International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 01 | Jan -2017 www.irjet.net p-ISSN: 2395-0072

A Paper on Web Data Segmentation for Terrorism Detection using Named Entity Recognition Technique

Ms. Pooja S. Kade[1] , Prof. N.M. Dhande[2]

[1]Student of Computer Science & Engineering, RTMNU University, A.C.E, Wardha, Maharashtra, India.
[2]HOD of Computer Science & Engineering,RTMNU University, A.C.E, Wardha, Maharashtra, India.

Abstract-*Terrorism has grown day by day, its roots quite deep in some parts of the world. With increasing terrorist activities it has become very important to control terrorism and stop its spread before certain time period. So as identified that internet is a major source of spreading terrorism through speeches, images and videos. Terrorist organizations use internet to brain wash individuals and younger's and also promote terrorist activities through provocative web pages that inspire helpless people and college student to join terrorist organizations. So here we propose an efficient web data mining system and segmentation technique to detect such web properties and mark them automatically for human review. Websites created in various platforms have different data structures and are difficult to read for a single algorithm so we use DOM Tree concept to extract the web data and SIFT feature for edge extraction that organized web data. Also we use Kmeans algorithm for segmentation and KNN for classification. In this way we may judge web pages and check if they may be promoting terrorism or not. This system proves useful in anti terrorism sectors and even search engines to classify web pages into the different category.*

Key Words: Data mining, Web mining, Patterns, DOMTree technique, object recognition, Segmentation.

1. INTRODUCTION

Terrorism has grown its roots quite deep in certain parts of the world. With increasing terrorist activities it has become important to curb terrorism and stop its spread before a certain time. So as identified internet is a major source of spreading terrorism through speeches and images. Terrorist organizations use internet to brain wash individuals and also promote terrorist activities through provocative web pages that inspire helpless people to join terrorist organizations. So here we propose an efficient web data mining system to detect such web properties and flag them automatically for human review. Data mining is a technique used to mine out patterns of useful data from large data sets and make the most use of obtained results. Data mining as well as web mining is used together at times for efficient system development. Web mining also consists of text mining methodologies that allow us to scan and extract useful content from unstructured data. Text mining allows us to detect patterns, keywords and relevant information in unstructured texts. Both Web mining and data mining systems are widely used for mining from text. Data mining algorithms are efficient at manipulating organized data sets, while web mining algorithms are widely used to scan and mine from unorganized and unstructured web pages and text data available on the internet. Websites created in various platforms have different data structures and are difficult to read for a single algorithm. Since it is not feasible to build a different algorithm to suit various web technologies we need to use efficient web mining algorithms to mine this huge amount of web data. Web pages are made up of HTML (Hypertext markup language) in various arrangements and have images, videos etc intermixed on a single web page. So here we propose to use DOM Tree concept to extracting text data from web pages and smartly designed web mining algorithms to mine textual information on web pages and detect their relevancy to terrorism. In this way we may judge web pages and check if they may be promoting terrorism. This system proves useful in anti terrorism sectors and even search engines to classify web pages into the category. Their relevancy to the field helps classify and sort them appropriately and flag them for human review.

2. LITERATURE SURVEY

1. TwiNER: Named Entity Recognition in Targeted Twitter Stream [1]

In this paper, NER system is used for targeted Twitter stream and is called as TwiNER to address the challenge of named entity recognition.

2. Named entity recognition in tweets: An experimental study [2]
In this paper, author addresses the issue of re-building the part-of-speech tagging and identifies the named entity from the tweets.

3. Recognizing named entities in tweets. [3]
This paper proposes a K-Nearest Neighbors (KNN) classifier and a linear Conditional Random Fields (CRF) to tackle the challenges of Named Entities Recognition (NER) for tweets lie in the insufficient information in a tweet.

4. Exacting Social Events for Tweets Using a Factor Graph. [4]
In this paper, author introduces the task of social event extraction for tweets, an important source of fresh events. One very important challenge is the lack of information in a single tweet, which is because of short and noise-prone nature of tweets. author propose to collectively extract social events from multiple similar tweets using a novel factor graph, to collect the redundancy in tweets, i.e., the repeated occurrences of a social event in several tweets.

5. Open Domain Event Extraction from Twitter. [5]
This paper describes TwiCal the rest open-domain event-extraction and categorization system for extracting the open events from the tweets of Twitter.

6. Entity-Centric Topic-Oriented Opinion Summarization in Twitter. [6]
In this paper, author presents an entity-centric topic-oriented opinion summarization framework, by using this framework author wants to extract opinions of user on different topics from the tweets of twitter.

7. Opinion retrieval in twitter. [7]
This paper considers the problem of finding opinionated tweets about any given topic. They automatically construct opinionated lexica from sets of different tweets which matching specific patterns to indicate opinionated messages.

8. Topic Sentiment Analysis in Twitter: A Graph-based Hash tag Sentiment Classification Approach. [8]
This paper proposes sentiment classification of hash tags in Twitter. Author believes this is important for sentiment analysis of topics since hash tags can be approximately viewed as user-commented topics.

9. Twevent: Segment-based Event Detection from Tweets. [9]
In this paper, author presents a novel event detection system for Twitter stream, called as Twevent to find out the dangerous impacts of tweets. Also this paper detects the segment based different events.

10. Design Challenges and Misconceptions in Named Entity Recognition. [11]
In this paper, important analysis of challenges and misconceptions takes place. In particular, author address issues such as the representation of text chunks.

11. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. [12]
In this paper, author uses, Gibbs Sampling method to Incorporating Non-local Information into Information Extraction Systems.

12. Named Entity Recognition using an HMM-based Chunk Tagger. [13]
This paper proposes a Hidden Markov Model (HMM) and an HMM-based chunk tagger, from which named entity (NE) recognition (NER) system is used, recognize and classify names, times and numerical quantities.

13. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. [15]
In this paper author address the problem of part-of-speech tagging for English data from the popular micro blogging service Twitter and develop a tagset, annotate data, develop features, and report tagging results nearing 90% accuracy.

14. To Link or Not to Link? A Study on End-to-End Tweet Entity Linking. [14]
In this paper, author proposes a structural SVM method to address the problem of end-to-end entity linking on Twitter.

3. PROBLEM STATEMENT

This paper presents an investigation of the real-time nature of Twitter that is designed to ascertain whether we can extract valid information from it. They propose an event notification system that monitors tweets and delivers notification promptly using knowledge from the investigation. In this research, they take three steps: first, They crawl numerous tweets related to target events; second, they propose probabilistic models to extract events from those tweets and estimate locations of events; finally, they developed an alerting reporting system that extracts earthquakes from Twitter and sends a message to registered users. Here, they explain our methods using an earthquake as a target event. This system needs a full time

administrator to look at system. It is restricted for twitter only.

4. OBJECTIVES

The main objectives of the study are listed below:

- Implement Text Clustering technique to make web data cluster.
- Implement Web Crawler technique to check online website data.
- To extract web page data efficiently use DOM Tree.
- To perform image processing to detect terrorism using SIFT algorithm.

5. PROPOSED SYSTEM

The proposed work is planned to be carried out in the following manner

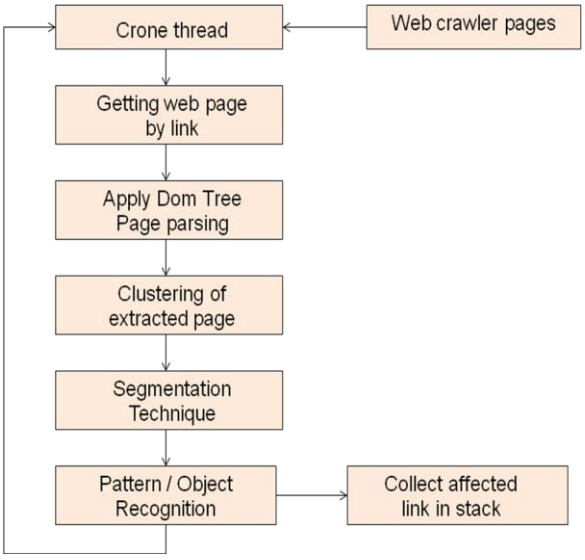


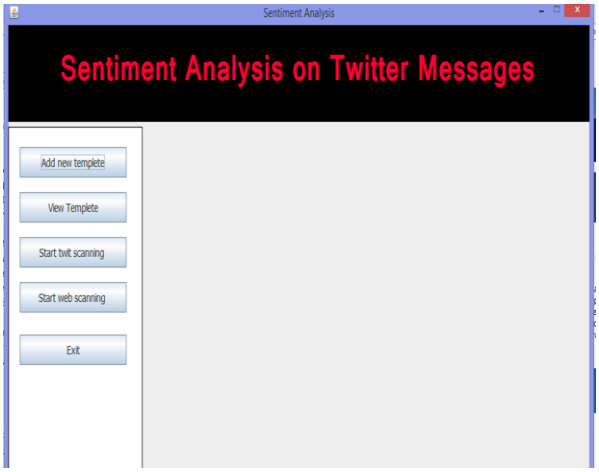
Fig 1: Basic System Architecture

Fig. shows the basic system architecture of proposed system, Firstly, We have a crone thread which accepts web page links from web crawler then apply parsing on that links using DOM Tree. After applying parsing we will do clustering of extracted pages using k-means clustering technique. Then as shown in architecture segmentation is applied on clustered data and lastly pattern matching is done for text data and object recognition is done for image objects. If any link is affected or promotes terrorism then we will store it in stack for further processing otherwise go back to crone thread and same process will follow for next website or web link.

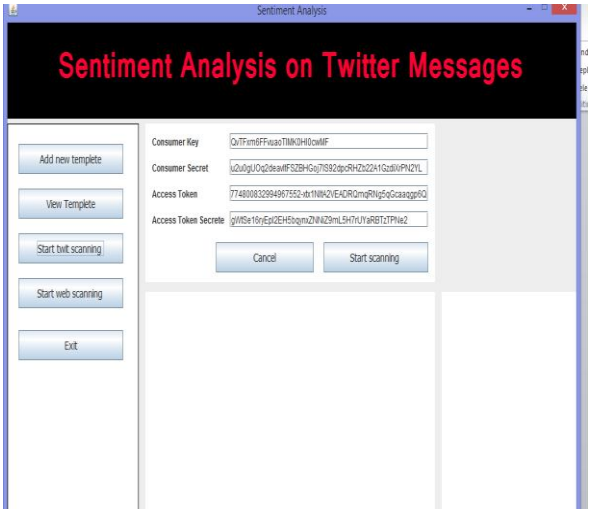
6. IMPLEMENTATION

Module1:

In first module we will doing the tweet scanning that means we will gather the twitter data after data gathering we perform segmentation on that data using the Named Entity Segmentation technique.



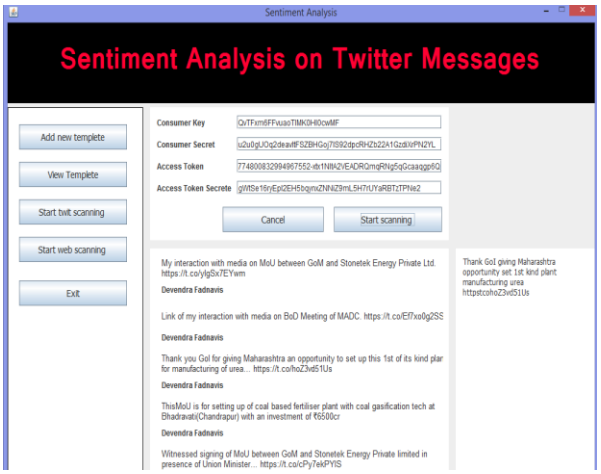
Screenshot 1



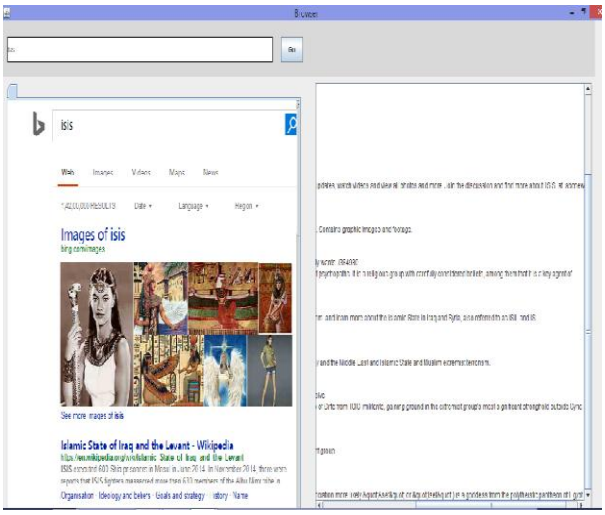
Screenshot 2

Module 2:

In second module we have do the web data segmentation. For which we first gather the web data then apply Dom Tree technique for extracting the text data. After this we had done the segmentation of text and then apply clustering technique on segmented data.



Screenshot 3



Screenshot 4

7. CONCLUSION

In this paper we have look out the different papers and from that we conclude that taking the reference of that papers we design the terrorism analysis system using some better and new techniques like DOM Tree for web data extraction, SVM algorithm for classification, SIFT algorithm for object recognition and extraction and kmeans algorithm for segmentation of text data.

REFERENCES

[1] 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.

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

[3] X. Liu, X. Zhou, Z. Fu, F. Wei, and M. Zhou, "*Exacting social events for tweets using a factor graph*," in Proc. AAAI Conf. Artif. Intell., 2012, pp. 1692–1698.

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

[5] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang, "*Entitycentric topic-oriented opinion summarization in twitter*," in Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2012, pp. 379–387.

[6] Z. Luo, M. Osborne, and T. Wang, "*Opinion retrieval in twitter*," in Proc. Int. AAAI Conf. Weblogs Social Media, 2012, pp. 507–510.

[7] K.-L. Liu, W.-J. Li, and M. Guo, "*Emoticon smoothed language models for twitter sentiment analysis*," in Proc. AAAI Conf. Artif. Intell., 2012, pp. 1678–1684.

[8] C. Li, A. Sun, and A. Datta, "*Twevent: segment-based event detection from tweets*," in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012, pp. 155–164.

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

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

[11] L. Ratinov and D. Roth, "*Design challenges and misconceptions in named entity recognition,*" in Proc. 13th Conf. Comput. Natural Language Learn., 2009, pp. 147–155.

[12] J. R. Finkel, T. Grenager, and C. Manning, "*Incorporating nonlocal information into information extraction systems by Gibbs sampling*," in Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics, 2005, pp. 363–370.

[13] G. Zhou and J. Su, "*Named entity recognition using an hmmbased chunk tagger*," in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, 2002, pp. 473–480.

[14] S. Guo, M.-W. Chang, and E. Kiciman, "To link or not to link? A study on end-to-end tweet entity linking," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technol.,
2013, pp. 1020–1030.

[15] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J.Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: annotation, features, and experiments," in Proc. 49th Annu. Meeting. Assoc. Comput.Linguistics: Human Language Technol., 2011, pp. 42–47.