Perfecting Structured Pattern Mining by Text Data through Decision Tree and Bitmap indicator Integration

S. Padmavathi*

Department of Computer Science, Auxilium College of arts and science for women, Pudukkottai, India
*Corresponding Author's e-mail: pdmhr1@gmail.com
doi: https://doi.org/10.21467/proceedings.173.10

ABSTRACT

This paper proposes Bitmap indicator- grounded Decision Tree (BIDT) fashion to estimate the rigidity of data in threat operations. Originally, the Bitmap indicator- grounded Decision Tree bracket is used to induce the knowledge tree which helps to ameliorate scalability. A bitmap indicator in BIDT is used to pierce large databases effectively. Rather of using a list of row ids in the BIDT bitmap indicator, it uses the primary crucial value in a table which is numbered in sequence with each crucial value (array of bytes). The BIDT algorithm along with the bit-wise logical operations AND applied through the select query in SQL. The BIDT fashion with the logical 'AND' driver which helps to combine queries and produce applicable result. Query results help to make a decision tree and to estimate the rigidity of colorful operations. The population frequenters are attained by simply counting the total number of "1" in the bitmaps constructed on the trait to prognosticate and estimate certain attributes like threat factor rate with the help of main table in decision tree for the root node. The algorithm is applied on colorful attributes similar as nonsupervisory false positive rate, threat rate and threat identification time.

Keywords: Bitmap Index-based Decision Tree.

1 Introduction

Data classification plays a vital role in the current growing scenario, and it is investigated by many data scientists, with the demand for a data classification set to continue growing in the future for several reasons: primarily it detects anti-social online behavior, anti-social users in a community, or who act strangely or even appear dangerous (Cheng et al., 2014); Secondly, the investigation of global social and information networks through classification is to gather special knowledge derived from hundreds of millions of users around the globe; thirdly, analyzing the group users about their locations, friends networks, hobbies, activities, and professions for media generated in social communities, including images, videos, sounds, and texts. The main goal is to select an instance that will be assigned and identified by a predefined class when the training set of instances with class labels is given. Classification methods of machine learning are unique data processing features that allow the multi-class text classification. Bitmap Index-based Decision Tree (IBDT) technique to evaluate the adaptability of data in risk operations. The knowledge tree is induced by the Bitmap Index-based Decision Tree classification which helps to improve scalability. A bitmap index in IBDT is used to effectively to access large databases.

1.1 Classification of Machine Learning Algorithms

Data extraction: The foremost goal of this stage is to select only required and related data fields to optimize the memory usage and process the data. This phase was carried as follows: The overall and review text fields are taken from the input dataset. The equal number of customer product-review records in each class (i.e. skewness method) will be collected.

Preparation of review texts: The main goal of this stage is to prepare review text fields for extraction of features (Fig. 2). This phase was carried as follows: Tokenizing each single word by white space or

Proceedings DOI: 10.21467/proceedings.173; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-970666-x-x

punctuation. All stop words have been eliminated (the NLTK website provided the stop word corpus). (Natural Language Toolkit Project)), which have been often used in any text area and the, but do not include specific information required to train this data model. Making all the capital letters in a lower case. Stemming (with Porter stemmer) and reducing different forms to a stemma form.

Bag of words: To construct bag of words the n-gram method as a sequence of written words of length is applied. The sentences are split into words and group them using a combination of n-grams. This stage was carried as follows: Based on the selected n-gram model bag of words (unigrams, bigrams, trigrams) are created from review texts that have passed previous stages and continuous text flow is in use instead of building n-grams from the sentences. Because of the task classifier isn't attempting to understand the meaning of a sentence, it basically creates the input to classifier with all features (tokenized terms, and term groups), and classifier builds the model that assigns the class as accurately as possible. Specific properties, using apostrophes, simple word, segmentation, phrases, parts of speech etc., might also include in N-gram models. The number of words is imported to a specially created hashing term-frequency vectorizer that counts the frequency in the set and assigns a unique numerical value as well as the weights needed for each word for the next classification stage. In other words, a term frequency is classifying how important a word is to a review in a corpus, i.e. in the given review set, the key as a word and value as the number of frequency.

The feature vector alters words in to the numerical value represented in the integer format, i.e. the value of frequency of the word and the numerical value to the given word.

