Effective Mining of Crime Patterns from Growing Volumes of Data Using Improved FP-Growth Algorithm

By
George Matto
PhD. Information and Communication Science and Engineering. Nelson Mandela African Institution of Science and Technology, 2019

Abstract

The spate of crimes in Tanzania, as in many other countries, has been on the increase in the last few years. The successes recorded by criminals have been attributed by improper mechanisms for crime detection, prevention and control. Proactive measures are needed to pre-empt further crimes. Frequent pattern mining stand to aid in finding emerging patterns, series, and trends in the crime data. This could eventually help Tanzania Police Force and other law enforcement agencies to understand crime trends and predict or forecast future occurrences and thus improve preventive measures against crimes. FP-Growth is the most effective and most widely used algorithm for frequent pattern mining and association rules generation. Unfortunately, studies have shown two main weaknesses associated with this algorithm. Inability of the algorithm to scale-up well with the growing volumes of data is the first weakness. Second weakness of FP-Growth is associated with the nature of crime data. Thus, this study aimed at improving FP-Growth algorithm for effective mining of frequent crime patterns in the growing volumes of data in Tanzania. Specifically, the study

- Explored ways in which frequent pattern mining can be useful in detecting crime patterns from available datasets in the country
- Proposed efficient FP-Growth scaling method for effective mining of frequent patterns of crime
- Proposed a generic framework for mining crime patterns from multiple sources of data and
- Developed pattern mining prototype basing on the proposed FP-Growth scaling method and proposed framework.

This study was conducted in Tanzania, data were collected through interview guides and secondary sources such as newspapers. The key respondents included police force and its department (intelligence, data and ICT).

Figure No. 1: Sample newspaper used for data extraction

The results regarding frequent Pattern Mining for Crime Detection

- Frequent pattern mining can be an effective tool for helping police and other law enforcement agencies to improve strategies for crime prevention in the country. This was revealed in this study where, through an employment of frequent pattern mining and specifically the FP-Growth algorithm, the study was able to discover a number of crimes that commonly occurred in the country and the magnitude of their occurrence. This result can be very helpful to the law enforcement agencies as it tells of which crimes are most commonly committed. The result can help them to strategize on crimes that require more and urgent attention. The study's results were based on the patterns that were extracted from news articles obtained from four selected daily published newspapers.

- A pattern mining model that was built on Rapid Miner was employed in the pattern's extraction. The study showed (through the generated association rules) how the mined crimes were related. Specifically, the study observed a strong relationship between killing and other crimes such as brutality, gunned crimes, explosives, and traditional armed crimes. In fact, killing was found to be a consequent of each of these other crimes which were established to be antecedents to it. Therefore, according to this result, it was fair to say most of other crimes that were committed at the time these data were collected resulted into killings, and therefore if such crimes are contained, killings will also be contained.

- Moreover, the mined crimes were mapped per regions of the country in which they occurred. According to the results, crimes were reported in almost half of all regions of Tanzania where Mwanza region was leading by having the highest number of crimes reported in the period data was collected. Some few regions had moderate number of

crimes reported, and nearly half of the region did not report any crime at the period in which data was collected. This finding can help the law enforcement agencies and other stakeholder including the general public to take necessary and reasonable preventive measures in the regions that report high crime rates.

The results regarding FP-Growth improvement method

- This study was specifically based on the FP-Growth algorithm for frequent patterns extraction. Unfortunately, this algorithm did not scale up well with the growing volumes of data. Besides, when the algorithm is to be used in crime datasets it suffers the rare item problem. The reasons are twofold. First, if the algorithms minimum support is set too low, huge amounts of crime patterns (especially with those crimes that are commonly committed) will be generated.
- On the other side, if it set too high lots of interesting pattern, including seasonal crimes patterns, will be lost. To tackle the challenge, this study proposed the use of multiple minimum supports approach. The proposed approach was based on Shannon Entropy.
- Unlike other existing MIS approaches that suffer some challenges such as difficulties in stating "good" MIS value for each item and constantly tuning the minimum support values to obtain the best values, the proposed approach scans the entire dataset and automatically assigns MIS values on each of the crime item in the dataset basing on how frequently it has occurred in the dataset. In this way the proposed approach handles not only those crime items that rarely appear in the dataset but also those that appear very frequently. In other words, the proposed approach handles each crime item separately depending on how frequently it has appeared in the dataset

The results on Generic Framework for Mining Frequent Patterns of Crime

- In the third specific objective of this study, a generic framework for the extraction of crime patterns from multiple data sources was proposed (Figure 2). The framework was proposed to confront the challenge that were observed in most of the existing similar frameworks. Some of those challenges were lack of input data flexibility, lack of clear procedures for cleaning and pre-processing the input data, as well as the rare item challenge in crime datasets.

