A DATA-DRIVEN INTELLIGENT APPLICATION FOR YOUTUBE VIDEO POPULARITY ANALYSIS USING MACHINE LEARNING AND STATISTICS

Wenxi Gao[1], Ishmael Rico[2], Yu Sun[3]

[1]University of Toronto, Toronto, ON M5S, Canada
[2]University of California, Berkeley, Berkeley, CA, 94720
[3]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

People now prefer to follow trends. Since the time is moving, people can only keep themselves from being left behind if they keep up with the pace of time. There are a lot of websites for people to explore the world, but websites for those who show the public something new are uncommon. This paper proposes an web application to help YouTuber with recommending trending video content because they sometimes have trouble in thinking of the video topic. Our method to solve the problem is basically in four steps: YouTube scraping, data processing, prediction by SVM and the webpage. Users input their thoughts on our web app and computer will scrap the trending page of YouTube and process the data to do prediction. We did some experiments by using different data, and got the accuracy evaluation of our method. The results show that our method is feasible so people can use it to get their own recommendation.

KEYWORDS

Machine Learning, data processing, SVM, topic prediction.

1. INTRODUCTION

Nowadays, with advanced technology, people can get the latest news quickly from the Internet. People try to catch on the trend in order to keep themselves from being left behind. On the Internet, people can broaden their horizons. They can keep abreast of current affairs and news, and get all kinds of latest knowledge and information online as well. For example, we can watch videos via YouTube [4]. The trending in YouTube shows the most popular videos in a period of time. It is a great platform for viewers. For people who create YouTube videos, YouTube can bring them not only popularity but also monetary encouragement to make creative videos. However, youtubers may sometimes have no idea of the video theme. Hence, this project is for youtubers who have trouble in deciding the content of the video. Youtubers can find similar trending videos through their own ideas to see if their topic is popular at that time. It provides ideas or materials for making new videos.

For some of the searching systems, like searching in YouTube, users can get popular videos based on the cumulative data of the whole site but not the recent data. Namely, people can get the videos that contain the keywords they entered but these videos might not be the hottest topic at that time. There is a list of trending videos but it will take some time to find what people want to get because they need to go through titles and contents.

David C. Wyld et al. (Eds): BIoT, DKMP, CCSEA, EMSA - 2021
pp. 73-78, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110807

Other techniques such as Twitter [5] can provide ideas but we cannot make sure that these ideas are advancing with the times. People share their thoughts and feelings on the platform with some tags, and different people have different ideas so sometimes they argue. For those who hesitate, it does not help, but instead makes them more entangled in choices of topics.

The difference between our method and other techniques is that we do both recommendation and trend at the same time while others can only meet one condition. In order to combine two functions, searching through trending videos and providing the new ideas, our method is to do searching and make analyses to trending videos. There are some advantages of our project. First of all, the result will only show one of the most related videos. It reduces many choices but provides the best one for reference. Secondly, the video for sure is trending. There is no doubt that the video is very popular at present because our analyses are all based on trending videos. And the third feature is that the result will provide the direct link to the related video so people can quickly get basic information. It saves the time of searching the video on YouTube. The efficiency is improved and our result is also the best.

In our experiment part, we are going to check the accuracy of our result: changing different datasets to check if the method is feasible and changing to another algorithm to compare the results between these two algorithms. First, we use different dataset as our training data and to test another class of data. Second, we make an experiment to check whether another algorithm is better than our method or not. After comparing the result with the fact, we get our percentage of accuracy. The higher of the percentages, the better of accuracy.

The rest parts of the paper are organized as follows: Next section gives the details that challenges and problem we met; Section 3 is about our solutions to the challenges we mentioned in the previous section; Section 4 shows the experiment we did to check how well our method works; Section 5 presents some related work; Section 6, which is the last part of the paper, gives the conclusion and discuss the future work.

