November 14, 2018

IDENTIFYING POLICY VIOLATING VIDEOS BASED ON USER BEHAVIOR

Corey Lane

Eileen Long

William Chambers

Guillermo Krovblit

IDENTIFYING POLICY VIOLATING VIDEOS BASED ON USER BEHAVIOR

A content item service (also referred to as "content item service platform" or "the platform") may allow users to upload media items (e.g., videos, audio, livestreams, etc.) on the content item service. Such media items may be streamed or otherwise provided or rendered to various types of users (e.g., adults, children, students, professionals, etc.). The media item may be rendered on different types of devices and platforms, such as, on desktop computers, laptops, phones, televisions, gaming consoles, different web browsers and operating systems, etc. Some media items may have particular ownership rights associated with the media item, such as, exclusive ownership rights by an entity, shared ownership rights, no ownership rights claimed, etc. The content service platform may have terms of service for users for media items (e.g., videos) that are uploaded to the content service platform. The content item service may set one or more policies with regards to media items on the content item service. For example, the policies may be related to content appropriateness for different types of users, ownership rights, quality of media items, device based factors, etc. An example policy, or set of policies, may be related to content appropriateness for children. The policies may indicate, for example, type of content that is appropriate for children. The one or more policies of the platform may govern what actions are taken for videos that violate the one or more policies. For example, in the instance of content appropriateness for children, videos that violate the policies (e.g., unsafe videos) may be removed from being available for children.

In many cases, preventive mechanisms, such as human reviews and computer algorithms, may be used to identify violations of the one or more policies associated with the videos. For example, various classifiers (e.g., machine learning algorithms) may be used to classify and mark videos, channels, and users as policy violating or not policy violating. Policy violating videos

may be filtered out from the content item service. The classification is generally performed based on metadata about the videos in isolation, without considering the behavior of users who consume the videos. Additionally, new types of usage and abuse may surface rapidly before the classifiers are updated, trained, or new classifiers are developed to address the new types of usage and abuse, in which case the existing classifiers trained using previously existing data may fail to generalize the data to new forms of usage and abuse. As such, under different instances, policy violating videos may still not be identified by these preventive mechanisms.

A mechanism is proposed for identifying policy violating videos based on user behavior such that additional policy violating videos are identified when existing preventive mechanisms fail to identify some videos as policy violating. The identification may be based on probability of videos to be policy violating and co-watch graph of the videos. For example, the content item service may identify a first video associated with a first probability value for the first video of violating one or more policies set within a content item service platform. For example, the probability value may be derived based on existing mechanisms that identify policy violating videos, such as using human reviews or computer algorithms. The content item service may identify a plurality of other videos that have been co-watched with the first video using a co-watch graph. In some examples, the content item service may identify a specified number of most co-watched videos with the first video. The content item service may calculate a score for the first video based on probability values for the plurality of other videos of violating the one or more policies and likelihood values for the plurality of other videos of being watched with the first video. In some examples, the score may be calculated based on a weighted average of the co-watched videos based on the probability values and the likelihood values of being co-watched. In some examples, the likelihood values of being co-watched may be based on historic

data of each of the plurality of other videos of being co-watched with the first video. The content item service may determine whether to initiate remedial actions regarding the first video based on the score. In some examples, the remedial action may be to remove the video from the content item service as related to the one or more policies it violates, and/or to further assess whether the video is policy violating using manual mechanisms, such as human reviews. In some examples, the determination may be based on comparing the calculated score to a threshold score set within the content item service.

Figure 1 depicts a flow diagram of a method for identifying policy violating videos based on user behavior. First, at step 101, a content item service may identify a first video associated with a first probability value for the first video of violating one or more policies set within a content item service platform. For example, the probability value may be derived based on existing mechanisms that identify policy violating videos, such as using human reviews or computer algorithms. In some examples, a database table may label videos as policy violating or not policy violating. In some examples, for each video of the content item service, a probability value maybe assigned that indicates the probability (e.g., expressed in percentage, fractions, etc.) that the video is policy violating. For example, the policies related to content appropriateness for children may identify a video as unsafe if the video contains content including, but not limited to, adult content, porn, abusive language, swearing, etc. The probability values may be based on frames of the video, metadata about the video such as title, description, etc. In an example, every video on the content item service may be associated with a probability value of being policy violating. For example, if a human review confirmed that a particular video is indeed policy violating, then the probability value of the particular video of being policy violating may be assigned as 100% (e.g., actual policy violating video). In another example, a machine learning

