iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Business Tax Fraud Detection in Big Data

Dr. Mohammed Abdul Waheed[1], Prema N Valkey 2
[1]Associate Professor, Department of Studies in Computer Science and Engineering, VTU CPGS, Kalaburagi, Karnataka, India
[2]M. Tech Student, Department of Studies in Computer Science and Engineering, VTU CPGS, Kalaburagi, Karnataka, India

Abstract: ***There is evidence that an increasing number of enterprises plot together to evade tax in an unperceived way. At the same time, the taxation information related data is a classic kind of big data. The issues challenge the effectiveness of traditional data mining-based tax evasion detection methods. To address this problem, we first investigate the classic tax evasion cases, and employ a graph-based method to characterize their property that describes two suspicious relationship trails with a same antecedent node behind an Interest Affiliated Transaction (IAT). Next, we propose a colored network-based model (CNBM) for characterizing economic beighaviors, social relationships and the IATs between taxpayers, and generating a Taxpayer Interest Interacted Network (TPIIN). To accomplish the tax evasion detection task by discovering suspicious groups in a TPIIN, methods for building a patterns tree and matching component patterns are introduced and the completeness of the methods based on graph theory is presented. Then, we describe an experiment based on real data and a simulated network. The experimental results show that our proposed method greatly improves the efficiency of tax evasion detection, as well as provides a clear explanation of the tax evasion behaviors of taxpayer groups.***

Keywords: ***Tax, MapReduce, Tax Payer Interest Interacted Network, colored network based model, Big Data.***

I. INTRODUCTION

Tax revenue collection is considered a top priority in every national and regional jurisdiction [4], [10], [16], [17], [18], [19]. Tax evasion is the way people evade tax by illegal and unfair means. They may claim lesser profit, gains or turnover than actual. Even if there is huge amount of tax to be paid, evaders get refund by making misrepresentations before tax authorities. Huge amount of revenue is lost through this way for government so that we cannot climb from economic stagnation. Most of the welfare activities for poor are put on hold due to lack of money while some people who can buy even the government with black money are growing daily, causes inflation and value erosion. The level of evasion tax also depends on the chartered accountants and tax lawyers who help companies, firms, and individuals evade paying taxes. Tax evasion is a crime in all major countries. Only 3 percent of Indian population pay Income tax over 1 billion people are estimated to pay taxes. Taxation information includes large amount of data with large number of tax payers. Data mining provide powerful techniques for tax administration to extract useful knowledge from huge database.

It was reported by the Chinese government that the rate of loss of tax revenue in China is above 22%. China Tax Administration Information System (CTAIS) was developed in 1996 and since 2000 it is in operation nationally boosting a revolution in the informatization supported tax administration and the data sharing between different provinces. The sharing of data has prepared the ground for deep mining and analysis of tax data. At the same time, three ways of tax inspection, manual case selection [23], computer-based case selection (data-mining-based methods [24], [7]), and whistle-blowing-based selection were adopted by China Taxation Administration in their daily operation of tax inspection. As a result of using these methods, the traditional tax evasion behaviors, such as writing false value added tax (VAT) invoices, fake invoices and the manipulation of accounts were restrained more than ever, and the number of tax evasion cases has been decreasing dramatically.

Tax data include large amount of data collected from many independent sources and perform data matching with information technology tools. This volume of data challenges traditional data mining based methods for detection of tax evasion. The reasons are: 1) Training data needs to be manually labeled, 2) the most important issue is that some of the covert relationships are not recorded in the database. 3) The tax authorities have limited resources, and traditional tax auditing methods are time consuming. There is pressing need to have further inputs to a tax avoidance database and additional information resources from big data, the hidden relation is encountered from big data base.

After analyzing classic tax evasion cases the heterogeneous information network is formed and their properties are analyzed by colored network based model for characterizing social relationship, economic behavior and IAT between tax payers. The tax evasion detection is a two phase process. The first phase is mining suspicious groups in order to identify the doubtful trading relationships from the suspicious groups which is built from the heterogeneous information network on the bases of CNBM. In second phase is identifying tax evasion in this traditional methods can be used on all transaction related to the suspicious trading relationships to detect tax evasion within the set of suspicious groups. The objective of using data mining technique in detection of

tax evasion to improves the efficiency of tax evasion detection, as well as provides clear explanation of the tax evasion behaviors.

II. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before improving the tools it is compulsory to decide the economy strength, time factor. Once the programmer‘s create the structure tools as programmer require a lot of external support, this type of support can be done by senior programmers, from websites or from books.