2. CHALLENGES

2.1. Challenge 1: Thinking of a Video Topic

Nowadays, high technology is developing rapidly but some challenges come with it. Take YouTube as an example, we know that YouTube is an online platform where people share videos. It is useful for users to search the video but for youtuber, there is no practical tool for youtubers to produce ideas. Here is the first challenge how youtuber can think of an idea for their videos. For example, new youtubers come to this platform and want to make the first video to expand their influence, the first thing is to choose the right topic. However, inspiration comes suddenly, people seldom have any ideas immediately. Even though there are many types of videos in YouTube, this will be a burden for youtubers because there are millions of videos waiting for them.

2.2. Challenge 2: Keeping up the Pace of Trending

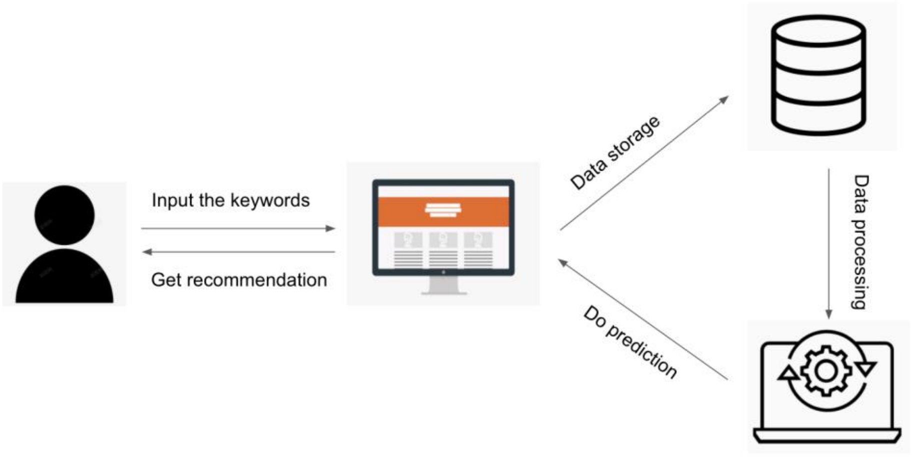
Media workers want to get a lot of attention from the public, but the content is hard to please all people. Sometimes, it is much more efficient to talk about hot spots. In our daily life, we can get news from social media but it changes frequently and fast, so we may have difficulty catching up. Some people may notice the beginning of the event on the news, and only some of them may know the end. Due to the fast speed of updating, people often miss important steps. Therefore, using a browser-like application is very important for people to understand and search popular trends. Moreover, this application has to be easy and concise so that people of all over the ages who know how to surf the Internet can operate easily.

2.3. Challenge 3: Trending Recommendation

There are a lot of videos introducing how to make your video topic on YouTube to attract more fans, they always tell you to choose a hot topic. But now it seems that few browsers can give ideas and catch up the trending simultaneously. For instance, there are many posts on the Internet that teach you how to make your first video but when you read the article, it may have been posted a long time ago. Those topics mentioned in the article might be popular when the article was just released, but now it is not a trend anymore. Another example is that we can see worldwide trending in Twitter but we cannot get recommendations for the contents of tweets. Therefore, the platforms can mostly only do the recommendations of contents or show trends but neither of them at the same time.

3. SOLUTION

This system works in two parts: one part is to search from the trending videos on YouTube, and another part is to analyze the data and give back the class that matches the input the most relatively. Users first type the main topic or some keywords of their video in the input box, and this system will use tfidfvectorizer to convert the text to a matrix with their frequency. Then, the support vector machine (SVM) [6, 7, 8, 9] will find the most related trending videos based on this matrix. Finally, it shows the result of the prediction which is the recommendation of the video.



From the overview diagram, we can see that these four steps are indispensable. Users input the keywords and get their recommendation in the web app. The web app needs YouTube scraping and SVM, and the SVM part needs YouTube scraping and preprocessor to filter the data.

There are four main components in this recommendation system. The web app, where users input their thoughts and get the result back, is one of the main components. In our webpage, there is a predict button, and users can click that button to get recommendations and also, users can see the trending video list. The result will be shown at the bottom.

