November 09, 2017

UTILIZING USER-LEVEL SIGNALS FOR EARLY SPAM PROCESSING

Vladimir Rychev

Varun Sharma

Recommended Citation

Rychev, Vladimir and Sharma, Varun, "UTILIZING USER-LEVEL SIGNALS FOR EARLY SPAM PROCESSING", Technical Disclosure Commons, (November 09, 2017)
http://www.tdcommons.org/dpubs_series/799

UTILIZING USER-LEVEL SIGNALS FOR EARLY SPAM PROCESSING

A content item service may allow users to upload content items (e.g., videos, songs, audiobooks, images, documents, etc.) on the content item service. Such content items may be streamed or otherwise provided or rendered to various users. The content item service may perform various processing actions with regards to content items when users start uploading content items. For example, one of the processing actions may include transcoding a content item. Transcoding may involve converting a content item from one coded format (e.g., MP3, MP4, MPEG-2, MPEG-4, AVI, JPEG, etc.) to another coded format. A content item may be transcoded into multiple formats to allow users to use it on various types of devices and platforms, such as, on desktop computers, laptops, phones, televisions, gaming consoles, different web browsers and operating systems, etc. The content item service may transcode the content item into formats with differing qualities (e.g., high resolution, low resolution, etc.) for users to choose from. The content item service may also spend resources on other types of non-transcoding processing, such as, generating meaningful thumbnails, analyzing content items to study content data, comparing with copyrighted content, etc.

Some users may upload spam items on the content item service. For example, spam items may include content that are undesired, vastly repetitive, misleading, deceptive, etc. Upon completion of a content item upload, the content item service may perform content analysis and detect that the content item is a spam item. Subsequently, the content item service may take a restrictive action according to its spam handling policies, such as, make the content item unavailable for users, deactivate the content item, remove it from the content item service, etc. However, by the time the content item may be detected as spam, significant amount of resources and cost may have been spent on performing the various processing actions. The cost and efforts involved with processing a spam item may be unnecessary, wasteful and strenuous given the content item may be ultimately restricted from use.

We propose a mechanism for early detection of a spam item by utilizing user-level signals associated with the user uploading the content item in advance of spending significant resources on processing the content item. For example, once the user initiates a content item upload, user-level signals associated with the user may be analyzed to determine whether the content item is a likely spam item. The user-level signals may include data at the level of the user uploading the content item (e.g., previous content items uploaded by the user, related user accounts, network address, web browser, platform, etc.), as opposed to data at the level of the content item (e.g., content of the file containing the content item). Since it is possible to analyze user-level signals early in the uploading process (e.g., immediately after initiation of upload, prior to upload complete, etc.), a determination may be made whether it is necessary to devote resources for performing full scale processing. If the user-level signals indicate that the content item may be a likely spam item, then the content item service may initially perform only limited processing of the content item and generate a single, cheaper version of the transcode. By doing so, wasteful processing costs may be avoided for a content item that is indeed a spam item and would be eventually restricted. Subsequently, full content analysis may be performed on the single transcode and a determination can be made as to whether the content item is in fact a spam item, based on which a further restrictive action may be taken.

Figure 1 depicts a flow diagram of a method for early detection of spam based on user-level signals. First, at step 101, a content item service may detect initiation of a content item upload by a user on the content item service. For example, a user may initiate uploading a content item on a content item service by selecting a content item from a device of the user (e.g., a smart phone, a PDA, a laptop, a personal computer, etc.). The user may be logged into the content item service using an account associated with the user. The content item may be a video file, an audio file, a word document, an image file, etc. The content item may be large, medium, small, etc. Portions of the content item may reach a server of the content item service in increments. The user's device may send an indication (e.g., a signal, a message, etc.) to the content item service when the user initiates the upload. The content item service may detect the indication immediately, or prior to or at the same time as receiving initial portions of the content item. The content item service may detect the initiation of a content item upload by the user as soon as either the indication or an initial portion of the content item is received. The initiation may be detected prior to completion of the content item upload.

Next, at step 102, user-level signals associated with the user may be received. For example, upon detecting initiation of the content item, the content item service may receive the identification of the user. The content item service may receive user-level signals (e.g., data available at the level of the specific user). For example, the user-level signals may include, but not be limited to, previous content items uploaded by the user, user's content item consumption patterns, profile information, related accounts, account metadata, etc. The user-level signals may be received in real time, such as, upon detection of the initiation of the content item. Some user-signals may be received in batch jobs that are run on a specified frequency.

