Discovering Advertisement Links by Using URL Text

Jing-Shan Xu[1, a], Peng Chang[2, b,*] and Yong-Zheng Zhang[2, c]

[1]School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, China

[2]Institute of Cyber Space Security, University of the Chinese Academy of Science, Beijing 100049, China

[a]xujingshan@iie.ac.cn, [b]changpeng@iie.ac.cn, [c]zhangyongzheng@iie.ac.cn

*Corresponding author

Keywords: Advertisement links, Text mining, SVM, Video capture.

Abstract. Capture of videos from websites is a basic work for searching video and analyzing video content. Discovery of advertising video URLs effectively and accurately is very important for video capture. It helps to improve the precision of video capture, optimize network utilization and reduce storage space. Currently, video content-based methods have a good ability to discover advertisement but in a limit speed, which do not meet the requirements of real projects. Since increasing achievements of URL based technologies have been made on classification subject of web pages, we utilize this technology to discover advertising video URLs. The method first produces a collection of URL segments. Next by applying N-gram feature selection, we get totally 2500 features. Afterwards, via combining the statistical information and selected word vector, the final features are generated. We use Naive Bayes, C4.5 Tree and SVM to train models. Ultimately, the experiment shows SVM is the most suitable model for discovery of advertising URLs discovery with 94% precision.

1. Introduction

Video on demand and share becomes increasingly popular in last few years. With its rich content, high speed of spread and easy operation, video on demand and share soon becomes an important part in many people's life. Take YouTube as an example, there are more than 13 billion active users in 2016. Owing to those developments, video related studies becomes more and more popular nowadays. In those studies, video capture is the foundation part especially when comes to web video search, network video content analysis, target video discovery etc.

The web video capture works in the following workflow. Initially, the web video capture will start by visiting the video pages. Further, it crawls the page content and communication information for analysis. In the end, it detects source URLs of videos and download them.

To discover the source URL of videos, we apply the methods below. It is readily to get static source URL by analyzing HTML Dom tree of a web page. Via utilizing browser plugin like chrome

extension, we are able to get the HTTP request and response from a page. By means of analyzing this information, the source URL about playing video is acquired. Furthermore, we capture the network traffic and restore HTTP or RTMP flows to find other source URLs that cannot be got by browser plugin. However, lots of ad video URLs are included in the result of video capture. We can set specific rule for specific website to sanitize them, but it is trivial and possibly gets omissions.

Many studies are aimed to solve the ads problem above. Rule-based method applies heuristic rules to low-level features to distinguish ad videos and other videos [1]. Logo-based method checks the key frames of videos through the presence or absence of station logos [2]. Furthermore, the machine learning based method uses multi-modal concepts to represent ad categories and is achieved with approximately 75% precision to identify ad videos [3]. It uses features like text, images and other resources extracted from videos to train models. Nevertheless, ads identification on capture of videos requires a higher accuracy and performance. Thus, they may be not appropriate for this problem.

URL based subject classification is widely used in many aspects. It only requires URL text for classification and quite efficient to filter web pages. In the medical-related web page classification, the method has about 87% precision [4]. Besides, there is 91% precision in classification of specific malicious URLs [5]. In this paper, we propose a URL subject classification based method to discover ad video links from the collection of video links. This method has 94% precision with quick speed to discover ad video links from thousands of URLs. The procedure of this method is shown in Fig. 1, and we will explain some critical processes afterwards.

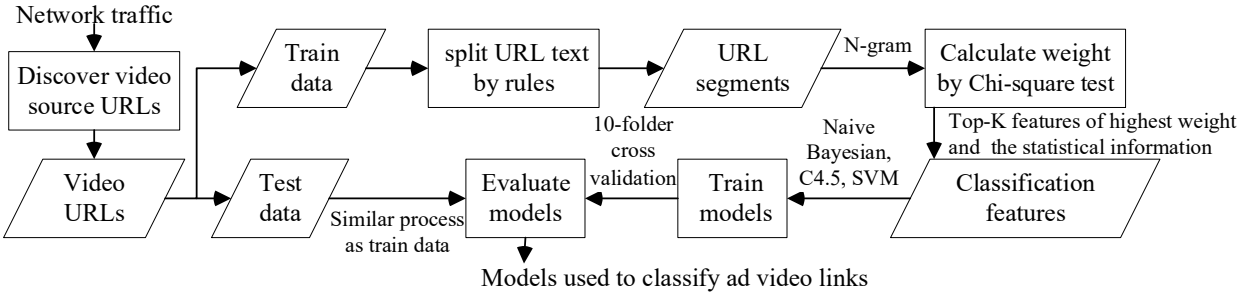

Fig.1 The flow chart of the URL based method

2. URL Feature Selection

URL Attributes. The URL attribute refers to some simple properties of the URL on the structure and content. It mainly contains three key parts below [6]. Firstly, the URL structure attribute is formatted by the structure "protocol://host:port/path?parameter#infomation-fragment". Secondly, the URL segment is constituted by a list of string after the URL string divided by some specific rules. Thirdly, the URL contains statistical information such as the length of the original URL, the quantity of URL segments, the number of parameters.