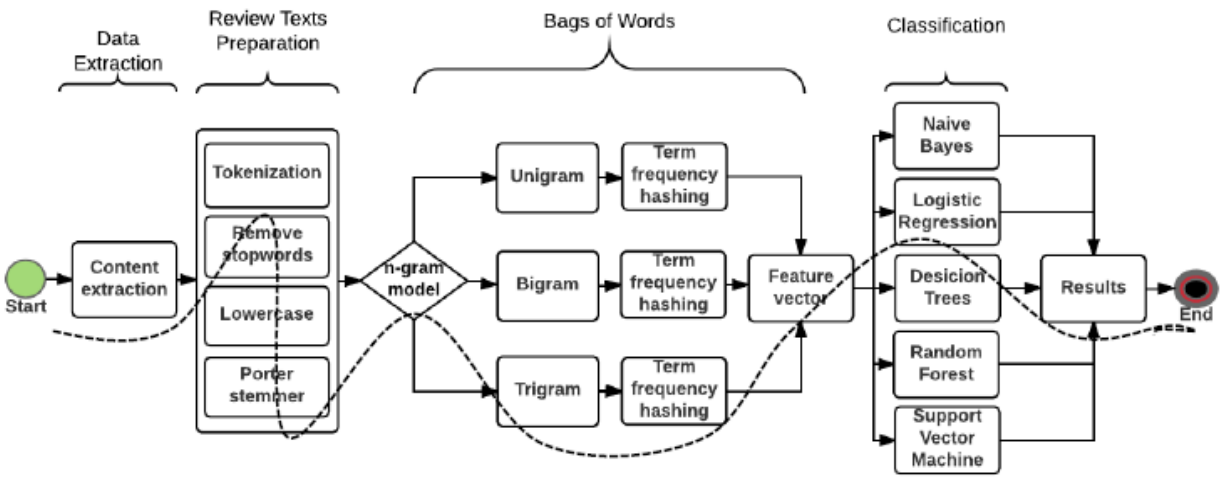

Figure 1: Classification of Machine Learning Algorithms

Classification: This stage was carried as follows: Training and testing of data were performed by the selected classification method using 10-fold cross-validation calculating the mean classification accuracy for the test data. The mean accuracy formula for multi-class classification can be presented as follow (Sokolova and Lapalme, 2009): Where are true positive classification examples, are false positive ones, are false negative ones, and are true negative ones, is the number of classes. By actual labels that are equal to predicted label divided by total corpus size in test data which shows the classification accuracy.

J48 Decision Tree: The classifier was trained by J48 Decision tree model and J48 model is used in weka for classification. The decision trees are made by C4.5 algorithm which uses release 8. (SefikIlkinSerengil, 2018). The C4.5 method is a more advanced variant of the ID3 algorithm, which was also created by Ross Quinlan and is used to generate decision trees (Wikipedia contributors, 2020). So, basically J48 is an advanced version of earlier version of ID3 algorithm which overcomes some of its shortcomings.The C4.5

Proceedings DOI: 10.21467/proceedings.173
ISBN: 978-81-970666-x-x

Series: AIJR Proceedings
ISSN: 2582-3922

algorithm's enhanced features include pruning to prevent overfitting, working with discrete and continuous data, and handling missing data (Sumit Saha, 2018). To train the model using the training data, the J48 algorithm's default parameter values were maintained. The following two images shows the screenshots of the attributes kept for the String to Word Vector filter and J48 decision tree respectively.A method called, bitmap index (BI) (Laxmaiah et al., 2013a, b) was designed to retrieve the information through the priority queue (PQ) and Ice Berg (IB) queries very quickly. The method was proved to be efficient in terms of different thresholds with execution time. However, reducing redundant bits remained unaddressed. This Bit Map Vectors (BMV) (Laxmaiah et al., 2013a, b) along with the data mining techniques were applied to minimize the redundant bits by introducing compacted number of AND operations. Fraud detection in credit card applications is reduced using Neural Network and Support Vector Machine (Zareapoo et al., 2012). This paper aims to design a technique called Bitmap Index-based Decision Tree (BIDT) which improves the adaptability rate. The contributions of BIDT include the following: BIDT evaluates the adaptability of data in risk operations. To determine the company's money laundering risk and improve the scalability efficiently. To evaluate the risk of adaptability using decision tree in more precise manner. To predict the money laundering briefly and assess the risk factor rate by simply counting the total number of integers in the bitmaps by obtaining the population frequencies.

2 Materials and Methods

2.1 Related Works

The greatest potential vehicles for money laundering in financial institutions have resulted in the largest form of cross-border money laundering. The opportunities and limitations related to anti-money laundering were studied through the reverse engineering methods (Moser et al., 2013). Money laundering under electronic payment (Weibing, 2011) is another method was designed with the motive of thwarting electronic money laundering crime and possibility of early detection (Pulakkazhy and Balan, 2013). Nevertheless, the financial computerization inclination remained unsolved. Due to advancement and development in the field of internet money laundering in banks, the volume of data is growing at a faster speed. Classification (Luo, 2014) based algorithm for the effective identification and detection of suspicious activities (Jayasree and Siva Balan, 2015) was designed. However, suspicious activities were detected at an earlier stage, it was done at the cost of time.

The relationships between attributes were identified by the indexing techniques and (Suresh and Thammi Reddy, 2014) graph theoretic approach was applied with the aid of apriori algorithm to reduce the retrieval time. One of the key problems when dealing with money laundering is the handling of voluminous financial information. The core drawbacks are due to the fact that there appears no same tactics with financial institutions to deal with the anti-money laundering. The verification of customer in an extensive manner and detect fraudulent activities (Jayasree and Siva Balan, 2013), digital forensics and database analysis were integrated (Flores et al., 2011). However, the evaluation of the digital forensic practices still challenged within the organization and the harmonization approach (Nikoloska and Simonovski, 2012) was applied for the timely detection of crime as the legalization still difficult. Adaptive join operators (Bornea et al., 2010) through multiple index nested loop reactive join resulted in the optimization of the customer activities. Based on the aforesaid methods, this work propose a technique called Bitmap Index-based Decision Tree for effective evaluation of risk factor on financial money laundering in banks.

