IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

CHIRRUP BIFURCATION ALONG WITH ITS APPLIANCE TO NAMED INDIVIDUAL ACKNOWLEDGMENT

D.M.CHITRA.,MCA.,M.Phil.,

Assistant professor,

PG & Research Department of Computer Science And Applications

Padmavani Arts and Science College for women, Salem-11

K.SATHYA.,MCA.,M.Phil.,

Assistant professor,

PG & Research Department of Computer Science And Applications

Padmavani Arts and Science College for women, Salem-11

ABSTRACT

Twitter is a social network and it can be used by more number of peoples. While open the normal browser many number of pages are opened .In that, number of suspicious pages and malicious pages are occur. It detects the unwanted pages based on the redirect chain and correlation of redirection chain features are extracted from the suspicious URL.Twitter is prone to malicious tweets containing URLs for spam, phishing, and malware distribution. Twitter spam detection schemes utilize account features such as the ratio of tweets containing URLs and the account creation date, or relation features in the Twitter graph. These detection schemes are ineffective against feature fabrications. Since it consume much time and more resources.

Conventional suspicious URL detection schemes utilize several features including lexical features of URLs, URL redirection, HTML content, and dynamic behavior. This system investigates correlations of URL redirect chains extracted from several tweets. Because attackers have limited resources and usually reuse them, their URL redirect chains frequently share the same URLs. It developmethods to discover correlated

URL redirect chains using the frequently shared URLs and to determine their suspiciousness. Evaluation results show that the classifier accurately and efficiently detects suspicious URLs and it provides high security against suspicious uniform resource location in social networks.

1. INTRODUCTION

MICROBLOGGING sites such as Twitter have reshaped the way people find, share, and disseminate timely information. Many organizations have been reported to create and monitor targeted Twitter streams to collect and understand users' opinions. Targeted Twitter stream is usually constructed by filtering tweets with predefined selection criteria (e.g., tweets published by users from a geographical region, tweets that match one or more predefined keywords). Due to its invaluable business value of timely information from these tweets, it is imperative to understand tweets' language for a large body of downstream applications, such as named entity recognition event detection and summarization, opinion mining, sentiment analysis and many others. Given the limited length of a tweet (i.e., 140 characters) and no restrictions on its writing styles, tweets often contain grammatical errors, misspellings, and informal abbreviations.

The error-prone and short nature of tweets often make the word-level language models for tweets less reliable. For example, given a tweet "I call her, no answer. Her phone in the bag, "she dancin", there is no clue to guess its true theme by disregarding word order (i.e., bag-of-word model). The situation is further exacerbated with the limited context provided by the tweet. That is, more than one explanation for this tweet could be derived by different readers if the tweet is considered in isolation. On the other hand, despite the noisy nature of tweets, the core semantic information is well preserved in tweets in the form of named entities or semantic phrases.

We focus on the task of tweet segmentation. The goal of this task is to split a tweet into a sequence of consecutive n-grams (n 1Þ, each of which is called a segment. A segment can be a named entity (e.g., a movie title "finding nemo"), a semantically meaningful information unit (e.g., "officially released"), or any other types of phrases which appear "more than by chance".

In fact, segment-based representation has shown its effectiveness over word-based representation in the tasks of named entity recognition and event detection. To achieve high quality tweet segmentation, we propose a generic tweet segmentation framework, named Hybrid Seg. HybridSeg learns from both global and local contexts, and has the ability of learning from pseudo feedback. Global context. Tweets are posted for information sharing and communication. The named entities and semantic phrases are well preserved in tweets. The global context derived from Web pages (e.g., Microsoft Web N-Gram corpus) or Wikipedia therefore helps identifying the meaningful segments in tweets. The method realizing the proposed framework that solely relies on global context is denoted by Local context Tweets are highly time-sensitive so that many emerging phrases like “She dancin” cannot be found in external knowledge bases.