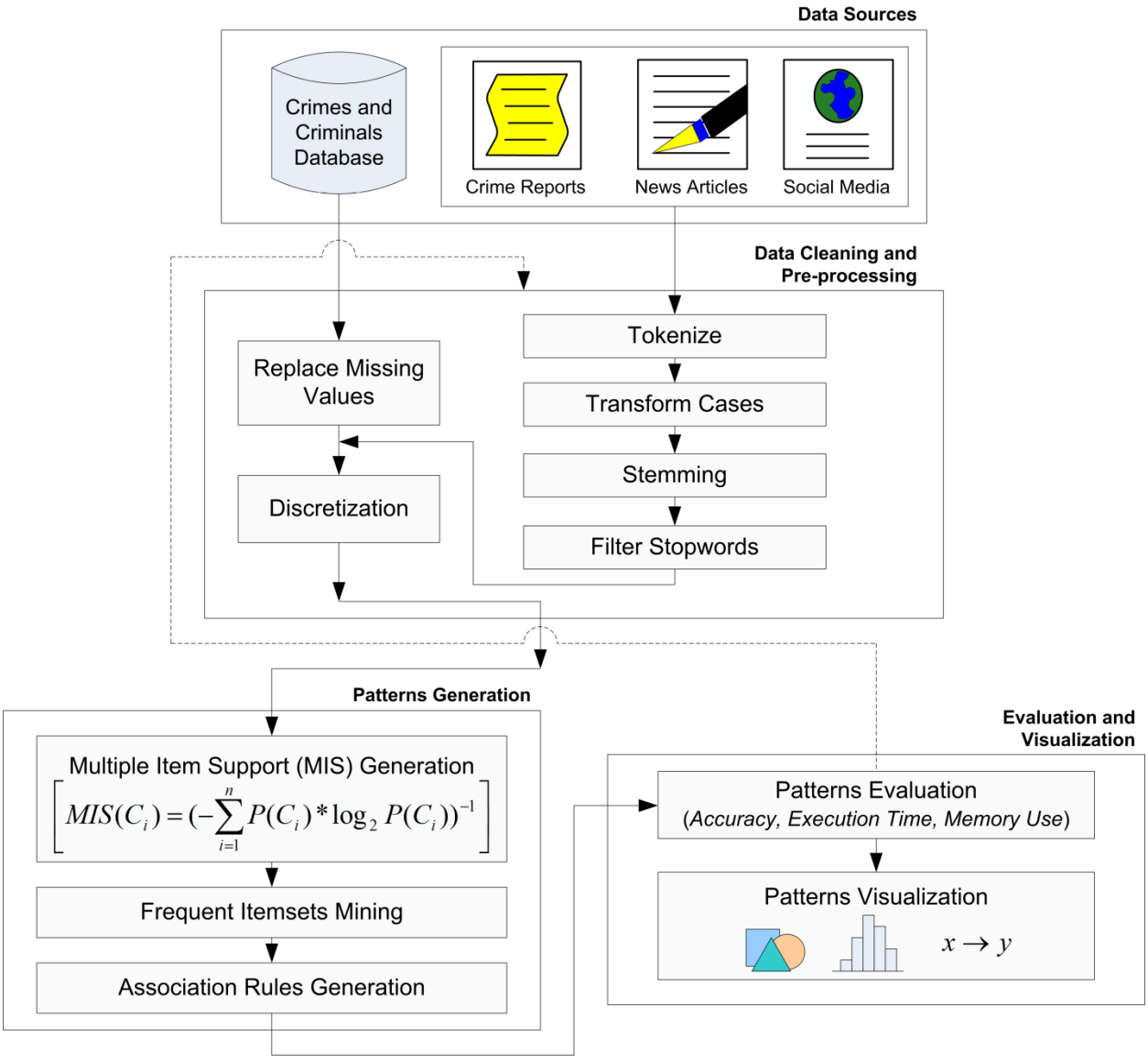

Figure 2: The proposed generic framework

- The proposed framework confronts these challenges in a four stages process. First stage is data collection. This stage allows flexibility for both structured and unstructured data to be employed in the mining process. Second stage is data cleaning and pre-processing. Depending on the type of input data, cleaning and preprocessing may involve replacing missing values, tokenizing, transforming cases, stemming, filtering stop words and discretization. Pattern generation is the frameworks third stage. In this stage patterns are generated by using our proposed multiple minimum supports FP-Growth algorithm. In the final stage, pattern evaluation and visualization, the generated patterns are investigated to eliminate irrelevant patterns and remain with only relevant ones and then put them in the form that make them easily understood.