algorithm may use various factors and criteria to estimate a probability that the video has a 40% probability of being policy violating. In another example, a video may be confirmed by a human review as not being policy violating, in which case the probability value is assigned as 0%. Thus, in some examples, the content item service may identify each video of the content item service as being associated with a probability value for the video of violating one or more policies set within a content item service platform.

Subsequently, at step 102, content item service may identify a plurality of other videos that have been co-watched with the first video using a co-watch graph. The content item service may keep track of data on actual, historic video consumption by users. The data may include which videos have been watched with which other videos on the content item service. A co-watch graph may identify relationship between two videos in terms of the two videos being watched together in the same session. In a co-watch graph, each node may represent a video. An edge connecting two nodes may be weighted and is associated with a likelihood value of the two videos (e.g., nodes) connected by the edge to be watched together. In some examples, the content item service may identify a specified number of most co-watched videos with the first video. For example, for each node, a specified number of most watched videos may be connected to the node by edges. In an example, the specified number may be 1000 videos. That is, for a given video node, the top 1000 most co-watched videos may be connected by edges to the given video node. The specified number may be a customizable and tunable parameter that can be modified as needed. Thus, every video on the content item service may be represented by a node, but every node of the co-watch graph may not be connected to every other node of the co-watch graph as the edges are placed for a specified number of most co-watched videos. Also, there may not be an edge between a video that is not watched with another video.

Next, at step 103, content item service may calculate a score for the first video based on probability values for the plurality of other videos of violating the one or more policies and likelihood values for the plurality of other videos of being watched with the first video. In some examples, the score may be calculated based on a weighted average of the co-watched videos based on the probability values and the likelihood values of being co-watched. In some examples, the likelihood values of being co-watched may be based on historic data of each of the plurality of other videos of actually being co-watched with the first video. Each of the nodes of a co-watch graph may be annotated with a probability value of the node (e.g., the video) of being policy violating. For example, probability values may be stored in and obtained from a database table for each video on the content item service and used for the annotation of the nodes. The probability values for the plurality of other videos may be comparable to the probability value of the first video in terms of how the values are obtained, stored, etc.

For calculating the score for the first video, every video node connected to the first video by the edges (e.g., indicating the videos have been co-watched with the first video) may be inspected. Each video connected to the first video may comprise the plurality of other videos. In an example, the score reflecting the weighted average of the co-watched videos may be calculated as follows: a product is derived by multiplying the probability value of each of the plurality of other videos by the likelihood value of being co-watched with the first video of the corresponding one of the plurality of other videos. A first sum of the product for each of the plurality of other videos is calculated. The first sum is divided by a second sum of the each of the likelihood values of being co-watched with the first video of the plurality of other videos. As an example, assume the content item service contains four videos A, B, C, and D. The probability value of violating the one or more policies may be expressed as, for example for video B, as

"vid_B probability_of_policy_violation" and the likelihood value of video B being watched with video A may be expressed as "vid_A to vid_B co_watch_likelihood." Thus, to calculate a score for video A, the formula may be expressed as:

Score for vid_A = ((vid_B probability_of_policy_violation * vid_A to vid_B co_watch_likelihood) + (vid_C probability_of_policy_violation * vid_A to vid_C co_watch_likelihood) + (vid_D probability_of_policy_violation * vid_A to vid_D co_watch_likelihood)) / (vid_A to vid_B co_watch_likelihood + vid_A to vid_C co_watch_likelihood + vid_A to vid_D co_watch_likelihood).

In an example, the values for each of the parameters may be obtained from a table of metadata as follows:

For the nodes of the videos:

video_id, probability_of_policy_violation

vid_A, 0.1

vid_B, 0.2

vid_C, 0.8

vid_D, 1.0

For the edges in the co-watch graph:

video_id_from, video_id_to, co_watch_likelihood:

vid_A, vid_B, 0.3

vid_A, vid_C, 0.9

vid_A, vid_D, 0.7

Using the above values and the formula, the score may be calculated as:

Score for vid_A = ((0.2 * 0.3) + (0.8 * 0.9) + (1.0 * 0.7)) / (0.3 + 0.9 + 0.7) = 0.7789.

