Detecting Copyright Infringement on YouTube Videos using YouTube Metadata

Student Name: Swati Agrawal

IIIT-D-MTech-CS-DE-11-034
April 02, 2013

Indraprastha Institute of Information Technology
New Delhi

Thesis Committee
Dr. Ashish Sureka (Chair)
Dr. Haimonti Dutta
Dr. Gaurav Gupta

Submitted in partial fulfillment of the requirements
for the Degree of M.Tech. in Computer Science,
with specialization in Data Engineering

This research was partially funded by Information Management & Data Mining group at Indraprastha Institute of Information Technology, Delhi.

Keywords: YouTube Copyright Infringement Detection, Social Media Analytics, Mining User Generated Content, Information Retrieval, Rule-Based Classification.

Certificate

This is to certify that the thesis titled **"Detecting Copyright Infringement on YouTube Videos Based on Ordinal Measures and YouTube Metadata"** submitted by **Swati Agrawal** for the partial fulfillment of the requirements for the degree of *Master of Technology* in *Computer Science & Engineering* is a record of the bonafide work carried out by her under my guidance and supervision in the Information Management & Data Mining group at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

Professor Ashish Sureka
Indraprastha Institute of Information Technology, New Delhi .

Abstract

YouTube is one of the largest video sharing website on the Internet. Several music and record companies, artists and bands have official channels on YouTube (part of the music ecosystem of YouTube) to promote and monetize their music videos. YouTube consists of the huge amount of *copyright violated* content including music videos (the focus of the work presented in this paper) despite the fact that they have defined several policies and implemented measures to combat copyright violations of content. We present a method to automatically detect copyright violated videos by mining video as well as older meta-data. We propose a multi-step approach consisting of computing textual similarity between query video title and video search results, detecting useful linguistic markers (based on a pre-defined lexicon) in title, mining user profile data, analyzing the popularity of the Uploader and the video to predict the category (original or copyright-violated) of the video. Our proposed solution approach is based on a *rule-based classification* framework. We validate our hypothesis by conducting a series of experiments on evaluation dataset acquired from YouTube. The empirical results indicate that the proposed approach is effective

Acknowledgments

First and foremost, I would like to thank to my supervisor *Dr. Ashish Sureka* for his valuable guidance and advice. He inspired me greatly to work on this project. Without his guidance and persistent help this dissertation would not have been possible. Thank you for believing in me and accepting me in your group without even knowing me. I would also like to thank my committee members *Dr. Haimonti Dutta* and *Dr. Gaurav Gupta* for agreeing to be in my committee.

I would also like to thank *Annapurna* for devoting her time in helping me with evaluation results and giving invaluable feedbacks. She has been more than a helper for this work. Thanks to *Ayushi* for helping me with lots of technical stuff. I am grateful for her constant support and help.

Finally, an honorable mention goes to my family; for their understanding, endless love and wishes for the successful completion of this project.

This thesis is dedicated to my loving brother and my godfather *Abhishek* who was always there for me when I needed encouragement, motivation and strength to do this project. He helped me unconditionally whenever I needed his help. Thank you Bhai!

Contents

List of Figures

List of Tables

As Pirates Run Rampant, TV Studios Dial Up Pursuit
-Wall Street Journal.

Chapter 1

Research Motivation and Aim

Over the past decade, social media (Facebook, Twitter, Myspace and YouTube) has transformed into a dynamic form of world-wide interpersonal communication. By 2012, 50 percent of all Internet users (nearly 2.5 billion) are signed on to at least one social platform [1]. In Web 2.0 terms this is all about creating connections between users (people) and enabling them to communicate and share. Visiting a social networking site is the 4th most popular activity over the web. Activities like streaming various video clips are very popular among the active internet users these days. A large number of video sharing platforms for example YouTube, Dailymotion and Videoweed and many more have come up. Among these, YouTube is the most widely used video sharing website. As more people capture special moments on video, YouTube is empowering them to become the broadcasters of tomorrow.[2]

1.0.1 YouTube

YouTube, started as a personal video sharing service, has become a worldwide entertainment destination.[3] It is the largest free video-sharing service that lets users upload and watch an unlimited videos online. Videos can be a song video, movies, animations, footage of public events, personal recordings of friends, virtually anything a user wants to post. These videos can be informative, entertaining, persuasive, or purely personal[4].

YouTube is free, though people who want to post videos or comments must register with the site, creating a profile. Users can share their videos publicly or restricted to members of specified contact lists or regions by using private sharing. These videos include title, relevant keywords in the tags, an appropriate category, and a brief description to help people discover it [5]. Figure 1.1 and 1.2 depict the basic interface and an example for uploading a video on YouTube.

The ease of watching and sharing videos, combined with the fact that the site is free, opens the experience of online video to a wide range of users. YouTube offers opportunities for expression through videoa new spin on the notion self-publishing, making content available for anyone interested in consuming it. The social networking tools further engage users, drawing them into an environment that encourages them to meet new people, read and share opinions, and be part

[1] http://www.relevanza.com/post.cfm/the-rise-of-social-media-in-its-first-real-decade
[2] UCLA School of Law, Los Angeles, CA. J.D. Expected May, 2007.
[3] Jason C. Breen; YouTube or YouLose? Can YouTube Survive a Copyright Infringement Lawsuit
[4] http://net.educause.edu
[5] http://www.youtube.com/t/about_getting_started

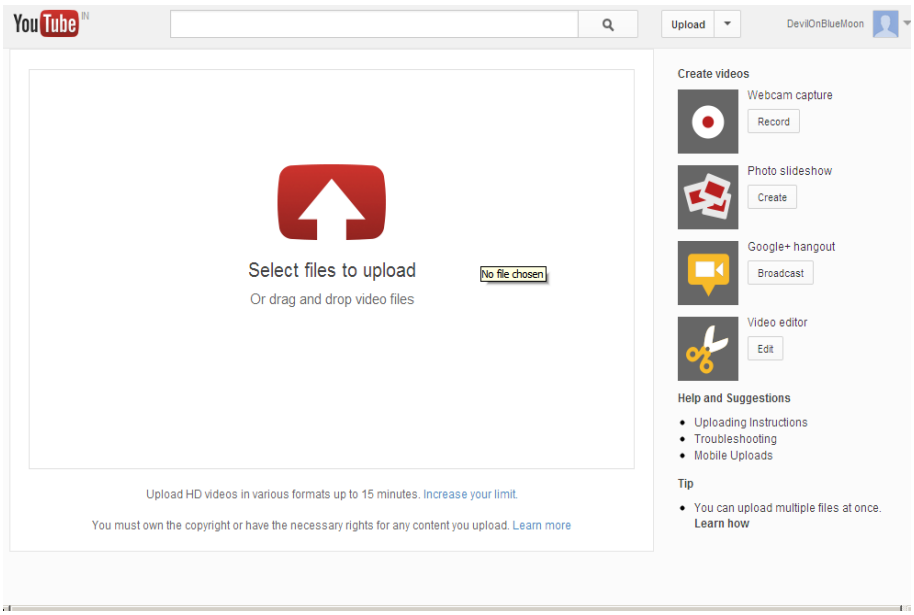


Figure 1.1: The screenshot illustrates an interface for Uploading a Video on YouTube

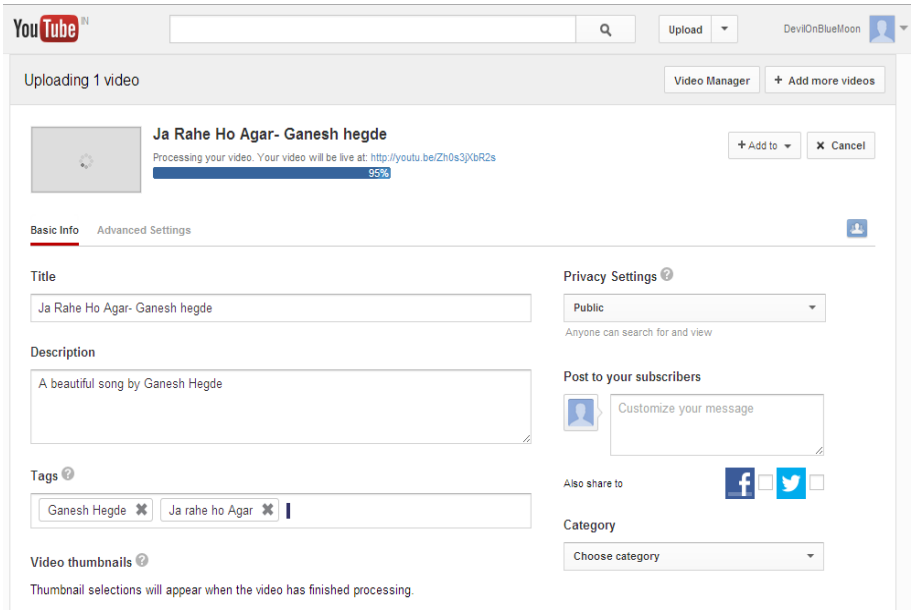


Figure 1.2: The screenshot illustrates an example of Uploading a Video on YouTube

of a community.[6] One of an emerging class of social applications, YouTube allows users to post and tag videos, watch those posted by others, post comments in a thread discussion format, search for content by keyword or category, and create and participate in topical groups. Users can view profiles of individuals who have posted or commented on videos, see their favorite videos, and contact them. Figure 1.3 shows the various social activities that can be performed on YouTube.

[6] 7 things you should know about...YouTube

Figure 1.3: The screenshot illustrates various social activities on YouTube

Social media-related YouTube statistics are just as impressive. According to Alexa 2013[7] YouTube is the most popular video sharing website on the web and it is worlds third most Visited social networking website. YouTube consumes the 10% of total internet traffic. There is an immense amount of videos on YouTube which will take 1700 years to watch them all. 24 Hours of Video are uploaded on YouTube Every minute which is equal to 150,000 full length videos every week. According to YouTube statistics[8], millions of subscriptions are made every day and 800 million of unique users join YouTube every month, while 100 million people do social networking on YouTube every week. Being such a popular video sharing website YouTube is providing enough pace for pirated and copyrighted infringed videos. Piracy costs the Music Industry and the Government millions of rupees each year. What has alarmed the entertainment business is the combination of growing access to high-speed internet and the popularity of a video streaming service like YouTube which cannot guarantee that the content has not been uploaded illegally. Uploading pirated movies on legitimate sites like YouTube has a real and lasting negative impact on revenues for content owners. According to an article in *The Economic Times*; Ashoka Holla, a director at digital video content provider in Berserk Media production house stated- "It's such a paradox. When digitisation came we thought it would cause an end to piracy in India. But with every film copy as good as the next one, we see a dramatic rise in piracy". So when *Son of Sardar* (a Hindi film starring Ajay Devgn and Sonakshi Sinha that released in 2012) was uploaded illegally on a Friday, it saw 15 lakh hits by Monday[9]. Figure 1.4 [10] demonstrates the statistics from economic times that the eventual increase in piracy on

[7] http://www.alexa.com/siteinfo/youtube.com
[8] http://www.youtube.com/yt/press/statistics.html
[9] http://articles.economictimes.indiatimes.com/2013-03-05/news/37469729_1_upload-videos-youtube-piracy
[10] http://articles.economictimes.indiatimes.com/2013-03-05/news/37469729_1_upload-videos-youtube-piracy

YouTube has affected the market value of right owners. And over the past decade music and movie industry has faced a loss of 5000 cr. ruppees in it's revenue and a cost of 50,000 jobs a year and it has declined the theatre collection from 95% to 60%.

