A Study on Search and Filtering Analysis of Unstructured Text in Financial Service Industry

Ashok Kumar. D[#1]
Department of Computer Science,
Government Arts College,
Trichy– 620 022, India
[1]akudaiyar@yahoo.com

Mohan Raja. D[*2]
Department of Computer Science,
PRIST University, Vallam,
Thanjavur – 613403, India
[2]drmrprist@gmail.com

Abstract — **Management of unstructured data is viewed as one of the major unsolved problems. Nearly eighty percentage of enterprise data resided in unstructured formats such as text files, email, customer profile information, external information, video documents and audio samples in various fields including search, prediction, business intelligence, financial service industry – FSI and in sematic web, which aims at converting web of unstructured data into a "web of data that can be processed directly and indirectly by machines." The reason behind is that the tools and techniques that have proved so successful in transforming structured and unstructured data in FSI and actionable information to resolve increasing complexity for the transacting database information. With the newly launched FSI regulations, the issue has drawn close attention from governments, financial institutions and research scholars. A few fundamental and important questions confront us in resolving the ever increasing complexities in this issue. However in FSI, there exists, many issues in the Suspicious Activity Report – SARs such as large value reporting, delivering, analysing processes and traditional investigations consuming large volume of man-hours.**

Keywords — Domain Index, Unstructured Text, SQL queries, Financial Service Industry

I. Introduction

The increase of economic globalization and market economy with advent of information technology, financial data are being generated and accumulated at an unprecedented scale. Such increased sources and quantity of unstructured information has created further need for categorization and interpretation of the content. Therefore there is a critical need for automated approaches to efficient and effective utilization of a massive financial data to support companies and individuals in strategic planning and investment decision-making in unstructured information retrieval.

Data mining is conceptualized to be able to uncover hidden patterns and predict trends and behaviors in financial markets. Data mining has been applied to a number of financial applications, including development of trading models, investment selection, loan assessment, portfolio optimization, fraud detection, bankruptcy prediction, real-estate assessment, and so on. However, previous researches showed that there exists huge space in enhancing efficiency of detecting SARs for financial applications [7]. In China, the detection efficiency was 14%, and in developed countries, the number is also small about 2%.

This paper aims to review the removal of constraints on unstructured data and the SARs of FSI and the criteria used in the proposed combination of Swarm Intelligence Technique and by the employment of an efficient traditional domain index. The proposed study is on the processes of unstructured data with redefined boost memory parameters in domain index and reengineered data mining algorithms in indexing which specifically used for the unstructured data and to enhance the performance and efficiency in online transaction database information.

The rest of the paper is organized as follows. Section II Introduces the basic concept of data interpretation involved in financial applications. Section III describes existing methods of identification. Section IV Indicated the proposal of this research and its criteria. In section V, we discuss challenges and area for scope of future research.

II. BASIC CONCEPTS DATA INTERPRETATION IN FINANCIAL APPLICATIONS

Systems used in financial service industry nowadays are mainly based on fixed rules or given thresholds which can be easily escaped and evaded by the money launders. These suspicious transactions prescribed in the administrative rules are so unclear that it is difficult to be used and quantified to help detect SARs, for example, 'high transferring frequency within a short time', 'abnormal transaction amount in recent days' and so on. In a financial application, the suspicious transactions are hidden in the normal customer transactions, but the suspicious transaction has its unique character, which is different from the normal transaction, and its main features are as follows:

(1) Multiple accounts of funds dispersed into the same account, or transfer out focally, or transfer in focally in the short term, or distributed, which is clearly incompatible with the customer identification in financial condition business. (2) Many bank account of one natural person have some unknown

funds (3) One natural person who has a different financial institution account under his name or the same financial institution account under the same username, in the short term, frequently moves funds among multiple accounts, or have unknown capital flowing in. (4) In the short term capital flows frequently (5) Long term fixed accounts suddenly excess a lot of money, but unknown origin (6) In accordance with single transactions or in the same day transactions volume accumulating more than the threshold limit of the customer daily transactions stipulated in – KYC Know Your Customer profile. (7) Equivalent foreign currency transactions of more than the country regulatory limit (8) Depositing case, Cheque book money, Cash exchange, settlement and sales, cash payment order (9) Cash bills releasing and other forms of cash receipts and payments and so on. [12].

