International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

iC3i-A Environment for Analyzing Customer Behaviour in Banking Sector

Ms. S. Archana [1]
UG Student
Department of Computer Science and Engineering
Manakula Vinayagar Institute of Technology,
Puducherry.

Ms. Y. Dhivyasri [2]
UG Student ,
Department of Computer Science and Engineering
Manakula Vinayagar Institute of Technology,
Puducherry.

Ms. R.Nivetha [3]
UG Student,
Department of Computer Science and Engineering
Manakula Vinayagar Institute of Technology,
Puducherry

Ms. R. Srividya [4]
UG Student,
Department of Computer Science and Engineering
Manakula Vinayagar Institute of Technology,
Puducherry

Mr. P. Anandajayam [5]
Assistant Professor,
Department of Computer Science and Engineering
Manakula Vinayagar Institute of Technology,
Puducherry

Abstract: **The volume of data gathered by bank is increasing speedily and provides moment for banks to conduct predictive analytics and boost its business. However handling large volume of data efficiently and developing insight with real business value which makes data scientists to face large challenges. In this paper, the Intelligent Customer Investigation for Identifying and Inquiring (iC3i) framework is provided to analyze banking customer performance through banking big data through analytical modeling procedure and approach framed for key business scenario. Combining big data processing power and IBM platform with personalized data analytic models, the iC3i solution to satisfy a bank's specific business need and data environment by provides deeper customer insights. In this case, iC3i helps generate insights for active customers based on their transaction behavior, using close to 20 terabytes of data.**

Keywords: iC3i, Data scientist, Analytic modeling, Big Insight.

INTRODUCTION

The period of big data has arrived [1-3]. As discussed in IBM corporation paper [4], big data is being generated by a vast range of devices and processes. For example, numerous digital process and social media exchanges produce data trails. Systems, sensors and mobile devices transmit data. Big data is arriving from multiple sources with an varying velocity, volume, and variety. Every day 2.5 quintillion bytes of data are created and 90% of the data in the world today was produced within the past two years as mentioned in paper [5]. In this big data period the amount of data stored by any bank is fast expanding, and the nature of the data has become more complex. These trends provide a big opportunity for a bank to enhance its businesses. Traditionally, banks have tried to extract information from a sample of its internal data and produced periodic reports to improve future decision making. Nowadays, with the availability of vast amounts of structured and unstructured data from both internal and external sources, there is increased pressure and focus on obtaining an enterprise view of the customer in a systematic way. This further enables a bank to conduct large-scale customer experience analytics and gain deeper insights for customers, channels, and the entire market. Integrating predictive analytics with automatic decision making, a bank can better understand the preference of its customers, identify customers with high spending potential, promote the right products to the right customers, improve customer experience, and drive revenue. However, data scientists are facing big challenges, including how to capture the massive amount of data in a cost-effective and efficient way, and how to sift through the big data to generate valuable business insights that translate into competitive advantages.

In this paper, the Intelligent Customer Analytics for Recognition and Exploration (iC3i) framework is presented as a method to efficiently analyze customer behavior using banking big data. The iC3i analytical models are customized and validated on the processed data according to specified business scenarios, so that they can provide valuable business insights that could not be generated by traditional data mining models. The iC3i models are deployed in a parallel computing manner to achieve high performance and low response time. The solution can be personalized to satisfy a bank's specific business need and data environment. The iC3i framework has been tested in a real case study involving a bank in southeast China. In this case, iC3i helped generate insights about the transaction behavior of active customers to develop data-driven marketing strategies.

The remainder of this paper is organized as follows. In the second section, some fundamental concepts are introduced, and then challenges in the big data era and the iC3i approach to solving the challenges are discussed. In the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

third section, the framework of iC3i is described in detail. Finally, the benefits of the iC3i framework in the fourth section. The conclusion is presented in the last section.