2.2 Bitmap Index-based Decision Tree

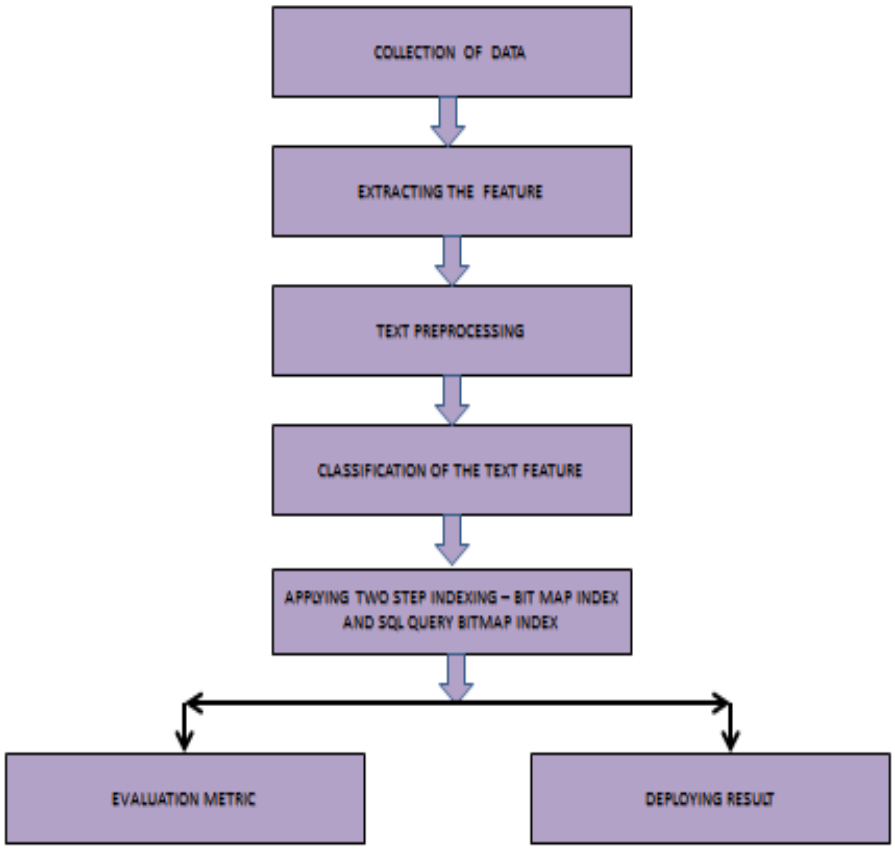

Figure 2: Evaluation and Deployment of Bitmap Index-based Decision Tree

Fig.2 shows the collection of data which is being processed and involved, comprises with many steps for evaluation. The features are extracted from the collection of data which is involved in text processing and the next step the classification of different label is predicted by analyzing text feature. Once the features are selected, it will be involved with two step indexing technique which is bitmap and sql query bitmap index to obtain better result through the indexing method.

The main objective of the proposed work Bitmap Index-based Decision Tree is to evaluate the risk factor of financial organizations using the indexing scheme. The indexing technique uses the rows and columns to store the information which improves the scalability rate. Transactions of various users are evaluated using the Bitmap Index-based Decision Tree technique. Decision tree is constructed by mapping of the bit in fuzzy form '0' and '1'. The determination rules in the BIDT technique is calculated based on the nodes and sub nodes. The pointers are used for purpose of providing indexing in BIDT technique to the rows in a table that contain given key values.

Proceedings DOI: 10.21467/proceedings.173
ISBN: 978-81-970666-x-x

Series: AIJR Proceedings
ISSN: 2582-3922

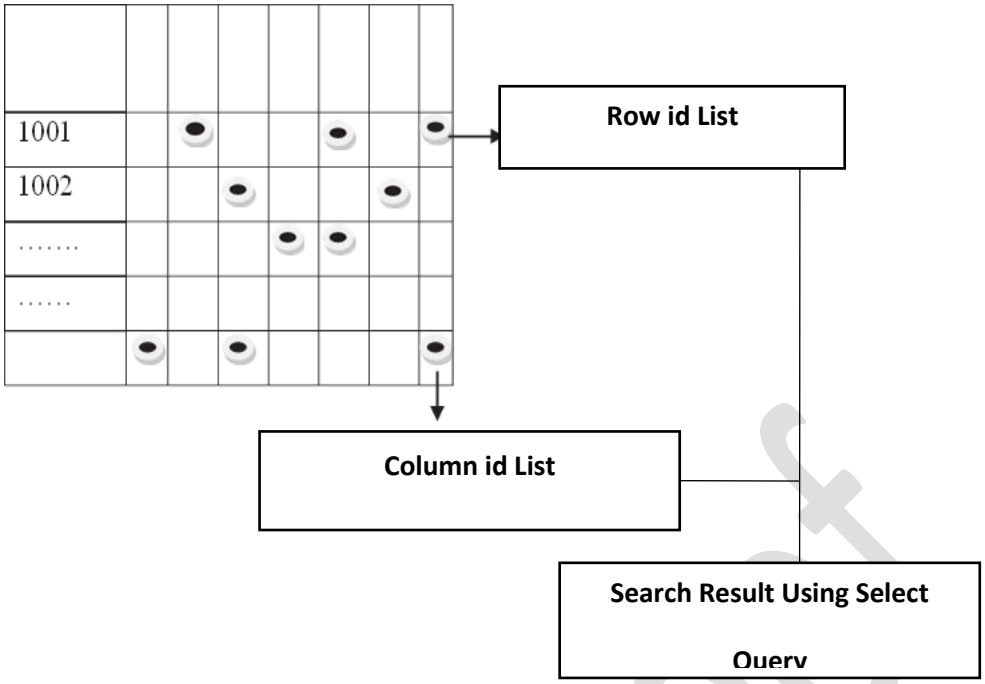

Figure 3: Bitmap Structure Representation

Fig.3 shows the bitmap structure representation. Bitmaps which use cardinality rows and columns effectively evaluates the risk factors involved in an account e.g. Money laundering account. BIDT technique easily evaluates the data in the database table by relating it effectively. Since the bitmap index evaluates easily through arrays and bitwise logical operation AND to obtain better results. In the figure 3, the horizontal distributed black dots represent the row ID list and vertically arranged dots indicate the column ID list. Considering the account number is 1001 in figure 3, the row id list is calculated according to the three horizontal black dots. Likewise, the black dots are distributed along with the column id list measured in a perpendicular way. The column id and row id list create the bitmap index easier.

