Effective mining of Crime Patterns from Growing Volumes of Data using improved FP-Growth Algorithm

By

George Matto - 2019

A Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy in Information and Communication Science and Engineering of the Nelson Mandela African Institution of Science and Technology

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 preempt further crimes. Frequent pattern mining stand to aid in finding emerging patterns, series, and trends in the crime data. This will 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 when used in the growing volumes of crime data. First is inability of the algorithm to scale-up well with the growing volumes data. Second weakness of FP-Growth is associated with the nature of crime data is that crime data consists of items (different crimes) that vary greatly in terms of frequencies of occurrence. Some of crimes (e.g. robbery) happen so frequently and thus frequently appear in the dataset while other crimes (e.g. killing of people with albinism, in the case of Tanzania) happens seasonally and therefore rarely found in the dataset. Classical FP-Growth algorithm extract frequent patterns by using single user-defined minimum support. This is the main source of the algorithm's challenge, especially when used in crime datasets. To tackle the challenges, this study have proposed a multiple minimum support FP-Growth algorithm that scans the dataset and automatically assigns minimum support values to each crime item basing on how frequently it has appeared in the dataset. The proposed solution is based on the Shannon Entropy in which an algorithm for obtaining multiple item support values was developed. In connection, the study developed a working prototype that was based on the proposed approach to help in recording crime data and extract crime patterns from data. The proposed approach was evaluated against classical FP-Growth as well as CFPGrowth algorithm on the varying sizes crime data. Evaluation results showed that the proposed approach is more efficient in mining frequent crime patterns, and more effective in terms of run-time and memory use. Basing on the results found, the study recommends for further experimentation of the proposed approach on streaming data and on distributed environments, among other recommendations as stipulated in the dissertation.