraine conflict and focus on the rapid collection of video data based on the seed to measure the video collection performance.

III. YouTube Content discovery Bot

The YouTube bot is purposefully constructed using a robust microservice pattern, strategically enhancing its operational flexibility and resilience. Embracing a cloud-based environment, the bot seamlessly accommodates multiple instances/agents, each dispersed across the cloud infrastructure. Each of these highly specialized agents is meticulously engineered to excel in distinct tasks, encompassing discovery, transcription, comment scraping, barcode conversion, and topic modeling. Notably, these agents operate autonomously as discrete microservices, thereby cultivating an environment of development, deployment, and scalability autonomy. This strategic approach significantly mitigates the risk of system-wide disruptions, thereby fostering an environment of dynamic iteration and innovation.

B. *Comment Collection*

Delving into a wealth of viewer sentiments and interactions, we transition to comment collection. Employing the YouTube API's *commentsThreads* endpoint, our endeavor is marked by both scope and efficiency, supported by a cadre of dedicated agents. Remarkably, this concerted effort accumulates an astounding 1,264,000 comments in a span of just 62 minutes, solidifying our dataset's breadth and depth.

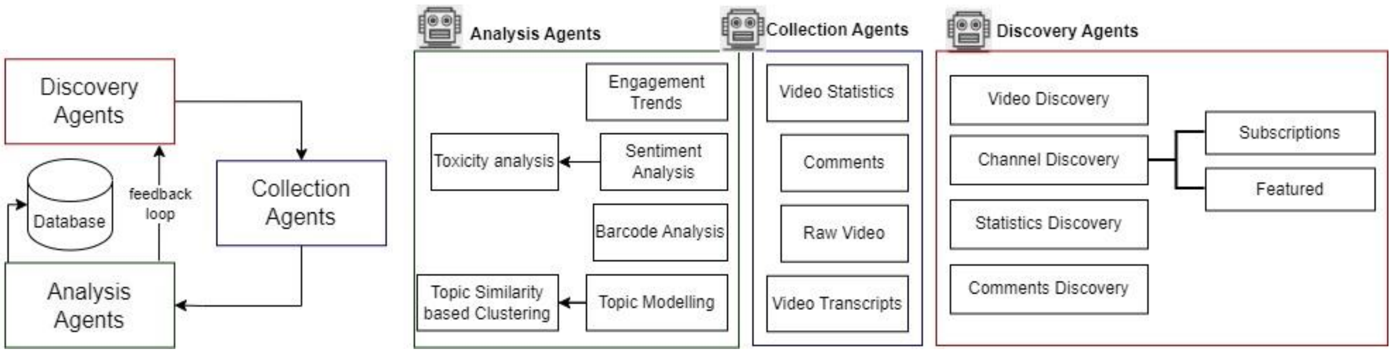

Fig. 1. System Architecture of YouTube Content Discovery Bot (left).

This microservice architecture further guarantees that the performance of any individual service remains insulated from any potential impact on others, thereby cultivating heightened levels of reliability and fault tolerance. Collectively, this meticulously crafted architecture empowers the bot to achieve peak operational efficiency and effectiveness, all while upholding a standard of unparalleled availability and responsiveness.

A. *Data Collection*

Utilizing the YouTube API as our primary data collection tool, our YouTube crawler employs a multifaceted approach to efficient video discovery. The recommendation API is used to swiftly unearth related videos and channels. Our discovery process is optimized through diverse exploration modes. These encompass video-to-video and video-to-channel discovery, dynamically harnessed through the related videos search API. Moreover, channel-to-channel discovery is seamlessly orchestrated via the featured channels and subscription endpoints, while channel-to-video exploration leverages the recent activity of the channels' API. This approach ensures a comprehensive dataset, encompassing critical video metadata such as titles, descriptions, view counts, likes, and comments.

At first the user is prompted to give an example or seed video or to enter a few keywords which are used to perform the first round of discovery. Afterwards the user is presented with analytics and has the option to choose the direction of the search based on content, topics or barcodes. The bot then expands video discovery based on the parameters the user has chosen. For our case study we have chosen the following video as the seed as shown in Fig 3.

Fig. 2 highlights the sequential stages of the bot's journey, ultimately amassing a collection of 2536 videos, 1640 unique, from 624 unique channels in a mere 25 seconds. Subsequently, 1774 featured channels and 589 subscription channels were discovered in 64 seconds.

C. *Audio and Video Collection*

To enable further analysis such as barcode analysis and transcription, we use the PyTube library to download both video and audio files of the collected video which are cached for analysis. The videos are converted to barcodes and the audio is used for transcription purposes when needed. After the analysis process, the videos are purged only keeping the barcodes and transcripts.

D. *Barcode Generation and analysis*

We use *Moviebarcode*[10] to summarize the colors in the frames of the video, we convert the collected videos into a barcode format. After generating the barcodes, we cluster the barcodes to form clusters of similar videos based on their color profile. The generated barcodes are also used for further analysis, such as video classification, similarity analysis, color-emotion analysis and categorization.