URL Segmentation. In order to get the classification feature, the first step is to divide the URL text into serval parts. The procedure is shown in Table. 1. At this point, we have a collection of URL segments, and next we can get the classification feature by using N-gram feature selection method.

Table. 1 The procedure of URL segmentation

Option Steps	Result	Description
Step1: Split whole URL	[http],[ips.ifeng.com],[video19.ifeng.com/video09/2016 /12/19/4423525-102-008-1300.mp4],[gid=F5uLEVlsyvi2]	protocol://host:port /path?parameter
Step 2: Split segments by specific mark	[http], [ips, ifeng, com], [video19, ifeng, com, video09, 2016, 12, 19, 4423525, 102, 008, 1300, mp4], [gid, F5uLEVlsyvi2]	Marks like "/", "_", "-", ".", "=", "&", "~" and so on
Step3: Participle segments	video19→video, 19; video09→ video, 09; Besides, mp4 will not be split, as it has meaning value in word segmentation dictionary	Use NLP methods and split string to words and digit
Step4: Count statistical infomation	4 segments in Step1, 18 segments in Step2, 20 in Step3. 9 pure digit strings, 11 pure strings only consist of letter. the extension is "mp4", the length of URL is 99, the number of parameter is 1	Count the quantity of segments in steps and quantity of pure digit and text finally

N-gram Feature Selection. The N-gram method divides the original string by a sliding window with length N to obtain a list of substrings. The original string here is a list of URL segments. For example, the original string equals {50, zera, com, swfs, 011, flv}. If N = 2, the sliding window works as Fig. 2, and we will produce a collection of 2-gram segments as
2-gram list equals {<50, zera>, <zera, com>, <com, swfs>, <swf, 011>, <011, flv>}

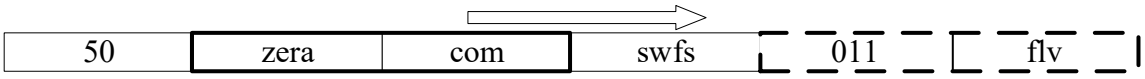

Fig. 2 the work instructions of the sliding window when N=2

Weight Calculation. When different values of N are used, the quantity of generated substring collections are quite different. These 1-gram, 2-gram segments will contain a large set of data, and also lead to a huge sparse matrix. To reduce the computational time and the complexity of space, we use an evaluation method to reduce the dimension of features. We calculate the weight of each segment in the set. Then, we sort them and get the top K features whose weight is the highest.

Some parts of URLs produced by hash function are meaningless, but have impact on precision of classification. As they share similar form in text structure, many of them are consist of letter and digit in random order. Thus, we check the segment set and set a basic low weight to this type of segments.

Finally, we use Chi-square test method to adjust the weight of features. The Chi-square test method analyzes the deviation of the actual and theoretical values. In the processing of selecting features, suppose the quantity of URLs is L, we divide the URLs into two sets, one is the set of ad video links as A, and another is the set of non-ad video links as B. We count the quantity of each set. After that, we count frequency AA as the quantity of a segment appears in A and frequency AB as it does not appear in A. Furthermore, we can get frequency BA and BB from B by the same way. Thus, we have Eq.1 to calculate the weight for each segment.

$$Q = \frac{L(AA \times BB - AB \times BA)^2}{(AA + BA)(AA + AB)(AB + BB)(AB + BA)} \tag{1}$$

3. Classification Experiment

Classification Algorithm. Based on the existing research, SVM and KNN algorithm works on to URLs classification effectively [7]. We will compare the performance of Naive Bayesian, C4.5 decision tree and SVM based on the classification result of ad video links.

Naive Bayesian is based on Condition Independence and Bayesian Theorem. It is less sensitive to missing data. It is very suitable for small-scale data collection and often used in text categorization.

Decision Tree is a method by extracting rules from a large number of cases and constructing a tree model to do classification. C4.5 is one type of the decision tree and uses the information gain rate as the attribute selection criteria. To prevent noise and loss of data, we use a post-pruning algorithm to enhance the model. We set the confidence factor C=0.25 and set the minimum instance value M=4.

SVM maximizes the distance between vectors and the hyperplane to find an optimal hyperplane. Besides, we use the polynomial function as the kernel function of SVM, and use SMO algorithm to optimize the model. SMO algorithm is a minimum optimization algorithm. Meanwhile it is a quadratic programming optimization algorithm with a faster processing speed. It has good performance in the text classification.