TABLE I
CONSTRAINTS IN FINANCIAL INSTITUTIONS

Constraints	

Physical Constraints	Money laundering infrastructure;
	Detection systems;
	Data incomplete problems and
	Data missing problems

Policy Constraints	Anti-Money laundering internal controlling mechanism; Changing regulations
Human behaviour Constraints	Anti-Money laundering motivation; Expertise, capability and training efficiency

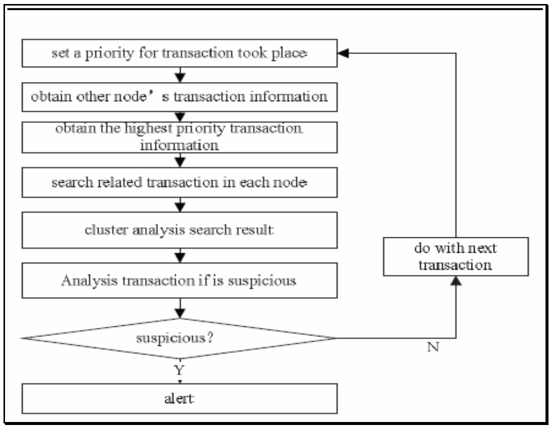
set a priority for transaction took place
obtain other node' s transaction information
obtain the highest priority transaction information
search related transaction in each node
cluster analysis search result
Analysis transaction if is suspicious
suspicious?
Y
N
alert
do with next transaction

Figure. 1 Outline of the suspicious report flow chart

III. EXISTING METHODS IN FINANCIAL APPLICATIONS

The approaches of flagging of suspicious activity, level of such detection rates, the ability to provide meaningful management information on them may involve huge time and cost over runs.

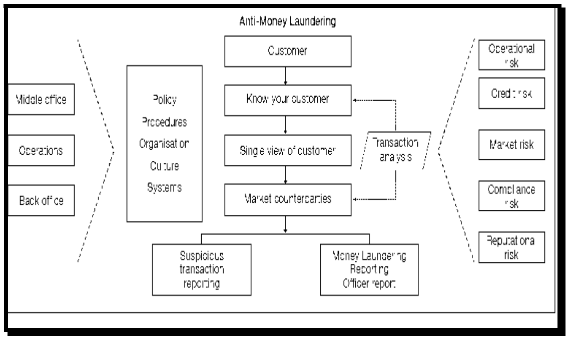

Figure. 2 A sample line diagram of financial application to detect suspicious transactions in the system

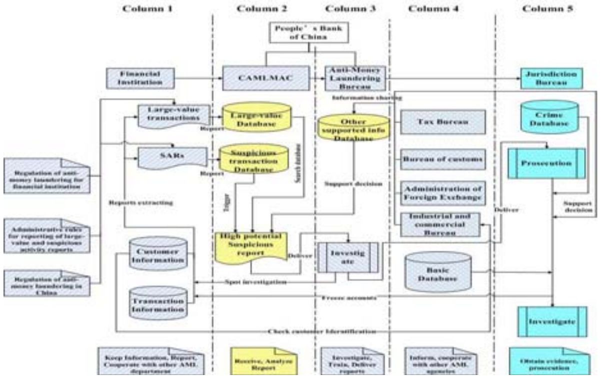
Column 1
Column 2
Column 3
Column 4
Column 5
People' s Bank of China
Financial Institution
CAMLMAC
Anti-Money Laundering Bureau
Jurisdiction Bureau
Information sharing
Large-value transactions
Large-value Database
Other supported info Database
Tax Bureau
Crime Database
SARs
Suspicious transaction Database
Bureau of customs
Prosecution
Report
Report
Search database
Support decision
Administration of Foreign Exchange
Regulation of anti-money laundering for financial institution
Reports extracting
Trigger
Deliver
Support decision
Administrative rules for reporting of large-value and suspicious activity reports
Industrial and commercial Bureau
High potential Suspicious report
Deliver
Investigate
Customer Information
Basic Database
Regulation of anti-money laundering in China
Spot investigation
Transaction Information
Freeze accounts
Investigate
Check customer Identification
Keep Information, Report, Cooperate with other AML department
Receive, Analyze Report
Investigate, Train, Deliver reports
Inform, cooperate with other AML agencies
Obtain evidence, prosecution

Figure.3 Financial application software business flow

During investigation through financial application software's, we could come across clear delineation of core business models related with commercial and I.T requirements, conducting workshops and training, stringent evaluation of stimulus response mechanism, vigorous shortlisting procedures etc., But these, inevitably depend on the quality of the data in consideration, for a quantifiable outputs and interpretations. Deficiencies or inconsistencies in existing financial service industry and KYC data can have large implications on the effectiveness and reliability of the information supplied, even by the most advanced transaction monitoring systems. Data management experience in helping prepare and cleanse data to ensure optimum data quality for the seamless construction of data feeds and correct manipulation of data during data model implementation reveals and requires for the necessity of new techniques Although the number of components placed in an overall detection of suspicious transactions, good technology will equip organisations with an improved level of defence in the fight against financial crime risks.

The list of actions referred above might be capable of being seamlessly processed, but yet the resultant derived outputs are not detected during the actual business transactions.

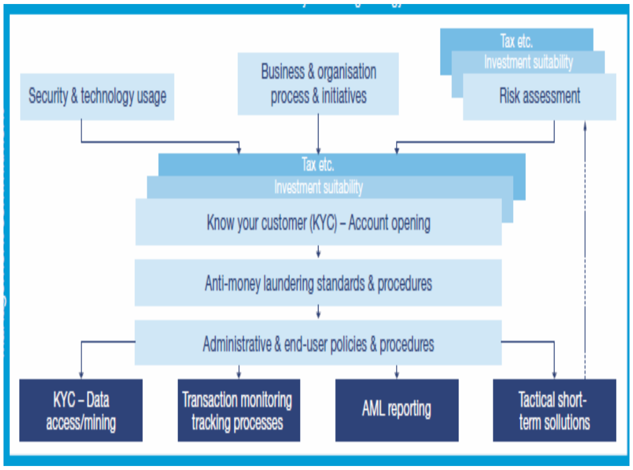
Security & technology usage
Business & organisation process & initiatives
Tax etc.
Investment suitability
Risk assessment
Tax etc.
Investment suitability
Know your customer (KYC) – Account opening
Anti-money laundering standards & procedures
Administrative & end-user policies & procedures
KYC – Data access/mining
Transaction monitoring tracking processes
AML reporting
Tactical short-term solutions

Figure. 4 Potential components of Financial Service Industry

The existing methods are outlined: (1) Transaction monitoring – scanning and analyzing data for potential suspicious transaction activity. (2) Automation of regulatory reporting, filing of suspicious activity reports (SARs), currency transaction reports (CTRs), or other statutory regulatory reports. (3) Suspicious Transaction Monitoring, identification and detection software, systems integration and data warehousing services; (4) a detailed audit trail to demonstrate compliance reports and action taken reports before the regulators. (5) Integrated data management solutions; Risk management assessment, processes and technology, including compliance enabling technique; and (6) Analytical engines.

There are some typical research method on mining suspicious transactions, such as: (1) Method based on multi-agent technology [9], (2) Method based on SAOP technology – Simple Access Object Protocol [10], (3) Method based on SIT- Swarm Intelligence Technique [8] and (4) Statistics based SARs detection [11]

Methods based on the data mining technology in distributed heterogeneous computing environment, which are incidentally in three discrete levels; the application support layer, the system control layer, resource encapsulation layer but the structure does not address real-time monitoring function. To overcome the constraints in the existing methods, this paper uses the Swarm Intelligence Technique with memory parameter boosters and reengineered data mining algorithm to analyse the customer transaction behaviour, and proposes the suspicious transaction detection for the financial service industry as real-time monitoring system.

IV. Proposed Research and its Criteria

The approach on unstructured text search and filtering classification in financial services industry proposed from the suspicious transaction monitoring and filtering can be tracked / traced right through the inception of the transactions in real time as event driven and event triggered. The basic tenet / concept / process of unstructured text search and filtering classification methods is the deployment of customer transactions and the behavioural issues in them. Though it may seem perverse, the per mutable approach approach proposed here is quite new and it shall run on real time basis.

This work is not for narrowed down solutions on specific area/or finite domain but it is to encompass a wide range of implementation in the unstructured text search and filtering classification. This could be the much needed approach in the financial applications for FSI, with an eye and scope for future/ further research and development such development shall be possible with a seamless interface with this present work. The main task of this study is to find emerging patterns and so as to make them support the already existing suspicious transaction monitoring efforts in such a way that re-invention of wheel is avoided. Further, it is to be said/mentioned that this work tends to redefine and reorient complex algorithms like SIT which are often/at times outperformed by less complex and faster algorithms.

SQL queries used over unstructured text databases coupled with memory parameter boosters/triggers in domain index and SIT (swarm intelligence technique) interleafing/interfacing as multilayer linkages are deployed/unveiled here for a better research kind.

It is quite normal that SQL queries over unstructured Text databases embed data resulting in ‘structured’ nature; while processing a text database with / through information extracting systems, one can come across / identify a variety of structured relations. Such relations may be interdependent, or mutually exclusive, both requiring further SQL queries issuances. Similarly processing SQL queries in text based scenario presents multiple challenges of varied kinds. One key challenge is efficiency with respect to time, since information extractions is a time consuming process in domain.

Another key challenge is result equality; extractions systems might yield/output erroneous information or misinformation, in effect that they should not have been captured. And further, processing efficiency related predetermined decisions to avoid large volume of unstructured documents may end up with compromised results.

Consider a text database D and n “base” relations R1, . . ., Rn defined over D. Each base relation Ri can be extracted from D using one or more information extraction systems. We assume that all base relations

R1, . . ., Rn share the same primary key K and no other attributes, and define a view V = K _ni = 1 Ri as the natural outerjoin of R1, . . ., Rn over the K attributes.

We consider SQL selection-projection queries over V with selection condition conjuncts of the form A = t, where A is a textual attribute and t is a constant.

Then, given such a SQL query, our goal is to identify an execution strategy that meets the desired efficiency and result quality requirements as closely as possible.
To evaluate a query Q over a text database D, we need to:

(1) Select an extraction system Eij for each base relation Ri, as well as a document retrieval strategy Xi for Eij.
(2) Use strategy Xi to retrieve from database D the set of text documents Pi that Eij will process.
(3) Process the documents in Pi with extraction system Eij to obtain a relation instance ri.
(4) Apply data cleaning techniques to the extracted relations, for record linkage, and eliminate data inconsistencies.
(5) Generate a candidate view v = K _n i=1 ri _, where ri _ is a “clean” version of ri.
(6) Execute Q over v and return the execution results.

Several collective behavior inspired algorithms have been proposed in the financial application area, these algorithms refer to well-studied optimization problems like NP-hard problems (Traveling Salesman Problem, Quadratic Assignment Problem, and Graph problems), network routing, clustering, data mining, job scheduling and so on.

