A DATA MINING BASED SPAM DETECTION SYSTEM FOR YOUTUBE

Prof. R. S. Yenape

Vitthal Shinde, Swami Shubham, Amol Funde, Kapil Yerekar

Student, Department of Computer Engineering [B.E], Sinhgad Academy of Engineering Savitribai phule Pune University, Maharashtra, India

Student, Department of Computer Engineering [B.E], Sinhgad Academy of Engineering Savitribai phule Pune University, Maharashtra, India

Student, Department of Computer Engineering [B.E], Sinhgad Academy of Engineering Savitribai phule Pune University, Maharashtra, India

Student, Department of Computer Engineering [B.E], Sinhgad Academy of Engineering Savitribai phule Pune University, Maharashtra, India

ABSTRACT

People are using social networking sites mostly than the front to front conversations. Thus the social media sites are very popular in these days. YouTube is very famous and popular website everyone using. YouTube is mostly used for uploading and sharing of videos. Users are taking advantage of this uploading videos on YouTube and sharing them on internet. As the popularity of YouTube is increased users are posting fake videos. Now, YouTube does not have any tool to handle the spams of video. To increase the famousity of videos different fake users uploading a fake video like uploading a different video and giving different names. The name and video content is not related with them. In this paper we collect different information such as likes, comments ,etc about videos then system will perform some data mining operations using different algorithms . We apply Data Mining Tools classify the video as either spam or legitimate. Our system's output will tell the videos is original or spam.

Keywords : *social network, spam detection, videos, rating, data mining.*

I. INTRODUCTION

In last few year social networking websites such as Facebook, Twitter and YouTube have made a dramatic growth in terms of popularity. Every second On YouTube 60 hours of video are uploaded every minute. Over 4 billion videos are viewed a day and Over 800 million unique users visit YouTube each month [1]. Amongst this growth, shared video contents of these sites become a regular part of life. The web becomes the main source to watch different videos and share them on internet. New type of interaction among users is supported such as video chats, mails and blogs. There are many services which are based on video functions which is used alternative to text functions such as video reviews of services. Video spammers post different videos to famous there videos and share on YouTube .Video spammers post different videos to famous there videos and share on YouTube . A number of spam detection techniques are available. They are mostly on text based spamming like email text spamming, comment spamming .While video spamming is very tough to detect the spam. In video spam number of comments, likes, rating are useful attributes to classify the spam. We will develop a system that uses the information of the video like comments, rating ,likes etc and we will give output so that it will tell the video is spam or not. The research shows the following

- Find the evidence of spam videos on YouTube
- Identification of video attributes that will distinguish spammers from legitimate ones.
- A collection of videos from YouTube, which will distinguish spam videos from legitimate videos.
- A video spam detection is based on classification algorithm such as ID3 and Naïve Bays, KNN algorithm.
- Predict the spam of videos from the data mining model.

II. YOUTUBE MEASUREMENTS

Our goal is to design a mechanism to classify users of social video sharing systems into legitimate and video spammers, using a set of their attributes and of their contributed videos. Towards this goal, we crawled through YouTube and retrieved data, one of the most popular social media networking sites today [3]. A test collection, including a sample of the crawled data, was then built and used to evaluate the effectiveness of our classification approach. Next section describes our crawling strategy followed by presenting the criteria used to select users for the test collection.

III. PROPOSED SYSTEM

In our proposed system, it will collect the video data through the API .Then we get different information like comment, ratings etc. We then characterize several video attributes view, like, comments, categories from our test data, after selecting these attributes that will detect spam in videos. Firstly user will copy the link from the videos and paste it on the system. Then with the help of url system will get all related information of the video. It checks attributes of the videos likes, comments, views etc. then does the operations on them with the help of three algorithms. Then system will collect 3 different outputs from algorithms. Then a comparison of the outputs is done and result is shown to the user.

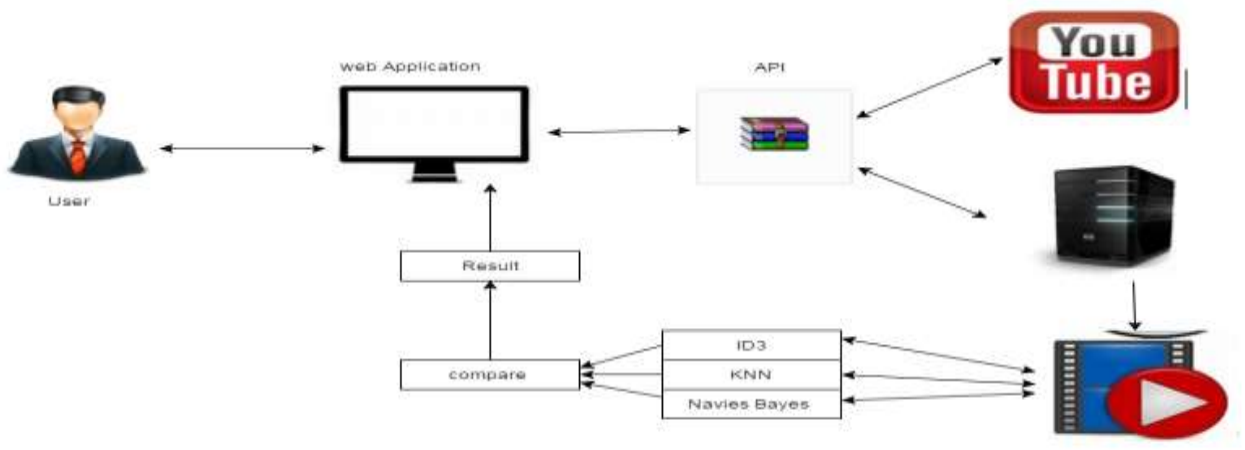

Fig.-1. Architecture of proposed system

METHODOLOGY

Steps:

- Upload the URL and Identify the Given URL Video in You tube
- URL to get you tube data using you tube API.
- Collect information related to you tube video according to those information we apply algorithms on it Implement algorithm of ID3 algorithm, k-NN algorithm and Naive Bayes Approach
- Display chart related to output of ID3 algorithm, KNN algorithm and Naive Bayes algorithm.
- Comparison of all 3 approaches and show the optimum results

IV. SPAMMER DETECTION MECHANISM

We are using following algorithms for detecting spam in videos

A) ID3

B) K-NN

C) NAVIE BAYES

A) ID3:

It is used to generate a decision tree from a dataset. In ID3 algorithm we select the attribute *S* which is giving highest information gain *IG(S*) or smallest entropy *H(S)* as the test attribute of current node.

Entropy [12]
Entropy *H(S*) is a measure of amount of uncertainty in the dataset S.
$H(S) = \sum p_i \ \log(1/p_i)$

Where,
S - Current dataset for which entropy is being calculated.
Pi – The proportion of the number of elements in class *i* to the number of elements in set S

When *H(S)* = 0, the set *S* is perfectly classified (i.e. all elements in *S* are of the same class)

Information gain[6]

Information gain *IG (A*) is a measure of the difference in entropy from before to after the set *S* is split on an attribute *A*. In other words, how much uncertainty in *S* was reduced after splitting set *S* on attribute *A*.
$IG\ (A, S) = H(S) - \sum P\ (t)\ H\ (t)$

Where,
H(S) – Entropy of set *S*
T- The subset created from splitting set S by attribute *A*.
P (t) - The proportion of the number of elements in *t* to the number of elements in set *S*.

Deal flow:

```
Split (node, {examples}):
1. A <- best attribute for splitting the {examples}
2. Decision attribute for this node <- A
3. For each value of A create child node
4. Split training {examples} to child node
5. for each child node / subset:
        If subset is pure: STOP
        Else: split (child node, {subset})
```

B) KNN:

It is a supervised learning method. In KNN data are represented in a vector space .For a given object E, get the top K dataset objects which are nearest to E by selecting distance measure .Then assign the class C to object E that represents the most objects after inspecting the class of these K objects .So for unknown tuple, KNN looks for pattern space for the k tuples which are closest to that tuple. These K tuples becomes the nearest neighbors of that unknown tuple.

To Find the Euclidean Distance between two points or tuples, the formula is given below

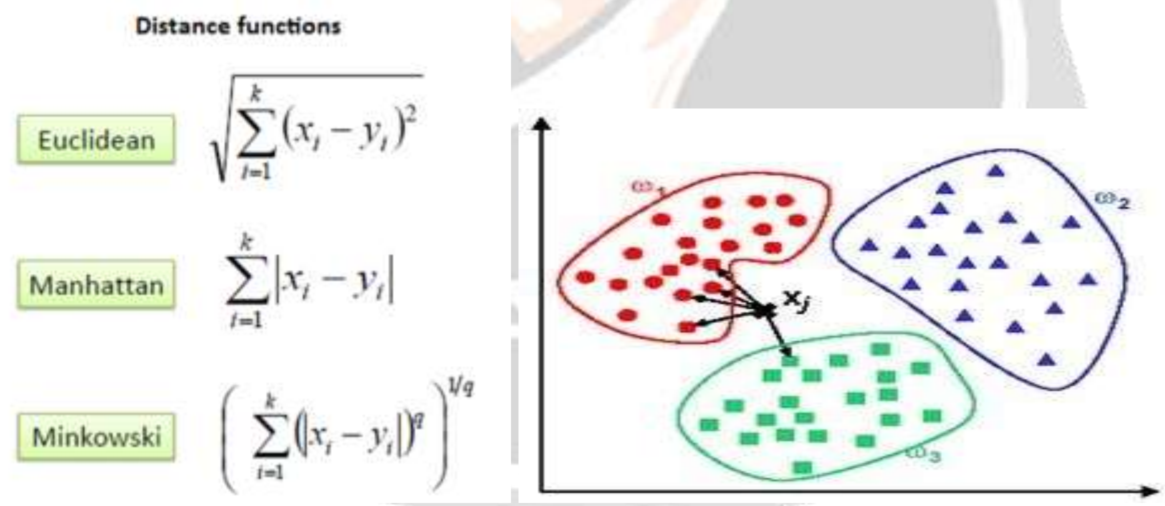
Distance functions
Euclidean
$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
Manhattan
$\sum_{i=1}^{k}|x_i - y_i|$
Minkowski
$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$
ω_1
ω_2
x_j
ω_3

