ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

# ONLINE LAND FRAUD DETECTION AND SAFETY THE PRIVACY

**Thamaraiselvi R[1], Mehanthivinothini G K[2], Sowntharya K[3], Shreedhivya N[4],**

[1]Assistant Professor, Department of Electronics & Communication Engineering,
Nandha College of Technology,
[2,3,4] UG Scholar, Department of Information Technology,
Nandha College of Technology
[1]thamaraiselviece11@gmail.com, [2]mehanthigopalan@gmail.com,
[3]sowntharyainfotech@gmail.com, [3]shreedhivyanatraj@gmail.com

## ABSTRACT

Collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. Fraud detection is a topic applicable to many industries including land registration banking and financial sectors, insurance, government agencies, telecommunication and law enforcement, and more . Fraud attempts have seen a drastic increase in recent years, making fraud detection is essential and more important than ever. Big data framework for this application aims to help people detect unexpected activity on their property that may be fraudulent. The Big Data properties will lead to significant system challenges to implement machine learning frameworks. This paper discusses the problems and challenges in handling Big Data classification using Naive Byes algorithm techniques to detect accurate fault with high performance. Tenure Security is a principle and critical factor in providing social and political stability. The major goal of this research is to assist the discovery of fraud patterns in land and property transactions using data mining techniques. The methodology starts with analysis of different fraud schemes to

Discover fraud patterns and their indicators in land records. A data simulator is developed to generate synthetic datasets. Different algorithms are then applied to detect fraud patterns in these datasets. Three major fraud schemes were used to validate the proposed approach (land grabbing in post conflict situations, Oklahoma Flip and ABC-Construction). The results have proven that Hadoop methods can identify fraudulent activities. However, these methods cannot be generalized to be used with all datasets. Finally, big data can facilitate the building of fraud detection models for land Transactions that can then be integrated with registration systems, and act as an alarm system.

## INTRODUCTION

This thesis identifies some of the illegal activities that occur in the trading of lands and other forms of real estate. It also examines the traceability of these activities in land Information management systems. In particular this research project explores the application of Hadoop methods to identify suspicious transactions in land records which may underlie illegal manipulation of land and property ownership. For Hadoop to be applied effectively, an understanding of the techniques used to commit illegal activities, such as land grabbing or exploiting the registration systems for personal gain, is vital to enable the tracking of these activities. Tracking land records for fraud and other forms of mischief or negligence is an important facet of tenure security management and this is the prime contribution of this thesis. Security of tenure is a critical factor underlying social and political stability. Secure tenure is one of mankind's most basic needs, and is a major contributor to the economic and cultural development of civilizations. To enhance land tenure security and Property ownership, research has been growing in two main directions. The first stream focuses on creating land administration models to capture the complexity of the different Situations which occur in developed and growing economies Fraud is an act of deception intended for personal gain or to cause a loss to another party. There are many words used to describe fraud: Scam, con, swindle, extortion, sham, double-cross, hoax, cheat, ploy, ruse, hoodwink, confidence trick. Fraud involves one or more persons who intentionally act secretly to deprive another of something of value, for their own benefit. Fraud is as old as humanity itself and can take an unlimited variety of different forms. However, in recent years, the development of new technologies has also provided further ways in which criminals may commit fraud. In addition to that, business reengineering, reorganization or downsizing may weaken or eliminate control, while new information systems may present additional opportunities to commit fraud. In different situational contexts, fraud can take somewhat different forms for example, Bribery Embezzlement, Securities fraud, Health care fraud, Money-laundering scams, Insurance fraud, Software piracy, Internet fraud, Telemarketing fraud, Identity theft. These have their own special characteristics. There are at least as many types of fraud as there are types of people who commit it. But in each instance, fraud involves deception. Someone knowingly lies in order to obtain an unlawful benefit, or an unfair advantage. Fraud Detection involves finding the "needles in a haystack," which requires methodologies and techniques that are unique to this application area. Often, there are instances of fraud that have not been detected, which adds to the challenge of training the computer to recognize fraudulent cases. A complete fraud detection solution uncovers patterns of suspicious behaviour and provides actionable alerts to the organization. The literature review is divided into two parts: a review of the fraud problem in land record systems and a review of Hadoop methods that can help solving the problem. This thesis presents the review of the fraud problem in land record systems and it also reports on interviews on the nature of the problem. Mainly, this thesis expands on the nature of the research problem discussed in fraud literature and examining interviews the author conducted with experts to better understand the problem. The aim in this chapter is to help the reader better understand the fraud problem. It also aims to give the reader an initial understanding of some fraud methods and how they are dealt with in some previous studies. Thereafter, it describes the problem of large land datasets with examples of how this data is being analyzed. The chapter concludes by describing major fraud types, methods and causes, and then examines two studies that deal with the fraud problem in land record systems.