Further to conceptualize this SIT, the (PSO) and Ant Colonies Optimization (ACO) are currently the most popular algorithms in the swarm intelligence domain. Particle Swarm Optimization (PSO) PSO is a population-based search algorithm and is initialized with a population of random solutions, called particles [1]. Unlike in the other evolutionary computation techniques, each particle in PSO is also associated with a velocity. Particles fly through the search space with velocities which are dynamically adjusted according to their historical behaviors. Therefore, the particles have the tendency to fly towards the better and better search area over the course of search process. The PSO was first designed to simulate birds seeking food which is defined as a ‘cornfield vector’ [2], [3], [4], [5], [6].

PSO learns from the scenario and uses it to solve the optimization problems. In PSO, each single solution is like a ‘bird’ in the search space, which is called ‘particle’. All particles have fitness values which are evaluated by the fitness function to be optimized, and have velocities which direct the flying of the particles. (The particles fly through the problem space by following the particles with the best solutions so far). PSO is initialized with a group of random particles (solutions) and then searches for optima by updating each generation. The data mining algorithm in the SIT domain will be reengineered in our future research scope.

Domain Index - A domain index that was used for fast retrieval of unstructured text and CTXCAT indexes work best when text is in “small chunks” may be a couple of lines maximum. Domain index supports most of the document formats and facilitates to 39 languages and efficient for searches within big collection of data.

Domain index is an option in relational database management systems and with no additional cost on licensing. Domain index used to search for large coherent documents and indexing small text fragments and related information to improve the mixed query performance and used to build a document classification application.

The domain index that was used for fast retrieval of unstructured text and this takes care of indexing, searching word and theme, viewing text and uses standards will reap a little. This throws a vision on our research issue, to enable in the entire financial trap. A small change over in the memory parameter boosters will result in big leap and produces larger results. But still it works on the core components of data mining algorithm. The proposed analysis in the existing and proposed task is to be reported, resulting a new algorithm.

Following the path outlined above, focus of the on-going research is to improve, reconcile and balance the skills in technology. Throughout, it must be kept in view that humans come to our systems with a broad repertoire of skills and knowledge. Any design that does not reflect this with a high degree of fidelity may attract a lack of respect, and eventually be discarded.

The inter-media index and execution of a search is built using USER_DATA_STORE. Manual partitioning has two sets of indices which are divided into three parts i.e., 90 days, 270 days and rest. These indices are kept on different columns of the document table. This improves index performance and manageability but on the other hand searches are more complex.

The second index is for security i.e., rebuilding and index consume more man hours. For example indexing of 1 million of records may consume 3 hours of time to complete the task. While indexing is triggered, the transactions are not allowed and the system is in blocked mode.

```
CREATE INDEX indexname ON tablename(col)
INDEXTYPE IS CTXCAT;
```

The following example creates a text index with degree 3:

```
CREATE INDEX myindex ON transactions
(concat_searchinfo) INDEXTYPE IS
ctxsys.ctxcat PARALLEL n;
```

A CTXCAT index is a “domain index”.
It supports the PARAMETERS clause.

A number of possible parameter settings are shared with CONTEXT indexes and they are: LEXER, MEMORY, STORAGE AND WORDLIST. The most important parameter is a new one : INDEXSET. INDEXSET defines the structured columns that are to be included in the CTXCAT index.

The text index structures are stored in the database and the index consists of four tables: referred to as the $I, $K, $N and $R tables respectively.

The $I table consists of all the tokens that have been indexed, together with a binary representation of the documents they occur in, and their positions within the documents. Each document is represented by an internal DOCID value.

The $K table is an index-organized table (IOT) which maps internal DOCID values to external ROWID values.

Each row in the table consists of a single ROWID/DOCID pair. The IOT allows for rapid retrieval of DOCID given in the corresponding ROWID value.The $R table is designed for the opposite lookup the $K table – fetching a ROWID when you know the DOCID value.