A. FP-tax: tree structure based generalized association rule mining

1) *Authors:* IkoPramudiono

The pattern growth mining paradigm based FP-tax algorithm, which employs a tree structure to compress the database. Two methods to traverse the tree structure are examined: Bottom-Up and Top-Down. Experimental results show that both methods significantly outperform classic cumulate algorithm, in particular Top-Down FP-tax can achieve two order of magnitudes better performance than Cumulate.

B. MapReduce: Simplified Data Processing on Large Clusters

1) Authors: Jeffrey Dean and Sanjay Ghemawa

MapReduce computation processes many terabytes of data on thousands of machines. Programmers find the system easy to use: hundreds of MapReduce programs have been implemented and upwards of one thousand MapReduce jobs are executed on Google’s clusters every day.

C. Mr-apriori: Association rules algorithm based on mapreduce

1) Authors: Xueyan Lin: Cloud computing is the development of distributed computing, parallel processing, and grid computing, which represents an emerging business computing model. Computing tasks are distributed by the cloud computing in a resource pool composed of a large number of computers. Through the cloud computing, various application systems can obtain computing capability, storage space and a variety of software services as required. The novelty of cloud computing is that it can provide almost unlimited cheap storage and computing ability

D. Inducing features of random fields

1) Authors: S. Della Pietra, V. Della Pietra, and J. Lafferty

We present a technique for constructing random fields from a set of training samples. The learning paradigm builds increasingly complex fields by allowing potential functions, or features, that are supported by increasingly large sub graphs. Each feature has a weight that is trained by minimizing the Kullback-Leibler divergence between the model and the empirical distribution of the training data. A greedy algorithm determines how features are incrementally added to the field and an iterative scaling algorithm is used to estimate the optimal values of the weights. The random field models and techniques introduced in this paper differ from those common to much of the computer vision literature in that the underlying random fields are non-Markovian and have a large number of parameters that must be estimated. Relations to other learning approaches, including decision trees, are given. As a demonstration of the method, we describe its application to the problem of automatic word classification in natural language processing

E. Incorporating non-local information into information extraction systems by Gibbs sampling.

1) Authors: J. R. Finkel, T. Grenager, and C. Manning

Most current statistical natural language processing models use only local features so as to permit dynamic programming in inference, but this makes them unable to fully account for the long distance structure that is prevalent in language use. We show how to solve this dilemma with *Gibbs sampling*, a simple Monte Carlo method used to perform approximate inference in factored probabilistic models. By using simulated annealing in place of Viterbi decoding in sequence models such as HMMs, CMMs, and CRFs, it is possible to incorporate non-local structure while preserving tractable inference. We use this technique to augment an existing CRF-based information extraction system with long-distance dependency models, enforcing label consistency and extraction template consistency constraints. This technique results in an error reduction of up to 9% over state-of-the-art systems on two established information extraction tasks.

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

III. SYSTEM ARCHITECTURE

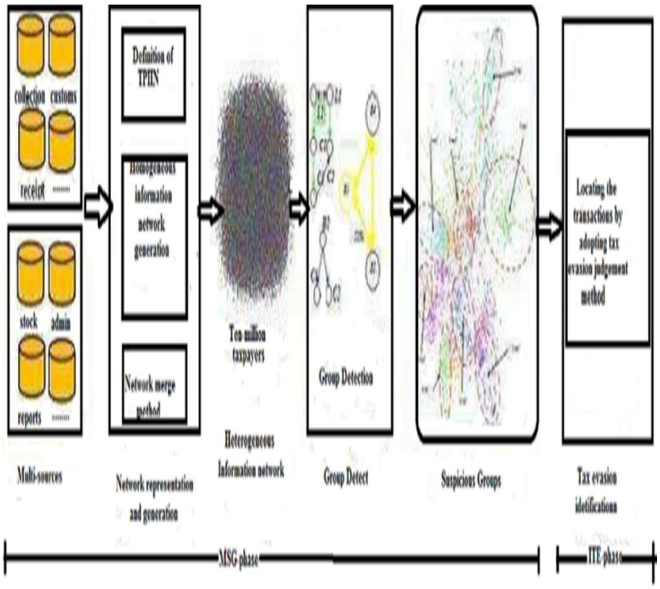


Figure1: System Architecture

To implement this concept, the most important step is to build a heterogeneous network and identify these suspicious groups via mining. For this data of each individual is collected like financial reports such as shares of companies or individual, profit of the company, receipt, admin, relationship with other company, an amount of tax paid by company, individual's property. Then this data is given to Taxpayer Interested Interacted Network (TPIIN), which is generated after multi-network fusion method has been adopted abstract various relationship from different information sources. For Example the relationships between the legal persons of the companies and tax-payers information forms a homogeneous network, which is represented using colored network based model (CNBM)[1]. These all homogeneous networks are grouped together to form heterogeneous network. Using heterogeneous network pattern will be generated and matched to find out suspicious tax evasion groups or detect the company or individuals who are avoiding tax. In the reminder of this paper, we first propose a colored network-based model (CNBM) for characterizing economic behavior [21], social relationships and IATs between taxpayers. Next, we show how this type of network is generated. To evaluate the effectiveness of the proposed method for the MSG-phase, a proof is given based on graph theory and experiments based on real data for all the nodes, most of the edges and a trading relationship simulated network are carried out. The experimental results show that our proposed method has ability to greatly improve the efficiency of detecting possible tax evasions in the MSG-phase, as well as provide a clear explanation of the tax evasion behaviors of taxpayer groups.