Fig. 3 shows the architecture diagram of Bitmap Index-based Decision Tree (BIDT) technique used for easy description of the risk evaluation. As illustrated in Fig. 3, the account details of a transaction sent as messages are noticed, converted into structured format, stored in a table and then the bitmap indexing procedure is applied. The bitmap index uses the rows and column id information on the large database. This paper proposes two steps indexing to identify the pattern with the help of the row and column in a table and decision tree are carried out in the design and implementation of BIDT technique. Select query option and logical AND operator is used in Indexing in BIDT. The results are used to for the purpose of constructing the decision tree, where the primary key of transaction is identified and risk occurrences are identified. Fraudulent activities occurring with high frequency are measured and processing is carried out with higher efficiency.

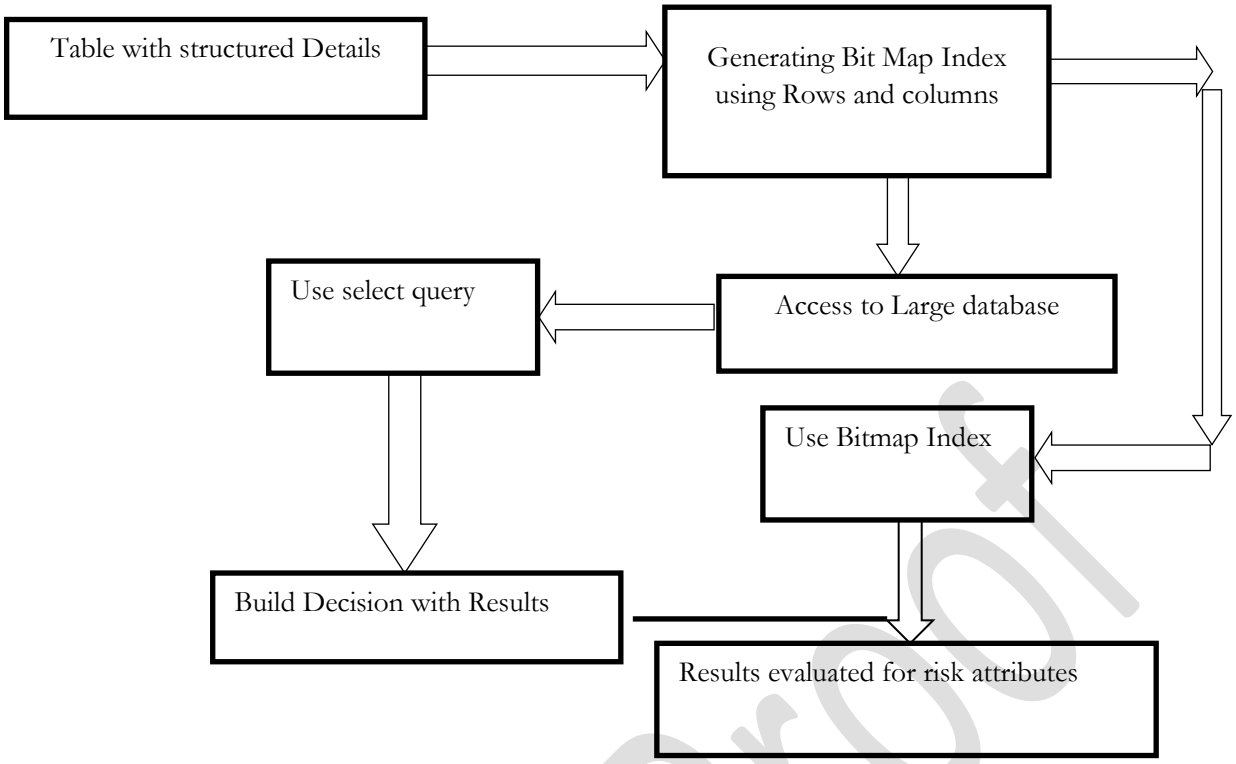

Figure 4: Evaluation of Risk Attributes

2.3 Bitmap indexing

The BIDT technique uses bitmap index, a data structure which is used to efficiently access larger bank databases involved in fraudulent risk factors such as money laundering etc. The proposed index work, provide pointers to the rows in the table where multiple transactions are carried out. The risk factor is analyzed easily using the given distinct key value on each transaction for a particular primary key value.

The bitmap index, list of row id and column id are used to identify whether a particular transaction is carried out on the specific primary key value. The map index contains the record in the table in a linear fashion. The proposed Bitmap index work produces the result in a fuzzy form (i.e., 0 or 1).

For each Primary key value bit mapping is used for easy risk evaluation on fraudulent accounts based on the fuzzy form. During the mapping operation, each bit in the bitmap corresponds to the existing row and column id.The rows and columns of a bitmap index are assigned unique values to facilitate risk assessment. Bitmap indexing method efficiently categorizes the rows and columns based on the transaction details that reduce the risk evaluation time, since key value is used in BIDT technique.

2.3.1 Select query structure

Bitmap indexing is used for efficient query-based risk evaluation on transactions with multiple different key-value databases. The bitwise logical operator is used to answer the query it is formularized as,

Select Bank Table where customer transaction is >> specified range AND count the transaction time.

Each operation takes bitmap indexing with the same size of primary key value and computes the desired tuple range. Bitmap indexing uses SQL queries are used on the table in the IBDT technique to efficiently perform special operations using combinations of multiple indices. The select query structure in the IBDT

Proceedings DOI: 10.21467/proceedings.173
ISBN: 978-81-970666-x-x

Series: AIJR Proceedings
ISSN: 2582-3922

technique is used to extensively provide data for risk analysis. To combine the queries and produce the desired result, the logical 'AND' operator is used in the IBDT technique.