## EXISTING SYSTEM

Data mining is a field which is concerned to understanding data patterns from huge datasets. We can say that the aim is to find out new patterns in data. Data mining to classify, cluster, and segment the data and automatically find associations and rules in the data that may signify interesting patterns, including those related to fraud. Expert systems to encode expertise for detecting fraud in the form of rules. Pattern recognition to detect approximate classes, clusters, or patterns of suspicious behavior either automatically (unsupervised) or to match given inputs. Machine learning techniques to automatically identify characteristics of fraud.Data mining techniques can be classified into two categories Supervised Learning for Fraud Detection: This method uses supervised learning in which all the available records are classified as “fraudulent” and “non-fraudulent”. Then machines are trained to identify records according to this classification. However, these methods are only capable of identifying frauds that has already occurred and about which the system has been trained. Unsupervised Learning for Fraud Detection: This method only identifies the likelihood of some records to be more fraudulent than others without statistical analysis assurance. Unsupervised learning is closer to the exploratory spirit of Data Mining as stressed in the definitions given above. In unsupervised learning situations all variables are treated in the same way, there is no distinction between explanatory and dependent variables.

## DISADVANTAGES OF EXISTING SYSTEM

Many applications suffer from the Data mining problem, including machine learning analysis, geospatial classification and business forecasting. New Type of fraud in online not to discover the data mining classification. One of the major limitations of this study is the lack of real datasets. As data is a major factor in the success of achieving the objectives, the existence of real datasets would have helped substantially in the progress of the research. A property transactions data simulator was developed to overcome the lack of data availability. While this simulator has advantages, it has some limitations and creates a finer scope for the type of data to be worked with.

## PROPOSED SYSTEM

Data framework for this application aims to help people detect unexpected activity on land property that may be fraudulent. Fraudsters may attempt to acquire ownership of a property through forging documents, impersonating the registered owner and convince mortgage lenders and solicitors that they own the property to detect and alert the owner and safe side forever. We propose a classification of naive Byes technique is a well supervised learning model and capture uncertainly fraud in classifications big data framework using this technique is a highly scalable and solve predictive problems in land registration. Able to perform classification processes on huge amounts of dataset, exploiting the benefits of working on effective patterns with the Hadoop framework. Land Record Systems are experiencing a data explosion, as are most other information systems. This study is concerned with the data held inside a system rather than the type of the system. In many parts of the developing world, much of the land data is still in paper format. However, studies are emphasizing the transition from analogue to digital form, and land data is being converted continually into digital form This part of the review was to understand the different cadastral models and how relations between people and land are conceptualised. This review was important because for any Hadoop method to work, the data infrastructure of the targeted system should be reviewed and understood. Many cadastral data models have been developed to support land record systems. All of these are trying to model the relation between land and people via rights. Some of these models include: Core Cadastral Domain Model (CCDM)

## ADVANTAGE OF PROPOSED SYSTEM

To predict the fraud detection and perform an appropriate action, before the fraud actually occurs. It is possible to overcome the weaknesses of all previous data mining models. It can provide continue correct Performance of fraud detection and previously alert the owner of land. This approach used to achieve the best effective and efficient for detecting fraud and reduce the execution time, Cost effective is also low. Performing computation on large volumes of data has been done before, usually in a distributed setting. What makes Hadoop unique is its simplified programming model which allows the user to quickly write and test distributed systems, and its efficient, automatic distribution of data and work across machines and in turn utilizing the underlying parallelism of the CPU cores. The Four reasons for why we are using Hadoop is, 1: Data exploration with full datasets, whether using R, SAS, Mat lab or Python, they always need a laptop with lots of memory to analyze data and build models. In the world of big data, laptop memory is never enough, and sometimes not even closes. A common approach is to use a sample of the large dataset, a large a sample as can fit in memory. With Hadoop, you can now run many exploratory data analysis tasks on full datasets, without sampling. Just write a map-reduce job, PIG or HIVE script, launch it directly on Hadoop over the full dataset, and get the results right back to your laptop. 2: Mining larger datasets, in many cases, machine-learning algorithms achieve better results when they have more data to learn from, particularly for techniques such as clustering, outlier detection and product recommenders. Historically, large datasets were not available or too expensive to acquire and store, and so machine-learning practitioners had to find innovative ways to improve models with rather limited datasets. With Hadoop as a platform that provides linearly scalable storage and processing power, you can now store all of the data in RAW format, and use the full dataset to build better, more accurate models. 3: Large scale pre-processing of raw data, as many data scientists will tell you, 80% of data science work is typically with data acquisition, transformation, and cleanup and feature extraction. This “pre-processing” step transforms the raw data into a format consumable by the naive byes algorithm, typically in a form of a feature matrix. Hadoop is an ideal platform for implementing this sort of pre-processing efficiently and in a distributed manner over large datasets, using map-reduce or tools like PIG, HIVE, and scripting languages like Python. For example, if your application involves text processing, it is often needed to represent data in word-vector format using TFIDF, which involves counting word frequencies over large corpus of documents, ideal for a batch map-reduce job 4: Data agility, It is often mentioned that Hadoop is “schema on read”, as opposed to most traditional RDBMS systems which require a strict schema definition before any data can be ingested into them. “Schema on read” creates “data agility”: when a new data field is needed, one is not required to go through a lengthy project of schema redesign and database migration in production, which can last months. The positive impact ripples through an organization and very quickly.

## EXPERIMENT ENVIRONMENT AND PROCESS

Fraud analysis has been one of the oft quoted use cases for Hadoop. We look at the topic further to explore usage of Hadoop ecosystem products. Per se, the fraud analytics can be divided into 3 further use cases:

1- Fraud detection: determining if a fraud is taking place or has occurred in the past and generating appropriate alert for it.

2- Fraud prevention: implementing controls and access to prevent fraud.

3- Fraud reduction: monitoring and predicting patterns to minimize chances of fraud occurrence

Listed below are some of the methods that can be implemented using Hadoop to ensure fulfilment of either of the 3 use cases above.

1- Reduplication -

a) Entity matching - This could include exact or similar matching of entities like name, father name or contact information (phone, e-mail id, street, city) or phonetic matches using the reduplication methods  Since this is a data intensive exercise and requires matching previously built index, there cannot be better technology fit than Hadoop.

b) Social network identity matching - Not very commonly used, but emerging off late, is a tendency to match social network profiles with customer identity. While this technique could be quite effective provided you have the right social network data feeds, please be aware of privacy laws that may be applicable.

2- Outlier detection -

A usual outlier will be a deviation from a common usage pattern of a customer or transaction set. Using custom machine learning algorithms or available libraries, we would tend to combine data to see any outlier points. Clustering, probabilistic distributions along with visualization techniques are more common methods to derive outliers.

These may be used in conjunction with techniques like path analysis, sessionization, tokenization and attribution. Regression, co-relation, averages and graph analysis may also be employed based on functional requirement.

3- Workflow -

Transaction streaming, monitoring, alert forwarding, alert disposal and transaction blocking could be among a few steps that a custom workflow may implement in fraud management system. Considering the massive volume of transactions, a custom DSL workflow may be implemented on top of Hadoop.

## CONCLUSION

This thesis identified different fraud activities in different situations. The effect of those activities on the underlying databases was examined and fraud indicators and patterns derived.

Using fraud indicators and patterns, it was possible to use Hadoop techniques to detect the fraudulent activities. So, using Hadoop should facilitate the building of fraud detection models for land transactions that can then be integrated in the registration systems and act as alarm systems.

## REFERENCES

[1] Outlier detection- Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques".

[2] Data mining tasks and Techniques" www.investopedia.com.

[3] Shashidhar, H.V. And S. Varadarajan, 2011. Customer segmentation of bank based on data mining-security value based heuristic approach as a replacement to k-means segmentation. Int. J. Comput. Appli., 19:13-18.

[4] Harmeet Kaur Khanuja, Dattatraya S. Adane, "Forensic Analysis for Monitoring Database Transactions", Springer, Computer and Information Science Volume 467, pp 201-210, 2014.

[5] Pradnya Kanhere "A Survey on Outlier Detection in Financial Transactions" Int. J Computer Applications (0975 – 8887) Volume 108 – No 17, December 2014.

[6] Land Registration and Land Fraud in the United States By J. David Stanfield, Jeff Underwood, Kirthimala Gunaskera of Terra Institute and Carl Ernst, Property Records Industry Association 21 October, 2008