We need YouTube data so YouTube web scraping is important. It needs to extract useful data from the YouTube website. We use "json" [10, 11] to get the HTML information and from those HTML [12] text, we can get video IDs, short descriptions and titles of the list of trending videos.

The third step is to vectorize the data. Because the raw data cannot be used directly, we need to process the data before our prediction. From scraping the YouTube web, our data is in string type but when we go further prediction, the trained data must be numbers. Therefore, we need to process the data before we use them. After that, our data is ready to make the model.

The last step is SVM. In this part, we are going to use the dataset in the previous part and make a model to predict.

In our web app, we use python flask to make a link to do prediction so that we can use it in our JavaScript. Also, we write DOCTYPE to build the webpage of our prediction program and a style CSS file to make it look functional and nice. There is a trending video list on the page for people to view the trending conveniently. The result will show under the prediction button. That is how our webpage looks like.

For YouTube data scraping, we use some libraries such as requests to get HTML text from YouTube website and json to convert JSON objects into the python dictionary. These two libraries allow us to successfully obtain the information we want from the trending list. We write two text files to save the HTML document and video links in the trending list. From the text files, we can find the information conveniently and cut them out for processing.

In our data processing part, we use TfidfVectorizer [13, 14] from sk-learn library [15] to convert the strings into numbers. Also, in order to make the data frame, pandas help us create the data frame into five columns: video ID, title, length seconds, keywords and short description.

Last but not least, our project needs SVM from the sk-learn library to make the model and do prediction. In this model, X data is vectorized short descriptions and Y data is vectorized titles of videos. When doing the prediction, the text that users enter cannot be used directly, so we need to vectorize the text as well. By these four parts, we can do the prediction.

4. EXPERIMENT

One of the most important things for this program is to give back an accurate result based on input content. To evaluate the accuracy, we need to test if the video ID is the most relevant to what we enter. Therefore, we conduct experiments with two directions: the first one is to test the data we have and the second is to change another algorithm.

For experiment 1, we use video descriptions and ID as our training data and video titles as our testing data. The number of training data in the experiment is the same as the number of training data in the original prediction. The reason why we choose titles as our test data is that each video title corresponds to an ID respectively, we can compare the fact to the outcome to check whether the prediction is correct or not. After that, we can get a percentage of the accuracy.

From the testing, the score that is based on video titles is about 0.6923. In other words, it is 69.23% accuracy to get the video ID if we enter video titles. This result shows that these predictions can still be justified.

In our second experiment, we use a new algorithm called the “KNN (K-Nearest Neighbor) [16]” algorithm. We want to compare the result of SVM and KNN which is more accurate. In our experiment, we take short descriptions and video id as training data and still use videos’ titles to predict. The training data and the test data are the same. We compute the percentage of correct results.

The result of experiment 2 reveals that KNN algorithm is better. From the outcome, we get the score of the prediction using KNN algorithm is about 0.7033. From testing of prediction using the SVM algorithm, we have 69.23% accuracy and for KNN one, it is about one percent more accurate (70.33%).

5. RELATED WORK

Paek, Hye-Jin, et al [1] showed a machine learning approach in the analysis of antismoking video contents on YouTube by using four characteristics of 934 antismoking videos. We are using only trending videos which have different contents on YouTube to analyze the popular video keywords or contents.

Covington, Paul, et al [2] presented a nice deep machine learning approach on YouTube videos recommendation system for viewers. In this work, we know how the recommendation system works. The difference between our project and theirs is that we apply recommendation systems for youtubers, but theirs is faced to viewers.

Victor Roman [3] explored some methods to make the model for the data in machine learning. In this article, we consider the optimization of the training model. Therefore, in our experiment, we tried another model to see if it is more accurate.