Fig-.2.knn classification

K nearest neighbor stores all available cases and classifies new test case based on similarity measure.

Deal Flow:
Step 1- integer k is given with new dataset
Step-2
 If K=1, select the nearest neighbour
 If K>1,select the most frequent neighbour.
Step3-The classification is given to the new sample.

C) Naive bayes:

It is also known as bayes rule. Bayes theorem is used to find conditional probabilities. The conditional probability isis probability of an event provided with another event has happened. This algorithm is less computationally intense than other.

P(X|Y) is the conditional probability of event X occurring for the event Y which has already occurred.

P(X|Y) = P(X and Y)/P(A)

An initial probability is called as priori probability which we get before additional information is obtained.

The probability is called as posterior probability value which we get or revised after any additional information is obtained.

Given a hypothesis h and data D which bears on the hypothesis:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- *P(h)* = prior probability of hypothesis h
- *P(D)* = prior probability of training data *D*
- *P(h|D)* = probability of *h* given *D*
- *P(D|h)* = probability of *D* given *h*

V. CONCLUSION AND FUTURE WORK

In this paper we have studied video spam in popular site YouTube. Our study is based on the attributes gathered from YouTube .By crawling method on YouTube to obtain all attributes of the video. Then we apply different algorithms like ID3, K-NN, Navie Bayes to classify the video as spam or legitimate. In our proposed system all the problems in existing system will be removed and the optimum result is provided to the user.

In future work we will improve our technique Identifying Spammers on Social media such as facebook , twitter etc..And we will also take a look on detail Knowledge about previous spammer adding video on YouTube.

VI. ACKNOWLEDGMENTS

We would like to thank our guide Prof. R.S.Yenape who assisted us throughout our Work and for showing us a right track and giving us a motivation.

REFERENCES

[1] http://www.jeffbullas.com/2012/05/23/35-mind-numbing-youtube-facts-figures-and-statistics-infographic/

[2] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri, “Know your neighbors: Web spam detection using the web topology”, In Int’l ACM SIGIR, pp. 423–430, 2007.

[3] Alexa. http://www.alexa.com, last accessed 24th August, 2013

[4] R. Crane and D. Sornette, “Robust Dynamic Classes Revealed by Measuring the Response Function of a Social System”, In: National Academy of Sciences, 105(41):15649-15653, 2008.

[5] Z. Gy¨ongyi, H. Garcia-Molina, and J. Pedersen, “Combating web spam with trustrank”, In International Conference on Very Large Data Bases (VLDB), pp. 576–587, 2004.

[6]] https://en.wikipedia.org/wiki/ID3_algorithm

[7] L. Gomes, F. Castro, V. Almeida, J. Almeida, R. Almeida, and L. Bettencourt, “Improving spam detection based on structural similarity”, In USENIX Workshop on Steps to Reducing Unwanted Traffic on the Internet (SRUTI), pp.85-91, 2005.

[8] G. Koutrika, F. Effendi, Z. Gy¨ongyi, P. Heymann, and H. Garcia-Molina, “Combating spam in tagging systems”, In Proceedings of the 3rd international workshop on Adversarial information retrieval on the web(AIRWeb), pp. 57-64, 2007

[9] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida and M.Gonçalves, “Detecting Spammers and Content Promoters in Online Video Social Networks”, In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 620627,2009

[10] H. Aradhye, G. Myers, and J. Herson. “Image analysis for efficientcategorization of image-based spam e-mail.” In Proc. of the Int’l Conf.on Document Analysis and Recognition (ICDAR), volume 2, pp.914-918, 2005

[11] C. Wu, K. Cheng, Q. Zhu, and Y. Wu, “Using visual features for antispam filtering”, In Proc. of 4th IEEE Int’l Conf. on Image Processing (ICIP), 2005