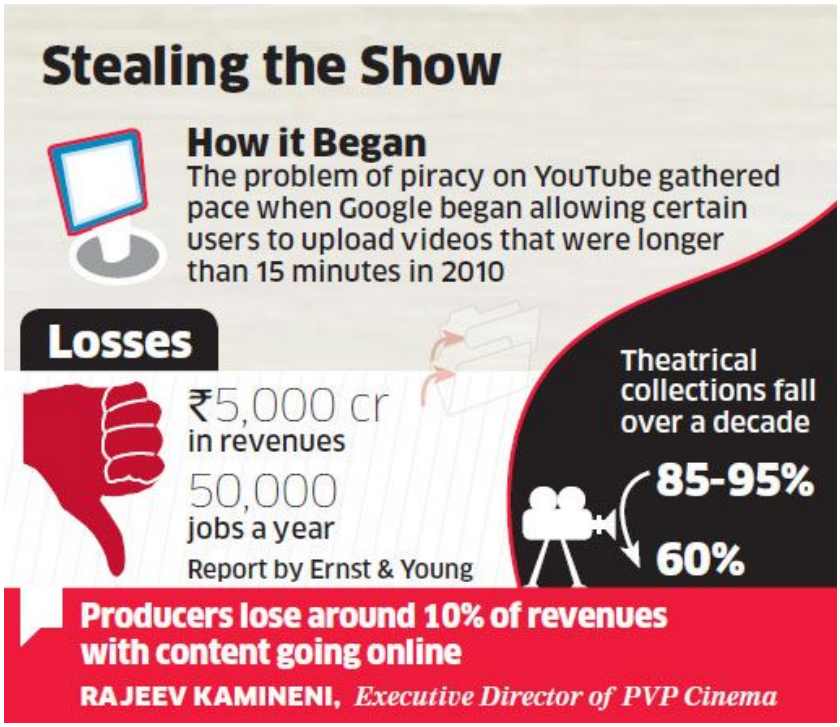


Figure 1.4: A snapshot illustrating a real and lasting negative impact on revenues for content owners because of YouTube Piracy.

1.0.2 Copyright Violation and YouTube

The three exclusive rights under U.S. copyright law that are most relevant to YouTube are the rights "to reproduce the copyrighted work", "to perform the copyrighted work publicly" and "to display the copyrighted work publicly". Thus, a YouTube user is required to obtain consent from a copyright owner, in the form of a licensing agreement, before being authorized to act within the scope of any of these exclusive rights.[11] YouTube is the largest video sharing website which allows us to upload any number of videos with low publication barriers, where we just need to have a YouTube account. Everybody is free to upload any content which influences piracy of content specially in the entertainment category of videos (Music Videos, TV serials, movies etc). The problem of piracy on YouTube was increased when Google began allowing certain users to upload videos that were longer than 15 minutes in 2010[12]. The owner loses the consumers and revenue to be earned by actual sale of his work. E.g. Hong Kong Film piracy on YouTube amounts to 308 million loss to original companies[13]. There are numerous applications available to download YouTube videos. Therefore once a video is public on YouTube, any user can download and re-upload it claiming as his/her own work. Users create their account on YouTube and pretend to be the actual right holder. These fake owners claim copyright on other people videos and monetize them through ads. There are many user channels on YouTube who

[11] http://www.law.cornell.edu/copyright/copyright.act.chapt1b.html
[12] http://economictimes.indiatimes.com/
[13] http://www.hollywoodreporter.com

Promote other pirated websites through dummy videos. They provide pirate link as a text overlay and in the description of the video. We can easily find thousands of such channels who have uploaded unlimited number of copyrighted videos, and those channels are very much popular too aswell. Fan-made channels and unofficial video blogs, video mix, uploads of latest HD music videos, TV series episodes and full movies in several parts are very common uploads by such users and those are the major contribution in copyright infringement over YouTube. Hence for the copyright owner on YouTube who wants to save his work from being pirated, YouTube has become "Defend Yourself" rather "Broadcast Yourself".

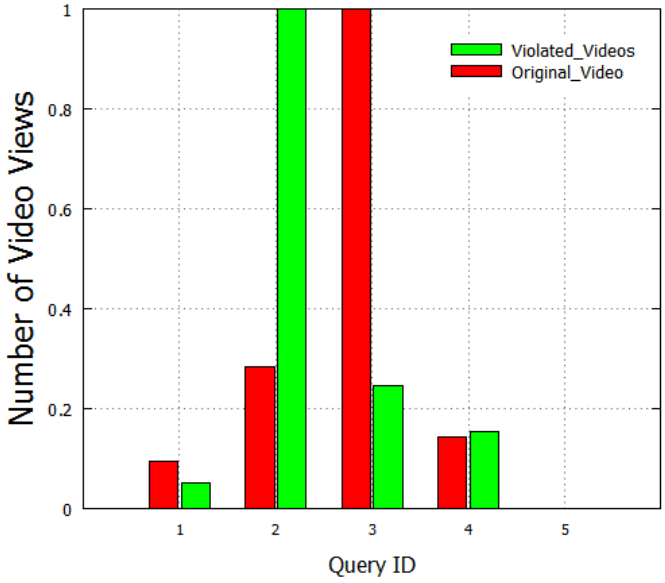


Figure 1.5: The screenshot illustrates variations in number of video views of an original and violated video for same query. (All values are normalized between 0 and 1, and query 5 has both values as 0 relatively to other videos.)

1.1 Research Motivation

In this theis project we are building a classifier to detect copyright infringed videos on YouTube based upon a user centric query. Effectiveness of our work is analyzed by a case study of hindi music videos in YouTube entertainment category. Work presented in our project is motivated by the following facts:

1. The search results for a video song on YouTube yield only a very small percentage of original videos, the majority being pirated videos which is clearly a case of copyright infringement. [14].

2. Copyright owners can list out the violated videos and therefore discourage the promotion of copyright infringed videos on YouTube. [15]

3. Enables the market analyst and rightful owners to analyze the popularity of their video which they uploaded in the past. These is determined based on the number of video views,

[14]Copyright infringement is unauthorized use of material having copyright

[15]Copyright provides the exclusive rights to the creator of an original work, and that piece of Art is the intellectual property of the beholder.

likes and comments shows the popularity of that video. These popularity measure either degrades or not correctly populated because of more hit counts on pirated copies uploaded in a short time span. We searched for 5 different song queries on YouTube and saved the number of video views on original and violated video among 10 search results. Figure 1.5 shows the variation in number of video views among original and violated videos for each query, which simply decrease the popularity of original video.

4. This problem of copyright infringement is further aggravated by the passive policies of YouTube to stop piracy. For e.g. YouTube only puts up a notice on the website advising the users not to upload pirated content thus making the users aware of the copyright infringement but has no barriers in place to stop uploading of a violated video.

1.2 Reserch Aim

The research aim of the work is presented in the following:

- Broad Objective: To make users more aware about copyright infringement problems on YouTube and propose effective solutions to detect the violated videos. (By mining video and user contextual data and identifying discriminatory features which can be used within a classification framework).
- Specific Objective: To analyze the application of a solution framework for detecting copyright infringed videos based on YouTube videos and user metadata and several ordinal measures. To implement a two class classifier for filtering relevant and irrelevant videos based a user centric query. To conduct a characterization study and empirical analysis on a real-world data set to measure the effectiveness of the proposed hypothesis.

Chapter 2

Technical Challenges

YouTube alone, the worlds largest repository of video film clips, contains over 12 billion videos and at the current rate, some 60 are added to its archives per minute, 24 hours a day, during all the days of the year. YouTube essentially represents an entire separate video based internet of its own in a way.[1] Due to the existence of such large scale networks and huge volume of data, infringement detection on YouTube is a challenging information extraction (or retrieval) task. The technical challenges that we had in this project are the following:

2.1 Human Intensive Task

YouTube has a huge amount of data in the form of videos and their up-loaders. This data is being updated in every hour with new user accounts and videos. This Data is complex, uncertain, unstructured and redundant. These videos and users have their own attributes (Video ID, Uploader id, watch counts, likes, dislikes, subscribers, etc.) that may be useful to implement a violation detection tool, but human involvement is still required with autonomous systems in order to deal with the unexpected.