2.3.2 Bitmap index frequency

A special type of Bitmap index uses the row and column id index frequency is used in the Bitmap index generation. The computation of frequency depends on the lower cardinality values, these values are used during the computation. The advantage of using a Low cardinality column in the IBDT technique is that it has a unique key value for easier evaluation of the risk factors. Multiple Bulk transactions of the primary key value are identified and risk that occurred in the transaction is analyzed in the IBDT technique. The bitmap index frequency using low cardinality column is computed as

$$\boldsymbol{Frequency} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{Ac}\,[\boldsymbol{\pi i}, \boldsymbol{\pi i + 1} \ldots . \boldsymbol{\pi i + n}]$$

From (2) analyze the risk factor by measuring the frequency point which is the overall sum of the attributes (i.e., type of attributes used to identify the customer way of the transaction). In (2), Ac represents the attribute column in the IBDT technique in which pi is the primary key value used for each transaction. Low cardinality is calculated in the IBDT technique to improve the regulatory risk rate.

2.3.3 Decision tree structure

The most popularly used data mining technique is the decision tree to analyze the risk factor. The IBDT technique uses a decision tree to partition the decision into smaller partitions for analyzing risk factors. IBDT technique takes the input as the set of objects with the predictive attributes 'A'. The various three sequential important attributes are location, business type, age, gender.

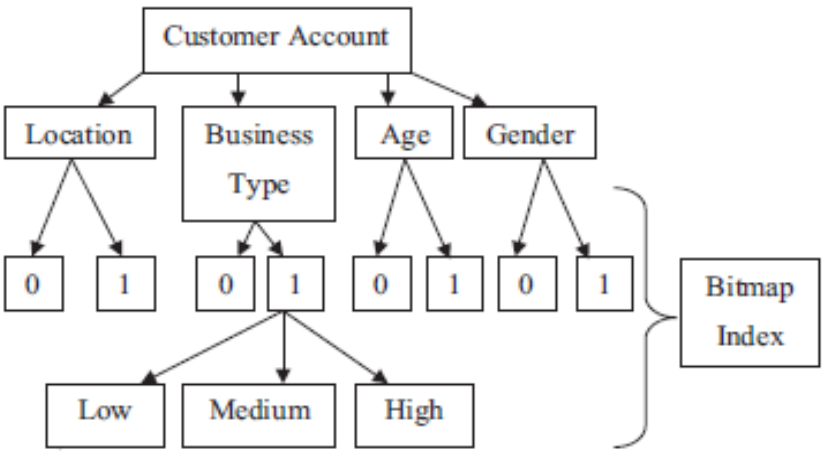

Figure 5: Analyzing Risk Factor with the Attributes using Decision Tree with BIDT

For each sub node, a new set of bitmaps is generated, each corresponding to a class of the node. The schematic form of the decision tree is shown in Fig. 5.The results of the sql queries coincide exactly in order to build a decision tree and to evaluate the adaptability risk factors. The attribute predicts and evaluates the risk factor rate. By applying Count and bitwise logical operations AND improves the Bitmap indices 0 and 1 performance. The population frequency of each class is to be determined, to obtain the root node of the decision tree, For the root node, The main value of the key value of

the table is used in a solutions tree, the population frequencies are obtained simply by including the total number of "1" in raster images. The number of '1' in the bitmap denotes the transaction carried out between the intermediate primary key values of the table. The intermediate and core transaction is noted and partitioned as low, medium, and high range.

Each primary key value belongs to the set of the mutually exclusive attribute classes. The risk factor evaluation rate can be obtained by successive construction of decision trees. An advanced iterative AID3 algorithm is used to analyze the risk factor effectively AID3 is used widely for the generation of the decision tree for risk evaluation. IBDT technique uses Occam's razor principle, where all the attributes are used for the entropy computation.

The processing step is described as,

Begin

Step 1: Senses all the attributes within the specified table

Step 2: for every Attribute in 'A'

Step 2.1: Entropy is computed with the utmost information

$$\boldsymbol{Entropy} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{\log_2 pi}$$

Pi is that the positive value range

Step 3: for each positive value on the transaction

Step 3.1: Add the new tree node below the root node of the decision tree to identify the bitmap index range

Step 3.2: Bitmap index value with the '1' is analyzed and the level of the transaction is noted

Step 3.3: Complete the decision tree till the leaf node with the target value range

Step 4: End For

Step 5: the positive and negative value is achieved by entropy

Step 6: Negative values are discarded, by minimizing the risk factor evaluation time

Step 7: Finding leaf data reduced the amount of test on pruning (i.e., risk evaluation)

Step 8: End for

End

4 Results and Discussions

4.1 Effect of Database Size on Query Processing Time

This experiment used the above data set to compare the relative performance of Query on non-indexed retrieval (SEQ) and Bitmap (IBDT) indexed retrieval. In order to observe how the methods to scale with respect to the database size, five data sets were generated and the size of database |D| was varied from 100 to 1000. Fig.6.1 to 6.5 shows the average processing time required by Query. It can be seen that the query processing time is proportional to the database size. As the database size increases, the query processing time also increases, as is expected. However, it can be seen that the indexed retrieval performs

Proceedings DOI: 10.21467/proceedings.173
ISBN: 978-81-970666-x-x
Series: AIJR Proceedings
ISSN: 2582-3922

better than non-indexed retrieval even when the database size is large, as indicated by the slope of the lines. Query processing time grows faster with the increase in database size in the case of non-indexed retrieval as compared to the increase in query processing time for indexed retrieval.

Table 6.1: Query Processing Time for non-indexed retrieval (SEQ) and Indexed Bitmap Retrieval (IBDT) - Dataset1.

