Systematic Literature Review on finding relevance YouTube videos using tag-based solution

Rahman, Md.Mostafizur	Afroz, Sadiya	Hassan M. Ashraful	Hassan M. Mahmudul
ID: 15-28616-1	ID: 16-32471-2	ID: 15-28410-1	ID: 15-28405-1
mostafiztusher404@gmail.com	sadiyaafroz119528@gmail.com	hassanmashraful@gmail.com	m.mahmudul77@outlook.com

Abstract— **With more than 2 billion active users [10], YouTube is the largest video sharing social media platform which currently not only helping sharing videos but also has become a large source of finance as well. But as the user increased the spam videos [9], [4] also increased which doesn't help neither the user nor the honest creator of videos of that domain.**

In this paper, we have performed and listed our observation based on reviewing several papers both conference and journal papers from IEEE, ACM, Springer Link and retrieve papers from Google Scholar and identified several research questions along with suggestive solutions. We have reviewed papers on different aspects including solution given comment based [2], content based, like dislike based video relevance. We have also tried to identify the datasets been used on those papers along with the used data science models and how those models and algorithms responded with the selected datasets. Our emphasis been given on a slight different domain, mostly how the algorithm and models can be used and improved for novice music video uploaders when an user try to find a new artist with new or unplugged music videos.

Index Terms— **Irreverent YouTube videos, misleading metadata detection, comment classification, tag relevance learning, re-ranking and relevance.**

I. INTRODUCTION

YouTube is the most popular social media for video sharing platform having nearly 2 billion [10] active users every month. YouTube is being used for browsing all sort of video contents teaching, entertainment, food blogging, travel blogging, makeup tutorial, and so on. A new trend also been followed in YouTube, publishing music videos done by independent artists which is helping finding new talents at an increasing length.

In this review paper, we have tried to identify the specific exploration been done on several different published papers along with how they approach a solution, which models been used and improvised, what datasets been used and how the data behaved on those models and algorithms. What more extend exploration can be done with those data, also keeping in mind our meaningless keyword search match using KNN algorithm rather than approaching relevance comment [2], like dislike data, tag-based solution [7], informative comments [5], comment classification solution. Several questions also been raised based on reviewing the papers, exploration extend needed in specific areas also been guided.

II. RESEARCH METHODOLOGY

A. Research Objective

We have done systematic literature review on several IEEE, ACM conference and journal papers along with several retrieved Google scholar papers. It was our aim to identify the current progress on YouTube video retrieval approaches, what algorithms, data models been identified and used on different datasets and how they responded along with the extended exploration can be approached. For efficient retrieval of relevance YouTube videos concerning unplugged music videos, we have reviewed several different papers, identifying several models provided on the papers, discuss about those models and how to improve those models.

B. Research Questions

Category	Research Question	Motivation
Target	Which dimensions of 'YouTube relevant videos' were explored and what are the outcomes of those studies?	Identify the articles according to their study focus and the depth of the studies in each focus area
	What dimension of 'relevant YouTube video' are explored under each study surface?	To determine the specific outcome under each study concentration
	Which learning model shows the better performance on training and test dataset?	To determine the model performances [1], [2], [7] and the performance level for each model
Approach	What approaches been followed in those previous studies?	To identify the general approach followed in those earlier studies
	What datasets or data sources of YouTube been used in those previous studies?	To identify the data sources e.g. Analyzing user comments [2], video retrieval by ranking and relevance [7] for YouTube relevance videos been
	What depth of research been explored for author based Meaningless keyword search [3]?	Determine the specific area that can identify and retrieve the specific music video of and specific artist
	What could be the best-suited dataset for showing relevant artist music video?	Determine the specific area that can influence the video listing on the platform
Target Group	What is the portfolio of datasets analyzed for previous studies and what are their domains?	To determine the response in specific domain according to their used datasets
	How an optimal relevance video can be shown from each or combined solution given in [2], [5], [9]	Determine the evaluation strategy on each of the proposed approach
Outcome	What contributions are made in literature to analyze the dependency in the solution given previous papers?	To explore what results are presented to enhance the result of relevance video listing
	How the result of the articles been validated?	To identify the approaches employed to evaluate the research approaches and study results

III. DISCUSSION

What could be the best-approached strategy for retrieving relevant videos?

On the articles, a number of possible models, algorithms been identified and applied for listing relevant YouTube videos using analyzation of user comment[2], identifying misleading metadata detection [3], informative comments [5], re-ranking and relevance feedback [7] .