2.2 Arbitrary Number of Irrelevant Video in Search Results

We define relevant video in the search result for a given query if the video (irrespective of whether it is copyright infringement or original) belongs to the same song or music as the query and irrelevant if the video is on a different song. We observe that YouTube search returns several irrelevant results for a given music title and hence identifying relevant videos (or filtering irrelevant videos) is necessary before detecting if the video in the search result is original or copyright violated (as it does not matter if a video is original or copyright violated if it does matches the information need and query of the user). We notice that the number of irrelevant videos returned in search result (for a given query) varies which makes the problem of recognizing relevant videos (before recognizing original or copyright violated) technically challenging. We conduct an experiment to demonstrate the technical challenge by creating a dataset of 50 queries (song titles) and top 20 videos in the search result. We manually label each search result as relevant or irrelevant. We count the number of irrelevant videos (20 - number of relevant videos) for each query. Figure 2.1 shows the histogram for the number of irrelevant videos for 50 queries. It reveals that the number of irrelevant videos varies from 0 to 15. We observe that for some

[1] http://webloggerz.com/video-content-evolved-into-important-web-design/

queries 75% of the search result consisted of irrelevant videos. The presence of irrelevant videos and wide variation in the number of irrelevant videos within Top K search result poses technical challenges to the relevant or irrelevant video classifier which is a step before the original or copyright-violated video classifier 4.1

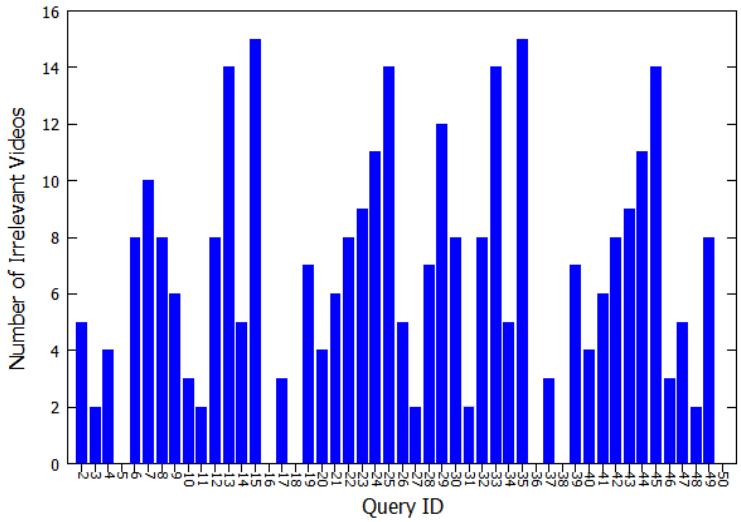


Figure 2.1: Variations in number of Irrelevant videos for every unique query

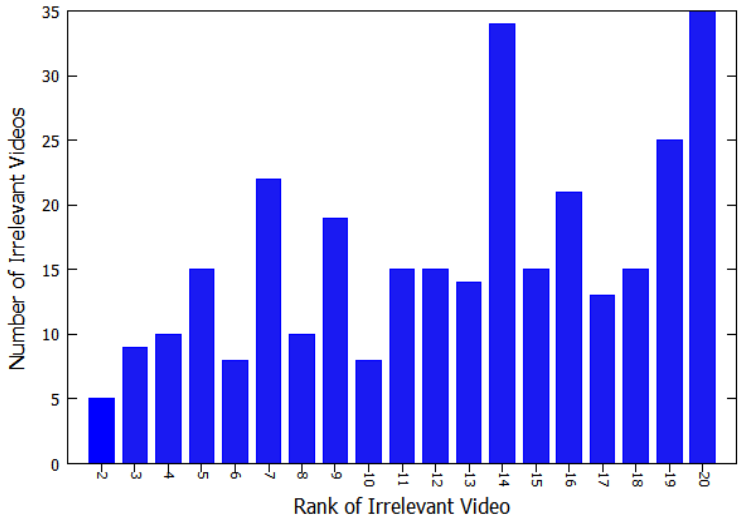


Figure 2.2: Variations in the rank of Irrelevant videos for every unique query

2.3 Non- Chronologically ordered irrelevance

We annotate the irrelevant videos in our experimental dataset (50 queries of song title and 20 search results) with their rank in the Top 20 search result. Figure displays the distribution of the number of irrelevant videos across 20 ranks (Rank 1 to 20). We observe that while irrelevant video does not show up as Rank 1 but are present at Rank 2 up-to Rank 20. Figure 2.2 reveals that there is a wide variation in the rank of irrelevant videos which poses challenges to the relevant or irrelevant video classifier. Our experiment reveals that the number of irrelevant

videos and their rank within the Top K search result varies across queries and makes the problem of relevant (to the given query) and original video detection challenging.

2.4 No baseline or benchmark available

The applications of machine learning techniques are limited by the absence of knowledge base or available measures. Hence it is hard to find relation between data and their attributes to detect copyright infringement. There is no specific measures to decide whether a video is violated or not except a few multi-media based techniques. But that doesn't help when a video is too long. Every time we need to do a manual search and classify a video as violated or original video.

2.5 Noisy Data

A video uploaded by a violated user might have a noisy data, for example the lack of information like *No Description or tags Available* or sometimes misleading information is there like a video uploaded by some other title while having a different content. Figure 2.3 shows an example of noisy data among several video results for a query. All three results represents three different kind of noisy data in titles.

Figure 2.3: This snapshot illustrates an example of Noisy data in several video titles according to user given query.

Figure 2.4 shows an example of video with misleading information in it's title. Originally the video is from a Hollywood movie of Actor Jackie Chan, while the movie was uploaded with the name of "Khiladi 786 movie", a Hindi movie of Actor Akshay Kumar. This video was published on YouTube only 13 days after official onscreen release of the movie. The snapshot also reveals the large number of views on this video because of the mis-leading information.

Figure 2.4: This snapshot illustrates an example of a video having misleading information in it's title and receiving large number of views.

Figure 2.5 shows the example of two video results for a song query. Both videos are of the two different songs, sung by same artist "Atif Aslam" as mentioned in query (to avoid ambiguity). Both the videos have almost same title. These types of queries make our problem challenging to filter relevant and irrelevant videos.

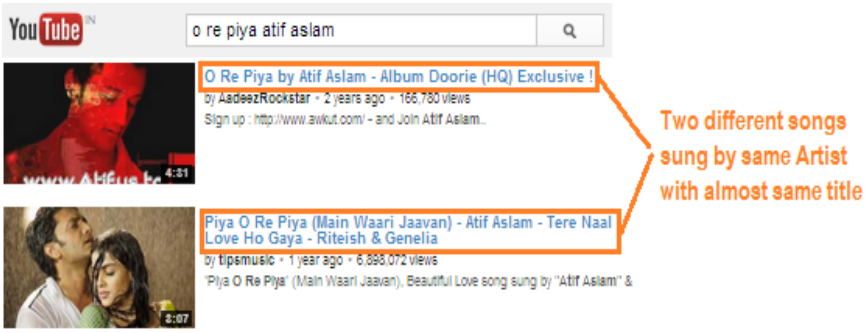


Figure 2.5: This snapshot illustrates an example of two videos of two different songs with almost same title and retrieved for one query.

Chapter 3

Related Work and Research Contributions

3.1 Related Work

In this section, we review the closely related literature associating with our work and the novel contribution in the context of existing work. Based on our analysis we divided the related work into two conceptual schemas: YouTube and Copyright Law, Detecting Content Redundant Videos using Multimedia and Contextual Features. Table 3.1 some of these papers.

3.1.1 YouTube and Copyright Law

This Survey is based onthe understanding of Govt policies and laws on copyright infringement given by several countries especially by the US Govt. In the United States, Copyright Law protects original works of authorship fixed in a tangible medium of expression, such as the videos, movies, and television shows that Viacom owns (Act 106).
George H. Pike highlights some disused services (Napster and Grokster) and their users that have been extensively sued by US Govt. for their role in facilitating the exchange of copyrighted content. It describes, even if YouTube has a different technical structure than Napster or Grokster, YouTube parallels the original Napster in its active role in the exchange of information between users. [1]
Jason C. Breen presents the services provided by YouTube and the *DMCA* (Digital Millennium Copyright Act) safe harbor. This paper focuses on case law that is applicable to YouTubes potential liability. How YouTube may fair under the concepts of direct and secondary infringement liabilities. It characterizes some policy implications that such decisions could result in. It also presents some basic *Terms of Use and Copyright Info* of YouTube.[1] [5]
Russ VerSteeg reviews the dynamic relationship between technology and copyright law over the years, as well as the principal legal theories that make up Viacoms complaint and YouTubes various defenses. It explores a few possible legal outcomes in the event of actual litigation and an ensuing judgment by the court. [3]
Eugene C. Kim the article depicts YouTube as a social entertainment network for individuals who want to share their "original" content- essentially, amateur videos created by users. It also describes the various laws, mechanisms, analysis and actions against YouTube copyright infringement issues. [4]

[1] http://www.youtube.com/t/terms

3.1.2 Detecting content redundant videos using multi-media and contextual features

We performed a review based on already existing approaches to detect copyright infringement and there are several such techniques which perform similar tasks and can be lead to violation detection like content redundancy techniques. These techniques are classified into multi-media, image processing, data mining categories.
Xiao Wu outlines different ways to cluster and filter out the near duplicate videos based on text keywords and user supplied tags. This paper proposed a hierarchical method to combine global signatures and local pairwise measure to detect clear near-duplicate videos with high confidence and filter out obvious dissimilar ones. [9]
Stefan Siersdorfer explains that content redundancy can provide useful information about connections between videos. It proposes different tag propagation methods for automatically obtaining richer video annotations. Additional information obtained by automatically tagging can largely improve the automatic structuring and organization of content. [7]
Hungsik Kim, Lee present Record Linkage Techniques and Multimedia based schemes to detect copied/altered/similar videos, comparing similarity among key-frame sequence, video signatures using MPLSH (Multi-Probe Locality Sensitive Hashing). [10]

Authors	Type	Objective
Xiao Wu 2007; [9]	CRMCF	Proposed Video tags and keywords based analysis to detect Nearly duplicate videos. %
Stefan Siersdorfer; 2009 [7]	CRMCF	Tags two videos together based on the redundancy in between their content
Hungsik Kim, Lee; 2008 [10]	CRMCF	Proposed a Multimedia scheme to detect copied/altered/similar videos.
Jason C. Breen; 2007 [5]	YTCL	Examines how YouTube would fare under the different theories of copyright infringement
George H. Pike; 2007 [1]	YTCL	Describes some legal issues of YouTube and Copyright infringement and a few obsolete mechanism and tools which detects copyright violation on small-scale data.
Russ VerSteeg; fall 2007 [3]	YTCL	Reviews the legal theories that make up Viacom's complaint and YouTube's various defenses on Copyright infringement.
Eugene C. Kim; 2007-08 [4]	YTCL	Describes the social networking on YouTube and various Govt. Laws, mechanisms, analysis and actions against YouTube copyright infringement issues.

Table 3.1: Literature survey of papers (chronological order) on YouTube Copyright Infringement. YTCL = YouTube and Copyright Law, CRMCF = Detecting Content Redundant videos using Multimedia and Contextual Features.

Video Copyright Violation has become a big issue in the last few years, a few approaches are being followed by YouTube itself. Based on our review we categorized these techniques into two parts: Infringement avoidance and detection techniques.

3.1.2.1 Avoidance techniques

1. *Digital Millennium Copyright Act* is applied to maintain an appropriate balance between the rights of copyright owners and the needs of users. The Digital Millennium Copyright Act is a law passed in 1998 governing copyright protection on the Internet initially divided into five titles.[2] DMCA for YouTube comes under the secons title i.e. "ONLINE COPYRIGHT INFRINGEMENT LIABILITY LIMITATION" section 512 of the Copyright Act to create four new limitations on the liability for copyright infringement by online service providers.

2. *YouTube help and support forums*[3]- the non-moderated forums for queries and opinions to make users aware about YouTube copyright policies.

