Big Data Initiative: Effective Caching in Online Video Platforms Draft*

EdN:1

Rongrong Bao Atabak Hafeez Tom Wiesing
Jinbo Zhang

October 30, 2015

Abstract

Data on the internet grows by 50 percent annually. More than 90% of the data has been generated in recent years. This is the time for big data. How can we effectively transfer this huge amount of data?

We want to investigate caching techniques used by online video platforms and in particular by YouTube. YouTube is a leading online video provider worldwide. Before 2012, video streaming in YouTube was done using Real Time Messaging Protocol (RTMP)-based servers. This requires a streaming server and a near-continuous connection between the server and user. Requiring such a streaming server can increase implementation cost and RTMP-based video streaming is at risk of being blocked by firewalls. In 2012, this was replaced by HTTP (Hypertext Transfer Protocol) based servers known as MPEG DASH (Dynamic Adaptive Streaming over HTTP). HTTP is the protocol used by websites to bring their content to the users. By using this technology it was possible to use existing optimizations in the form of HTTP-Caching. This capability decreased total bandwidth costs associated with delivering the video since videos would be served from web-based caches rather than the origin server. This improved quality of service, since cached data is generally closer to the viewer and more easily retrievable.

The essay will explain and discuss different kinds of caching techniques, optimizations, data analysis and prediction techniques used by YouTube, including their advantages/disadvantages and potential social impacts.

*EdNote: Remove draft status

Contents

1 Introduction

[2] EdN:2

[2] EdNote: Write introduction

2 What is Caching?

[3] EdN:3

[3] EdNote: Write technical

3 Advantages of Caching

[4] EdN:4

[4] EdNote: Write advantages

4 Disadvantages of Caching

While caching videos in the users web browser has many useful effects as discussed above it also has several bad side effects. In this section we will first explain some of the technical side effects and then continue discuss some social implications[5]. EdN:5

4.1 Technical side effects

In order to take advantage of caching of videos the browser needs to create caches for all the videos that the user plays. Additionally every time a video is played the browser has to check if a cached version of the video already exists. If this is the case the video can be played from the cached version. If this is not the case however the browser has to either create a cached version of the video or play a direct version from the server. Both of these decisions have a negative performance impact since the video has to be downloaded from the server which is slower then playing a local version.

Furthermore a cached version might not always be up-to-date. It is conceivable that when a news station has uploaded a video to YouTube and later on finds an error in their report that they might update their video in order to provide the most correct and up-to-date information. If this video has been cached by a viewer of the video they might no longer have the newest version available in their cache. Thus in order to use caches properly every time the cache is used the web browser needs to check if it is still up-to-date. This process can be further complicated if only parts of the video are cached. Furthermore if the cache is invalidated (meaning it is found to no longer be up-to-date) it is inefficient to reload the entire video if only parts of it have changed.

It is also possible that some videos that are cached even though the user does not want to watch them. This is for example the case when adverts are played before the video. They are not desired by the user and might nonetheless be cached. Sometimes the users might also watch a few seconds of the beginning of the video and then decide they do not want to watch the remainder. This can cause both a waste of bandwidth and local hard drive space.

Since caching technically copies the videos from the server to the hard disks of the local user the video is no longer stored in only one central location. This means that even if the video has to be deleted (for example for legal reasons) it might still be available locally. This might mean legal hurdles if a license has to be obtained in order to show the video.

Additionally some of the videos the user watches might remain cached on the users hard drive even after they have watched the video. While a single video does not take too much space this can cause a problem if there

[5] EdNote: Better wording, maybe a longer introduction and link to the previous section

are many of these cached videos. It is especially important for mobile users as these typically have less space available on their devices. Furthermore the user might not want a record of the videos they watched for privacy reasons.

4.2 Social implications

When using predictive caching YouTube (or other online video providers) are using data from the user. They try to analyse the videos the user has watched in the past and want to predict which videos the user might watch in the future. Some users might not desire getting videos predicted for several reasons. Especially when on a connection with limited bandwidth the user might object to having this bandwidth used up with videos that he might not watch anyways.

Additionally several users might not want their video history to be recorded. This can easily be considered an intrusion into the users privacy. This information could also be used against the user [6] especially when the wrong videos are predicted by the algorithm. EdN:6

YouTube sells anonymised customer data to advertisers [7]. If sold to advertisers, this data is commonly used to predict products the user might be intersted in. Related adverts are then played on the next video the user watches. This brings money to the content provider but does not give any significant advantage to the user. EdN:7

[6] EdNote: Where?
[7] EdNote: Citation needed

5 Conclusion

6 References