6. CONCLUSION AND FUTURE WORK

In conclusion, this recommendation web app is for people who want to make YouTube videos but do not have any ideas. After people type their keywords in the box, the website starts to do the prediction. The web app first collects the data from YouTube, and saves the data into files. The data needs to be processed to use as training data. The last step is to do prediction by the data we collect and to display the information so that people can see the trending video and get the recommended trending video. And we apply our method to do the experiment with different data and algorithms, the result is acceptable. These good results of our experiments demonstrate the good accuracy of our method and it is feasible to search through the trending list and do recommendation at the same time by our project.

Because the system is based on the limited data of trending videos, the keywords entered by people may not match the data at all, resulting in inaccurate results. However, we have no training data if we start a brand-new recommendation system. Moreover, people who want to make YouTube videos always struggle with the content of the video and our recommendation system is for those who do not have ideas on making videos. They can use the system well. Tfidf calculates the word frequency which very relies on the database, but the disadvantage is that keywords do not appear frequently and it cannot reflect the importance of the word in context. Therefore, the optimization of using tfidf is needed to improve our system.

Since the trending video list is changing every time, we could save the data of trending in a period of time so that we have more data to recommend. We can add some numbers to do the classification so that the result by tfidfvectorizer will be more accurate.

REFERENCES

[1] Paek, Hye-Jin, et al. “Content Analysis of Antismoking Videos on YouTube: Message Sensation Value, Message Appeals, and Their Relationships with Viewer Responses.” OUP Academic, Oxford University Press, 5 Oct. 2010, academic.oup.com/her/article/25/6/1085/660720.

[2] Covington, Paul, et al. “Deep Neural Networks for YouTube Recommendations.” RecSys '16: Proceedings of the 10th ACM Conference on Recommender Systems, Sept. 2016, doi.org/10.1145/2959100.2959190.

[3] Roman, Victor. “How To Develop a Machine Learning Model From Scratch.” Medium, Towards Data Science, 2 Apr. 2019, towardsdatascience.com/machine-learning-general-process-8f1b510bd8af.

[4] Burgess, Jean E. "YouTube." Oxford Bibliographies Online (2011).

[5] Murthy, Dhiraj. Twitter. Cambridge: Polity Press, 2018.

[6] Jakkula, Vikramaditya. "Tutorial on support vector machine (svm)." School of EECS, Washington State University 37 (2006).

[7] Wang, Lipo, ed. Support vector machines: theory and applications. Vol. 177. Springer Science & Business Media, 2005.

[8] Noble, William S. "What is a support vector machine?" Nature biotechnology 24, no. 12 (2006): 1565-1567.

[9] Ma, Yunqian, and Guodong Guo, eds. Support vector machines applications. Vol. 649. New York, NY, USA: Springer, 2014.

[10] Nurseitov, Nurzhan, Michael Paulson, Randall Reynolds, and Clemente Izurieta. "Comparison of JSON and XML data interchange formats: a case study." Caine 9 (2009): 157-162.

[11] Pezoa, Felipe, Juan L. Reutter, Fernando Suarez, Martín Ugarte, and Domagoj Vrgoč. "Foundations of JSON schema." In Proceedings of the 25th International Conference on World Wide Web, pp. 263-273. 2016.

[12] Duckett, Jon. HTML & CSS: design and build websites. Vol. 15. Indianapolis, IN: Wiley, 2011.

[13] Kumar, Vipin, and Basant Subba. "A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus." In 2020 National Conference on Communications (NCC), pp. 1-6. IEEE, 2020.

[14] Subba, Basant, and Prakriti Gupta. "A tfidfvectorizer and singular value decomposition-based host intrusion detection system framework for detecting anomalous system processes." Computers & Security 100 (2021): 102084.

[15] Feurer, Matthias, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. "Auto-sklearn: efficient and robust automated machine learning." Automated Machine Learning (2018): 113-134.

[16] Peterson, Leif E. "K-nearest neighbor." Scholarpedia 4, no. 2 (2009): 1883.

AUTHOR

Wenxi Gao now is an undergraduate student in University of Toronto Scarborough in Canada. She studies in the specialist program in Mathematics - Statistics Stream.