BACKGROUND AND MOTIVATION

TRADITIONAL BANKING CUSTOMER BEHAVIOR ANALYTICS

Today, with online banking and credit card and mobile payment systems, banks have access to a large amount of customer information. Adding to the complexity, the data is also increasingly available as Bsoft information instead of Bhard information. For example, social media data can shed light on brand sentiment and brand loyalty. Hard data is usually recorded as numbers and easy to store and transmit in impersonal ways [6], whereas soft information is mostly communicated in text with the content dependent on the collection process. Thus, it becomes more difficult to process and analyze the soft information. However, it can help build a comprehensive understanding of the customer behaviors that can lead to new and important insights. However, most traditional customer behavior analytic techniques are only focuses on hard information.

As discussed elsewhere[11-14], traditional customer behavior analytics includes four dimensions: customer identification, customer attraction, customer retention, and customer development. Among them, customer identification is most fundamental and widely implemented in the banking industry. Customer identification includes customer segmentation and targeting. After customer identification, companies adopt appropriate marketing strategies to attract specific customers. An important element of the customer behavior analytics is improving customer retention and identifying the cause of customer attrition. The last dimension includes customer lifetime value analysis, up-selling/cross-selling, and affinity analysis for consistent expansion of transaction intensity, transaction value, and individual customer profitability.

Each component of customer behavior analytics mentioned above is linked to some standard data mining techniques. In customer identification, classification and clustering methods are usually used to target a specific customer group based on the business objective. In addition, regression techniques are applied to predict new potential customers. Almost all frequently used data mining techniques can be applied to better understand customer loyalty, including classification, clustering, sequence discovery, association, and regression. Association techniques are often used in customer development in affinity analysis to find the relationship between different products that are bought by a given customer over his or her lifetime.

Numerous solutions have been developed and studies done on traditional customer behavior analytics. For example, a framework for analyzing customer value and segmenting customers based on their value was proposed in S. Y. Kim, T. S. Jung, E. H. Suh, and H. S. Hwang paper. Data on about 2,000 customers was used to train the segmentation model. Transaction history data of 150,000 credit card users were used. However, complexity of the data handled in most of the previous research is limited.

Challenges in the big data Generation:

Compared with usual customer behavior analytics methods most used in banking, there are two major challenges in the big data generation. The first challenge involves how to handle the huge amount of difficult data in a cost-effective and useful way. The availability of data has grown in importance, speed, and capacity. As the number of channels that generate data has increased, so the number of transactions and needed storage for related data. Moreover, valuable but unstructured registered data is also collected from online banking system as a type of soft information. Traditionally, solutions to manage the large amount of data were unable to provide reasonable response times in handling expanding data volumes, leaving few options-either to run the analytics models on the large volume of data over days or perform gradually transactions for a more reasonable response time. Therefore, a bank needs to ensure the real-time response for huge amount of data, which requires new knowledge in the data management and the latest systems management methods. Additionally, new data analytical models are required to capture the value behind the increasing amount of unstructured, soft information.

Effective generation of business values from the analytics obtained and create advantages for banks is involved in the second challenge. Big data has a most critical impact that is how its decisions are made. The people have to make decisions directly based on even unreal or based on experience whenever the data are scarce, very expensive or unavailability in digital format. Large amount of data is produced and transferred in the era of big data. Combination of problem understanding and problem-solving techniques will improve decision making and brings real value of business to a bank. According to McKinsey [18], trillions of dollars and Euros can be generated in value worldwide on the basis of big data. For an example, on an annual basis, 250 billion Euros in European government, $300 billion dollars in American healthcare, and more than 100 billion dollars in global personal location data services can be earned. Two charts in [18] are used to compare the values expected form the big data usage of different sectors and the ease at which big data can be obtained. From the chart we can understand that getting data from the financial sector is quite difficult due to the volume of data and the intensity of transaction. But it has the potential to unlock the new streams of revenue. As a result of these challenges, a solution iC3i has been designed and it is presented to

study the banking behavior of customers by using analytical methods in banking big data that are designed for important business scenarios. From dozens of terabytes of data, a real case iC3i has been established to generate insights on transaction behaviors for active customers.

The iC3i framework design

The architecture diagram of iC3i frame work is given in Figure1. There are four stages in the solution: Data procurement, data arrangement, data designing and various field applications. Using various IBM software platform and data converting power, combine data from various

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

sources, and developing analytical model for various field applications, iC3i can provide deeper banking customer understanding to benefit bank. Each stage is described in following sections.