What could be the effective datasets for influencing the retrieval of the relevant videos?

On the reviewed papers, models and algorithms are developed based on the datasets containing comments [2], like dislike response, comment response [5], views of the videos, spam detection framework [8] or title based. However, more in depth Meta data exploration could be explored and used. According to the updated explored data, more appropriate models and algorithms can be explored.

How effectively can the videos of a specific artist be identified?

Most of the papers either followed a specific domain e.g. YouTube's educational videos [6] or tried to provide the list of videos based on different tag-based solution [7], [1], analyzing user comments [2]. However, on none of the reviewed papers followed on how can a specific author or artists content be listed.

IV. FUTURE RESEARCH DIRECTION

Provide efficient way to find relevant videos for meaningless keyword search

Here [1] using a method called TBVR with social tag relevance learning considering the imbalance of the tag occurrence frequency. Which helps to find the synonyms of searching keyword and then implemented KNN for finding neighbor nodes. This performs well with meaningful searching keyword. Here we will use meaningless searching keywords and find the occurrences of same word in video title, description, hashtags. Using KNN, we try to find the best-fitted keywords.

Appropriate dataset for meaningless keyword search

Here [2] dataset is giving solution based analyzing user comment. Comment classification and comment summarization may assist well in coding video tutorial contents but for music videos such model or algorithm may be a solution but we must identify the correct dimension related to the user search criteria along with relevant artist.

V. CONCLUSION

This literature review paper has reviewed several papers, papers included given solution with tag based, TBVR [1]

method, user comment analyzing [2], re-ranking and relevance feedback [7], finding misleading metadata [4]. We have tried to identify the extension of work been done on those papers, their data collection and used data, which algorithms and models been used. Several research questions also been proposed in regarding the current progress of those papers with specific motivation along with what more can be explored on the current progress specially on the segment for novice music video artists.

REFERENCES

[1] H. Takeda, S. Yoshida and M. Muneyasu, "Tag-based Video Retrieval with Social Tag Relevance Learning," 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE), Osaka, Japan, 2019, pp. 869-870, doi: 10.1109/GCCE46687.2019.9015338.

[2] E. Poché, N. Jha, G. Williams, J. Staten, M. Vesper and A. Mahmoud, "Analyzing User Comments on YouTube Coding Tutorial Videos," 2017 IEEE/ACM 25th International Conference on Program Comprehension (ICPC), Buenos Aires, 2017, pp. 196-206, doi: 10.1109/ICPC.2017.26.

[3] Palod, Priyank & Patwari, Ayush & Bahety, Sudhanshu & Bagchi, Saurabh & Goyal, Pawan. (2019). Misleading Metadata Detection on YouTube.

[4] Bajaj, Payal & Kavidayal, Mridul & Srivastava, Priyanshu & Akhtar, Md & Kumaraguru, Ponnurangam. (2016). Disinformation in Multimedia Annotation: Misleading Metadata Detection on YouTube. 53-61. 10.1145/2983563.2983569

[5] Choi, S., Segev, A. Finding Informative Comments for Video Viewing. *SN COMPUT. SCI.* **1,** 47 (2020). https://doi.org/10.1007/s42979-019-0048-2

[6] Saurabh, S. & Gautam, Sanjana. (2018). Modelling and statistical analysis of YouTube's educational videos: A channel Owner's perspective. Computers & Education. 128. 10.1016/j.compedu.2018.09.003.

[7] T. Fujii, S. Yoshida and M. Muneyasu, "Video Retrieval by Reranking and Relevance Feedback with Tag-Based Similarity," 2018 IEEE 7th Global Conference on Consumer Electronics (GCCE), Nara, 2018, pp. 1-2, doi: 10.1109/GCCE.2018.8574510.

[8] Samsudin, Nur'Ain & Mohd Foozy, Cik Feresa & Alias, Nabilah & Shamala, Palaniappan & Othman, Nur & Din, Wan. (2019). Youtube spam detection framework using naïve bayes and logistic regression. Indonesian Journal of Electrical Engineering and Computer Science. 14. 1508-1517. 10.11591/ijeecs.v14.i3.pp1508-1517.

[9] Obadimu, Adewale & Mead, Esther & Hussain, Muhammad & Agarwal, Nitin. (2019). Identifying Toxicity within YouTube Video Comment Text Data. 10.13140/RG.2.2.15254.19522.

[10] https://www.oberlo.com/blog/youtube-statistics