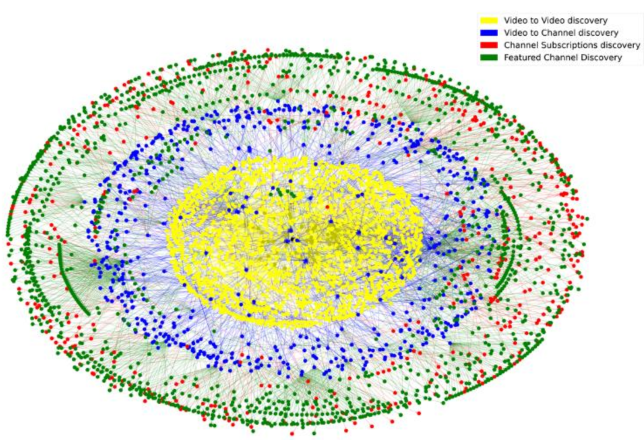

Fig. 2. Videos and Channels discovery path from seed video.

Fig. 3 Example Seed Video for case study

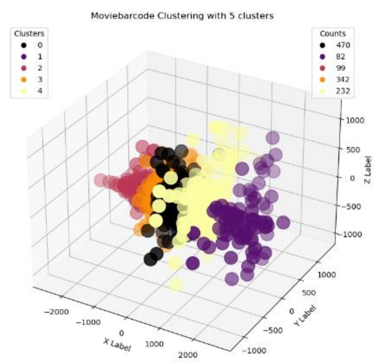

Fig. 4 (a). Clustering of video barcodes

Fig. 4 (b). A video barcode

E. *Transcription and Topic Modelling*

This section outlines the transcription and topic modeling process, which follows the comment collection phase. Leveraging the YouTube API and Whisper AI, we transcribe audio from videos lacking default YouTube transcripts. We use multiple agents to parallelly collect and process data. Processing involves using topic models, autonomously detecting the optimal topics, to classify each of the transcripts which are vectorized then clustered based on their dominant topics.

Topic ID	Word 0	Word 1	Word 2	Word 3	Word 4	Word 5	Topic Count
3	ukrainian	yeah	army	city	air	russians	442
7	yeah	ukrainian	china	army	life	states	142
9	yeah	life	mean	okay	ukrainian	yes	135
6	china	yeah	queen	army	united	ukrainian	118
4	yeah	yes	okay	army	mean	ukrainian	108

Fig. 5. Top 5 words of the top 5 topics

In Fig. 5 we can see the top 5 words of the 5 dominant topics extracted from the transcripts out of a total of 10 topics. Despite having common topic words the type of videos in each topic vary, for example Topic 3 contain videos from official news channels whereas Topics 7 & 9 contain videos of personal interviews or videos of people commenting on the subject. The *Topic Count* column represents the number of videos in the topic *i.e.* Topic 3 has 442 videos whereas Topic 4 has 108 videos. To visualize topic evolution over time, line charts portray topic distribution per video, juxtaposing video publish dates and topic distributions. This affords insight into changing video topics across time intervals. The evolution of topics is further elucidated through a topic stream visualization as seen in Fig 6.

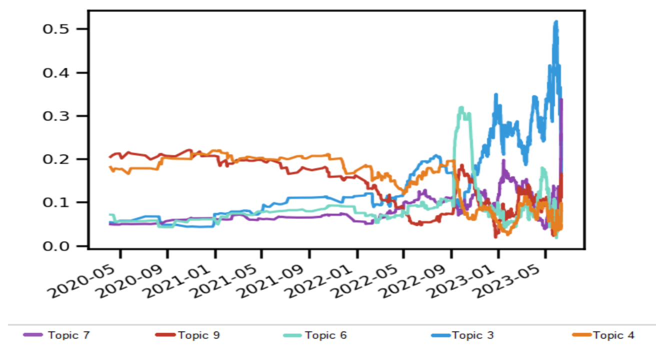

Fig. 6. Topic Stream visualization of top 5 topics

The transcripts are clustered to find similar videos based on the video content. Fig. 7. Shows 2D representation of the transcript clusters. The graph suggests Topics 8 and 4 vary in terms of content compared to the rest of the topics in the dataset.

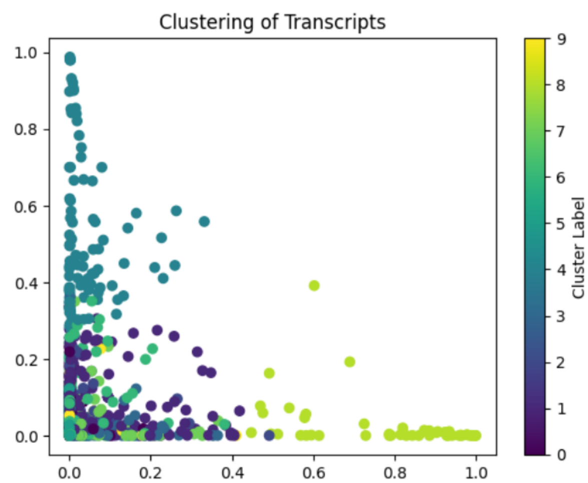

Fig. 7. 2D clustering of Transcripts.

