ISSN 2348 – 8034
Impact Factor- 3.155

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

AN ANALYTICAL MODEL FOR THE SOCIAL INFLUENCE OF ONLINE VIDEOS

Vishnu K
Assistant Professor, Department of Computer Science and Engineering, Cochin College of Engineering and Technology, Malappuram, Kerala

ABSTRACT

A novel approach is proposed to analyze how a popular video is propagated in the cyberspace, to identify if it originated from a certain sharing-site, and to identify how it reached the current popularity in its propagation. In addition, influences are estimated across different websites outside the major hosting website. Web video is gaining significance due to its rich content. When a video receives some degree of popularity, it tends to appear on various websites including not only video-sharing websites but also news websites, social networks or even Wikipedia. As a result, it is becoming more difficult to determine how the propagation took place - was the video a piece of original work that was intentionally uploaded to its major hosting site by the authors, or did the video originate from some small site then reached the sharing site after already getting a good level of popularity, or did it originate from other places in the cyberspace but the sharing site made it popular. Existing study regarding this flow of influence is lacking. Literature that discusses the problem of estimating a video's influence in the whole cyberspace also remains rare. Therefore a novel framework is introduced to identify the propagation of popular videos from its major hosting site's perspective, and to estimate its influence. So Unified Virtual Community Space (UVCS) is used to model the propagation and influence of a video, and devise a novel learning method called Active Noise Control method (ANC) to effectively estimate a video's origin and influence.

Keywords: *ANC, UVCS.*

I. INTRODUCTION

Web mining is the one of the application of data mining techniques to discover patterns from the web. They can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the Internet. Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. Web content mining is the mining, extraction and integration of useful data, Information and knowledge from Web page content [1].Video is a visual presentation, typically a moving picture which is accompanied by sound. Along with the technical progress in Internet technology and broadband connections a booming number of web masters include videos in their websites [2]. As Internet users are less and less willing to read text, video is an excellent way to bring one's message across. Be it a testimonial or demonstration of a software, a filmed lecture or just something funny or intriguing, people enjoy watching movies.
Web Videos are either filmed videos taken with a camera (click here to view an example) or are screen-capture video where the action on a computer screen is recorded (click here for an example) [3]. As shown in the above figure the video is get from the web page and the features are extracted and then it is propagated in the social network. The URL of the video is detected and the origin and influence of the video is estimated [4]. A common activity on these networks is sharing of content.

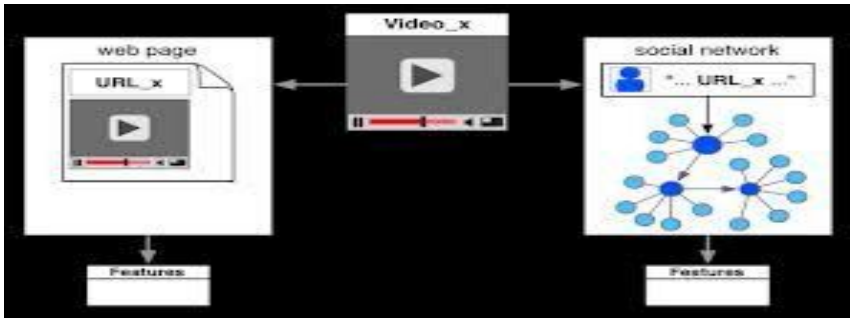

Fig 1: Propagation of online videos

ISSN 2348 – 8034
Impact Factor- 3.155

In such context, it is utterly important to identify how the propagation took place, i.e., to determine if a popular video on a video sharing website actually originated from that website, or it is merely a projection of influence from somewhere else of the cyberspace. Particularly, in this study we primarily focus on the identification of the propagation patterns of the web videos. We also study their influence in the entire cyberspace. The problem we aim to solve is partially similar to the analysis a user's friends and the identification of his/her influence in a social network, but there are some key differences. In influence analysis in social network, all users, or nodes from a network's perspective, are normally considered to be in a single website, in which a user's influence can be identified with existing approaches by analyzing the friend relationships and interactions with other users. In such case, the concept of origin for a user does not exist. However, the problem becomes more difficult if we consider an online video's propagation and influence as in this case multiple websites need to be examined. Due to the open nature of the Web, an online video's influence often exhibits a bi-directional fashion. On the one hand, a video's existence on a hosting site may be affected by some emerging events from other websites. On the other hand, a video originating from a hosting site makes its way to the most popular video inside the site, and then draws dramatic attention from other websites.

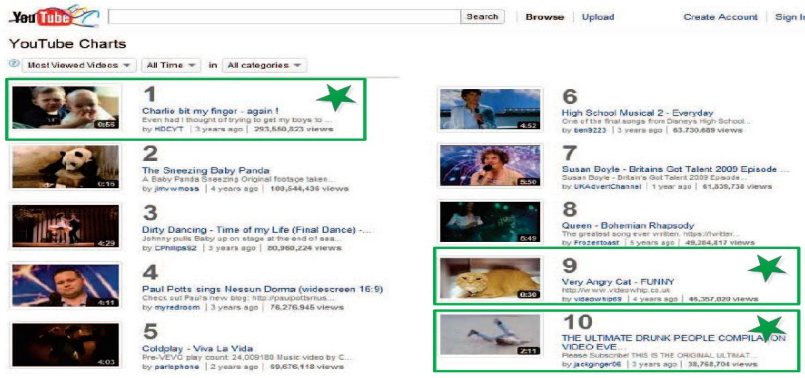

Fig 2: Popular online videos

Fig. 1 shows the most viewed ten videos in all time from the largest online video-sharing site YouTube.com. After close investigation of the videos' propagation in cyberspace, we conclude that video 1, 9 and 10, which are marked with stars, are the origins of other duplicate videos in the cyberspace. These videos originate from YouTube.com, and are then propagated to the cyberspace via other websites. During the process, they have drawn remarkable public attention from both YouTube.com and other websites.

II. SYSTEM MODEL

To model an online video's propagation and influence in the cross-community cyberspace, we define a Unified Virtual Community Space (UVCS) that captures the propagation history of an online video. The UVCS records key information of an online video, such as the video page's ranking in the search results for a text query with the video's title on search engines, and the information about the video page's inbound and outbound links, etc. UVCS is used as the raw feature for our algorithm to classify the propagation and rank the influence of an online video. A video's UVCS is independent from another video's UVCS.

We propose an advanced learning method called Noise-reductive Local-and-Global Learning (ANC) to effectively estimate a video's origin, influence and origin of video. The method should be able to reduce noise. The UVCS feature is a combination of multiple sematic components; the significance of each component is not specified in the raw feature. Fields of the UVCS feature may be missing for some feature vectors due to the diversified nature of web pages. Overall the feature is regarded very noisy.

The method should be able to learn a classification/ ranking model with a relatively limited number of learning data, and then apply the obtained model on out-of-learning-set data. Consequently, it should be able to preserve the structure of the learning set, but we also need to control the risk of over-fitting. With this model we are able to predict the propagation pattern and influence ranking for any new online video.

ISSN 2348 – 8034
Impact Factor- 3.155

III. FRAMEWORK

Formulation: *Given a set of videos V = {vi}, establish corresponding distinctive features X = {xi} to describe their patterns of propagation and overall influence, classify their propagation patterns into C = {ci} and evaluate their influence scores S = {si}.*

Fig. 3 outlines the proposed framework. The framework starts with the most popular videos retrieved from a particular sharing site that we aim to analyze. Those videos receive the most attention, which is reflected by their view counts, ratings and discussions, so they are collected as candidates. For each video candidate, its title text is used as search terms to search on search engines. The reason why we used this technique is threefold: firstly it involves dramatic effort to crawl all possible web pages and identify duplicate imbedded is videos to identify relevant pages; secondly some of the pages only have text reference to the video but have no links or actual embedded video on them; finally in our investigation we find that most of the relevant pages, with or without video links, could be accessed through text search. Missing a video's true origin in the text search engine results is highly unlikely. Hence we argue that this technique is effective enough for our task.

The pages returned by the search engines are analyzed, based on which a corresponding feature vector in the Unified Virtual Community Space (UVCS) is constructed. The UVCS is a feature space that consists of elements relevant to a video's propagation and influence, including the link relations of relevant pages, and the tracking of its presence on other websites, e.g. Twitter, Wikipedia, the Blogsphere, and the news websites. We formulate the feature for the candidate video so that the features are used in the ANC algorithm. At last, a web interface is illustrated to present the results of the analysis.

