July 2020

Method to Identify Change based on Content-based Embedding

Luca Chiarandini

Joe Hung

Method to Identify Change based on Content-based Embedding

Content sharing platforms allow users to access and consume media content, as well as allow users to store and share media content with other users. The media content may include video content, audio content, or other forms of media. The content sharing platform may provide users with one or more personal channels, which users can customize by uploading content or linking media (e.g., videos). Users can visit other users' personal channels to view the content hosted by others and express opinions on the content (e.g., likes, dislikes, comments, etc.). Additionally, users can subscribe to other users' personal channels, which can result in the subscriber receiving notifications when the owner of the subscribed-to channel uploads new content.

Content sharing platforms may track how many users view each content item and how many users subscribe to each personal channel, and based on these numbers may provide users with monetization opportunities. For example, a content sharing platform may enable advertisements to be displayed on a personal channel that has over a certain number of subscribers, or may enable advertisements to be displayed alongside a content item that has over a certain number of views. A user who uploads a content item that has advertisements enabled, or a user whose personal channel has advertisements enabled, may generate income via ad-based revenue streams. In order to generate income via ad-based revenue streams, users may apply to join the monetization program. In the application process, the user may be provided with an explanation of what the program entails and what kind of examination may be performed. If a channel of a channel owner satisfies predefined requirements and/or policies of the content sharing platform, the channel owner may receive access to monetization features (ad-based revenue streams, subscription fees, sale of merchandise, etc.) for a respective channel. A channel that is associated with enabled monetization features is referred to herein as a "monetized channel."

If the user prefers not to be subjected to the inspection detailed in the agreement, the user may choose not to enter the monetization program.

Some users who are accepted into the monetization program may exploit the policies related to a monetized channel. For example, policies related to a monetized channel may require a user not to upload certain disallowed content, such as content that infringes copyrights, or content including inappropriate material. To exploit this policy, a user may create a personal channel, upload appropriate content until the channel meets the requirements and policies to become a monetized channel, and then upload disallowed content that no longer meets said policies after having been accepted into the monetization program. The disallowed content may be popular among viewers (e.g., a copyrighted movie), resulting in increased ad-based income to the user uploading the disallowed content. However, as a consequence of the content switch, advertisers are unintentionally sponsoring content that may not be otherwise allowed. Identifying this upload pattern may reduce the amount of bad actors in the monetization program and make more advertising money available to users who are following the policy requirements.

Aspects of the present disclosure address the above and other deficiencies by proposing a method capable of identifying changes based on content-based embedding. Specifically, the proposed method detects changes in types of content uploaded to a user's personal channel by comparing the videos uploaded before the channel was approved for monetization to the videos uploaded after the channel was approved for monetization. The method described herein is not limited to monetization of content sharing platforms, but may be applied to any situation that requires identifying changes in behavior before and after a certain date and/or time.

The method described herein may automatically identify channels in which the types of content items uploaded before the channel was last reviewed are different from types of content items uploaded after the channel was last reviewed. This change in uploaded content types may indicate content-switcher abuse, which should be sent for additional administrative review. The method may use a parametric approach to identify changes in a user's channel. The method may start by identifying the types (or categories) of content items uploaded to a user's personal channel before the channel was reviewed and the types (or categories) of content items uploaded to the user's personal channel after the channel was last reviewed. The content types or categories may include, for example, disallowed content (e.g., content that infringes copyrights, content including inappropriate material) and allowed content (any other type of content). The method may then compute a channel risk score, which measures the amount of change in content since the last time the channel was reviewed and accepted into the monetization program. The higher the risk score, the more likely it is that the user has changed the types of content items uploaded to the user's personal channel, which may indicate content-switch abuse.

Figure 1 illustrates a flow diagram of a method for identifying change based on content-based embedding. At block 102, the method computes the content-based similarity of content items uploaded by a user to the user's personal channel. This can be based on a number of similarity measures. For example, the method may look at consumption patterns across users and content items. The method may estimate the likeliness of two content items being co-watched by the same users using content-based embedding. Content-based embedding may be described as grouping together content items that share similarity factors. For example, each content item can be placed along a vector based on a similarity factor. One example of a similarity factor is based on the users who consume the content item. The content items that are viewed by one group of people can be considered to be of the same type and be placed in one area along the vector, while content items that are viewed by another group of people may be considered to be of a different type and be placed somewhere else along the vector. Alternative similarity measures include metadata similarities, content similarities, or content ID matches based on the content ID provided by the content sharing platform. For illustration purposes, the similarity between a content item *a* and a content item *b* may be obtained from the following function: $sim(a, b) \rightarrow [0, 1]$.