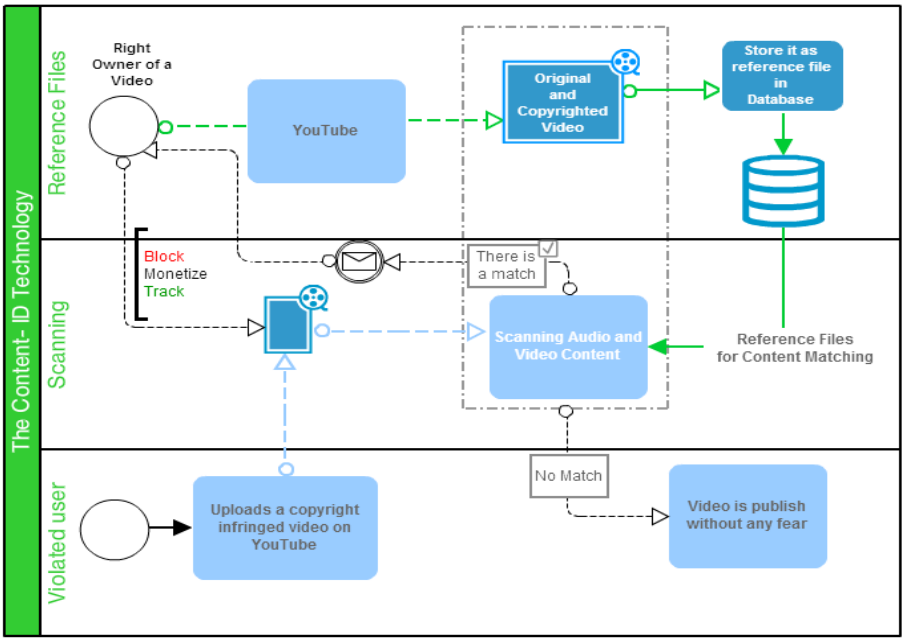


Figure 3.1: Work-Flow of content- id to detect copyright infringed videos over YouTube.

3.1.2.2 Violation Detection technique

YouTube uses Content-ID[4] to detect copyright infringed videos. A general framework for content-id is shown in Figure 3.1. It is a multi-media tool to detect if a newly uploaded video is an infringed copy of any copyrighted work. A database of full abstract image of copyrighted videos is maintained, called referenced files. Any video unmarked as copyrighted, having the complete or partial content (audio or video) against these reference videos is considered to be a violated video. YouTube uses fingerprinting approach to match these contents.
The content - id detects 100 years of videos everyday, But apart from this content-id has several limitations. (a) Piracy is still there on YouTube. (b) We can find millions of channels who are uploading others copyrighted videos without any fear and many of them are majorly

[2]http://support.google.com/youtube/bin/answer.py?hl=en&answer=141810
[3]http://support.google.com/youtube/?hl=en
[4]http://www.youtube.com/t/contentid

violated channels, e.g. KaBulHDvideoCenter, KhanHDMusicVideos, 911rtr, sominaltvtheater.[5] (c) Content-ID is unable to detect fake owners.[6]

3.2 Research Contributions

In the context of existing literature, the work presented in this paper makes the following novel contributions: An approach to detect copyright infringed videos using YouTube open source data set. Violation detection among videos is based on a user centric query. Proposed solution approach detects violated videos on YouTube for a particular video, user asks for.

1. To the best of of our knowledge, this is the first empirical study to detect copyright infringement in videos using YouTube meta data (Title and YouTube ID of the video, Uploader of video, Number of subscribers of video and popularity of the channel over the web.). We use a publicly available YouTube data set accessible using an API and web service and perform empirical analysis to find a violation among videos.

2. In comparison to previous work on infringement detection, we use textual data for filtering original and violated video unlike other multi-media techniques. We show that the features proposed in this paper are reliable and effective.

3. Empirical analysis and user based evaluation of the proposed solution is done on publicly (& currently) available data set of YouTube (videos and their uploaders). Efficiency and effectiveness of the solution approach are highlighted by performing manual in-depth experimental analysis (validation and statistical hypothesis testing).

[5] http://www.youtube.com/user/*username*

[6] Fake owner are the users who pretend to be the actual right holder and make money by claiming copyright on other peoples videos.

Chapter 4

Proposed Solution Approach

Figure 4.1 presents a general framework for the proposed solution approach. We divided the copyright infringement video detection into two sub-problems: Irrelevant v's Relevant Video Detection (IRVD) and Original vs Copyright Infringed Video Detection (OCIVD). Both IRVD and OCIVD are two class classification problems performing rule-mining classification. As shown in Figure 4.1 the proposed method is a multi-step process primarily consists of Data Extraction, Relevant v's Irrelevant Classifier and Original vs Copyright Violated Classifier cited as phase 1, 2 and 3 respectively.

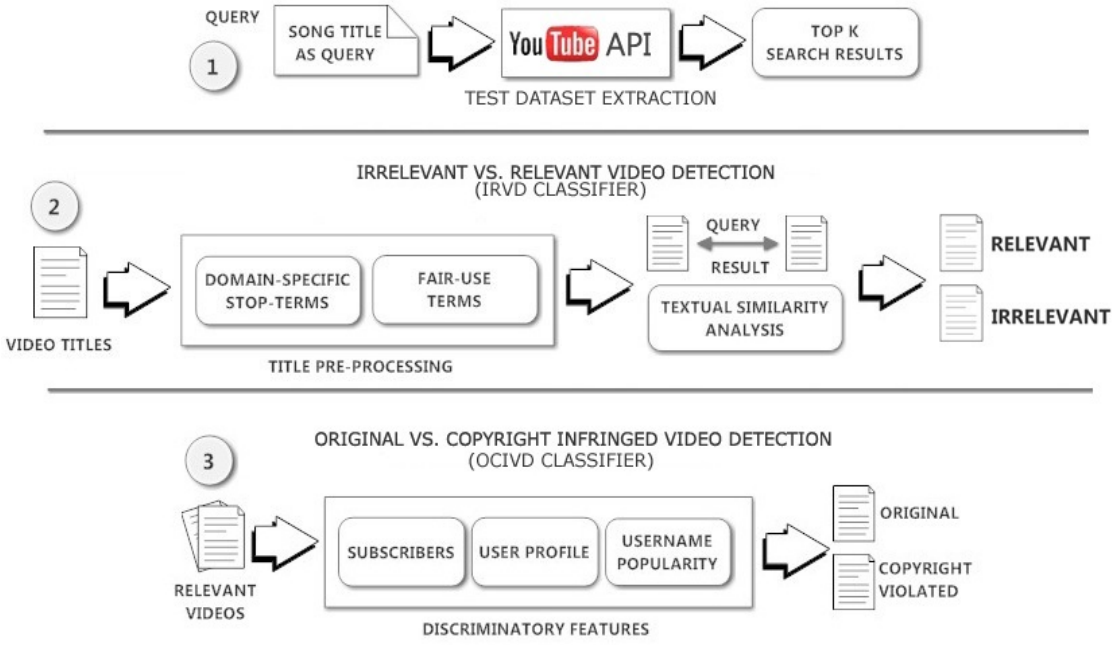


Figure 4.1: A general framework for our proposed solution approach

4.1 Phase 1: Test Dataset Extraction

This phase takes input from the user in the form of a song title query. The first step consists of retrieving *k=20* video search results based upon that query *q* using YouTube search API.[1] We retrieve the titles of all available search results for onward processing.

4.2 Phase 2: Solution approach to classify Relevant and Irrelevant Videos

An irrelevant video is any video that is not expected in search results by the user at the time of query. For example the cover version of a song video.[2] Figure 4.2 shows some irrelevant videos for a user centric query. The aim of this phase is to classify all irrelevant videos from k search results retrieved in phase 1. As shown in figure 4.1 phase 2 is again divided into sub-phases: Title Pre-Processing and Textual Similarity.

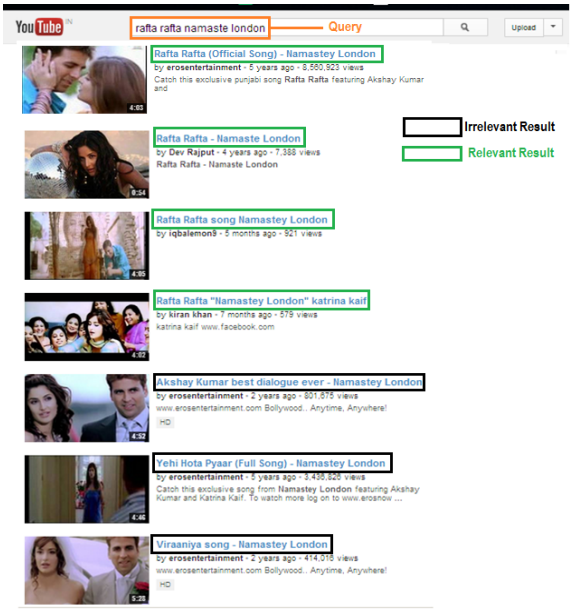


Figure 4.2: The Snapshot depicts the relevant and irrelevant videos for a query on YouTube.

4.2.1 Title Pre-Processing

To classify relevant and irrelevant videos the first step is the Title Pre-processing (TPP) based on some features. This phase uses three basic linguistic discriminatory features: Title of the video and query, Lexicons of domain specific stowords (DS2W) and fair use keywords (FKW). Title Preprocessing involves tokenization and stop word removal. Tokenization is the task of chopping the title up into pieces (or words), called tokens , at the same time throwing away certain words,[3] mentioned in the specific token list. These stop words are ineffective to find the

[1] https://developers.google.com/youtube/

[2] Any singer who appeared to be copying an already successful version of the song would be viewed with disfavor. (From Wikipedia)

[3] http://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html

(a) Domain Specific Stopwords Cloud

(b) Fair-use keywords Cloud

Figure 4.3: Words Cloud for Title Pre-Processing

relevance of a video but their presence affects the performance of the algorithm. A bunch of such stop words is depicted in figure 4.3 (a).
Figure 4.4 shows an example of removing stopwords from video titles.

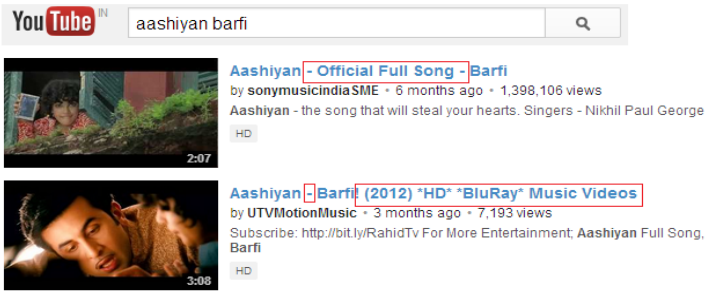


Figure 4.4: This snapshot illustrates an example of removing domain specific stopwords from a video title.