At step 103, the content item service may determine, based on the user-level signals, that the content item is a likely spam item. For example, the received user-level signals may be analyzed to identify the likelihood of the user uploading a spam item. The content item may be classified as a likely spam item or not a likely spam item based on the analysis of the user-level signals. In one example, the content item may be classified as a likely spam item. The analysis of the user-level signals may take various factors into consideration when classifying a user's content item as a likely spam item. For example, with regards to previous content items uploaded by the user, if the user-level signals indicate that the user has been uploading a large number of spam items, it may be likely that the user's next upload may also be a spam item. In other examples, with regards to user's content item consumption patterns, a user who has a regular pattern of viewing content items or interacts with content items may be less likely to upload a spam item, whereas a user with empty throwaway account may be more likely to upload a spam item. User's account usage information may be useful to identify behavior matching normal human behavior or with that of spam engines. A spam engine may display certain characteristics, such as, not consuming any content, creating a number of accounts, etc.

In other examples, with regards to related accounts, a new account opened by a user prone to upload spam items may share certain information, such as an adsense account for monetization purposes, a phone number for validation, etc. Examples of account metadata for likely spam classification may include the user's network address, location of the closest data center where the user's uploads are going through, time zone, language, web browser, platform, editing software, user-agent for uploading the content item, etc. For example, the content item may be uploaded from a certain location, IP address, an API, etc. that spammers are known to use. In another example, a user prone to upload spam items may use a certain editing software to insert spam item into a video content item, clone a spam audio content item in thousands of copies, etc. The user-signals and criteria listed above are only some examples and do not include an exhaustive list.

The analysis of the user-level signals may be performed by running a user-level classifier. The user-level classifier may specify a certain threshold or criteria for classifying the user's content item as a likely spam item. For example, if 99% of the user's previously uploaded content items were spam items, the classifier may identify the current content item as a likely spam item. In other examples, the user-level classifier may use a machine learning algorithm (e.g., a neural network) that has been trained on a set of user-level signals. When the machine learning algorithm is provided with available user-level signals, it may provide a determination whether the content item is a likely spam item or not. In one example, the user-level classifier may determine that the content item is a likely spam item based on user-level signals associated with the user.

Subsequently, at step 104, the content item may be converted into a single file. For example, once the content item is determined as a likely spam, the content item service may limit the processing of the content item. The limited processing may include converting (i.e., transcoding) the content item into a single file instead of transcoding the content item into multiple files with various formats. In one example, the single file may be a low resolution file. The limited processing may also include abandoning various other transcoding actions that may have already started prior to determining the content item as a likely spam item, deprioritizing or abandoning any other non-transcoding processing of the content item, such as, comparing with copyrighted materials, etc. The single file may be published on the content item service fast and made available for users. The single file may be a format that is usable by majority of devices and platforms used by various users.

At step 105, content analysis may be performed on the single file. For example, uploading of the content item may be completed when the content item is converted into the single file and published. The content of the single file may be analyzed. For example, the content item service may analyze the content of a video file to determine whether the file includes vastly repeated portions throughout the file, or portions matching with identified spam materials, etc.

At step 106, the content item service may determine, based on the content analysis, that the content item is an actual spam item. For example, based on the results of the content analysis of the single file, the content item may at this point be classified as an actual spam item rather than a likely spam item.

Subsequently, at step 107, availability of the content item may be restricted on the content item service. For example, the content item service may deactivate the published content item, make the content item unavailable to regular users, only keep the content available for third parties, remove the content item from the content item service, etc. In an example, the content item may be kept on the content item service for a specified length of time. After the specified time, the content item may be removed. The content item service may allow the user uploading the content item an option to dispute that the content item is a spam. Further actions on the content item may be decided based on spam handling policies of the content item service.

In an implementation, if the content item is identified as a not likely spam based on user-level signals, the content item processing may continue as is. That is, the content item may be transcoded into various formats, and any additional non-transcoding processing may continue to be performed. Content analysis may be performed on all of the transcoded formats of the content item or any one of the transcoded formats (e.g. on a transcode that meets a minimum resolution, etc.). Appropriate (e.g., restrictive, non-restrictive) actions may be taken based on the result of content analysis (e.g., whether actual spam or not actual spam).

In another implementation, after the content item is determined as a likely spam item and has been converted into a single file, if it is determined based on the content analysis that the content item is not an actual spam item, further processing steps may be taken. The further processing steps may include transcoding the content item into various desired formats, and starting or resuming other non-transcoding processing actions, such as data analysis, copyright validation, etc. The content item may be fully processed and available to various users on widespread formats. In an example, no immediate restrictive actions may be necessary if it is determined that the content item is not a spam item.

The mechanism described herein allows early detection of spam by utilizing user-level signals associated with the user uploading the content item in advance of spending significant resources on processing the content item. Because the mechanism allows limited processing of content items that is marked as a likely spam item, valuable resources are not spent on full processing of the content item and the content item is not transcoded into various formats when there is a likelihood that the content item may ultimately be restricted or removed from the content item service. Additionally, since the content item still goes through some minimal processing and is transcoded into at least a single file, this mechanism does not introduce any latency issue. The content item can be also deleted sooner in the process because the user-level signals can indicate an item is likely spam sooner in the process.