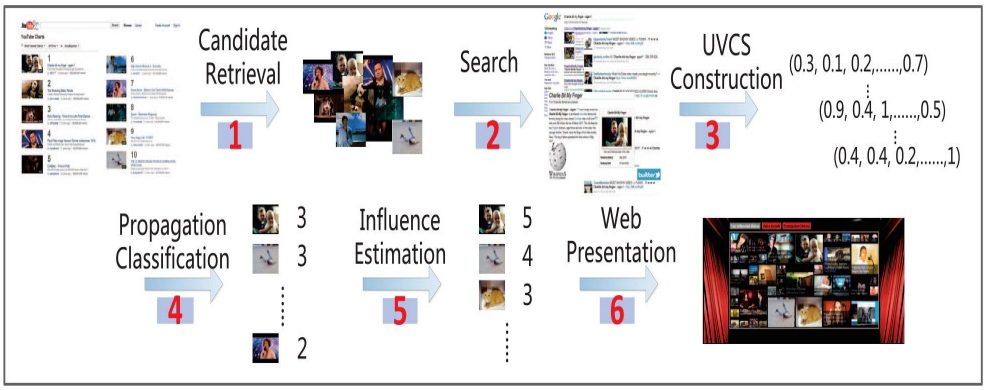

Fig 3: Framework

IV. UNIFIED VIRTUAL COMMUNITY SPACE

To reasonably represent the relevant factors of a video's propagation in cyberspace, we define the concept of the Unified Virtual Community

Space (UVCS) which contains the web pages relevant to the video. In this UVCS, all the pages have relevance to the video, and they have some intrinsic properties:

1) Each page has a time stamp, indicating the publishing time or modification time of the page. The temporal relationships among pages can be explored. An older page naturally shows a larger likelihood to be the origin of propagation of a video.

2) Each page receives a set of inbound links. The link relations among pages inside the UVCS and pages outside of UVCS are ignored. That means we count only the links from the pages in the UVCS to the the pages in the UVCS. A graph with pages that are related to the same video is then established.

3) Each page's rank in the UVCS is known. Provided by the search engine, this factor mainly reflects the importance of the website that hosts the page. Combined with the previous item, this item describes how important this page is from a Page Rank type of view. Nevertheless this factor does not provide much information about the direction of influence. Note that each video has its own individual UVCS. In other word, we describe a video's propagation with a separate UVCS in which it lives. Each video and its correspondent UVCS is independent from other videos and their UVCS. Such characteristics entitles the whole framework another advantage, that any learning model we trained will be able to be applied on out-of-sample videos, given that the new UVCS are constructed for the out-of-sample videos. Next we show how to construct the UVCS for a given video.

ISSN 2348 – 8034
Impact Factor- 3.155

A. FEATURE REPRESENTATION

Given an online video, its title is used as the search term to perform web searches on general web search engines (like Google web search) and specialized web search engines (Google blog search, Twitter search, etc.). After the top-ranked pages related to these terms are retrieved, they form the UVCS, which is then processed and analyzed. The properties of the UVCS are extracted and combined into a formally defined UVCS feature vector as follows:

$x = [nt, tr, r, ir, ni, ipr, nip, L1 \circ W1, L2 \circ W2, NT, NB, NN, wi]$.

The features can be considered as having three parts.

• [*nt, tr*] The first part preserves the temporal order in which all the related pages are published with *nt* and *tr*. It is an important indicator of the video page's roles during the propagation. *nt* and *tr* record the time stamp and the rank of the video page's time stamp in the UVCS. It describes the temporal relations between the video page and other related pages.

• [*r, ir, ni, ipr, nip,NT,NB,NN, wi*] The second part contains the search engine's rank of the video page as well as the related references in various communities of the UVCS. The components included here effectively reflect how popular the video is in other places of the cyberspace outside one sharing site. For example, *ir* and *ni* directly describe the within-UVCS linkage pattern for the page without considering other weights or importance of host websites.

• [$L1 \circ W1$, $L2 \circ W2$] The rest of the feature vector helps to evaluate both the propagation and the influence of the online video. The involvement of $L1 \circ W1$ and $L2 \circ W2$ is critical to the cases where the publishers themselves use the online video website as a primary channel to spread the video. Each component of the feature vector is normalized differently. Overall we use the rank of that component in the UVCS to normalize it relatively. For instance, if a page has a time stamp that says it's the 10th latest page in the UVCS which has 100 pages all together, then it is normalized to 0.9.