Database Size	SEQ (ms)	IBDT (ms)
200	30	10
400	45	16
600	67	29
800	80	31
1000	110	56

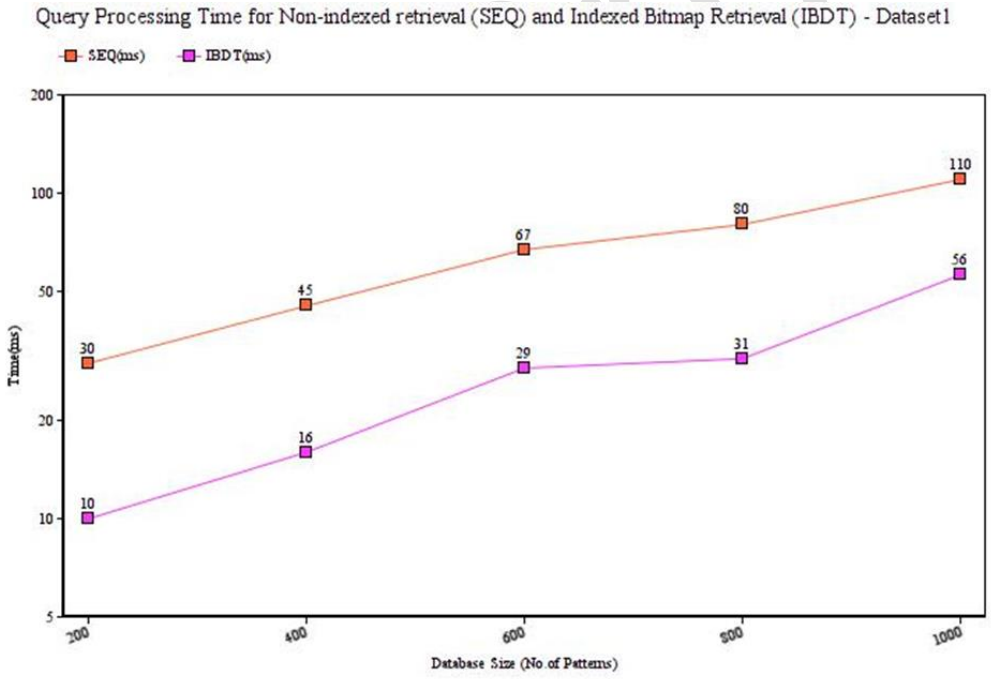

Figure 6.1: Query Processing Time for non-indexed retrieval (SEQ) and Bitmap Indexed Retrieval (BIDT)

Table 6.2: Query Processing Time for non-indexed retrieval (SEQ) and Indexed Bitmap Retrieval (IBDT) – Dataset2

Database Size	SEQ(ms)	IBDT (ms)
200	25	10
400	44	15
600	65	29
800	77	30
1000	99	52

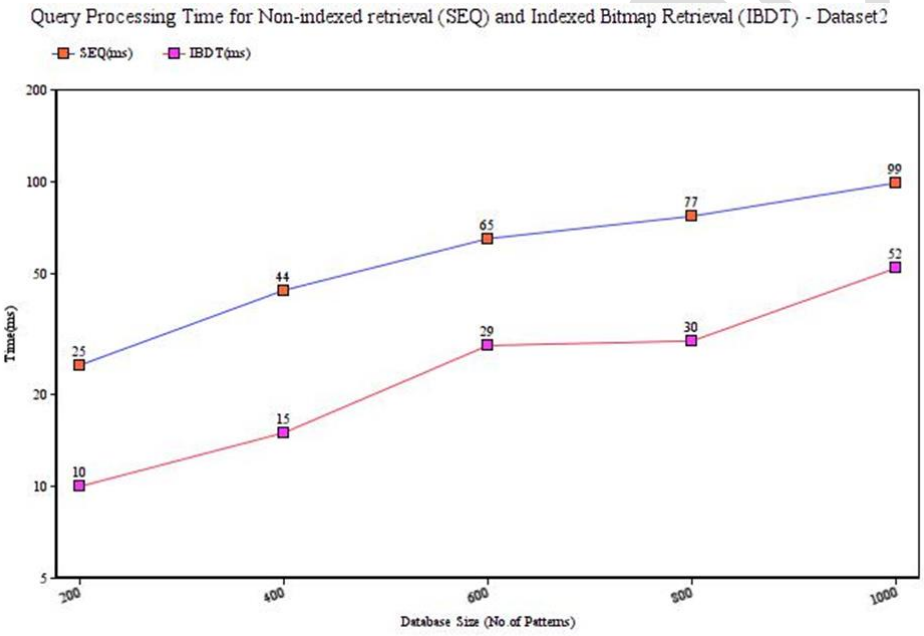

Figure 6.2: Query Processing Time for non-indexed retrieval (SEQ) and Indexed Bitmap Retrieval (IBDT) – Dataset2

Proceedings DOI: 10.21467/proceedings.173
ISBN: 978-81-970666-x-x

Series: AIJR Proceedings
ISSN: 2582-3922

Table 6.3: Query Processing Time for non-indexed retrieval (SEQ) and Indexed Bitmap Retrieval (IBDT) – Dataset3

Database Size	SEQ(ms)	IBDT (ms)
200	27	22
400	30	19
600	55	40
800	79	50
1000	101	70

