Predicting Popularity of YouTube Video
Midway Report

Anonymous Author(s)
Affiliation
Address
email

1 Revised Plans

As a result of our feedback on the proposal, as well as a meeting with our TA, we have changed the scope of our project significantlythough we ensured that we could still use our web-crawler data, since the crawler has been running for weeks, now.

Our previous plan was to generate a pandora-radio-like playlist based on a user input seed video, where the list was supposed to appeal to a viewer who enjoyed the seed video. However, since viewer opinion is highly subjective and since evaluating our results would be extremely expensive, we found that this was no practical. As well, were we to merely cluster videos by similarity and evaluate our clustering without considering user enjoyment, the only easily available labeling we could use to evaluate is the set of YouTube-recommended links for each videoand if that is our standard of success, we would at best be mimicking an existing functionality but with far less data to help us.

Our new goal is to successfully predict the popularity (number of views, percentage of likes, and percentage of dislikes) of a youtube video given the rest of its metadata this is both simpler and more easily evaluated than our old goal, and yet should still involve plenty of machine learning and make use of topics covered during the course.

2 Evaluation Metric

We will evaluate our results for each of our three outputs number of views, percentage of likes, and percentage of dislikes against the true numbers associated with each at the time of our crawler's visit. Because views increase over time, we will separately build predictors for different time scales (one day old, one week old, one month old, one year old, older) we believe that this will help significantly, because certain topics of videos are more likely to continue attracting new views over time than others, and because to view them together would be difficult. For the number of views, we will set our loss function equal to the square of the difference between the log of our prediction and the log of the observed value. This will ensure that order of magnitude is more important than precise number (there is more of a difference between 5 and 1000 than there is between 1,000,000 and 1,100,000, in terms of how popular we would say a video is, and we do not want the few extremely popular videos to drown out all differences in the others). For likes and dislikes, we will simply consider a loss given by the square of the difference between the predicted percentage of viewers to like/dislike a video and the observed percentages.

2.1 Goals

Our minimum goal will be to achieve an average loss of no more than 1 for the number of views (we hope to get the correct order of magnitude) and of no more than 0.5

Our stretch goal will be

3 Data Set

We have, for several weeks, been collecting data by crawling YouTube. We initialize the crawler with a video and it explores all other videos that YouTube suggests as being related; periodically we restart the crawler anew with a new starting destination, to ensure a broad sampling. (We began that practice relatively late, so some video categories are more fully explored than others). For each video, we grab the title, uploader, description, upload date, number of views/likes/dislikes, video length, and a number of other attributes, as well as the list of the first 40 videos YouTube recommends as being similar.

There is one additional piece of information that we decided our crawler should collect: the number of subscribers for each user. Since this was not needed for our original proposal, we will need to go back and gather this information, which may take some time considering the vast quantity of videos crawled.

4 First Steps Taken

The data gathering required significant time, and is a big part of the project, but apart from this we had to change most of our plans regarding first steps. Therefore, what we will present in this report is the results of our first attempt to predict the popularity of a video. For this, we have considered the data from just five days of crawling, and we have reduced the number of fields considered to the title, uploader, video length, and upload data. We here describe the process in detail before commenting on our observations and our plans to adjust our technique before running the full dataset.

For all data, we first build a dictionary mapping the uploader to the number of videos they have uploaded and the total number of views there videos have. Before each time we actually use these numbers to make predictions for a video, we subtract the values of the current video from these numbers this is because we do not want videos produced by an uploader who uploaded only one (or very few) videos to be predictable based on this field alone: the idea is to use only such information as would have already been available before the video is uploaded to predict its future popularity.

We then randomly select 80

Features extracted via a bag-of-words model on the title, using Tf-Idf.

- The # of videos uploaded by the uploader, minus 1.
- total # of views for uploader's videos views for this video
- log(total # of views for uploader's videos views for this video)
- (total # of views for uploader's videos views for this video) / (# of videos uploaded by the uploader, minus 1) Special care taken when the denominator is zero: for now, we output the average number of views per video, averaged over all videos.
- log((total # of views for uploader's videos views for this video) / (# of videos uploaded by the uploader, minus 1)) Special care taken when the denominator is zero: for now, we output the average number of views per video, averaged over all videos.
- The runtime of the video, in seconds.
- The age of the video at the time of crawling, in days.

Once the training is finished, we run the other 20

5 Results

6 Next Steps

7 Related work

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

References follow the acknowledgments. Use unnumbered third level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to 'small' (9-point) when listing the references. **Remember that this year you can use a ninth page as long as it contains *only* cited references.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D. S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609-616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer-Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.