At block 104, the method may identify the type(s) of content items uploaded to a user's personal channel before the channel was last reviewed (called the PRE grouping) and the type(s) content items uploaded to a user's personal channel after the channel was last reviewed (called the POST grouping). If content items in the PRE grouping are of similar type(s) and the content items in the POST grouping are of similar type(s), but the type(s) of content items in the PRE grouping are different from the type(s) of content items in the POST grouping, this may indicate content-switcher abuse.

Figure 2 illustrates an example of content-switcher abuse. In figure 2, the X's represent content items uploaded to a user's personal channel before the channel was last reviewed (before time T), while the O's represent the content items uploaded to the user's personal channel after the channel was last reviewed. As represented in the illustration, the X's are similar to each other, and most of the O's are similar to each other, however the group of X's and the group O's are not similar to each other. This may indicate that the user changed the type(s) of the content items uploaded to the user's personal channel.

The method provides for a number of ways to determine which content items to group in the PRE grouping (i.e., the content items uploaded before the channel was last reviewed, illustrated as X's in figure 2), and which content items to group in the POST grouping (i.e., the content items uploaded after the channel was last reviewed, illustrated as O's in figure 2). For example, the PRE grouping may contain the N most recent content items before the channel was last reviewed, and the POST grouping may contain the N most recent content items after the channel was last reviewed (where N is an integer larger than 0). In another example, the PRE grouping may contain the N most recent content items before the channel was last reviewed while the POST grouping may contain the oldest content items after the channel was last reviewed. In another example, the PRE grouping may contain the content items with the highest consumption time (e.g., watch time) that sum up to at least a threshold percentage of total watch time of the channel before the channel was last reviewed, while the POST grouping may contain the content items with the highest watch time that sum up to at least a threshold percentage of total watch time of the channel after the channel was last reviewed. In another example, the PRE grouping may contain the content items with the highest number of views that sum up to at least a certain percentage of total views before the channel was last reviewed, while the POST grouping may contain the content items with the highest number of views that sum up to at least a certain percentage of total views after the channel was last reviewed. In another example, the PRE grouping may contain the M content items before the channel was last reviewed in descending order of pairwise similarity (e.g. take the two most similar content items, then take the next two most similar content items, etc.), while the POST grouping contains the M content items after the channel was last reviewed in descending order of pairwise similarity (where M is an integer higher than 0). There may be many various ways to form the PRE and POST groupings other than those listed above.

At block 106, the method may compute the overall similarities of the PRE and POST groupings. There may be many ways to compute the overall similarities, including but not limited to computing the average pairwise similarity, the median pairwise similarity, and/or the max similarity (or minimum distance).

The average pairwise similarity considers every possible combination of content items in PRE and content items in POST and computes the average similarity. Average pairwise similarity may be expressed by the following equation:

$$\frac{\sum_{a \in A} \sum_{b \in B} sim(a, b)}{|\mathrm{A}| \cdot |\mathrm{B}|}$$

where A represents the PRE grouping, and B represents the POST grouping.

The median pairwise similarity may be the same as the average pairwise similarity but using median instead of the average. Using median instead of average may reduce the contribution of outliers, which may result in a more reliable calculation.

Another example for computing the overall similarities between the PRE and POST groupings is computing the maximum distance between the content items that are most dissimilar. For example, with regard to Figure 2, instead of considering all possible pairwise comparisons between all the X's and all the O's, this example computes the maximum distance between the two groups. In figure 2 for example, this example would compare the X on the very left (210) to the O on the bottom right (212). The max similarity may be express by the following equation:

$$max_{a \in A, b \in B} \, sim(a, b)$$

At block 108, the method may compute the channel risk score based on the overall similarities of the PRE and POST groupings. The method may compute the channel risk score using the following equation:

$$risk(PRE,\ POST) = \frac{sim(PRE,PRE)sim(POST,POST)}{sim\ (PRE,POST)^2}$$

A high risk score indicates that the channel's content items in the PRE and POST sets are very self-similar, but at the same time the PRE set is far apart from the POST set (as illustrated in figure 2, for example). A high channel risk score may indicate that the content items uploaded before the channel was last reviewed are different from the content items uploaded after the channel was last reviewed.