IV. METHODOLOGY

Generally, manual case selection [23], computer-based case selection (data-mining-based methods [24], [7]), and whistle-blowing-based selection are three frequently used ways of tax inspection. However, many researchers believed that manual case selection and whistle-blowing based selection are time-consuming and tedious, while data mining techniques used by tax administrations to detect tax fraud are considered to be the most promising approaches [7]. Mechanisms, such as neural networks, decision trees [8], logistic regression, SOM (Self organizing map), K-means, support vector machines, visualization techniques, Bayesian networks, rough set [3], K-nearest neighbor, association rules [24], fuzzy rules, Markov chains, time series, regression and simulations [2], have been used to check tax evasion [23], [12]. For example, Wu et al. [23] used a data mining technique and developed a screening framework to filter possible noncompliant value-added tax (VAT) reports that may be subject to further auditing. Chen and Cheng [3] proposed a hybrid model, which combines the Delphi method with rough sets classifier approaches, for intelligently classifying the vehicle license tax payment (called VLTP) to solve real-world problems faced by taxation agencies. Antunes et al. [2] claimed that the method of simulation with multiple agents provides a strong methodological tool to support the design of public policies. To address the tax compliance problem [11], they adopted a mean of exploring the link between micro-level motivations leading to and being influenced by macro-level outcomes, to study the complex issue of tax evasion.

Volume 5 Issue V, May 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

V. RESULTS AND DISCUSSION

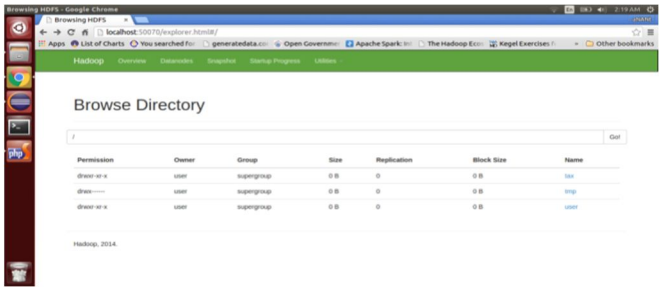


Fig2 Shows details of Hadoop browse directory. Tax file uploaded to the source code. The file holds all customer details like name, address, salary, transaction details of Customers.

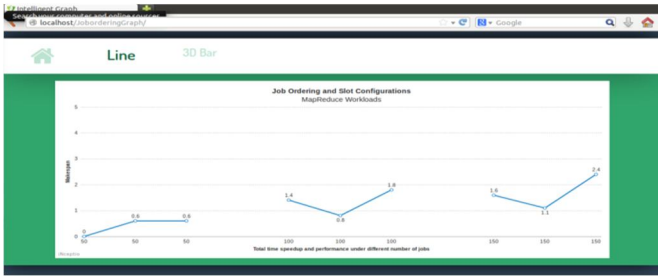


Fig3 Shows map reduce workload for processing uploaded tax file with total speedup and performance under different number of jobs.