Clustering provides an overview of the videos collected and a high-level view of the different types of videos collected based on their contextual content. In sum, our comprehensive approach, spanning data collection, discovery, barcode analysis, and transcription, empowers our YouTube crawler to proficiently gather and analyze videos pertaining to our focal subject. The user can utilize this information to select an area of interest to further propagate video discovery while leveraging the analytics to gain an understanding of the data without having to watch most of the videos.

IV. Performance

The software amassed a collection of 2536 videos, 1640 unique, from 624 unique channels in a mere 25 seconds. Subsequently, 1774 featured channels and 589 subscription channels were discovered in 64 seconds with 2 hops of video discovery given a single seed video. 1,264,000 comments were collected from these videos providing a treasure trove of data and viewer insights. Of the 1640 videos, 1321 videos were transcribable, and a total of 33.2 GB of audio and video data was collected in the cache during the run. The total run time was 358 minutes to collect videos of 13,265 minutes of videos which amounts to 2.69% of the total playback time.

Data Fields	Time Taken	Data Collected
Video metadata	25s	1640 videos
Channels	64s	2987 channels
Comments	3733s	1,264,000
Video and Audio	14,844s	1420 videos
Transcripts	2851s	1321 videos

Table 1. Data collection performance

V. Conclusion

In summary, this paper introduces an innovative framework for video collection within an analytics-driven, multi-agent autonomous system. This approach accelerates data accumulation significantly, benefiting researchers by expediting video data collection while avoiding exhaustive manual assessment. The framework's adaptability empowers users to guide data exploration, yielding a refined search space for relevant videos. By integrating analytics, the system enhances data comprehension. This framework revolutionizes video discovery, saving time and enhancing insights, and sets the stage for advanced research and knowledge acquisition in video data.

Acknowledgments

This research is funded in part by the U.S. National Science Foundation (OIA-1946391, OIA-1920920, IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of the Under Secretary of Defense for Research and Engineering (FA9550-22-1-0332), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2675, N00014-17-1-2605, N68335-19-C-0359, N00014-19-1-2336, N68335-20-C-0540, N00014-21-1-2121, N00014-21-1-2765, N00014-22-1-2318), U.S. Air Force Research Laboratory, U.S. Army Research Office (W911NF-20-1-0262, W911NF-16-1-0189, W911NF-23-1-0011), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059), Arkansas Research Alliance, the Jerry L. Maulden/Entergy Endowment at the University of Arkansas at Little Rock, and the Australian Department of Defense Strategic Policy Grants Program (SPGP) (award number: 2020-106-094). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support. The authors would like to express sincere gratitude to Tenzing Briggs and Tanzila Sultana Maria for proofreading and editing the manuscript.

References

[1] Alexa. (n.d.). YouTube.com traffic, demographics and competitors. Retrieved from https://www.alexa.com/siteinfo/youtube.com

[2] Negara, E. S., Andryani, R., & Amanda, R. (2021). Network analysis of YouTube videos based on keyword search with graph centrality approach. Indonesian Journal of Electrical Engineering and Computer Science, 22(2), 780-786. https://doi.org/10.11591/ijeecs.v22.i2.pp780-786

[3] Kready, J., Hussain, M. N., & Agarwal, N. (2020, May 22). YouTube data collection using parallel processing. Paper presented at the IEEE Workshop on Parallel and Distributed Processing for Computational Social Systems (ParSocial 2020), New Orleans, Louisiana, USA.

[4] Cakmak, M., Okeke, O., Spann, B., & Agarwal, N. (2023, May 15-19). Adopting parallel processing for rapid generation of transcripts in multimedia-rich online information environment. Paper presented at the 7th IEEE Workshop on Parallel and Distributed Processing for Computational Social Systems (ParSocial 2023) co-located with the 37th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2023), St. Petersburg, Florida, USA.

[5] Oberlo. (n.d.). 10 YouTube stats every marketer should know in 2023 [Infographic]. Retrieved from https://www.oberlo.com/blog/youtube-statistics

[6] Malik, H., & Tian, Z. (2017). A framework for collecting YouTube meta-data. Procedia Computer Science, 113, 194-201. https://doi.org/10.1016/j.procs.2017.08.347

[7] EarthWeb. (n.d.). YouTube, data & statistics, resources: How many hours of video are uploaded to YouTube every minute in 2023? Retrieved from https://earthweb.com/how-many-hours-of-video-are-uploaded-to-youtube-every-minute/

[8] Google Developers. (n.d.). YouTube Data API. Retrieved from https://developers.google.com/youtube/v3

[9] Google Developers. (n.d.). YouTube Data API (v3) - Quota Calculator. Retrieved from https://developers.google.com/youtube/v3/determine_quota_cost

[10] Erol, R., Rejeleene, R., Young, R., Marcoux, T., Hussain, M. N., & Agarwal, N. (2020). YouTube video categorization using moviebarcode. In The Sixth International Conference on Human and Social Analytics (HUSO 2020), Porto.

[11] S. Agarwal and A. Sureka, "Topic-Specific YouTube Crawling to Detect Online Radicalization", International Workshop on Databases in Networked Information Systems, pp. 133-151, 2015.