Intuitively, SVM or Least Square Regression can be used to classify the video's propagation and evaluate its influence. However, they will be quite ineffective due to the following reasons. Although the UVCS feature representation contains vital and discriminative information needed for the tasks, the downside of the vast range of information it carries is that the noise and bias are also introduced. For some videos, *r* may be dominant, while for others, *tr* and *nt* may be of greater importance. Other components in the feature vector also have the chance of being more significant than others for some videos. Meanwhile, the manual annotation for the training data may also introduce inconsistency. Naturally we consider using feature selection or dimension reduction to minimize the impact of the noisy feature space.

V. THE ANC ALGORITHM

The propagation of a video's influence may show different patterns. Naturally the video propagation analysis problem can be formulated as a classification problem. And this is the primary objective of the study. In addition, we also show the estimation of video influence can be modeled as a ranking problem, to demonstrate that ANC is capable of similar applications. In this chapter, we propose the ANC algorithm to unify the two problems into a single learning framework.

ANC is designed to fit in our application scenario. In our application, we will gather a collection of UVCS features for the popular online videos, however only a small portion of them will be annotated by expert annotators due to limited human resources. Whatever learning method used should be able to fully utilize the manually annotated data, and ideally it should use the manually annotated data to infer the un-annotated ones, and finally it should be able to generate a model to predict the labels or ranks for the out-of-sample data.

A.LABELLING THE UN-ANNOTATED SAMPLE DATA

Before we start to explain the ANC, we first obtain a training dataset with n total videos and $n\_$ manually annotated videos ($n\_ < n$), as $X = [x1, \ldots, xn]T \in Rn \times d$, where xi is the d-dimensional UVCS feature vector for the i^{th} video in the dataset. Now we denote the annotated videos as $Xl = [x1, \ldots, xn\_]T$. We use $Yl = [y1, \ldots, yn\_]T \in Rn\_ \times c$ to denote the class lables or ranks for the manually annotated videos Xl, where c is an integer greater or equal to To learn a classification model to predict the label of new videos, a straightforward method is to minimize the empirical error for the following regression model:

$$\underset{Wo}{\text{Min}} \sum_{i=1}^{n'} ||WoTxi - yi||2F,$$

ISSN 2348 – 8034
Impact Factor- 3.155

where $Wo \in Rd \times c$ is the classifier which is learnt from the manually annotated data and is able to predict the labels for the un-annotated data. The performance of this classifier is closely related to the number of manually labeled videos, namely n'. However, the availability of labeled videos is rather limited, as excessive human effort is needed for manually labeling a large-scale dataset. To leverage the unlabeled data for a better performance, it is desirable to design an algorithm which is able to predict the labels of the unlabeled videos by exploiting the data distribution of X.

B. UTILIZING THE ANNOTATED DATA

In a local learning method is proposed for cross media retrieval and it shows robust results in data ranking. The basic idea is to train a local linear regression model to predict the ranking score of each datum and its k-nearest neighbors, and then all the local linear regression models are optimized globally. We propose to employ a group of local linear regression models to predict. Suppose $Nk(xi)$ is a set that contains xi and its k-nearest neighbors in training set.

C. DIMENSION REDUCTION METHOD

Despite the high accuracy, a limitation is that it is not able to predict the labels of the videos outside the training set. To this end, we propose to simultaneously learn the predicted label of the training data and a classifier which can be used to predict the label of a video outside the training set. Meanwhile, the representation of the feature vector in the UVCS could be noisy, which may degrade the performance. Previous research efforts have shown that dimension reduction can remove irrelevant, redundant and
noisy information from a dataset, keeping only the informative, relevant or important information. Meanwhile, it has been shown in that it is beneficial for supervised learning to project the data into a low dimensional subspace when training the classifiers. Suppose there is a linear transformation, which transforms the video data X in feature space to a more compact and accurate representation
$Z \in Rn \times d'$, where $d' \leq d$ is the reduced dimensionality. The transformation between X and Z can be formulated as:

$$Z = XQ$$

where $Q \sum Rd \times d'$ is the transformation matrix. If Q is properly trained, the classification based on Z would result in better performance. While most of the existing classification algorithms rarely consider the correlation between the low dimensional subspace and a classifier during the training, we argue that such information is helpful for semi supervised learning. Therefore, we integrate training data label prediction, dimension reduction and classifier inference into a joint framework and propose to minimize the final objective function of ANC.