The $N table consists a list of deleted DOCID values, which is used and cleaned up by the index optimization process. It uses a paltry 2 megabytes of memory for indexing. If not specified in the index creation process and implementation, the system parameter DEFAULT_INDEX_MEMORY is consulted for the system and it is 12M which leads to infinite time – and there is no specific time is estimated during index creations. The amount of memory cannot exceed system parameter MAX_INDEX_MEMORY. This allows the system to disallow outrageous of index memory. With the limited predefinitions, the text search and filtering criteria would be met. With these memory parameter boosters the indexing of the unstructured text will be entertained at shorter time line.

V. RESULTS AND DISCUSSIONS

The solution in the study is based on the time factor. Even though time reduces by incorporating domain index and fine tuning done in memory parameter boosters, it consumes sufficient amount of time and the relevant output are:

TABLE III
DATA VOULME IN TIME VS MEMORY UTILIZATION

Data Volume in Million	Time in sec	Physical Memory in gb
1	14400	100
2	32400	200
20	43200	300
39	604800	500

During the index creation the system will not allow the transactions for read / write and the amount of time was very huge and some test scenarios it went to infinite time and the test case not in a position to judge that the test progress is completed or not.

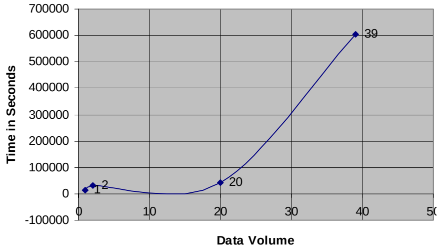

Fig. 5 Data Volume Vs Time in Seconds

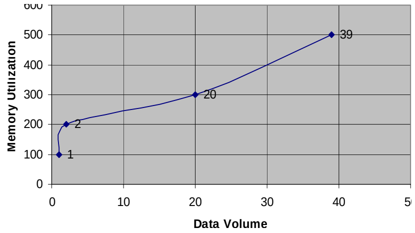

Fig. 6 Data Volume Vs Memory Utilization

Updates and changes are done in memory parameter and space constraints, the time was reduced and the total test case process is completed within a day i.e., 75600 seconds while the previous test case stipulates with unlimited time utilization and cannot be predicted. The progressive and a giant leap in time utilization achieved and the following diagram shows the results.

TABLE IV
DATA VOULME IN TIME VS MEMORY UTILIZATION

Data Volume in Million	Time in sec	Physical Memory in gb
1	3600	25
2	7200	50
20	36000	100
39	75600	220

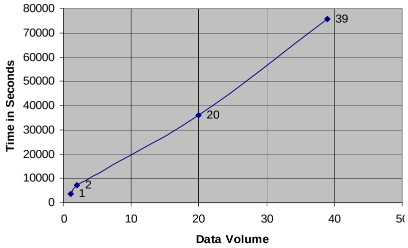

Fig. 7 Data Volume Vs Time in seconds

To overcome this situation a reengineered few data mining algorithms in SIT, SQL Query over unstructured text database and memory parameters boosters in domain index, are proposed and to compete the scenario. Once the data mining algorithms are reengineered and frozen, the Financial Service Industry systems approach will be in different approach and it will a vital role even if regulatory changes are done frequently on country / data wise. This paper aims to review the constraints, and it is possible to eliminate or reduce the device of those constraints and to improve the system efficiency in unstructured text search and filtering classifications at higher level.

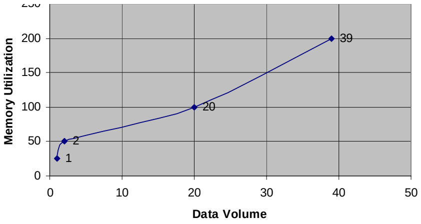

Fig. 8 Data Volume Vs Memory Utilization

ACKNOWLEDGMENT

Thanks to Fellow Co-author and PRIST University for providing me the lab and library facilities to run the data scenario test. Thanks to my wife and kids who made many helpful comments on the manuscript. The research discussed in this paper has been placed against the challenges in the financial community where the financial service industry a deserving study trend of today’s need and in force.