Figure 4.5: This snapshot illustrates an example of detecting some fair use kwywords in a video title.

The second step in title pre-processing is to find the existence of fair use key words Key_{fair} in titles. On the basis of FKW, if any of FKW is present only in the query but not in that video

title or vice versa, that video is classified as irrelevant. A few of such keywords are shown in depicted in figure 4.3 (b). Figure 4.5 shows an example of detecting fair use keywords in the video titles. These words are verified from our annotators and users.

4.2.2 Textual Similarity

Another concurrent filtering on titles is performed by using textual similarity among video title and query. As we mentioned above, noisy data could be there, and users titles the videos in non-standard writing or misspells them. Therefore this comparison is performed on both word and character level using well known string matching algorithm[4]. In textual similarity we use two types of title comparison. First we compute the overall matching of title with the query as a whole string. In the second step we look for the total number of words of the query that match with the title comparing on both word and character level. Figure 4.6 illustrates an example for such similarity.

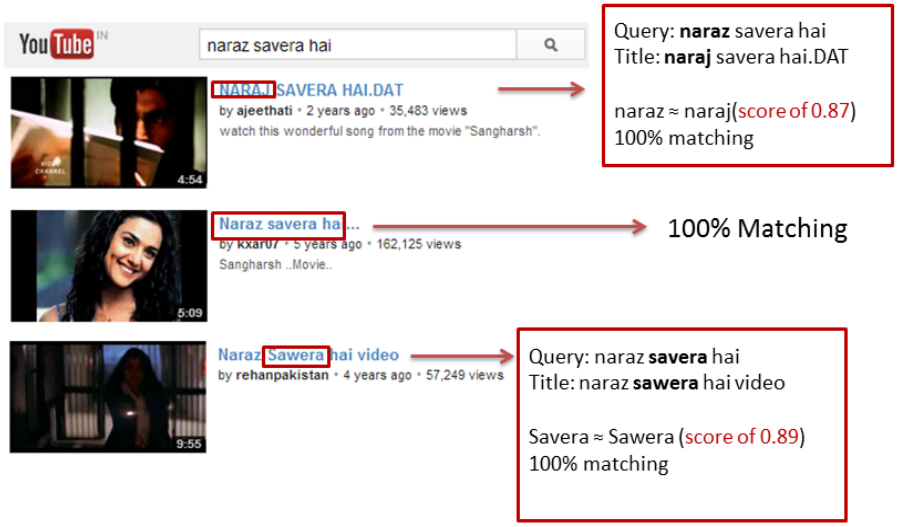


Figure 4.6: This snapshot illustrates an example of Textual similarity of video title to query.

4.2.3 Classification

In the final step, we perform rule based classification on all titles based on their scores in all above comparisons. If the video title satisfies a specified threshold in all sub-phases, then it is classified as a relevant video. Higher the score in all phases, higher the chances to be a relevant video.

[4]http://en.wikipedia.org/wiki/Jaro-Winkler_distance

4.3 Phase 3: Solution approach to classify Original and Copyright Infringed Videos

A copyrighted infringed videos is the video, shared or re-produced illegally without the permission and knowledge of the copyright holder. Figure 4.7 shows some original and violated videos in YouTube search results for a random song title. The aim of this phase is to classify such violated videos from original and legitimate videos.

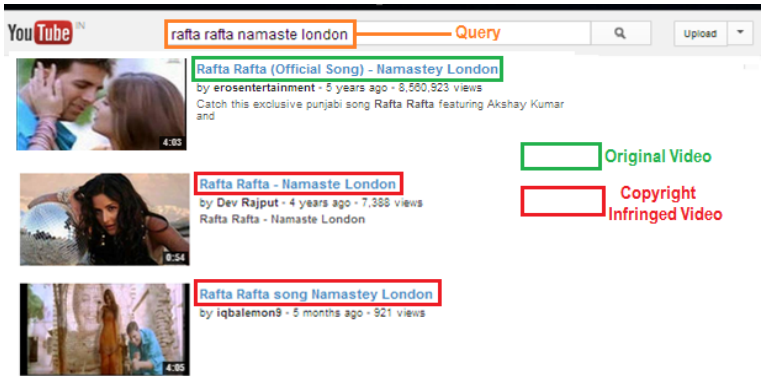


Figure 4.7: The Snapshot depicts the Original and Copyright Infringed videos for a query on YouTube.

As shown in figure 4.1, phase 3 i.e. Original vs Copyright Violated video classifier takes the input of all relevant videos classified in phase 2. A video is classified as an infringed video, if the uploader of that video is unauthorized. Therefore the information about uploader plays an important role to detect violated videos. We searched for few official and unofficial channels on YouTube to analyze the behavior of various features for different channels. Figure 4.8 shows a word cloud of a few official channels on YouTube related to hindi movies and music. [5]

Figure 4.8: A cloud of a few legitimate channels on YouTube (Hindi Music, Movies and entertainment category).

Here for classifying original vs violated videos we are using three Discriminatory Features: Number of subscribers, number of google page hits and YouTube profile information of the uploaders of relevant videos. These features are the quantitative measures used in the classification process

[5] With the best of our knowledge, these user channels are verified from their official websites and all feature values are correct as per March 2013 statistics.

and are selected by applying in-depth manual analysis on hundreds of user channels and their metadata. We obtained these features using YouTube and Google custom search API.

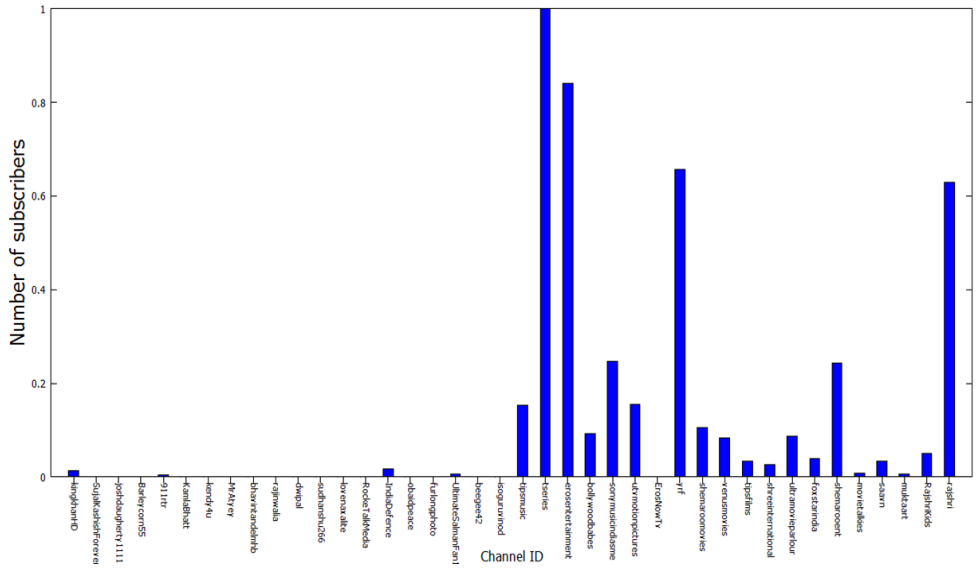


Figure 4.9: Variations in number of subscribers for Original and Violated Channels

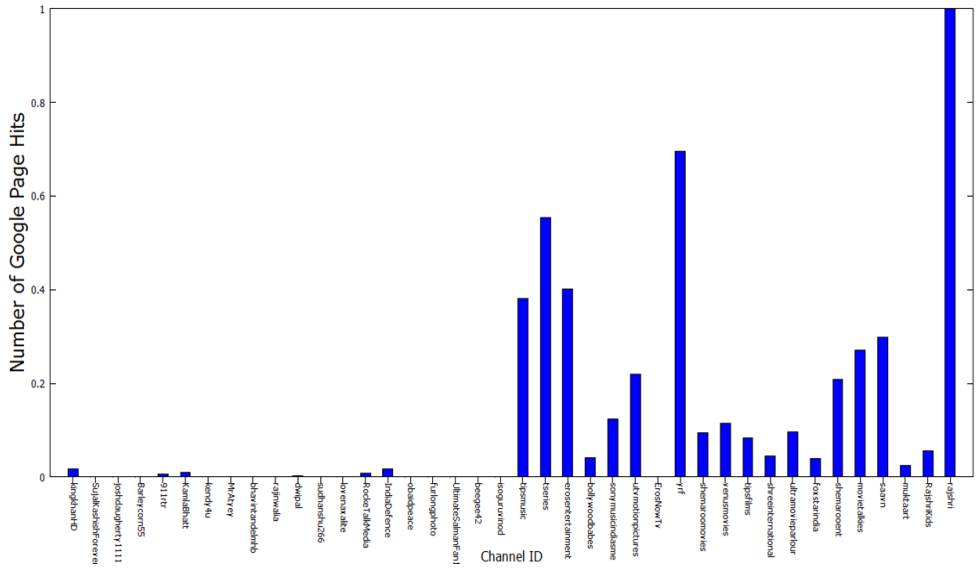


Figure 4.10: Variations in number of search hit counts for Original and Violated Channels

Figure 4.9 illustrates the variations of number of subscribers for original and violated channels. X-axis shows the ids of user channels where first 20 channels are violated and last 20 channels are original ones. The graph shows the number of subscribers (shown on Y-axis) of original channels are much more in comparison to infringed channels. Figure 4.10 illustrates the variations of number of google search counts for original and violated channels. On X-axis first 20 channels are violated and last 20 channels are official ones.The graph shows the number of search counts (shown on Y-axis) of original channels are much more in comparison to infringed channels. (All values are normalized between 0 and 1).