VI. EXPERIMENTS

A.EVALUATION

The evaluation for ANC follows the K-fold cross validation, where K=5 in our experiments. That means the 500 manually labeled videos are randomly partitioned into 5 folds, each of which contains 100 videos. The tests repeat 5 times. For each repeat, one of the folds is selected as the test data and the remaining 4 folds are used as part of the training data. The average performance of the 5 tests is reported. As for performance metric, the Area Under the receiver operator characteristic Curve (AUC) has been frequently used to evaluate the effectiveness of classification, model selection, etc. The receiver operator characteristic curve is a plot of the true positive rate as a function of the false positive rate of a classifier, and the AUC functions as a performance metric for misclassification costs of classifiers. Since our application stands as a multi-class classification, we use both microAUC and macroAUC [7] to measure the performance of our method. For the public influence ranking task, we use the average precision.

B. PERFORMANCE ANALYSIS

Four reference methods are used for comparison for the video propagation classification task, i.e. Least Square Regression (LSR), Support Vector Machine (SVM) [5], Classification with Dimensionality Reduction (CDR) [8], and LSR with Manifold Regularization (LSRMR) [1]. The ranking versions for these reference methods (such as in[4])are used to compare with ANC on the public influence task. For each reference method, their parameters have been carefully tuned and selected to achieve best performance. We use the same experimental procedure to evaluate their performance for the PC and IE tasks. The 5-fold cross validation is applied, and the average results are

ISSN 2348 – 8034
Impact Factor- 3.155

reported. Our method outperforms all 4 reference methods in all cases. For the PC task, the micro and macro AUC values of our method reach 0.86 and 0.736 respectively, while none of LSR, SVM or LSRMR has higher micro AUC value than 0.75, or macro AUC values higher than 0.64. Similarly, for IE the performance of ANC surpasses the others' by a great margin.

VII. CONCLUSION

Online videos are so popular nowadays that they begin to change people's way of daily entertainment greatly. The study of how online videos propagate and how influential they are outside a video sharing site is an increasingly significant research problem. The identification of an online video's origin and propagation patterns, from the video sharing site's perspective, is crucial to its business models as well to its partner's decision making for their marketing strategies. In this article, we propose a novel learning model for the classification of video propagation and demonstrate it can also be applied to the estimation of video influence. We determine if a popular online video originated from the video sharing site, or from somewhere else of the Internet. We determine how it got popular through its analyzing life cycle. We also give a rough estimation of its general influence on the Internet and also find the source of web videos.

REFERENCES

1. *M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," J. Mach. Learn. Res., vol. 7, pp. 2399–2434, Nov. 2006.*
2. *F. Benevenuto et al., "Understanding video interactions in youtube," in Proc. ACM Multimedia, New York, NY, USA, 2008, pp. 761–764.*
3. *M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. B. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in Proc. IMC, San Diego, CA, USA, 2007, pp. 1–14.*
4. *O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with SVMs," Inform. Retr., vol. 13, no. 3, pp. 201–215, 2010.*
5. *Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.*
6. *R. Ji, X. Xie, H. Yao, and W.-Y. Ma, "Mining city landmarks from blogs by graph modeling," in Proc. ACM Multimedia, Beijing, China, 2009, pp. 105–114.*
7. *D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in Proc. NIPS, 2003, pp. 595–602.*
8. *S. Ji and J. Ye, "Linear dimensionality reduction for multi-label classification," in Proc. 21st IJCAI, San Francisco, CA, USA, 2009, pp. 1077–1082.*
9. *J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," J. ACM, vol. 46, no. 5, pp. 604–632, 1999.*
10. *Y. Lee, H.-Y. Jung, W. Song, and J.-H. Lee, "Mining the blogosphere for top news stories identification," in Proc. 33rd SIGIR, New York, NY, USA, 2010, pp. 395–402.*
11. *C. X. Lin, B. Zhao, Q. Mei, and J. Han, "Pet: A statistical model for popular events tracking in social communities," in Proc. 16th KDD, Washington, DC, USA, 2010, pp. 929–938.*
12. *J. Liu et al., "Near-duplicate video retrieval: Current research and future trends," ACM CSUR, vol. 45, no. 4, p. 1, 2013.*