At block 110, the method may determine whether to send the channel for review based on the channel risk score. The higher the channel risk score, the more likely the content uploaded to the channel has changed since the last review. The method may compare the channel risk score to a threshold in order to determine whether to flag the channel for additional review. Additionally or alternatively, the method may rank the channels by their risk score and flag the top M channels that have the highest risk scores, where M is an integer higher than 0. The channel may be reviewed by an automated system, a human reviewer, or a combination of the two.

Abstract

A method for identifying change based on content-based embedding is disclosed. The proposed method computes content similarity of content items uploaded to a user's personal channel. The method then computes a channel risk score, which identifies whether the type(s) of content uploaded to a user's personal channel has changed after a specific point in time. A channel in which the type(s) of content has changed after a specific point in time, for example after the channel was reviewed for a monetization program, may indicate potential abuse of the monetization program. Based on the risk score, the method may flag potentially abusive channels for administrative review, which may be reviewed by an automated system, a human reviewer, or a combination.

Keywords: content switch abuse, monetization abuse, content sharing platforms, channels, inappropriate content, co-watch similarity, administrative review, channel risk score, content-based embedding

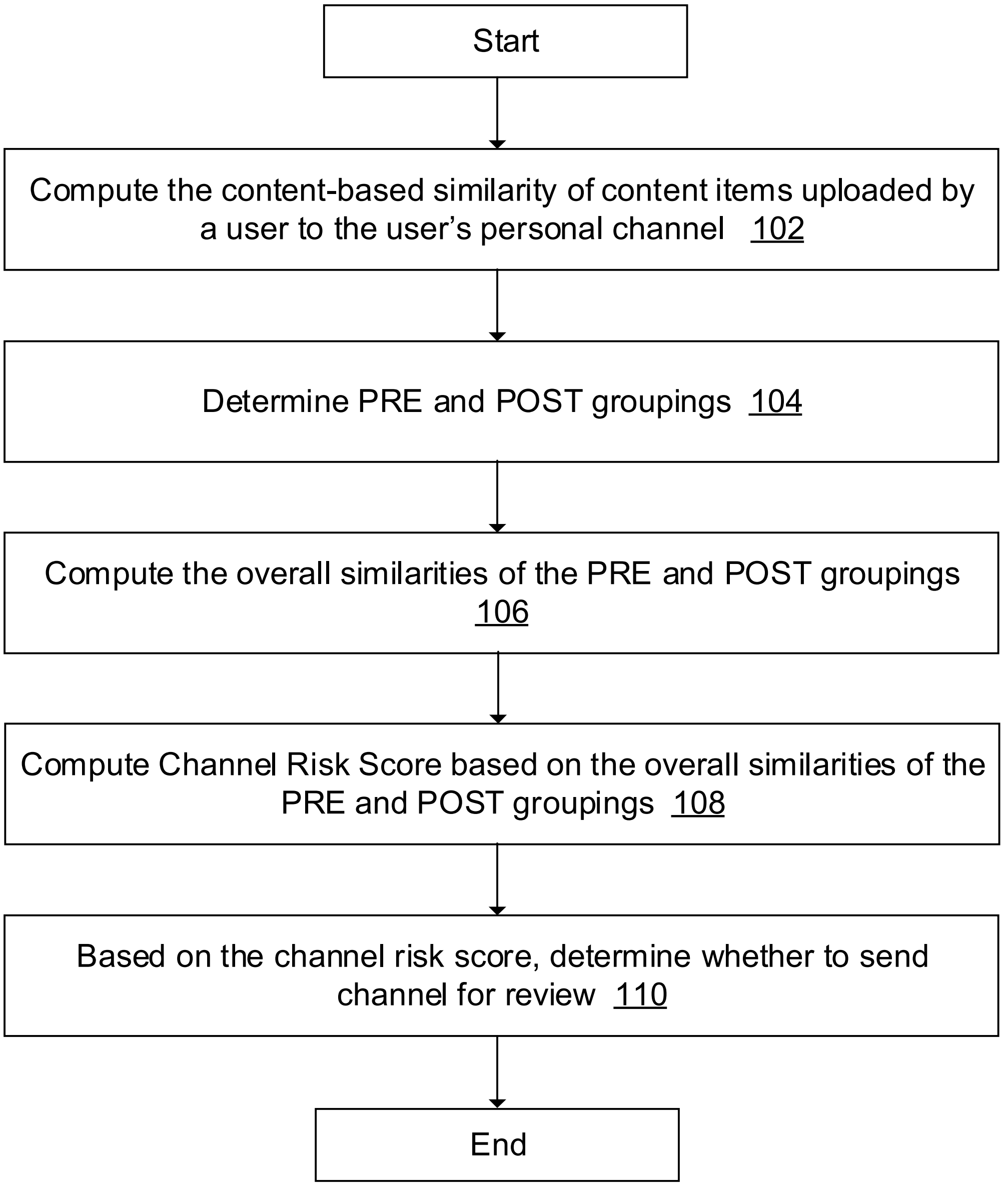
Start
Compute the content-based similarity of content items uploaded by a user to the user's personal channel 102
Determine PRE and POST groupings 104
Compute the overall similarities of the PRE and POST groupings 106
Compute Channel Risk Score based on the overall similarities of the PRE and POST groupings 108
Based on the channel risk score, determine whether to send channel for review 110
End