REFERENCE

[1] Hu X, Shi Y, Eberhart RC Recent Advences in Particle Swarm, In Proceedings of Congress on evolutionary Computation (CEC), Portland, Oregon, 90-97, 2004.

[2] Kennedy J, Eberhart RC Particle Swarm Optimization. In Proceedings of IEEE International Conference on Neural Networks, Perth, Australia, IEEE Service Center, Piscataway, NJ, Vol.IV,. 1995.

[3] Kennedy J Minds and cultures: Particle swarm implications. Socially Intelligent Agents. Papers from the 1997 AAAI Fall Symposium. Technical Report FS-97-02, Menlo Park, CA: AAAI Press, 67-72, 1997

[4] Kennedy J The Behavior of Particles, In Proceedings of 7th Annual Conference on Evolutionary Programming, San Diego, USA, 1998.

[5] Kennedy J The Particle Swarm: Social Adaptation of Knowledge. In Proceedings of IEEE International Conference on Evolutionary Computation, Indianapolis, Indiana, IEEE Service Center, Piscataway, NJ, 303-308, 1997.

[6] Kennedy J Thinking is social: Experiments with the adaptive culture model.Journal of Conflict Resolution, 42, 56-76, 1992.

[7] Liu X., P.Z. Zhang, “Research on Constraints in anit-money laundering (AML) business process in China based on theory of constraints”, IEEE conference, Proceedings of the 41th Hawaii International Conference on System Sciences, 2008

[8] Liu Bo, Pan Jiuhui, “Distributed Data Mining Method Based on Swarm Intelligence”. Computer Engineering. 8th ed,vol. 31, pp. 145-147,2005

[9] Wang Liming, Chai Yumei, Huang Houkuan, “Model for distributed data mining based on muti-agent”. Computer Engineering and Application, 9th ed,vol. 40, pp. 197-199,2004.

[10] Xia Hongxue. Shui Junfeng, Zhong Luo, Ma Zhijun, “Design of distributed data mining system based on SOAP”. Journal of Wuhan University of Technology. 1st ed,vol. 25, pp. 188-190, 2003.

[11] Xuan Liu, Pengzhu Zhang, A scan statistics bases suspicious transactions detection model for anti-money laundering (AML) in financial institutions, IEEE, 2010.

[12] Zhang Cheng-hu, Yue Xin, Yue Hui, “Client transaction behavior pattern recognition based on clustering method”. Computer Engineering and Applications, 10th ed, vol. 43, pp. 195-198, 2007.

AUTHORS BIOGRAPHY

First Author: D. Ashok kumar did his Master degree in Mathematics and Computer Applications in 1995 and completed Ph.D., on Intelligent Partitional Clustering Algorithm’s in 2008, from Gandhigram Rural Institute–Deemed University, Gandhigram, Tamilnadu, INDIA. He is currently working as Senior Grade Associate Professor and Head in the Department of Computer Science, Government Arts College, Trichy–620 022, Tamilnadu, INDIA. His research interest includes Pattern Recognition and Data Mining by various soft computing approaches viz., Neural Networks, Genetic Algorithms, Fuzzy Logic, Rough set, etc.,. Corresponding details: Fax:+91–0431-2520805, Phone:+91–0431-2520259(O), +91–9443654052 (CELL), E-mail: akudaiyar@yahoo.com

Second Author: MohanRaja. D, is a Senior Technical System Engineer at Sivaa Software Park a research and development organization. He graduated with a M.Phil., in Computer Science from Madurai Kamaraj University during 2008, a Master’s Degree in Computer Science from Madurai Kamaraj University during 2003. He has used his various database expertise in banking domain where the Anti-Money laundering systems plays a major role in the domain in combating ML, for a multi-national UK based bank. Their major transactions are in Asia Pacific and Middle-East countries. Can be reached at drmriprist@gmail.com.