Another discriminatory feature i.e. user profile identifies a violated chhanel based upon a few

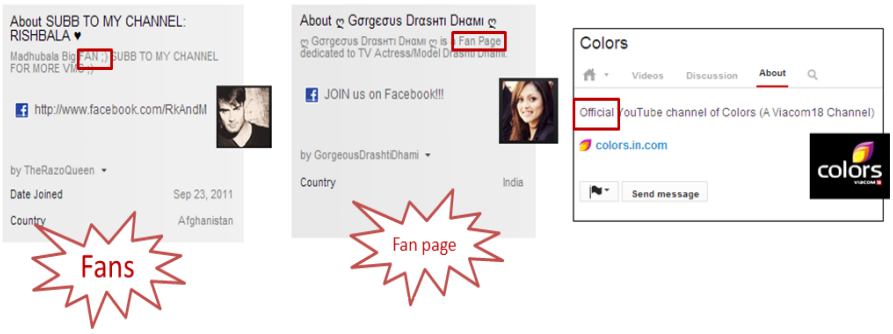


Figure 4.11: Snapshots are an example of Profile information of Original and Violated Channels

keywords existing in information. For example fanpage, fan made, unofficial etc. But this information doesn't identify if a channel is original. We can not set some specific keywords for legitimate channel because any violated user write them in it's profile information. Figure 4.11 shows an example of profile informations of three different channels. Where first two channels are violated and last snapshot is of original chhanel. First two channels has words fan page and fans in their profile information therefore we consider them as violated while the third channel has word "official" in it's information but we don't classify it as original at first step. It may happen that there is no information available in neither of official nor violated channel. Figure **??** shows an example where original channel (verified by artist himself) has no information available while violated channels have specific keywords in their information.

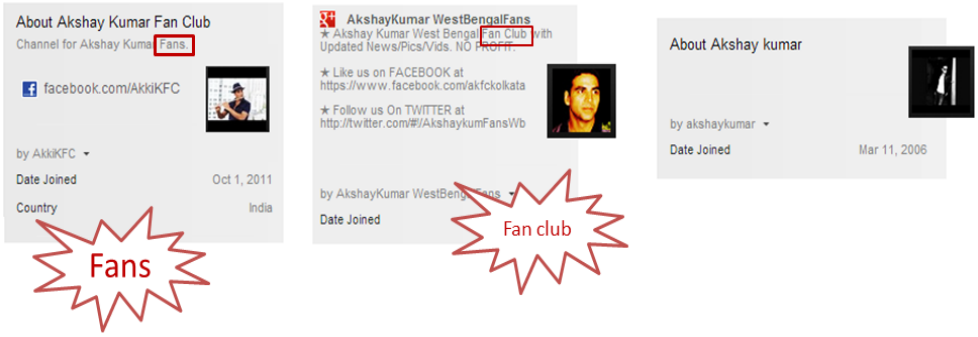


Figure 4.12: Snapshot illustrates an example of missing metadata for official channel but available for violated users.

4.3.1 Classification

If the Uploader of any video is already listed in our lexicon then we classify this video as original video immediately. For the rest of the videos we check their user meta data and perform classification based upon their score for respective discriminatory features. If any video satisfies the evaluated threshold measure for all features; it is classified as original video otherwise a violated video.

4.4 Detailed Description About Classifier

4.4.1 IRVD Classifier

In two class classification problem, either the video is relevant or irrelevant compared to the user's query. The algorithm does the similarity computation of video titles with query to recognize as relevant or irrelevant. In this section, we describe the classifier we have developed to detect irrelevant video in search results for a query.

Algorithm 1: Relevant vs Irrelevant Video Classifier

Data: Domain Specific Stopwords $D_{st} \in LexiconL1$, Fair Use Keywords $Key_{fair} \in L2$, User Query q, Video IDs $V_{id} \in V$, $t_{NumChar}$, $t_{NumWords}$

Result: V', Video IDs of Relevant Videos

1 V_{title} ←V.title(); $qlen$ ←q.length();
2 **for** ***all*** $t \in V_{title}$, q **do**
3 **if** $(\exists D_{st} \in t) \vee (\exists D_{st} \in q)$ **then**
4 t ←t.remove(D_{st});
5 q ←q.remove(D_{st});
6 **for** ***all*** $t \in V_{title}$ **do**
7 **if** $((\exists Key_{fair} \in t) \wedge (\exists Key_{fair} \in q)) \vee ((\nexists Key_{fair} \in t) \wedge (\nexists Key_{fair} \in t))$ **then**
8 $Score_1$ ←1;
  else
9 $Score_1$ ←0;
10 **for** ***all*** $t \in V_{title}$, q **do**
11 **if** *(t.contains(q))* **then**
12 $Score_2$ ←qlen;
  else
13 $match$ ←qword∩tword;
14 **for** ***all*** $words \in qword \setminus tword$ **do**
15 $flag \leftarrow JaroWinkler$(qword, $tword$);
16 **if** $(flag \geq t_{NumChar})$ **then**
17 $match \leftarrow match.add$(qword);
18 $mlen$ ←match.length();
19 $Score_2$ ←mlen;
20 **if** $(Score_1 \leq 0) \vee (Score_2 \leq t_{NumWords})$ **then**
21 $Class$ ←Irelevant;
  else
22 $Class$ ←Relevant;
23 $V' \leftarrow V_{id}$;
24 **return** V';

Algorithm 1 shows our approach of classifying the relevant and irrelevant videos. The result of the algorithm shows the video IDs V' of all relevant videos. This classification approach is based on similarity computation between video title t and query q.

Steps 2 to 5 represent the first sub-phase of title pre-processing 4.1. It removes all domain specific

stop words D_{st} from query and video title including punctuations and special characters. Figure 4.4 shows an example of a title pre-processing.

Steps 6 to 9 represent the second sub-phase of title pre-processing 4.1. This phase looks up the existence some specific fair use keywords in query and video title. If they exist in only one of the query or video title, respective video gets a score of 0 and 1 for otherwise.

$$Score_1 = \begin{cases} 0, \ if \ \exists Key_{fair} \in t \wedge \forall Key_{fair} \notin q \\ \qquad\qquad\qquad\qquad Orviceversa \\ 1, \qquad\qquad\qquad\qquad otherwise \end{cases} \tag{4.1}$$

Steps 11 to 19 represent the textual similarity between t and q. Title contains the whole query as a substring gets the score of equal to query length. Otherwise we perform a character level comparison for word string matching (Steps 13 to 17).
If two words each from query and title matches up to a specified threshold- considered to be similar [6] After this comparison video title gets a score of equal to the number of words that matched with the query.

$$\boxed{Score_2 = \text{match.length(); where match=}qword \cap \text{tword}} \tag{4.2}$$

Threshold values for Word and character level matching are presented in equations 4.3 and 4.4 respectively.

$$\boxed{t_{NumWord} = 0.7} \tag{4.3}$$

$$\boxed{t_{NumChar} = \text{qlen} * 0.6} \tag{4.4}$$

Steps 20 to 23 represents the classification procedure and labeling of videos as relevant or irrelevant.

4.4.2 OCIVD Classifier

Again Original vs Copyright Infringed Video Detection is a two class classification problem, either the video is original or copyright infringed with respect to a user's query. The algorithm performs the rule based classification on YouTube video meta data to recognize it as a legitimate or violated video. Algorithm 2 shows proposed solution approach for classifying these videos. The result of the algorithm shows the video titles of all violated videos. The classification approach is based on a discriminatory features compilation and evaluation.

Steps 1 to 7 performs the metadata collection for the test set.
Steps 1 to 3 extracts all YouTube user channel name or Uploader ids U_{id} of all relevant videos V'. Steps 4 to 7 extracts all required meta data of all U_{id} i.e. number of subscribers, number of google page hit counts and user profile information represented as N_{subs}, $N_{pagehits}$ and profile respectively. These features are retrieved using YouTube[7] and Google Custom Search[8] APIs.

$$t_{sub} = \frac{\sum_{i=1}^{n} User_{off_i.subscribers}}{n} \tag{4.5}$$

[6]This margin is kept because typing mistakes could be there.
[7]https://developers.google.com/youtube/
[8]https://developers.google.com/custom-search/v1/overview

Algorithm 2: Original vs Copyright Infringed Video Classifier

Data: Video ID $V'_{id} \in V'$, official YT channels $User_{off} \in LexiconL3$, User profile keywords $Key_{unoff} \in L4$, Subscriber Count threshold t_{sub}, pagehit threshold $t_{hitcount}$

Result: V''_t, Title of Infringed Videos

1 **for** ***all*** $V'_{id} \in V'$ **do**
2 $U_{id} \leftarrow$UploaderName of video;
3 $U' \leftarrow$U'.$add(U_{id})$;
4 **for** ***all*** $U_{id} \in U'$ **do**
5 $N_{subs} \leftarrow$SubscriberCounts();
6 $N_{pagehits} \leftarrow$SearchPageCounts;
7 $Profile \leftarrow$UserInfo;
8 **for** ***all*** N_{subs}, $N_{pagehits}$, $Profile$ **do**
9 **if** *(*$\exists Key_{unoff} \in$*Profile)* **then**
10 $Class \leftarrow$Violated;
11 V_t" $\leftarrow$Title of V'_{id};
12 **if** *(*$\nexists User_{off} \in$U')$\vee$*(*$(N_{subs} \geq t_{sub})\wedge(N_{pagehits} \geq t_{hitcount})$*)* **then**
13 $Class \leftarrow$Original;
else
14 $Class \leftarrow$Violated;
15 V_t" $\leftarrow$Title of V'_{id};
16 **return** V_t";

$$t_{hitcount} = \frac{\sum_{i=1}^{n} User_{off_i.pagehits}}{n} \tag{4.6}$$

Threshold values for both the parameters are as follows:

$$\boxed{t_{sub} = 6415 \text{ and } t_{hitcount} = 1051} \tag{4.7}$$

Steps 8 to 13 performs the classification procedure.
In steps 8 to 10 we look up for some keywords in the user profile information for example fan-page, fan-based, unofficial etc. User channels having such keywords in the profile are actually illegal channels. Therefore we classify the respective video as violated.
Steps 11 to 13 compare the value of N_{subs} and $N_{pagehits}$ with their respective threshold values (obtained from the training set).

Chapter 5

Empirical Analysis and Performance Evaluation

The aim of this section is to demonstrate the experiments and analysis set up for proposed solution approach. We present the characterization of the related features for each sub-problem i.e. relevant vs irrelevant video detection and original vs copyright infringed video detection. Apart from empirical analysis, in this section we present the performance results and system effectiveness.

5.1 Experimental Dataset

For IRVD classifier we performed a manual analysis on YouTube for the 50 random songs videos in entertainment (Hindi Television and Bollywood) category and created a lexicon of domain specific all possible stopwords D_{st}. As we are working on Hindi music videos we can not general english stopwords therefore we looked for our domain specific stopwords for example HD, full length, latest, BluRay, part etc. See Figure 4.2. During the stopword selection we also observed some fair use keywords in the video titles whose relative existence directly implies whether the video is relevant or irrelevant to the query. We again searched for 50 more queries and collected a few of such fair use keywords for example piano, guitar, karaoke, cover etc. See Figure 4.2. We etracted top 20 search results therefore we had total of 2000 videos for 1000 queries.