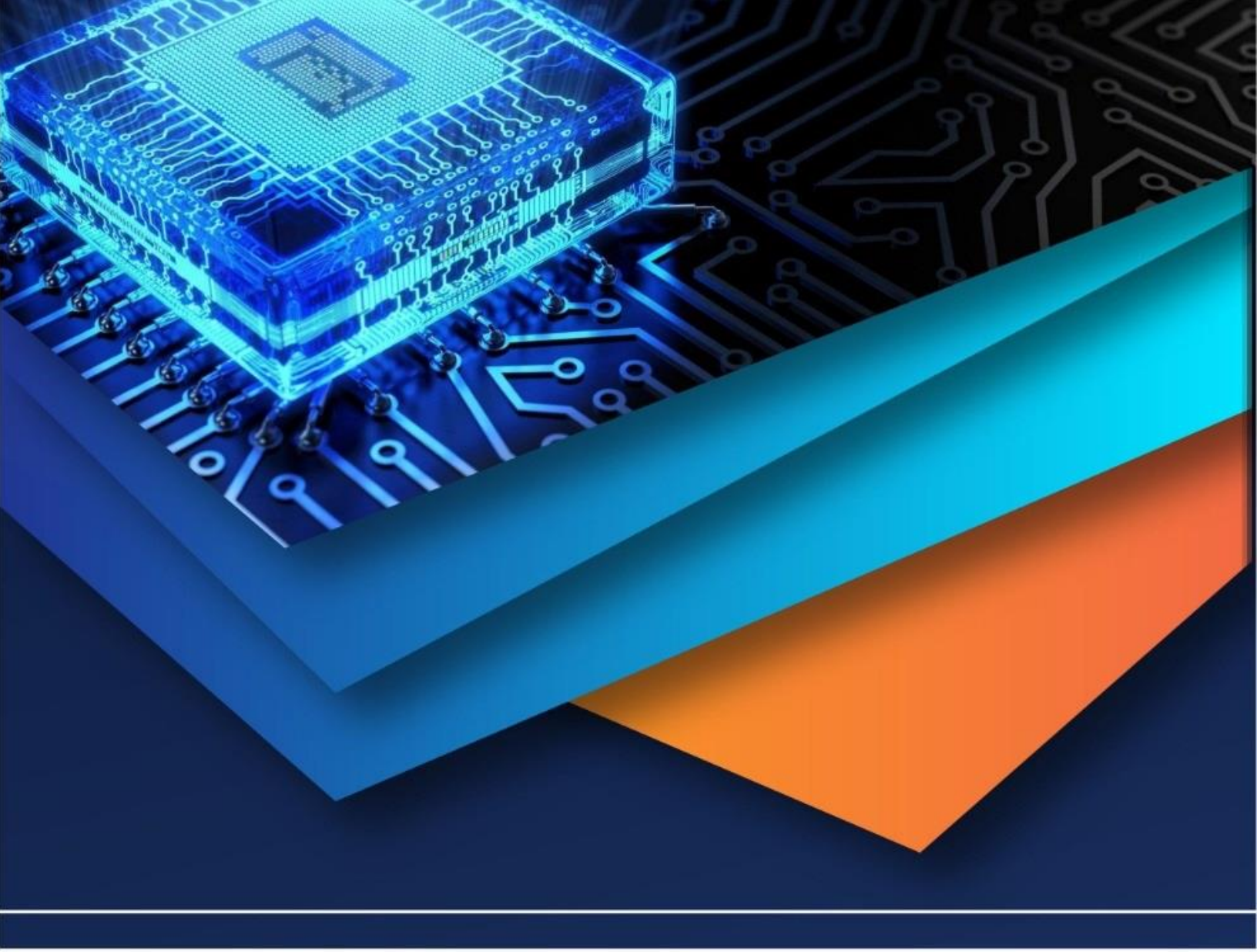
Fig4 Shows suspicious tax evasion in 3D bar graph. Shows percentage of paid/unpaid tax details month-wise.

VI. CONCLUSION AND FUTEURE SCOPE

The proposed method adopts a heterogeneous information network to describe economic behaviors among taxpayers and casts a new light on the tax evasion detection issue. It not only utilizes multiple homogeneous relationships in big data to form the heterogeneous information network, but also maximally utilizes the advantage of trail-based pattern recognition to select the suspicious groups. Through multi-social relationships fusion and reduction, we simplify the heterogeneous information network into a colored model with two node colors and two edge colors.

REFERENCES

[1] D. S. Almeling, “Seven Reasons Why Trade Secrets are Increasingly Important,” Berkeley Technology Law Journal, vol. 27, no. 2, pp. 1092–1118, 2012.
[2] L. Antunes, J. Balsa, and H. Coelho, “Agents that Collude to Evade Taxes,” in Proc. 6th Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS), pp. 1263–1265, May. 2007.
[3] Y. S. Chen and C.-H. Cheng, “A Delphi-Based Rough Sets Fusion Model for Extracting Payment Rules of Vehicle License Tax in the Government Sector,” Expert Systems with Applications, vol. 37, no. 3, pp. 2161–2174, Mar. 2010.
[4] M. J. Ferrantino, X. Liu, and Z. Wang, “Evasion Behaviors of Exporters and Importers: Evidence from the U.S.-China Trade Data Discrepancy,” Journal of International Economics, vol. 86, no. 1, pp. 141–157, Jan. 2012.
[5] W. F. Fox, L. Lunab, and G. Schau, “Destination Taxation and Evasion: Evidence from U.S. Inter-State Commodity Flows,” Journal of Accounting and Economics, vol. 57, no. 1, pp. 43–57, Feb. 2014.
[6] Z. Gao, “Transfer Price-Based Money Laundering: A Transit Trader’s Perspective,” in Proc. 4th Int. Conf. Wireless Communications, Networking and Mobile Computing (WiCOM), pp. 1–5, Oct. 2008.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY