Review - Using Entropy and Mutual Information to Extract Threat Actions from Cyber Threat Intelligence

The objective of this paper is to develop a computational framework to "engineer" an unstructured report into a structured list of threat actions, as the key information of the report for the purpose of further analysis by machine and timely defensive strategy. Specifically, this approach uses some Information-Theoretic Metrics combined with the basic NLP technique to extract the malicious actions from the shared threat reports.

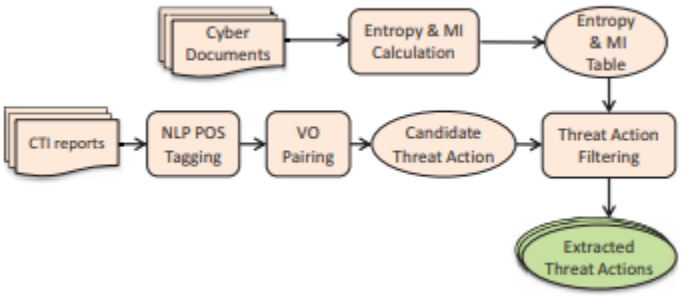

Fig. 2: The Architecture of *ActionMiner*

Three main contributions

- A framework, ActionMiner, to reduce each unstructured CTI report to a structured list of threat actions for the purpose of future analysis.
- Used entropy and mutual information, two metrics from Information Theory, combined with some basic NLP techniques, to accurately extract each cyber threat action, represented as a VO (verb-object) combination.
- A valuable insight into the common language used in this community and help to better understand future reports.

NLP techniques used

NLP POS tagging tool labels each word in the Cyber documents and CTI report with a POS tag, such as VB (verb base form), VBD (verb past tense), NN (singular or mass noun), NNS (plural noun), NNP (singular proper noun) and NNPS (plural proper noun), making the reports ready for further processing.

Entropy and mutual information

Entropy, a core concept in Information Theory. It measures the degree of uncertainty or disorder of information. Entropy is calculated as

$$H(X) = -\sum_{x \in X} p(x) log_2 p(x)$$

Information Theory is mutual information (MI), which is defined by two random variables X and Y. MI measures the reduction of the entropy of X due to the knowledge about Y.

$$MI(X;Y) = \frac{H(X) - H(X|Y)}{\sqrt{H(X) \times H(Y)}}$$

Entropy is used to understand the specificity levels of the action verbs used in CTI report and mutual information to test if extracted VO pairs co-occur frequently enough to be reasonably regarded as a valid pair.

Dataset used

- 18,000 Wikipedia articles broadly related to computer and information technology and specifically related to cyber threat topics like worms, viruses, Trojan Horses.
- 2,200 malware reports which include various cyber threat types, such as Trojan, Spyware, Ransomware, RATs, Backdoors, etc.

Expected Improvements

- Extend the approach of viewing a threat action as a VO representation to other syntactic blocks, because not all threat actions are described in the format of VO.
- A computational approach to automatically parse all kinds of threat action expressions, and extract them as the key information of a report for further analysis by machine.
- Automate the process of capturing recent information including articles and news related to cyber security, which improves the VO pair table and thereby the system.

Group Members :

Abdu Musowir U	38221002
Midhun N	38221032
Muhammed Mahir P C	38221035