June 2021

Clustering of Curve Types using Similarity Scores

Aviv Reznik

Bart van Delft

Clustering of Curve Types using Similarity Scores

ABSTRACT

Curves of trends of content interaction events (e.g., views, shares, likes, etc.) versus time are of interest to social media and content-sharing services. Clustering trends can be used to classify, profile, and understand content creators or influencers. Clustering can also be used to recognize abusers, e.g., users that attempt to generate fraudulent views or likes in an effort to boost advertising revenue. This disclosure describes techniques to define similarity functions, or equivalently, distance measures, to enable the clustering of trends into trend types.

KEYWORDS

- Trend clustering
- Trend analysis
- Creator profile
- Similarity function
- Similarity score
- Distance measure
- Curve type
- Machine learning
- Statistical classification

BACKGROUND

Social media posts, videos, and other online content can experience a variety of growth trends. For example, a particular content item may enjoy a rapid growth followed by a plateau; it may experience a slow but steady growth; it may experience an initial spike in growth and after a time gap (e.g., a year), another spike due to renewed interest in its content; etc. Curves of trends

versus time are of interest to social media and content-sharing (e.g., video, photo, blog, etc.) services.

Clustering trends, e.g., identifying the trend type and assigning a given curve to a trend type, can be useful to classify, profile, and understand content creators or influencers and can also be used to recognize abusers, e.g., users that attempt to generate fraudulent views, likes, or other content interactions in an effort to boost advertising revenue. For example, as abusers seed fraudulent clicks with the help of bots, growth trends for their content can exhibit telltale and recognizable signatures. A given abuse trend can help identify a particular abuser, enabling the content hosting/sharing service to act upon all instances or userids of the identified abuser. Abusers with a certain view pattern can be clustered together, which can be helpful for abuse classification.

DESCRIPTION

This disclosure describes techniques to define a similarity function, or equivalently, a distance measure, to enable the clustering of trends into trend types.

Mapping curve types to vectors of keywords: Per this technique, a trend type is labeled, e.g., as "sharp increase," "steady, slow increase," etc. In doing so, a curve can be divided into sections in time, each section being labeled by an integer depending on its growth rate, such that a curve is mapped to an *N*-length vector.

