Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

# Effect of Sentiment Analysis on YouTube SEO

Prof. Y.K. Sharma
Department of Computer Engineering
V.I.I.T, Pune, India

Atharva Ashtaputre
Department of Computer Engineering
V.I.I.T, Pune, India

Anvay Joshi
Department of Computer Engineering
V.I.I.T, Pune, India

Shah Ali Raza Zaidi
Department of Computer Engineering
V.I.I.T, Pune, India

Vishal Jaiswal
Department of Computer Engineering
V.I.I.T, Pune, India

***Abstract*— Videos are one of the main components of Web 2.0. Videos are depiction of information in a graphical format. YouTube is one of the main platforms on which videos are viewed on Web 2.0. People share their experiences, knowledge and views with the help of such sites. However, it is not easy to fetch valuable information from the various videos in available time, which normally is very short. In this paper, a new approach of ranking the videos on YouTube based on various factors like user interest, views, likes/dislikes, comments etc is introduced. This new method of content curation will improve the knowledge experience of the user. The purpose of this paper or our project is to help students like us who need to get the knowledge from the social media platforms like YouTube to get required data swiftly. Today the problem is that large amount of information is available on the web, but getting right information is difficult. In our case of videos, while getting required content on a video much of the time gets wasted. With our system, what we want is with the help of certain features of a video which are mentioned above, we are aiming to design a system which will provide user top ranked videos of a specified query.**

***Keywords—Content, Curation, Sentiment, YouTube***

## I. INTRODUCTION

YouTube channels are like online diaries handled by an individual, maybe a person or even an organization. The videos share knowledge and information across a wide audience. YouTube videos allow one or more individuals to upload about things they want to share with others. YouTube, as a new platform possesses big differences compared to other social media platforms on the aspect of information updating frequency, organization structure, user connection etc., which has an astonishing power of convergence and penetration. People usually create a video as a hobby to share their information and experience on a subject. The domain of the video uploaded ultimately depends on the user. Entries, on the actual YouTube website are ranked based on a multitude of factors.

The exact technical details of the YouTube algorithm are a closely guarded secret. But some speculations can be made by reverse engineering the algorithm. Initially, YouTube only paid attention to the play button as a factor of determining the ranking of a video. However, views were visualized only by the amount of times the video loaded which ultimately rewarded YouTubers whose videos received a high amount of clicks without any concern of the actual time spent watching the video. Hence, YouTube tweaked their algorithm by rewarding creators whose videos actually had more amount of engagements which provided more accurate measures. Today videos are prioritized not by quantity, but by quality. Statistics such as average watch time better determines the success of a video rather than its view count. Along with the addition to SEO, viewer feedback plays a crucial role in the algorithm. The key thing to remember is that YouTube isn't in the business of judging whether a video is good or not. The YouTube algorithm which is used for ranking gives more focus on the audience interaction by using AI techniques such as neural networks to learn the activity and preferences of each user in addition to their feedback. Using these techniques, it serves the right results at right time. In broad terms, that audience feedback includes:

- What they do (and don't) watch
- How much time they spend watching a video (watch time)
- How much time they spend watching videos during each visit (session time)
- Likes, dislikes, and 'not interested' feedback

*Watch Time*

- When YouTube decided to mothball the view metric, they tweaked the algorithm by considering the factor of the duration of time a video is watched. However, one shouldn't be fooled into thinking that improving your watch time is as simple as creating longer videos. A 30-second video that people watch from beginning to end will rank better than a 10-minute video that only gets watched for a couple of minutes.

## II. PREVIOUS WORK

### *A. Blog Searching and Curating*

Taking in factor of the flaws in current systems, an innovative approach is proposed which results in better searching experience of the user. The proposed model and consists of four major modules: Search Manager, Curator, Personalized Module, and Rating Engine. Working of the proposed model

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

is similar to a mining activity, which mines the blog posts as per the need of the user, and thus the name of the model Blog Miner. The model extracts the blog posts from various blog websites. The beginning module of the proposed model is the model's very own interface named as Blog Miner Interface, through which the rest of the modules are connected. The interface can be used to insert a search query to track blog posts.

*1) Blog Miner Interface:* The beginning module of the proposed model is the model's very own interface named as Blog Miner Interface, through which the rest of the modules are connected. The interface can be used to insert a search query to track blog posts. To interact with the user, Blog Miner Interface consists of four sub-interfaces: Curator Interface, Login Interface, Keyword rate, comment and share the blog posts results of curator module. To access the services of the system, the user has to login with his/her respective id and password. On Login Interface, there are two fields: user name and password. There is one more option of sign up. A search bar is designed which takes a query as an input from the user. Two search bars are given on the home page; one is for local blog posts search result, and second is for the global blog posts search results. The blog posts search results are shown using Post Visualize.