HybridSegNGram segments tweets by estimating the term-dependency within a batch of tweets. Pseudo feedback. The segments recognized based on local context with high confidence serve as good feedback to extract more meaningful segments. The learning from pseudo feedback is conducted iteratively and the method implementing the iterative learning is named Hybrid SegIter. HybridSegNER is less sensitive to parameter settings than Hybrid SegNGram and achieves better segmentation quality.

MODULES

- Tweet collection Module
- Crawling tweets from Twitter Module
- Twitter Search API Module
- Filtering tweets using machine learning Module
- Semantic Analysis on Tweets Module
- Earthquake reporting System Module

Tweet collection Module

In this module, we develop our system by posting tweets by the users. It is necessary to collect tweets referring to an earthquake from Twitter.

This process includes two steps:

1. Crawling tweets from Twitter
2. Filtering out tweets that do not refer to the earthquake.

For crawling and filtering tweets, script programming languages is used.

Crawling tweets from Twitter Module

To collect tweets or some user information from Twitter, one must use the Twitter Application Programmers Interface (API). Twitter API is a group of commands that are necessary to extract data from Twitter.

Twitter has APIs of three kinds:

- Search API
- REST API
- Streaming API

Filtering tweets using machine learning Module

We collected data from tweets including keywords related to earthquakes, such as earthquake, shake. Those tweets include not only tweets that users posted immediately after they felt earthquakes, but also tweets that users posted shortly after they heard earthquake news, or perhaps they misinterpreted some sense of shaking from a large truck passing nearby.Creation of a classifier to categorize crawled tweets into positive tweets and negative tweets, using Support Vector Machine: a supervised learning method.

Semantic Analysis on Tweets Module

Semantic Analysis on Tweet Create classifier for tweets use Support Vector Machine (SVM) Features (Example: I am in Japan, earthquake right now!) Statistical features (7 words, the 5th word) the number of words in a tweet message and the position of the query within a tweet Keyword features (I, am, in, Japan, earthquake, right, now) the words in a tweet Word context features (Japan, right) the words before and after the query word

Earthquake Reporting System Module

In this module, the users will be altered if the earthquake occurs based on their location and the tweets. Effectiveness of alerts of this system Alert E-mails urges users to prepare for the earthquake if they are received by a user shortly before the earthquake actually arrives.

METHODOLOGY

Twitter Social Network

The Twitter is a social networking site just like Facebook and MySpace except that it only provides a microblogging service where users can send short messages (referred to as tweets) that appear on their friends' pages. A Twitter user is only identified by a username and optionally by a real name. User X who is "followed" can follow back if she so desires. Tweets can be grouped using hashtags which are popular words, beginning with a "#" character. When a user likes someone's tweet, she can "retweet" that message. As a result, that message is shown to all her followers. A user can decide to protect her profile. By doing so, any user who wants to follow that private user needs her permission. Twitter is the fastest growing social networking site with a reported growth rate of 660%.

Random Walk Algorithm

The Random Walk algorithm Personalized Page Rank vectors (PPVs) consists on a ranked sequence of WordNet.Synsets weighted according to a random walk algorithm. Taking the graph of WordNet[10], where nodes are Synsets and axes are the different semantic relations among them, and the terms contained in a tweet, A Synset is the basic item of information in WordNet and it represents a "concept" that is unambiguous. Most of the relations over the lexical graph use Synsets as nodes (hyperonymy, synonymy, homonymy and more). SentiWordNet returns from every Synset a set of three scores representing the notions of "positivity", "negativity" and "neutrality". Therefore, every concept in the graph is weighting according to its subjectivity and polarity. The last version of SentiWordNet[28] (3.0) has been constructed starting from manual annotations of previous versions, populating the whole graph by applying a random walk algorithm.

Classification of Suspicious URL

Features for classifying suspicious URLs on Twitter are based on two types:

- Features derived from correlated URL redirect chains length
- Features derived from related tweet context information

All statistics were checked in every 10,000 tweets and only consider URLs had appeared more than once.