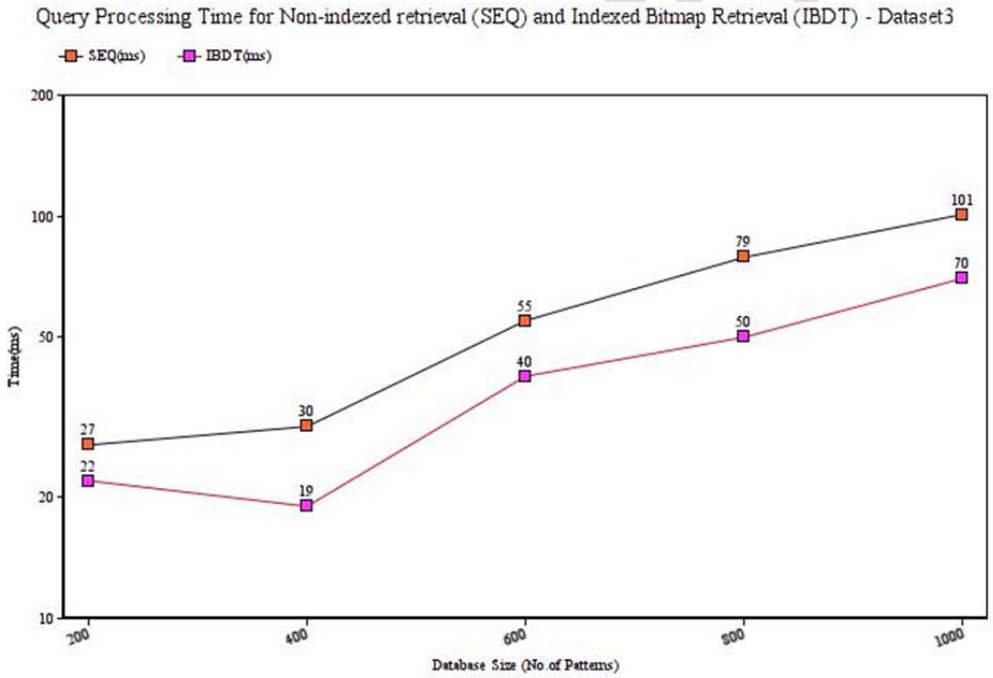

Figure 6.3: Query Processing Time for non-indexed retrieval (SEQ) and Indexed Bitmap Retrieval (IBDT) – Dataset3

Table 6.4: Query Processing Time for non-indexed retrieval (SEQ) and Indexed Bitmap Retrieval (IBDT) – Dataset4.

Database Size	SEQ (ms)	IBDT (ms)
200	29	11
400	40	20
600	60	31
800	70	30
1000	100	48

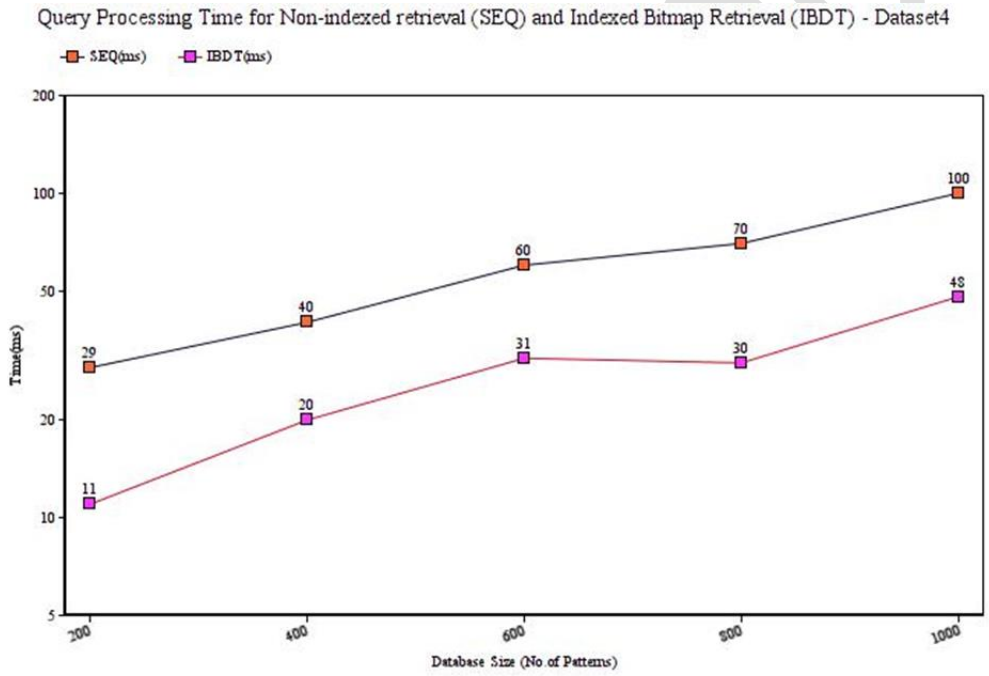

Figure 6.4: Query Processing Time for non-indexed retrieval (SEQ) and Indexed Bitmap Retrieval (IBDT) – Dataset4

Proceedings DOI: 10.21467/proceedings.173
ISBN: 978-81-970666-x-x

Series: AIJR Proceedings
ISSN: 2582-3922

Table 6.5: Query Processing Time for non-indexed retrieval (SEQ) and Indexed Bitmap Retrieval (IBDT) – Dataset5

Database Size	SEQ (ms)	IBDT (ms)
200	27	10
400	40	12
600	57	21
800	68	20
1000	98	40

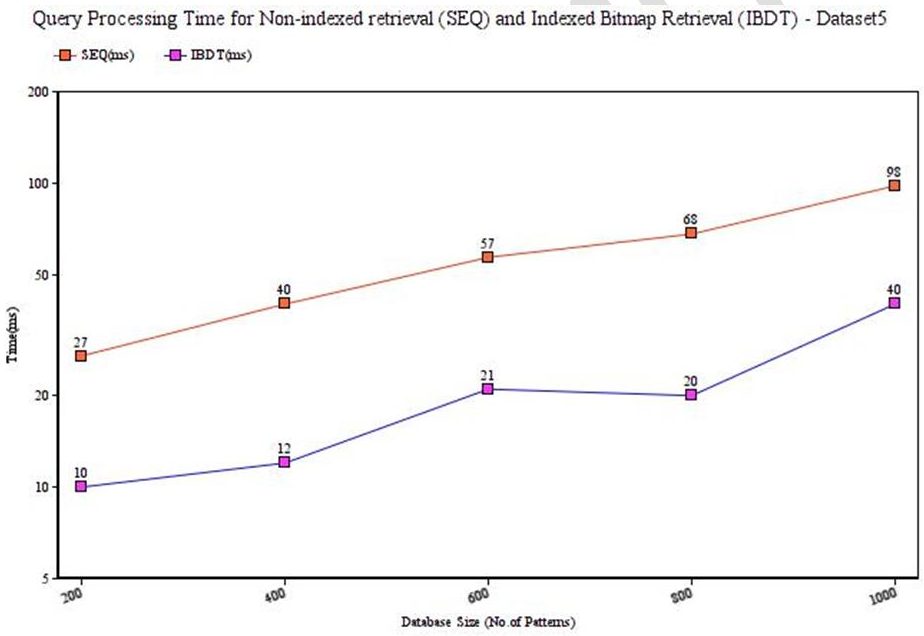

Figure 6.5: Query Processing Time for non-indexed retrieval (SEQ) and Indexed Bitmap Retrieval (IBDT) – Dataset5