*2) Login and personalization:* As name show, the user logs in using this module. After logging in, the user is provided with functionalities such as sharing the post, adding a comment, rating the blog etc. When user performs any operation on bog post, then his/her name also displays with that blog post. There is an option for Forgot Password and New User. Mail id of user is taken for verification purpose.
After login, user can share, add, comment, rate and modify the blog posts. When user performs any operation on bog post, then his/her name also displays with that blog post. There is an option for Forgot Password and New User. Mail id of user is taken for verification purpose.

*3) Search Manager:* The module is comprised of all the sub processes needed to search the blogs. There are two ways a search can be done: Internally and globally. After accepting a query from the user, the query is broken down in keywords and the search begins for blogs relevant to the keywords. If a user wants to retrieve blogs from a local site, then the internal search process is carried out. The global search is operational if the user wants to retrieve results from all around the world wide web.
Following the search operation, the 'Result Aggregator' module presents the retrieved blog posts. The search process can be carried out on all sorts of data such as structured, unstructured or semi-structured. An algorithm similar to Guoliang Li et al. (2008) is designed which works well on such form of data.

*4) Curator:* The Curator module is basically a method which brings forth the most related posts according to the query of the user. In this system the content curation process is done automatically. The refined posts are stored in the system's database and presented to the user using the visualization module which is the Blog Miner Interface.

*5) Rating Engine:* User can rate the local blog posts. A numerical value is generated based on the ratings received on a post. According to the values generated, the results are sorted in a manner of descending order i.e. best to worst. Data structure used for storing all blog posts is simply a queue. An algorithm is used to sort the blog posts in the descending order. The blog posts are displayed to the user under the tag Top Rated.

*6) Working of blog miner:* General working of proposed model, Blog Miner, is divided into three parts, first is to search the content, second is to curate it, and finally to present it to the user.

### *B. Official YouTube Research Paper:*

*1) System Overview:* The system is built using two neural networks: one which is used for candidate generation and the other for ranking. The candidate generation network uses input from the user's YouTube activity history and returns a small subset of videos. These videos are intended to be relevant to the user. Using collaborative filtering the candidate network provides broad personalization. Using features such as IDs of users who watched video, demography of the users and search tokens similarity between users is derived. Fine-level representation is required to present the best recommendations among the candidates with high recall.
This task is accomplished by the network used for ranking by assigning a score to each video using various features describing video and user. According to the score obtained, the videos are ranked and presented to the user. This approach using two stages helps YouTube to provide recommendations from a huge set of videos while guaranteeing the top results are personalized and engaging to the user. However, for the final determination of the effectiveness of an algorithm or model, we rely on A/B testing via live experiments.

*2) Candidate Generation:* According to the relevance of user, the huge set of videos is shrunk down to a mere hundred videos during candidate generation. The recommendation system's predecessor used a matrix factorization technique and was trained under rank loss. Early iterations of our neural network model mimicked this factorization behavior with shallow networks that only embedded the user's previous watches.

*3) Ranking:* The usage of impression data to specialize and calibrate candidate predictions is the primary role of ranking. Access to a lot of features is available during ranking which specify the user's relationship to the video because the amount of videos being scored are low due to the score obtained in candidate generation. A deep neural network similar to the architecture of candidate generation is used to assign a score to each video using logistic regression. According to the score obtained, a list of videos is then sorted and returned to user. Our final ranking objective is constantly being tuned based on live A/B testing results but is generally a simple function of expected watch time per impression.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

## III. IMPLEMENTATION

Our target was to rank the YouTube videos based on the various parameters such as likes, dislikes, views and sentiment analysis of the comments under the videos so that users get maximum satisfaction by watching them and their time is saved. Our domain was restricted to education only so videos related to education were only considered. For assigning a score we needed a target variable in the form of score. For that we conducted a survey with students and took input from them regarding the likeability of a video and how helpful it was from them. Our dataset consisted of 500 data points. Using this dataset we trained our Machine Learning algorithm to predict future scores of videos based on their values.

As for our algorithm we decided to go with Random Forest Regression as it performed the best compared with other algorithms. Performance can be increased by increasing the training data. The results of all algorithms are discussed later.