ABSTRACT

A mechanism for early detection of a spam item by utilizing user-level signals associated with the user uploading the content item in advance of spending significant resources on processing the content item is disclosed. Upon detection of initiation of a content item uploaded by a user on a content item service, user-level signals associated with the user uploading the content item may be retrieved. Based on the user-level signals, it may be determined that the content item is a likely spam item. Subsequently, limited processing maybe performed by converting the content item may into a single file and content analysis may be performed on the single file. Based on the content analysis, it may be determined that the content item is an actual spam item and the availability of the content item may be restricted on the content item service.

Keywords: video, content, user-level signals, transcode, content analysis, spam.

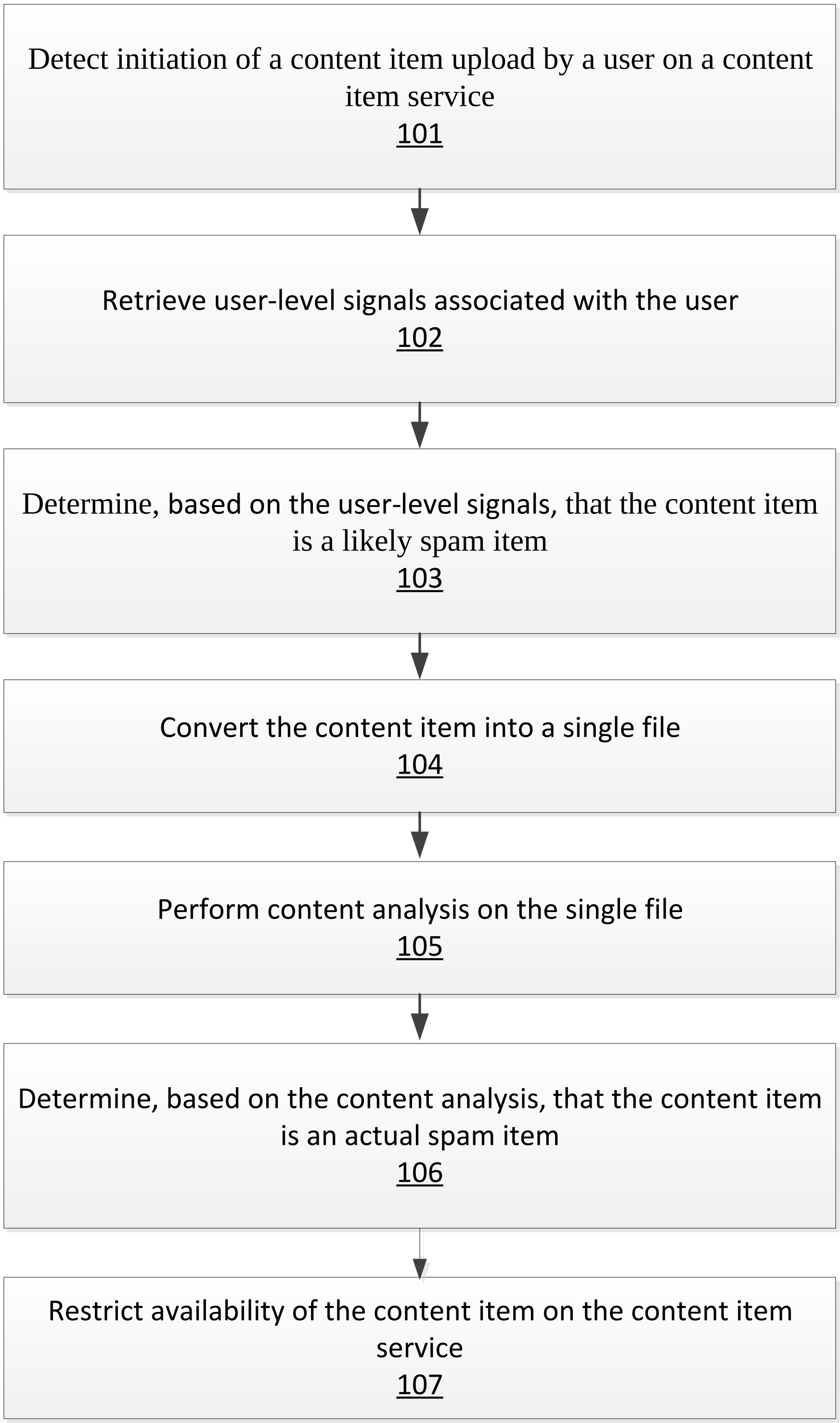
Detect initiation of a content item upload by a user on a content item service
101
Retrieve user-level signals associated with the user
102
Determine, based on the user-level signals, that the content item is a likely spam item
103
Convert the content item into a single file
104
Perform content analysis on the single file
105
Determine, based on the content analysis, that the content item is an actual spam item
106
Restrict availability of the content item on the content item service
107

FIG. 1