4.2 Impact of Adaptability Rate

The adaptability rate of crime using the IBDT technique is the ability of the service provider (i.e., bank) to adjust changes in services based on customers' requests during the transaction. Higher the adaptability rate, the decision will be quick. The measure of adaptability using the IBDT technique, SEQ.

Table 6.6: Measurement of Adaptability Rate

Technique	Adaptability Rate (%)
IBDT	75
SEQ	60

The adaptability rate using the IBDT technique is higher when compared to the existing method SEQ the rate of adaptability in the IBDT technique improves with the application of decision tree structure that efficiently partitions the decision into smaller partitions.

5 Conclusion

It is challenging to evaluate the risk factor of the financial organizations since the proposed work Bitmap Index-based Decision Tree is to evaluate the risk factor of financial organizations using the indexing scheme with most popularly used method called decision tree structure in data mining techniques. The two steps indexing is used to identify the pattern with the help of the row and column id in a table and decision tree are carried out in the design and implemented with BIDT technique which is identify fraudulent activities occurring with high frequency are measured and processing is carried out with higher efficiency. The indexing technique uses the rows and columns to store the information which improves the scalability rate. By measuring frequency point the overall sum of the attributes for analyzing the risk was evaluated. Bitmap indexing method efficiently categorizes the rows and columns based on the location, business type, age and gender details of the customer that reduces the risk identification time and greatly improves the adaptability rate. As the database size increases, the query processing time also increases, so five datasets were generated to know how the index creation time varies with the increase in the size of the database so as to rule out erratic behavior as the database size grows very large. The algorithm is applied and concentrated on various attributes such as regulatory false positive rate, risk rate and risk identification time.

6 Declarations

6.1 Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Bornea, A. Mihaela, Vasilis, Vassalos, Yannis, Kotidis, "Adaptive join operators for result rate optimization on streaming inputs". *IEEE Transactions on Knowledge and Data Engineering*, 22 (8),2010.

[2] Castello´ n Gonza´ lez, Pamela, Vela´ squez, D. Juan, "Characterization and detection of taxpayers with false invoices using data mining techniques". *Expert Systems with Applications*,2013.

[3] Eldin Helmy, Tamer Hossam et al, "Design of a monitor for detecting money laundering and terrorist financing". *International Journal of Computer Networks and Applications*, 1 (1). 2014.

[4] A. Flores, Denys, Angelopoulou, Olga, Self, Richard, "An antimony laundering methodology: financial regulations, information security and digital forensics working together". *Journal of Internet Services and Information Security*, (JISIS) 3.

[5] Jayasree, Vikas, Siva Balan, R.V., "A review on data mining in banking sector". *American Journal of Applied Sciences*, 10 (10), 1160–1165, 2013.

[6] Jayasree, Vikas, Siva Balan, R.V., Money laundering identification on banking data using probabilistic relational audit sequential pattern. Asian Journal of Applied Sciences, 1996–3343,2015.

[7] M. Laxmaiah, et al,"A compressed bitmap vector method to assess aggregate queries competently",*International Journal of Advanced Research in Computer Science and Software Engineering*, 3 (11),2013a.

[8] M. Laxmaiah, etal, "An approach to evaluate aggregate queries efficiently using priority queue technique". *International Journal of Emerging Trends & Technology in Computer Science*, 2 (3),2013b.

[9] Luo, Xingrong, "Suspicious transaction detection for anti-money laundering",*International Journal of Security and its Applications*, 8 (2),2014.

[10] Moser, Malte, Bohem, Rainer, Breuker, Dominic, "An inquiry into money laundering tools in the bitcoin ecosystem". e-Crime Researchers Summit, IEEE, 2013.

[11] Nikoloska, Svetlana, Simonov ski, Ivica, "Role of banks as entity in the system for prevention of money laundering in the Macedonia". *Procedia- Social and Behavioral Sciences*, 2012.

[12] Phua, Clifton et al, "Resilient identity crime detection",*IEEE Transactions on Knowledge and Data Engineering*, 24 (3). 2012.

[13] Pulakkazhy, Sreekumar, R.V.S., Balan, "Data mining in banking and its applications-a review",*Journal of Computer Science*, 9 (10), 1252–1259,2013.

[14] Roberto Cortinas et al, "Secure failure detection and consensus in trusted pals". *IEEE Transactions on Dependable Secure Computing*, 9 (4). 2012a.

[15] Roberto Cortinas et al, "Secure failure detection and consensus in trusted pals",*IEEE Transactions on Dependable Secure Computing*,2012b.

[16] C.H. Suresh, K.Tammi Reddy, "Graph based approach to identify suspicious accounts in the layering stage of money laundering". *Global Journal of Computer Science and Technology*. 1 (1), 81–87. 2014.

[17] Weibing, Peng, "Research on money laundering crime under electronic payment background". *Journal of Computers*, 6 (1).2011.

[18] Zareapoo, Masoumeh, K.R Sreeja, M. Afshar Alam, "Analysis of credit card fraud detection techniques: based on certain design criteria". *International Journal of Computer Applications*, 52 (3).2012.