Figure 1

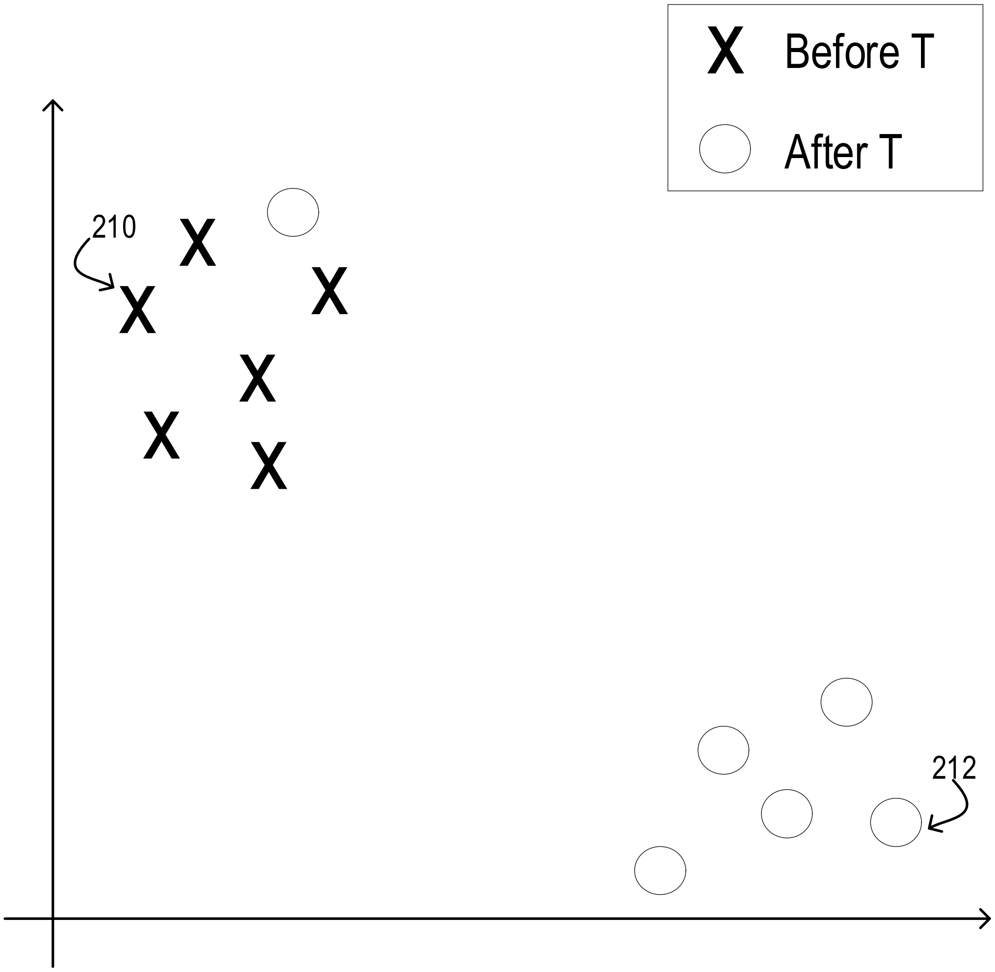

Figure 2