Internet Data Analysis Methodology for Cyberterrorism Vocabulary Detection, Combining Techniques of Big Data Analytics, NLP and Semantic Web

Iván Castillo-Zúñiga, Instituto Tecnológico del Llano, Aguascalientes / Instituto Tecnológico de Aguascalientes, Aguascalientes, Mexico
Francisco Javier Luna-Rosas, TecNM/Instituto Tecnológico de Aguascalientes, Aguascalientes, Mexico
Laura C. Rodríguez-Martínez, Tecnológico Nacional de México/I.T. Aguascalientes, Mexico
Jaime Muñoz-Arteaga, Universidad Autonoma de Aguascalientes, Aguascalientes, Mexico

DOI: https://orcid.org/0000-0002-3635-7592

ABSTRACT

This article presents a methodology for the analysis of data on the Internet, combining techniques of Big Data analytics, NLP and semantic web in order to find knowledge about large amounts of information on the web. To test the effectiveness of the proposed method, webpages about cyberterrorism were analyzed as a case study. The procedure implemented a genetic strategy in parallel, which integrates (Crawler to locate and download information from the web; to retrieve the vocabulary, using techniques of NLP (tokenization, stop word, TF, TFIDF), methods of stemming and synonyms). For the pursuit of knowledge was built a dataset through the description of a linguistic corpus with semantic ontologies, considering the characteristics of cyber-terrorism, which was analyzed with the algorithms, Random Forests (parallel), Boosting, SVM, neural network, K-nn and Bayes. The results reveal a percentage of the 95.62% accuracy in the detection of the vocabulary of cyber-terrorism, which were approved through cross validation, reaching 576% time savings with parallel processing.

KEYWORDS

Big Data Analytics, cyberterrorism, Internet Data Analysis, Machine Learning, Natural Language Processing, Parallel Processing, Semantic Web, Text Mining

INTRODUCTION

The accelerated growth of the Internet, the use of social networks, cloud computing, have led to the generation of large volumes of data, in which the opportunity to commit a crime is latent. One of the greatest threats to society in the world today is cyber-terrorism, a new way of engaging in violence, which is executed by terrorist groups on the Internet, seeking to harm people, groups or nations (Alqahtani, 2015).
Allister et al. (2010), indicates that cyber-terrorism is the convergence of terrorism and cyberspace for unlawful attacks and threats aimed at damage to individuals, groups or nations, via ICTs; Sánchez (2015), mentions that it is a violent action that instills terror carried out by one or more people on the Internet or through the improper usage of communications technologies. From the perspective of Poveda & Torrente (2016), indicate that cyberterrorism is the deliberate usage of technologies related to computer science for threaten or attack people, as well as to property and infrastructure, in order to instill terror to achieve a political, ideological, social or religious purpose. Finally, Salellas (2012), describes that cyberspace is being used by terrorist groups such as the Al Qaeda, ETA in Spain, neo-Nazi groups from Belgium and the Netherlands, Supreme Truth in Japan, and KKK in United States, to carry out propaganda, financing, recruitment, collection and exchange of information.

In general information is essential against this threat, and also prevention measures will be determined by the difference of information between victims and cyberterrorists (Schenone, 2014). However, the process to analyze the enormous amount of data generated on the Internet and the possibility of identify possible cyberterrorism vocabulary has been addressed from different approaches. With this perspective there are programs, developments, algorithms and processes that are not well defined, implementing partial solutions that have not been fully accepted by the scientific community. For these reasons, this research proposes a new approach to process large volumes of data with the aim to identify cyberterrorism vocabulary.

To gain knowledge of the information and that this represents a value, it is necessary to carry out effective and administration data and to apply different processing techniques that allow you to handle large volumes of information, with a speed of acceptable response. It is also possible to analyze a variety of complex data, semi-structured and unstructured as documents, images, videos, music, among others (Joyanes, 2013), with the purpose of obtaining accuracy in data on a theme in particular, including the Big Data characteristics in this way (Chawda & Thakur, 2016). On the other hand semantic problems should be considered as an obstacle for the interpretation of words or meanings, integration of scattered and unrelated information. Further the recovery of data that has problems of synonymy, polysemy and multilingualism (Pastor, 2013).

Recent reports Allister et al. (2010), Kolajo & Daramola (2017), Bosques & Garza (2016), Pu et al. (2015), Semberecki & Maciejewski (2016), Weir et al. (2016), and Sarnovsky & Vronc (2014), have aimed to the treatment of large volumes of data, the analysis of information on the Semantic Web and Natural Language Processing (NLP), which are summarized in Table 1.

Some of the cited research presented models geared to the pursuit of knowledge with large volumes of data-oriented distributed processes (Kolajo & Daramola, 2017; Sarnovsky & Vronc, 2014), others develop processes and programs to locate specific information on the web (Weir et al., 2016), also get value-added data through text mining techniques (Allister et al., 2010; Bosques & Garza, 2016; Pu et al., 2015; Semberecki & Maciejewski, 2016; Weir et al., 2016), in the same way is considered with NLP techniques information retrieval (Bosques & Garza, 2016; Weir et al., 2016) and its analysis with Semantic Web methods (Pu et al., 2015).

The first reason for this research project is to explore the benefits that you can get of the information that circulates on the Internet. Second to discover patterns, unknown correlations, additional information that can be very useful, and finally giving support to the decision-making process in the solution of problems.

Our research is based on the combination of Big Data analytics techniques (genetic strategy, supervised learning algorithms, cross validation) with parallel processing, semantic web (ontologies, vocabularies, linguistic corpus) and NLP (tokenization, word separation; cleaning, separation of HTML code information and empty words; vocabulary, synonyms, stemming, word root; frequency of term (TF); frequency of term combined with inverse frequency of document (TF-IDF)), with the purpose to generate valuable and value-added information obtained from the analysis of a large number of pages which are downloaded from the Web, we also build a semantic ontology that classifies information, and allows the identification of the vocabulary, organizing information through validated concepts
Genetic-Fuzzy Programming Based Linkage Rule Miner (GFPLR-Miner) for Entity Linking in Semantic Web
[www.igi-global.com/article/genetic-fuzzy-programming-based-linkage-rule-miner-gfplr-miner-for-entity-linking-in-semantic-web/206257?camid=4v1a](www.igi-global.com/article/genetic-fuzzy-programming-based-linkage-rule-miner-gfplr-miner-for-entity-linking-in-semantic-web/206257?camid=4v1a)

Elementary: Large-Scale Knowledge-Base Construction via Machine Learning and Statistical Inference
Feng Niu, Ce Zhang, Christopher Ré and Jude Shavlik (2012). *International Journal on Semantic Web and Information Systems* (pp. 42-73).
[www.igi-global.com/article/elementary-large-scale-knowledge-base/74339?camid=4v1a](www.igi-global.com/article/elementary-large-scale-knowledge-base/74339?camid=4v1a)