In the example, Video A is identified as being highly co-watched (e.g., 0.9 or 90% likelihood of being co-watched with video C and 0.7 or 70% likelihood of being co-watched with video D) with policy violating video C with 0.8 (e.g., 80%) probability of policy violation and video D with 1.0 (e.g., 100%) probability of policy violation. Given that video A is highly co-watched with these policy violating videos, it receives a high score of 0.7789, or over 77%. Thus, the score for video A may indicate that video A itself is at risk of being a policy violating video.

Subsequently, at step 104, content item service may determine whether to initiate remedial actions regarding the first video based on the score. A remedial action may be initiated if it is determined that a video violates a policy. In some examples, the determination may be based on comparing the calculated score to a threshold score set within the content item service. For example, a threshold score may be set at 10% probability for being a policy violating video. If the calculated score is over 10%, then the video may be identified as a policy violating video. In some examples, the calculated score may be within a specified range of the threshold value. In that case, the video may be under a watch list for future remedial actions. In some examples, there may not be enough data available for a particular video. For example, a video may not have been co-watched with more than a specified number (e.g., 15) of other videos. In that case, in some examples it may be decided that the video is policy violating. In some examples, the lack of data may not impact the determination of policy violation.

In some examples, the remedial action may be to remove the video from the content item service as related to the one or more policies it violates. For example, if the policy violation is related to a content appropriateness for children, then the video may be removed from being available to children on the content item service. In some examples, the removal may be

performed automatically by the content item service when it is determined that the video violates a policy based on the score. In some examples, the automatic removal may be based on a different threshold score specified in the platform from the threshold score used to identify policy violating videos. For example, a first threshold score of 10% is used to identify policy violating videos and a second threshold score of 20% maybe used to automatically remove videos. That is, if the calculated score is over 20%, then the video may be automatically removed. The remedial action may be to further assess whether the video is policy violating using manual mechanisms, such as human reviews. In some examples, the calculated score can be used to determine whether a manual review is appropriate. For example, if the calculated score is over 10% (e.g., first threshold) and equal to or below 20% (e.g., second threshold score) using the previous example, then the remedial action may be a human review of the video. That is, different actions (e.g., automatic removal, queuing for manual policy review, etc.) can be taken using different threshold levels.

In some examples, the process may be run at a regular interval (e.g., daily, every other day, etc.), for all videos in the content item service. In other examples, the process can be run on an ad hoc or as needed basis.

The mechanism described herein allows for efficiently identifying new, hard to identify policy violating videos. The mechanism provides for identifying policy violations that may not be missed by existing mechanisms. The mechanism can bolster existing video-level policy violation classifiers by providing additional means of using co-watch factors to identify additional violations. The mechanism provides for identifying additional policy violation that is not based on video-level factors for the video in question, rather considers metadata for other videos co-watched with the video in question. As a result, even if the video in question is not

originally associated with a high, or unsafe, level of probability of violation in isolation, the mechanism provides for identifying the video as unsafe based on data about other co-watched videos. The mechanism uses a weighted average for the co-watched video, thus balancing the impact of each of the co-watched videos.

ABSTRACT

A mechanism is proposed for identifying policy violating videos based on user behavior. A content item service may identify a first video associated with a first probability value for the first video of violating one or more policies set within a content item service platform. The content item service may identify a plurality of other videos that have been co-watched with the first video using a co-watch graph. The content item service may calculate a score for the first video based on probability values for the plurality of other videos of violating the one or more policies and likelihood values for the plurality of other videos of being watched with the first video. The content item service may determine whether to initiate remedial actions regarding the first video based on the score.

Keywords: video, media, co-watch, graph, policy, terms of service, children, kids, machine learning.

