Chapter 72
Exploring Video Sharing Websites Content with Machine Learning

Nan Zhao
Télécom ParisTech, France

Löic Baud
DREV, Hadopi, France

Patrick Bellot
Télécom ParisTech, France

ABSTRACT

This article studies the characteristics of content on video sharing websites. A better understanding on online video content can help to analyse Internet users' behaviour and improve the video-sharing service. We improved an existing graph-sampling algorithm so that it could be more adapted to sample over the video sharing websites. A newly category system is defined in this paper, which can be applied on many different video sharing websites for content analysis. We also implement machine learning to realize the content re-classification with the newly defined category system. The efficiency reaches at 90%. From the classified content analysis, we find the content category distribution is not constant, and nowadays, cultural goods content take about 70% over all the sampled videos.

1. INTRODUCTION

Video sharing is a type of web services, which allows people to upload, share, distribute or store video content on the Internet. The type for video content can vary from a short clip to a full film. The service normally generates an embedded code for the uploaded video content, which provides user to share their video content in many ways as mail, blog or the social network. In the last decade, the video sharing service turns to one of most active web services, which brings a great raise of the traffic volume over Internet according to the study result of ipoque (Schulze & Mochalski, 2009). As the increase of the

DOI: 10.4018/978-1-5225-1759-7.ch072

bandwidth by the ISPs grows, the Internet users can have a better on-line video performance. Thence, comparing to download video content, the Internet users prefer to enjoy the content on video sharing websites immediately. Meanwhile, the video sharing service can also provide a large space for storing the video clips free of charge or with a fee very low.

Therefore in the recent years, the video sharing service has drawn a lot of interest to Internet researchers. There are several studies with certain video sharing websites as traffic characteristics analysis (Gill et al., 2007) and some properties researches (Cha et al., Halvey & Keane, 2007, Cheng et al., 2008, Kaiser, 2012, Mitra et al., 2011). Those first studies of the video sharing services are very important because they give the first opinions for exchanges of Internet traffic and consummation of video sharing service by the Internet users. However, the sampling algorithm in those prior studies can cause bias to popular videos, and as a consequence their results may be also biased. What is more, there are not many deeper studies on the content types and the distribution of shared video content on those websites. Therefore, in this paper, we present our recent study on the video sharing websites. Our study mainly concerns about video content characteristics based on a different video-sampling algorithm from those used in the existing studies. We try to figure out what kinds of videos are uploaded on the video sharing websites, how are those uploaded videos consumed by Internet users and the distribution of video duration and video count of views. The study results could be helpful for content resource management over the video sharing websites and making better video sharing service. Our study is composed by two parts. The first part work is implemented from January to May in 2013. In this part work, we collected videos from two video sharing websites YouTube and DailyMotion to analyse the video content distribution and other video characteristics on the two video sharing websites. The second part work is implemented from March to May in 2014. In this part work, we apply machine learning on the newly collected videos from YouTube. We aim to figure out an efficient machine learning classification algorithm to classify a large quantities of videos on video sharing websites. The highlights of our work could be summarized as below:

- We use a graph-sampling algorithm based on Random Walk and suggested videos supplied by the video sharing websites, which to some extent, can reduce certain effect caused by popular videos and correlations between videos.
- We then define a new category system, which is sufficiently uncorrelated and independent, and it can replace the defaulted category system given by the video sharing websites. This new system can easily be adopted by video sharing websites. With this category system, it becomes easier to compare content category distribution and popularity among different video sharing websites, which is helpful to understand users' behaviour on video sharing.
- We use software Weka to realize the machine learning. We choose J48 algorithm as decision trees building algorithm. With ten different decision trees corresponding to different categories, we succeed to classify 91.60% of sampled videos. The average classification accuracy of the ten decision trees is about 95%.
- Finally, Comparing with the results from the two-part work, we find that during the one-year time, the proportion of content categories change a lot. There are more cultural goods on YouTube in 2014 than in 2013. However, for the distribution of content popularity, there is no so much changing. The popularity of most content increases, especially for Series content, grows to the second most popular category among all the content.