For OCIVD classifier we acquired 100 official YouTube channels (from non-popular to most popular) and 100 of unofficial or violated channels of Hindi music, movies[1], entertainment and television channels[2]. We downloaded required discriminatory features for these 200 channels and computed the average of their values as a threshold for respective attribute.

We collected our experimental data in the form of 100 unique queries taken from 20 different users. They provided us 5 unique queries (Hindi song titles). We extracted total of 2000 video (search results extracted using YouTube Search API) for these 100 songs, 20 results each. We selected title of those videos for IRVD classifier and their profile information and popularity measures for OCIVD classifier. Algorithm 3 shows our apporach of collecting experimental dataset for the same (using YouTube and Google Custom Search APIs). First we analyzed each video to be relevant or irrelevant and classified 1490 and 510 videos as relevant and irrelevant respectively. Again based on our manual analysis we classified 1533, 151 and 316 videos as

[1] http://en.wikipedia.org/wiki/List_of_Indian_film_production_houses
[2] http://en.wikipedia.org/wiki/List_of_Hindi-language_television_channels

original, copyright infringed and fair-used respectively. For all relevant videos we classified 1369 and 121 videos as violated and original.
A video is said to be an original video if it is relevant to the user's query and satisfies all benchmark values. All the evaluations are performed by the annotators (all 20 users and every query is analyzed by 3 annotators.)

Algorithm 3: Experimental Data Set Collection

Data: List of 100 official YouTube channels $User_{off} \in LexiconL3$
Result: Experimental data for OCIVD classifier

```
for all U_off ∈ L3 do
    Subs_ed ←SubscriberCounts();
    pagehits_ed ←SearchPageCounts;
    Profile_ed ←UserInfo;
```

	Training Dataset	**Test Dataset**	**Total Dataset**
IRVD	2000 Videos	2000 Videos	4000 Videos
OCIVD	2000 Videos & 200 user channels	2000 Videos & approx 1500 channels	4000 videos & approx 1700 channels

Table 5.1: Experimental Dataset

Table 5.1 shows the final set of experimental data used to detect copyright infringed videos for a user centric query.

5.2 Evaluation Metric

We measured effectiveness of our approach using standard information retrieval techniques i.e. precision and recall. As for a two class classifier each instance can only be assigned one of two classes: Positive or Negative. The precision of a class Y is the fraction of the number of videos classified as Y to the total number of videos predicted to be in Y. Recall is the ratio of a class Y is the number of videos classified to the total number of videos available for classification. Both precision and recall can be represented in form of a confusion matrix depicted in Table 5.2.

Assuming that 'a' represents the number of correctly classified Violated videos, 'b' represents the number of missclassified violated videos. Here 'c' represents the number of correctly classified original videos and 'd' represents the number of missclassified original videos. System accuracy depends upon T_P and F_N.

- True Positive (T_P) =a/a+b
- True Positive (T_N) =d/c+d
- True Positive (F_P) =b/a+b
- True Positive (F_N) =c/c+d

$$Accuracy = (a+d)/(a+b+c+d) \tag{5.1}$$

		Predicted		
		Violated	Original	Total
Actual	Violated	a	b	a+b
	Original	c	d	c+d
	Total	a+c	b+d	

Table 5.2: Illustrates the standard confusion matrix for two class classifier.

5.3 Empirical Analysis

In this section we present various features of a YouTube video. We analyzed different YouTube meta data and their effectiveness for each sub-problem. Based on the results we priortized these features to be used as discriminatory features.

5.3.1 IRVD Classifier

For IRVD classifier we used only contextual features of a video to find out its relevance to users' query Table 5.3 shows a list of such features.

S.No.	**Feature Title**	**Remarks**
1.	Video Title	It explains what is the video all about. Most important feature to search any video. (Directly relates to the query)
2.	Description of the video	Lack of information is there. Based on our analysis on music videos on YouTube, now 40% of the uploaders use same description of every video, if they copied it (See Figure 5.1). Relevant descriptions are available only in 30% of the videos and rest of the videos have no descriptions.
3.	Tags	YouTube API and videos don't support tags retrieval anymore.
4.	Existance of Fair use keywords	Terms like cover, karaoke, piano, live performance etc.
5.	Existance of Query in title	A video is considered to be relevant if it has same title as the query.

Table 5.3: Table illustrates all features for IRVD classifier.

5.3.2 OCIVD Classifier

To classify original and violated videos we have used Popularity and contextual based features. Table 5.4 shows a list of such features.

We extracted the feature values (number of subscribers, user profile and Google page hits) for several original and violated user channels and variations in their values are depicted in Figure 5.2. All values are normalized between 0 and 10. Q1 and Q3 represent the first and third quartile of clusters. Box "A" represents the number of subscribers for original channels. Points at A represnts the outlier channels which have very large number of subscribers. and 3/4th of the channels have similar number of subscribers. Box "B" represnts the gogole page hit counts for original channels. It shows that almost all official channels have a large number of search page hits at constant pace. But there exists some channels which have small number of hit

Figure 5.1: Snapshot illustrates some examples of descriptions available in 4 different videos for one song query

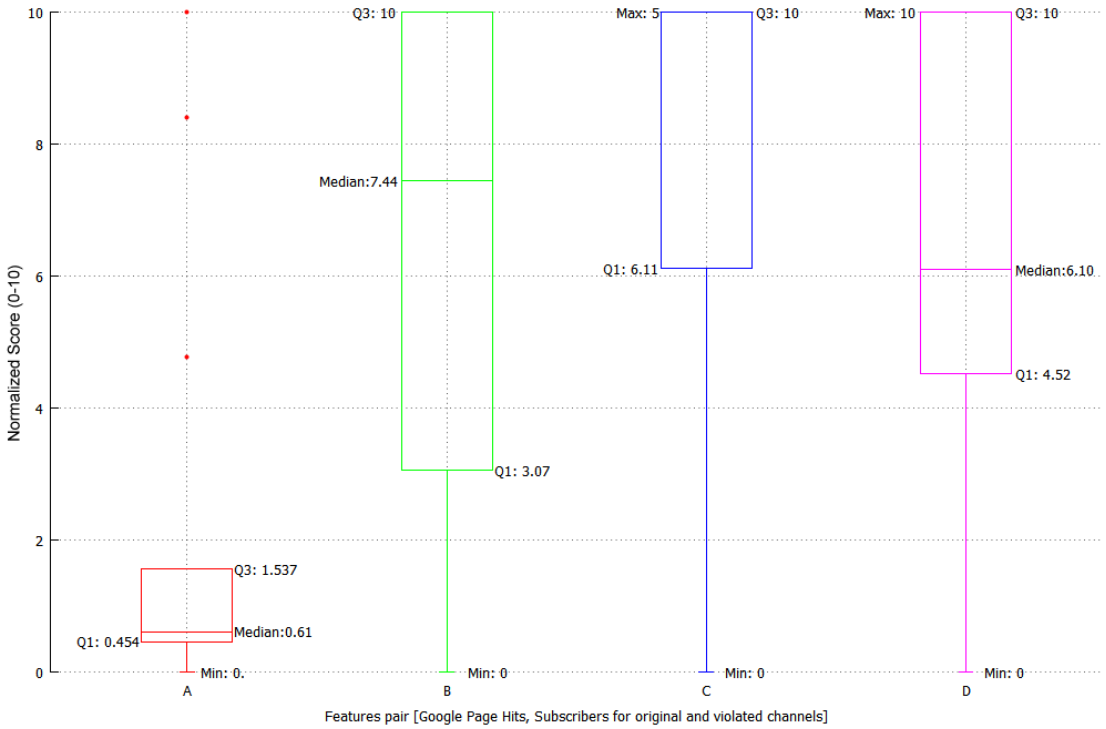


Figure 5.2: Variations in Feature values for Original and violated channels on YouTube. A, B: Number of subcribers and google page hits for original channels, C, D: Number of subscribers and page hits for violated channels

counts lying in first quartile Q1. Boxes "C" and "D" represent the number of subscribers and google page hit counts for copyright infringed videos respectively. For violated channels number of page hits and sunscribers are changing very frequently. These users lie in all ranges and 3rd quartile shows the major violated channels over YouTube.

S.No.	Feature Title	Remarks
1.	Number of Video Views	This feature is highly dependent on the published date of a video. And if there is no original video on YouTube then Violated videos will have more number of views.
2.	Number of subscribers	It's a popularity measure and A channel having a lot more number of subscribers is considered to be a legitimate channel.
3.	Google Page Hits	A legitimate channel has more number of hit counts than a violated channel.
4.	Existance of some keywords in Profile Information	Terms like Fan-Based, Fan-page, unofficial channel etc.
5.	Number of subscriptions	Part of the future work.
6.	Number of Uploaded Videos	Part of Future Work.

Table 5.4: The table illustrates some YouTube features relevant to Original vs Copyright Infringement Video detection. For feature 5 and 6 YouTube API doesn't allow to retrieve actual values for all user channels(fall January 2013)

Summary

Figure 5.4 shows the variations in number of relevant, irrelevant, original and copyright infringed videos for 100 unique queries. It shows that number of relevant and irrelevant videos are arbitrary for each query. This figure shows that the number of violated videos is much larger than the number of original videos. Figure 5.3 represents that there are many such queries, for which only violated videos are uploaded. It illustrates that among 50 random queries only 30% of them have original video reults on YouTube.

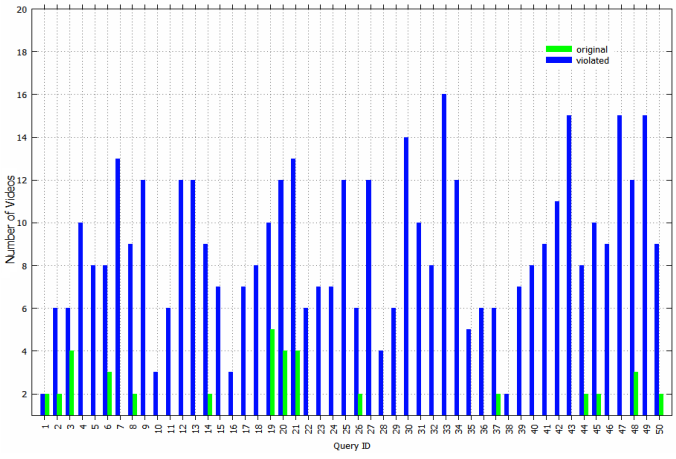


Figure 5.3: The snapshot illustrates the ratio of original and violated videos for 50 unique queries.