The findings on Crime Pattern Mining Prototype

- A fourth objective of this study was to develop a working prototype for crime patterns

extraction. This prototype, named Crime Reporting and Pattern Extraction System (CRaPES), was developed based on three main reasons. First, to enable police and other the law enforcement agencies in Tanzania to register reported crimes and to extract useful patterns of crimes from stored datasets.

- The second reason for developing CRaPES was to evaluate effectiveness and applicability of the proposed FP-Growth scaling method. And a third reason for coming up with this prototype was to partly validate the framework for extraction of frequent patterns of crime from multiple sources of data that was also proposed in this study.
- System requirements for the prototype were obtained from the TPF, head office. The prototype was up and running and was able to not also allow reporting of crimes but also extract useful patterns from both stored in CRaPES database and in other external sources. Results on the evaluations that were performed by information systems development experts and system users showed that the system is generally user friendly and is capable of not only reporting crimes but also extracting patterns.

Generally, the study concluded that;

- Frequent pattern mining is an effective tool for helping police and other law enforcement agencies to improve strategies for crime prevention in the country. However, this study observed that TPF still relies on traditional and semi-automated methods to record and analyze crime data. As a result, some useful crime insights that could be discovered if frequent pattern mining could be employed might be missed out.
- In connection, the study concurs with other existing literature on the effectiveness of FP-Growth algorithm in crime patterns mining. However, it was established that classical FP-Growth not only suffers the rare item problem but also does not scale up well with the growing volumes of data. This challenge is attributed mainly by the algorithm user-defined minimum support approach. In fact, a single user-defined minimum support is not appropriate for massive amounts of crime data.
- This study, therefore, proposed an FP-Growth scaling method that employs the Shannon Entropy method to find the minimum supports of each of the crime item in the dataset. The proposed approach was tested on static crime data and established to be reasonably effective. The proposed approach was also evaluated through the developed prototype and found to be user friendly and capable of extracting useful crime patterns in the growing volumes of data. In connection, the study concludes that a crime pattern mining framework that consider multiple sources of input data is more ideal for crime pattern mining.

General recommendations drawn by this study;

Following the findings from this study and the general conclusion that have been stated, the study puts forward the following general recommendations.

- ✓ TPF should embark on data (inside and outside their boundaries) for the fight against crime.
- ✓ In addition to considering storing and managing its crime data electronically, TPF should employ frequent pattern mining as an effective way to extract useful crime patterns from available crime datasets.
- ✓ For the purpose of this study, the crime pattern mining model extracted patterns from news articles that were collected in the period of two months only. It was recommended

that other researches who will be interested in doing similar research should consider collecting same data in a longer period of time so as to observe if there is seasonality of patterns mined. It was equally recommended for involvement of more other data sources

- ✓ The proposed FP-Growth scaling method was tested and validated on static datasets. It is recommended other researchers to extend experimentations of the same method on streaming data.
- ✓ The study, furthermore, recommends that other researchers extend experimentations of the proposed methods on distributed environments such as Hadoop/Map Reduce framework.
- ✓ Since CRaPES was developed as a desktop application, the study recommends for further studies by other scholars to extend it into both web-based and mobile applications.
- ✓ Although the proposed framework considered input data from structured and unstructured sources, validation of the framework was done on structured data only. It is therefore recommended for further validation of the framework on more other sources of crime data as described in the framework.

Future Works proposed in this study;

- ✓ More Datasets and More Data Sources

Although experimental results on crime patterns detection from news articles can create basis for improving strategies for crime prevention, such experiments were based on only one source of crime data (i.e. newspapers) in which data was also collected in a relatively short period of time (i.e. two months). There is a need therefore to extend this work to collect such data in a longer period of time to see if similar results will be obtained. In connection, future work should also involve more other sources of data from the media. The use of such different sources of data will also help to further validate the proposed framework.

- ✓ Live, Streaming and Big Data Processing

Another area for future work was extension of experimentations on the proposed FP-Growth improvement method on data that is generated continuously by different sources such as mobile and web generated data as well as cloud data. This future work should also involve employment of distributed computing and Big Data processing.

- ✓ Mobile and Web-based Applications for Crime Patterns Mining

Future work can also be carried out to develop a mobile and web-based system for crime pattern mining basing on the prototype that was developed in this study. In order for the developed extended system to be useful to the TPF and other law enforcement agencies, it should be capable of helping such law enforcement agencies to report crimes and extract patterns from stored crime data. It is also important that the improved systems especially the mobile application to automatically identify the location of someone reporting crime and record that to the system. By so doing, crime reporting system will be more intelligent in plotting locations with high crimes for proper Interventions.