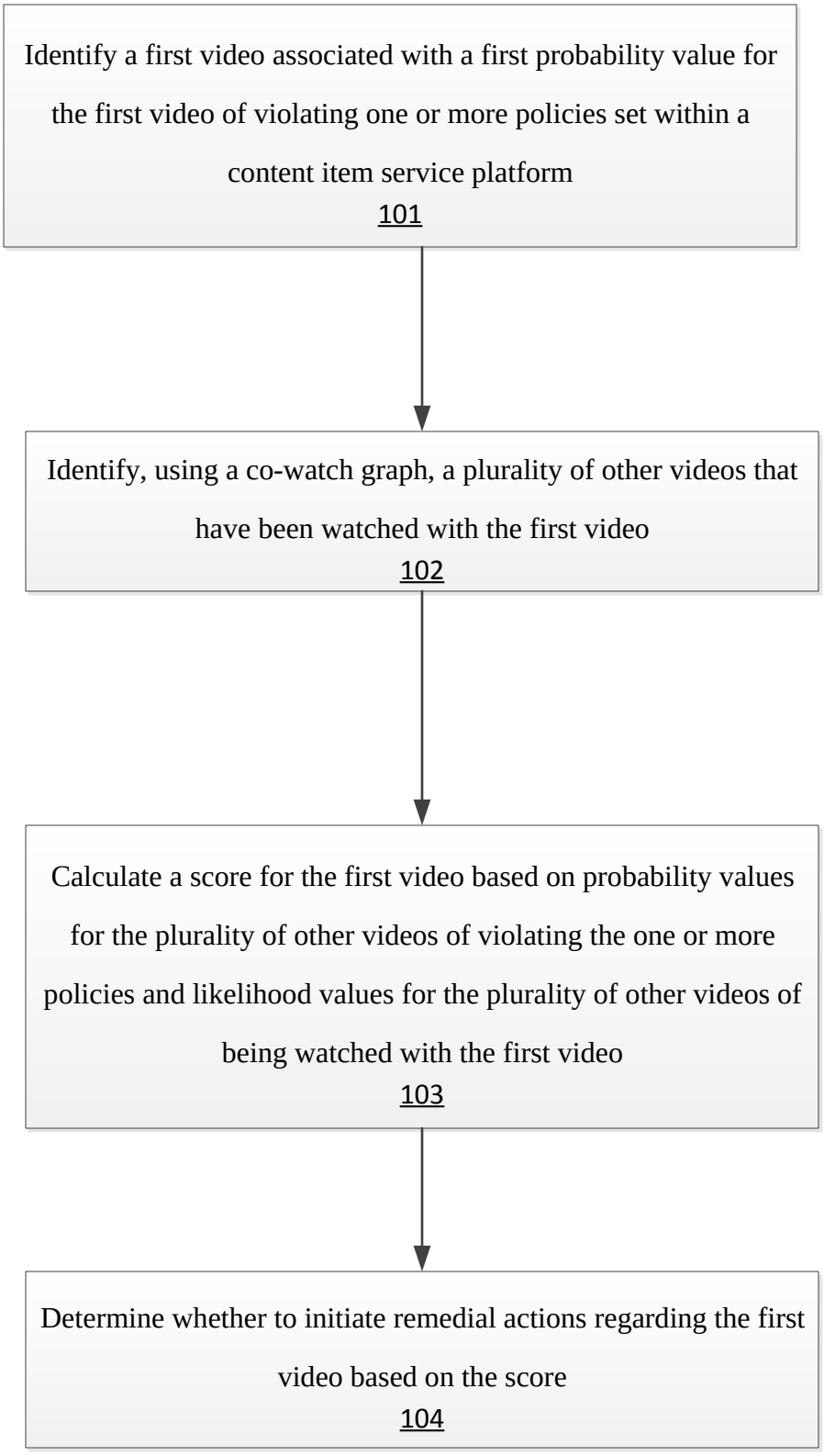
Identify a first video associated with a first probability value for the first video of violating one or more policies set within a content item service platform
101
Identify, using a co-watch graph, a plurality of other videos that have been watched with the first video
102
Calculate a score for the first video based on probability values for the plurality of other videos of violating the one or more policies and likelihood values for the plurality of other videos of being watched with the first video
103
Determine whether to initiate remedial actions regarding the first video based on the score
104

FIG. 1