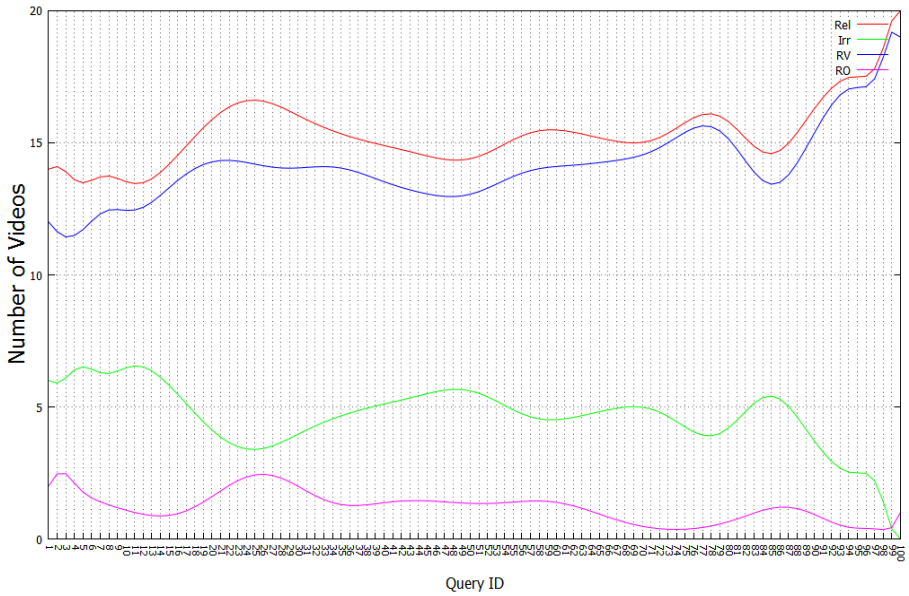


Figure 5.4: The snapshot illustrates the variation in number of irrelevant, relevant, original and violated videos for 100 unique queries.

5.4 Classifier Accuracy Results

For IRVD classifier we experiment with 1, 4 and 5 features, shown in Table 5.3. We apply Algorithm 1 to classify Irrelevant and relevant videos among k=20 search results for a given query. For relevant videos we apply 2, 3 and 4 features, shown in Table 5.4 to classify them as original or violated.

IRVD Classifier			
		Predicted	
		Rel	Irr
Actual	Rel	75.10% (1119/1490)	24.9% (371/1490)
	Irr	17.45% (89/510)	82.55% (421/510)

OCIVD Classifier			
		Predicted	
		Vio	Orig
Actual	Vio	94.9% (1455/1533)	5.1% (78/1533)
	Orig	9.27% (14/151)	90.73% (137/151)

Combined Classifier			
		Predicted	
		Vio	Orig
Actual	Vio	66.24% (919/1413)	33.76% (494/1413)
	Orig	52% (63/121)	48% (58/121)

Table 5.5: Confusion Matrix for (a) Irrelevant vs. Relevant Video Classifier, (b) Original vs Violated Video Classifier and (c) Combined Classifier

Table 5.5 shows the value confusion matrix for each classifier i.e. IRVD, OCIVD. Combined classifier is the complete solution approach after combining both classifiers into a pipelining.

The performance results are represents in form of precision, recall and F-score. F-score is evaluated to maintain the balance of accuracy. For our problem both precision and recall are important. Because user wants to retrieve only relevant results and he doesn't want to miss any violated videos among all search results. F-score is being evaluated by giving equal weigh to both precision and recall.

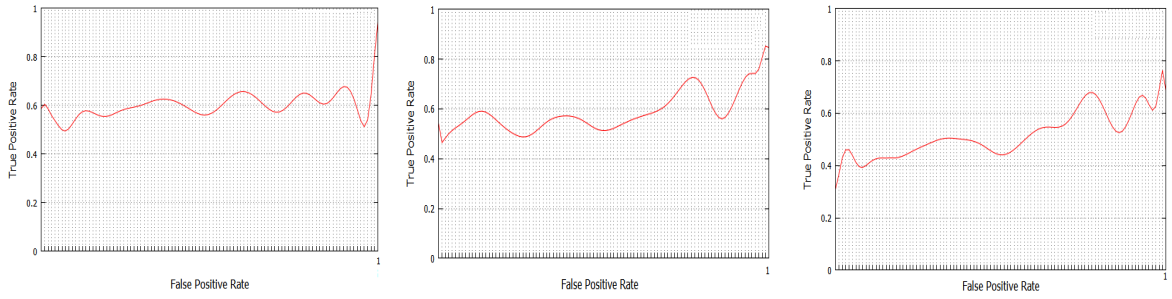
Phases	TPR	FNR	FPR	TNR	Accuracy	Precision	Recall	F1 Score
IRVD	0.77	0.23	0.20	0.80	76.8%	81%	75%	0.78
OCIVD	0.95	0.05	0.09	0.91	94.68%	91%	95%	0.93
Combined Approach	0.66	0.34	0.52	0.48	65%	56%	66%	0.61

Table 5.6: Precision rate for the test data set based on manual analysis.

- Precision (P) =a/a+c
- Recall (R) =a/a+b
- F1 Score (F) =PR/xP +yR

Accuracy Results shows that the given solution approach classifies original vs violated videos for all relevant videos with the accuracy of 65% where IRVD classifier classifies irrelevant vs relevant videos with the accuracy of 76.8% and OCIVD classifier classifies original vs violated videos with 94.68% accuracy.

Figure 5.5 represent the ROC curve for all three classifiers. X and Y axis represent the False Postive Rate and True Positive Rate respectively. All values of TPR and FPR are normalized between 0 and 1.



(a) IRVD (b) OCIVD (c) Combined Approach

Figure 5.5: ROC Curve for IRVD, OCIVD and combined Classifier Results

Chapter 6

Conclusion

Today millions of people around the world upload their own videos to YouTube. And sometimes, though it's against YouTube and Copyright laws. In this Thesis project we present an approach to detect original and violated videos on YouTube. Proposed method uses rule based classification, when applying on the test data set, it shows that the technique is effective and efficient to detect copyright infringement among YouTube videos.
In thesis thesis project we concluded that the number of subscribers and hit counts can be used to classify copyright infringed videos. User profile information is useful to detect violated channels but not the original and legitimate channels. Conclusion also states that if we filter irrelevant videos from search results then classification of original vs violated videos is a time efficient approach. We performed experiments on publicly available real-world dataset using two open source APIs (YouTube and Google Custom Search) and evaluated the overall performance of the proposed model. The Results (measured in the form of confusion matrix) indicate that YouTube video and users contextual data can be used to classify original and violated videos up to a reasonable accuracy.

Chapter 7

Future Work

The future work of this project is to improve the accuracy of both proposed classifiers (IRVD and OCIVD). For IRVD, semantic analysis and Natural Language Processing will be integrated for title comparison. To check the relevance of a video keywords in user comments will be used. For OCIVD classifier future work will be to improve performance results using more contextual and trust features.
The work presented in this thesis report is limited to the area of copyright infringed videos detection based on a user centric query. The future work will be to detect violated users uploading pirated music videos on YouTube.

Bibliography

[1] George H. Pike *Legal Issues: Google YouTube Copyright and Privacy.* Information Today. Vol 24, number 4, Page 15 April 2007, University of Pittsburgh - School of Law.

[2] Library of Congress, *How to Investigate the Copyright Status of a Work*, United states copyright office, Washington, DC 20559, January, 1991

[3] Russ Versteeg1 *Viacom V/S YouTube: Preliminary Observations* North Carolina Journal Of Law & Technology, Volume 9, Issue 1, Fall 2007

[4] Eugene C. Kim *YouTube: Testing the Safe, Harbors Of Digital, Copyright Law* 17 S. Cal. Interdisc. L.J. 139 (2007-2008)

[5] Jason C. Breen, *YouTube or YouLose? Can YouTube Survive a Copyright Infringement Lawsuits*, UCLA School of Law Year, Texas Intellectual Property, Journal 16.1 (2007): 151-182. Available at: http://works.bepress.com/jasonbreen/1

[6] Avery Li-Chun Wang *An Industrial-Strength Audio Search Algorithm* ISMIR 2003, 4th Symposium Conference on Music Information Retrieval, page 7–13. (2003)in , S. Choudhury and S. Manus, Eds., The International Society for Music Information Retrieval. http://www.ismir.net: ISMIR, October , pp. . Online. Available: http://www.ee.columbia.edu/ dpwe/papers/Wang03-shazam.pdf.

[7] Stefan Siersdorfer, Jose San Pedro, Mark Sanderson, *Automatic Video Tagging using Content Redundancy* 32nd international ACM SIGIR conference on Research and development in information retrieval, Pages 395-402, July 2009

[8] *Copyscape Premium, Copyscape Indigo Stream Technologies, L t d . 2009, http://www.copyscape.com/*

[9] Xiao Wu, Alexander G. Hauptmann, and Chong-Wah Ngo. *Practical Elimination of Near-Duplicates from Web Video Search*, Proceedings of the 15th international conference on Multimedia, Pages 218-227, ACM New York, NY, USA, 2007

[10] Hungsik Kim, Jeongkyu Lee, Haibin Liu, Dongwon Lee *Video Linkage: Group Based Copied Video Detection*, CIVR '08 Proceedings of the 2008 international conference on Content-based image and video retrieva,l Pages 397-406, July 2008.

[11] Sakrapee Paisitkriangkrai, Tao Mei, Jian Zhang, and Xian-Sheng Hua. *Scalable Clip-based Near-duplicate Video Detection with Ordinal Measure*, CIVR '10 Proceedings of the ACM International Conference on Image and Video Retrieval, Pages 121-128, 2010

[12] Junge Shen, Tao Mei, Xinbo Gao*Automatic Video Archaeology: Tracing Your Online Videos*, WSM '10 Proceedings of second ACM SIGMM workshop on Social media, Pages 59-64, 2010.

[13] Guangyu Zhu, Ming Yang, Kai Yu, Wei Xu, and Yihong Gong. *Detecting Video Events Based on Action Recognition in Complex Scenes Using Spatio-Temporal Descriptor*, Proceedings of the 17th ACM international conference on Multimedia, Pages 165-174 , October 2009

[14] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg*A Comparison of String Distance Metrics for Name-Matching Tasks*, Proceedings of IJCAI-03 Workshop on Information Integration, page 73–78. (August 2003)

[15] Educause Learning Initiatives *7 things you should know about...YouTube* September 2006

[16] US copyright Office *THE DIGITAL MILLENNIUM COPYRIGHT ACT OF 1998* U.S. Copyright Office Summary, December 1998