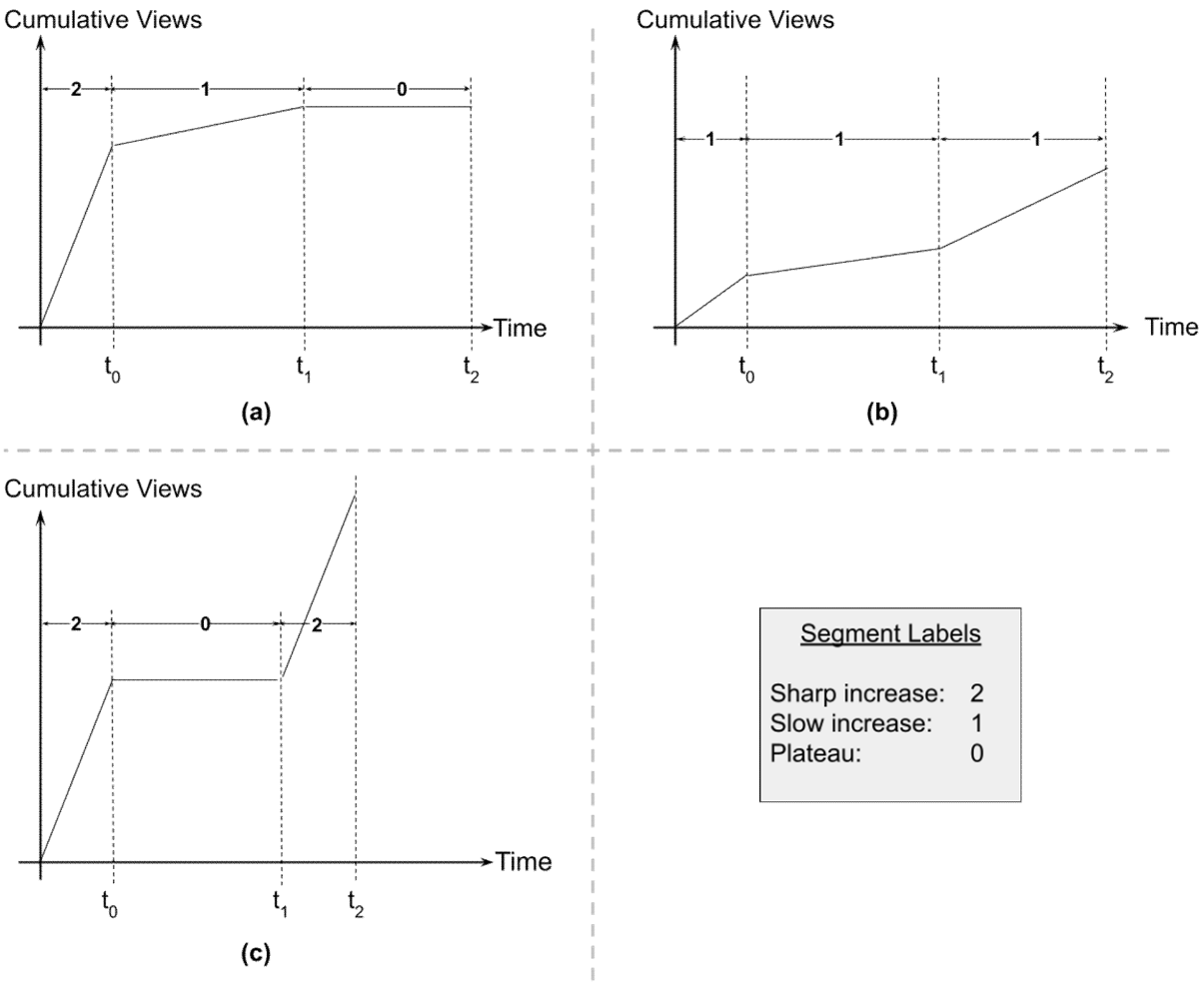

Fig. 1: Mapping trends to *N*-length vectors

For example, as illustrated in Fig. 1, which uses $N = 3$, a sharp increase might be assigned the label '2'; a slow increase, '1'; a plateau, '0'; etc. A trend of a sharp increase, a slow increase, and a plateau, as in the example of Fig. 1(a) is assigned the vector [2, 1, 0]. A trend of continuous slow increases, as in the example of Fig. 1(b) is assigned the vector [1, 1, 1]. A trend of a sharp increase, a plateau, and a sharp increase, as in the example of Fig. 1(c) is assigned the vector [2, 0, 2]. The curves can be time-averaged and normalized suitably along the X- and Y-axes prior to labeling. The points t_0, t_1, …, t_{N-1}, where the curves exhibit change in slope can be automatically determined using standard mathematical techniques.

In this manner, a curve or a chart is turned into a vector of keywords. The vector of keywords effectively serves as a similarity (or distance) measure. Machine learning, pattern recognition, or statistical classification techniques can then be used to cluster trends. For example, the Euclidean distance between vectors corresponding to trend types can be used to cluster trends.

Direct training of a machine learning model

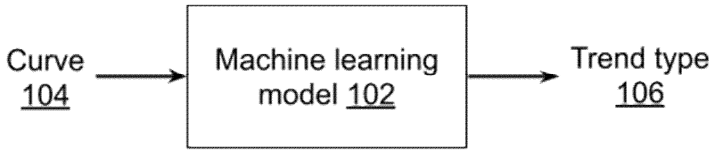

Fig. 2: Trend classification by training of a machine learning model

Per this technique, illustrated in Fig. 2, a corpus of curves (104) is provided as input to a machine learning model (102), which is trained to classify the curves into types of trends (106). The training of the machine learning model can be performed via supervised learning. Training data used to train the model can include human-assigned labels that identify curves that are similar.

Alternatively, training can be unsupervised, e.g., curves can be clustered automatically by the machine learning model. For example, to construct the ground truth, the views-chart on two videos of the same channel uploaded close in time to each other can be considered similar (potentially with a simple heuristic that ensures there is indeed some similarity). Two arbitrarily selected videos can be considered dissimilar. This procedure effectively defines a similarity function. When training for the similarity function, embeddings can be produced. The similarity function and the embeddings can be used to cluster curves with standard machine learning techniques.

By clustering similar curves, it is possible to identify abuse trends, as well as to identify different content creator profiles that help the content hosting/sharing service to better understand the users that contribute content items to the service.

CONCLUSION

This disclosure describes techniques to define similarity functions, or equivalently, distance measures, to enable the clustering of trends into trend types.

REFERENCES

[1] Al-Rfou, Rami, Dustin Zelle, and Bryan Perozzi. "Systems and Methods for Determining Graph Similarity." U.S. Patent Application Publication Number 20200334495A1, published October 22, 2020.

[2] UW Interactive Data Lab, "GraphScape: Modeling Similarity & Sequence among Charts" accessed Jun. 2, 2021.