Steps:

1] Accept a query from user.

2] Pass this query to a function which uses the YouTube API to retrieve videos from the YouTube video corpus based on keyword-based searching.

3] Fetch all the meta-data of respective video. Meta data includes likes, dislikes, views and comments.

4] Perform sentiment analysis on the comments using the NLTK module and calculate the compound rank of that specific comment. Process on all comments and finally generate a DataFrame which has the features Views, Likes, Dislikes, Sentiment score of each video respectively.

5] Perform data pre-processing on the DataFrame generated such as managing missing values, feature scaling etc.

6] Pass this DataFrame as a vector to the Machine Learning model and generate scores of each video.

7] Present these videos in a descending manner of scores so that the best ranked videos are featured first.

## IV. RESULTS

For evaluating our results, we once again conducted a survey regarding the satisfiability of the order videos are presented to the end user. We got a satisfactory response from the students who expressed that their time was saved as a result of proper ranking of videos.

Below are the performance measures of various algorithms which we tried and tested.

| Sr.no | Algorithm | R2 score |
|---|---|---|
| 1 | KNeighbors Regressor | 0.4572 |
| 2 | Decision Tree Regressor | 0.3762 |
| 3 | Huber Regression | 0.3220 |
| 4 | Support Vector Regression | 0.4784 |
| 5 | Random Forest Regressor | 0.6421 |

Dataset generated from survey: -

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | | View Coun | Likes | Dislikes | Sentiment | Score |
| 2 | 0 | 5352077 | 47830 | 1421 | 0.618841 | 5 |
| 3 | 1 | 613729 | 10884 | 228 | 0.182721 | 3 |
| 4 | 2 | 44819 | 1076 | 29 | 0.357731 | 3 |
| 5 | 3 | 4220694 | 42409 | 2131 | 0.289769 | 4 |
| 6 | 4 | 67450 | 876 | 71 | 0.456059 | 4 |
| 7 | 5 | 422096 | 3337 | 129 | 0.415488 | 3 |
| 8 | 6 | 261838 | 4264 | 98 | 0.188085 | 3 |
| 9 | 7 | 45702 | 1565 | 20 | 0.464236 | 4 |
| 10 | 8 | 108750 | 986 | 34 | 0.388299 | 3 |
| 11 | 9 | 61112 | 432 | 38 | 0.448986 | 3 |
| 12 | 10 | 26038 | 528 | 8 | 0.332632 | 4 |
| 13 | 11 | 60974 | 467 | 41 | 0.447343 | 3 |
| 14 | 12 | 28833 | 807 | 13 | 0.099986 | 3 |
| 15 | 13 | 32295 | 512 | 20 | 0.521138 | 3 |
| 16 | 14 | 25820 | 182 | 21 | 0.269189 | 2 |
| 17 | 15 | 44259 | 465 | 13 | 0.432389 | 3 |
| 18 | 16 | 541829 | 5633 | 224 | 0.238839 | 4 |
| 19 | 17 | 10275 | 150 | 20 | 0.541027 | 2 |
| 20 | 18 | 219384 | 2270 | 51 | 0.256638 | 4 |
| 21 | 19 | 25513 | 308 | 10 | 0.334062 | 3 |
| 22 | 20 | 398176 | 4222 | 198 | 0.58293 | 5 |
| 23 | 21 | 20799 | 285 | 68 | 0.271617 | 2 |
| 24 | 22 | 28616 | 268 | 32 | 0.218809 | 2 |
| 25 | 23 | 36175 | 126 | 13 | 0.263825 | 2 |

## V. CONCLUSIONS

YouTube comes under the category of important web tools. Nowadays, major part of knowledge and recent activities are shared using YouTube. To enhance the performance of YouTube, a new approach has been discussed in this paper, which will surely improve the information searching and knowledge experience of the user.

## REFERENCES

[1] Harsh Khatter Brij Mohan Kalra" A New Approach to Blog Information Searching and Curating".J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Sergiu Chelaru, Claudia Orellana-Rodriguez, and Ismail Sengor Altin- govde" Can Social Features Help Learning to Rank YouTube Videos?" Conference Paper November 2012.K. Elissa, "Title of paper if known," unpublished.

[3] Deep Neural Networks for YouTube Recommendations by Paul Covington, Jay Adams, Emre Sargin Google Mountain View , CA.

[4] Vincent Simonet , Classifying YouTube Channels : a Practical System, May 13-17, 2013.