Features Derived from Correlated URL Redirect Chains length

Attackers usually use long URL redirect chains to make investigations more difficult and avoid dismantling of their servers. Therefore, when an entry point URL is malicious, its chain length may be longer than those of benign URLs.

Frequency of entry point URL: The number of occurrences of the current entry point URL within a tweet window is important. Frequently appearing URLs that are not whitelisted are usually deemed suspicious.

Features Derived from Tweet Context Information

The Features derived from the related tweet context information are variations of previously discovered feature. Preparing a large number of dissimilar Twitter accounts for distributing spam URLs becomes a burden to attackers; therefore, similarity checking is effective.

Relative number of different source applications: Sources are applications that upload the current entry point URL to Twitter. Attackers usually use the same source application as maintaining a number of different applications is difficult. Benign users, however, typically use various Twitter applications, such as TweetDeck and Echofon. Therefore, the number of different sources may be small when the current entry point URL is suspicious.

Suspicious URL Detection

A number of suspicious URL detection schemes[2] have also been introduced. They use static or dynamic crawlers, and they may be executed in virtual machine honey pots, such as Capture-HPC, HoneyMonkey, and Wepawet, to investigate newly observed URLs. These schemes classify URLs according to several features including lexical features of URLs, DNS information, URL redirections, and the HTML content of the landing pages.

Spam Detection

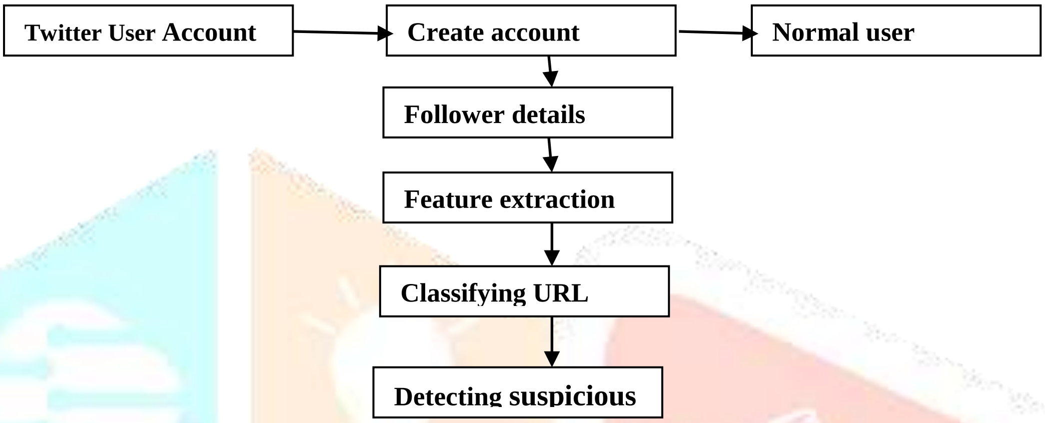
The feasibility of applying a supervised learning algorithm along with the attributes for the task of detecting spammers on Twitter. In this approach, each user is represented by a vector of values, one for each attribute. The algorithm learns a classification model from a set of previously labeled (i.e., pre-classified) data, and then applies the acquired knowledge to classify new (unseen) users into two classes: **Spammers and Non-Spammers**.

Spam and non-spam accounts extract the features that can effectively distinguish spam from non-spam accounts. Twitter stream apply **Kullback–LeiblerDivergence (KLD)** between their respective language models. **KLD** is an asymmetric divergence measure originating in information theory.

Twitter Spam Detection

To cope with malicious tweets, several Twitter spam detection schemes, have been proposed. These schemes can be classified into **Account feature-based, relation feature-basedand Message feature-based schemes**.