Dataset. In this experiment, we manually collect 40,000 video URLs from the major web video websites. In this dataset, there are 5,700 ad video links and 34,300 non-ad video links. The ad video links take 14.3% of the total dataset. When the pre-process is finished, we have 56 million segments of URLs. After merging the same segments, we approximately have 47,630 of 1-gram segments. We use N-gram method to get the 2-gram data. At last, we have 42,870 of 2-gram segments.

Evaluation. In this page, we use 10-folder cross validation method to test data. Meanwhile, we also apply F1-Meature value, which is an evaluation value based on the combination of precision and recall rate, to compare the accuracy of these classification models. The equations to calculate precision rate, recall rate and F1 value are displayed as Eq.2. TP is quantity of the URLs, which are ad video links and correctly classified as ad video links. FP is the quantity of URLs, which are not the ad video links but classified as ad video links. FN is the quantity of URLs, which are ad video links but classified as non-ad video links. T is the quantity of all URLs.

$$Precision = \frac{TP}{TP+FP} \quad Recall = \frac{TP}{TP+FN} = 1 - \frac{FN}{T} \quad F1\text{-}Meature = \frac{2TP}{2TP+FP+FN} \tag{2}$$

In Fig.3, we set different dimension value of K to see how it will affect the precision of classification. As a result, we know when K=2500, the classification has the best precision.

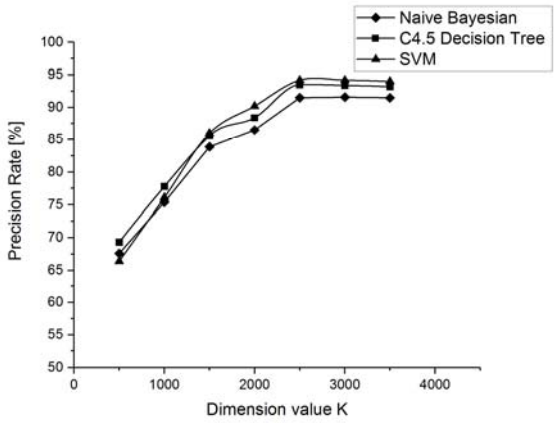

Fig. 3 Precision Rate under Different K

Table.2 the Result of Classification as K=2500

Result / Algorithm	Accuracy [%]	Recall [%]	F1-Meature [%]
Naive Bayesian	91.2	70.7	79.7
C4.5 Decision Tree	93.5	73.8	82.5
SVM	94.2	75.3	83.6

In Table.2, we display the final classification result when K=2500. From this table, we know that SVM has 94.2% precision among three algorithms. Thus, the method can help to reduce the download of useless resource, and will benefit for video capture.

4. Summary

This paper presents an URL based method to discover ad video links among lots of video links. The method first divides the URLs into serval segments and uses N-gram feature selection method to get the top-K highest weight features. Then it combines statistical information of URLs and selected features as the final features for classification. Furthermore, the paper compares the precision of Naive Bayesian, C4.5 decision tree and SVM on the classification of ad video links, and finds out a best model for ad videos discovery. According to the experiment, we know that when K=2500, SVM algorithm has best 94% precision to identify ad video links. It has better precision than the method based on video content using multi-modal features with 75% precision.

Acknowledgements

This work was supported by National Natural Science Foundation of Youth Fund (No.61303261).

References

[1] R. Lienhart, C. Kuhmunch, and W. Effelsberg. On the Detec- tion and Recognition of Television Commercials. Int. Conf. on Multimedia Computing and Systems, 1997, pp. 509-516.

[2] A. Albiol, etc. Detection of TV Commercials. Proc.ICASSP'04, Montreal, 2004, pp. 541-544.

[3] Wang J, Duan L, Xu L, et al. TV ad video categorization with probabilistic latent concept learning[C]//Proceedings of the international workshop on Workshop on multimedia information retrieval. ACM, 2007: 217-226.

[4] Rajalakshmi R. Identifying Health Domain URLs using SVM[C]//Proceedings of the Third International Symposium on Women in Computing and Informatics. ACM, 2015: 203-208

[5] Darling M, Heileman G, Gressel G, et al. A lexical approach for classifying malicious URLs[C]//High Performance Computing & Simulation (HPCS), 2015 International Conference on. IEEE, 2015: 195-202

[6] Jebari C. A Pure URL-Based Genre Classification of Web Pages[C]//2014 25th International Workshop on Database and Expert Systems Applications. IEEE, 2014: 233-237.

[7] Rajalakshmi R, Aravindan C.Web page classification using n-gram based URL features[C]//2013 fifth international conference on advanced computing (ICoAC). IEEE, 2013: 15-21.