ARCHITECTURE



CONCLUSION

Thus it concluded that comparing the existing methods of suspicious URL detection utilizes much resources and it consuming more time to detect the suspicious URL.It used account feature-based, relation feature-based and message feature-based schemes. However, malicious users can easily fabricate these account and messagefeatures. The relation feature-based schemes rely on more robust features that malicious users cannot easily fabricate such as the distance and connectivity apparent in the Twitter graph.

So, in proposed system a new suspicious URL detection method was used. It used supervised learning algorithm to detect and classify suspicious URL.It extracts feature vectors such as URL redirect chain length, IP address, and domain name. It also addresses dynamic and multiple redirections. The final goal of the proposed method is it introduced some newfeatures on the basis of these correlations, and the system's accuracy and performance was increased.

REFERENCES

[1]C. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee,"Twiner: Named entity recognition in targeted twitter stream," inProc. 35th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2012,pp. 721–730.

[2] C. Li, A. Sun, J. Weng, and Q. He, "Exploiting hybrid contexts fortweet segmentation," in

Proc. 36th Int. ACM SIGIR Conf. Res.Develop. Inf. Retrieval, 2013, pp. 523–532.

[3] A. Ritter, S. Clark, Mausam, and O. Etzioni, “Named entity recognition in tweets: An experimental study,” inProc. Conf. EmpiricalMethods Natural Language Process, 2011, pp. 1524–1534.

[4] X. Liu, S. Zhang, F. Wei, and M. Zhou, “Recognizing named entities in tweets,” inProc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Language Technol, 2011, pp. 359–367.

[5] X. Liu, X. Zhou, Z. Fu, F. Wei, and M. Zhou, “Exacting socialevents for tweets using a factor graph,” inProc. AAAI Conf. Artif.Intell, 2012, pp. 1692–1698.

[6] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang, “Discover breaking events with popular hashtags in twitter,” inProc. 21st ACM Int. Conf. Inf. Knowl. Manage, 2012, pp. 1794–1798.

[7] A. Ritter, Mausam, O. Etzioni, and S. Clark, “Open domain eventextraction from twitter,” in

Proc. 18th ACM SIGKDD Int. Conf.Knowledge Discovery Data Mining, 2012, pp. 1104–1112.

[8] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang, “Entity-centric topic-oriented opinion summarization in twitter,” inProc.18th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining.

[9] Z. Luo, M. Osborne, and T. Wang, “Opinion retrieval in twitter,”inProc. Int. AAAI Conf. Weblogs Social Media, 2012, pp. 507–510.

[10] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, “Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach,” inProc. 20th ACM Int. Conf. Inf. Knowl. Manage,2011, pp. 1031–1040.

[11] K.-L. Liu, W.-J. Li, and M. Guo, “Emoticon smoothed languagemodels for twitter sentiment analysis,” inProc. AAAI Conf. Artif.Intell, 2012, pp. 1678–1684.

[12] S. Hosseini, S. Unankard, X. Zhou, and S. W. Sadiq, “Location oriented phrase detection in microblogs,” in Proc. 19th Int. Conf. Database Syst. Adv. Appl, 2014, pp. 495–509.

[13] C. Li, A. Sun, and A. Datta, “Twevent: segment-based event detection from tweets,” inProc. 21st ACM Int. Conf. Inf. Knowl. Manage,2012, pp. 155–164.

[14] L. Ratinov and D. Roth, “Design challenges and misconceptions in named entity recognition,” inProc. 13th Conf. Comput.Natural Language Learn., 2009, pp. 147–155.

[15] J. R. Finkel, T. Grenager, and C. Manning, “Incorporating non-local information into Information extraction systems by Gibbssampling,” inProc. 43rd Annu. Meeting Assoc. Comput. Linguistics,2005, pp. 363–370.

[16] G. Zhou and J. Su, “Named entity recognition using an hmm-based chunk tagger,” inProc. 40th Annu. Meeting Assoc. Comput.Linguistics, 2002, pp. 473–480.