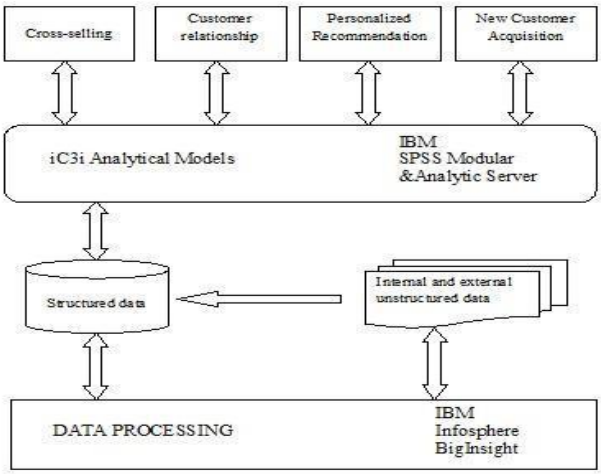

Figure 1
Architecture of iC3i framework

Data procurement:

With the advancement of new banking services, bank database are expanding to modify depending on the field needs. As a result, these databases have become complex. Hence commonly structured data are saved in tables, there may occur more opportunity for increased complication; for example, for the new business upgrade a new table is added to the database or for business upgrade a new database can be replaced by another. Away from internal data source there are many other structured data from external source. For having the consistency and accuracy of data, a standard input format is defined in iC3i framework for the structured data. Before providing input to the big data storage platform all data must be transformed to the standard format defined. This will require clear study about the banking business.

Moreover, apart from structured data the growth of unstructured data creates more complication. Some of the unstructured data can be obtain from inside the bank, including web log files call record and more from the external source. The unstructured data are stored in form of files, not as in form of tables. Many files with hundred terabyte of data can be effectively handled in Big Insight platform, it provide new way of managing vast amount of data.

Data arrangement:

Because of using a standard format for handling data, an additional work must be done to transfer the unstructured data into a regularized structure before designing. IBM SPSS analytic server provides big data analytical power, and also provides support for integrating the unstructured predictive analytics from the Hadoop surrounding. It is also easy in obtaining and querying data stored in BigInsight., its provide easiest way for obtaining data without move data and operating on vast data. By means of the AS tool, the unstructured data can be normalized without means of using any complex coding and scripts.

Also the structured data needs additional work to improve the quality of data on BigInsight by means of big SQL. It is used in handling incomplete, irrelevant data. Additionally some statistical methods are used in Big SQL for reducing the brunt of noise in data. For example, some unwanted data can be detected and eliminated; some features can be normalized and provided with ranks. This help in removing noise from big data analytics.

After all data are prepared and cleansed, integration of data in carried out in BigInsight. Data from multiple sources are integrated and all data are stored in data warehouse. Because of this the different table relationship are known and data clash due to multiple source are resolved. Join operation with billions of data object can be done in BigInsight within a minute, which takes hours in other computing technology.

Based on data warehouse, each customer will be with hundreds of values, and overall view of customer is generated. The iC3i model will be built on the data.

iC3i analytic model:

Based on the business objectives and consolidated data, the iC3i analytic model is build. There are two advantage of using iC3i analytic model. One of the advantages is it provides parallelized model that help in achieving better computational performance. For example to determine model parameter most data mining algorithm scans the training data. For accessing data frequently it needs intensive computing, this is impossible in single-processor computer for large data. The second advantage is method such as statistical and machine learning are customized in such a way to support business scenarios. For example while using customer retention model, design the new interactive decision tree using the domain model to optimize the model. from the model itself some business strategies will be automatically generated to improve the data driven decision making process. To increase the efficiency parallel programming method is used [19]. For this reason all algorithms are designed in such a way to follow Map reduce programming model.

The cost and time spend on model building and model evaluation is low in case of parallel computing. The next section illustrates the example of analytical model build in iC3i and its advantages.

An iC3i analytical model example: A personalized and parallelized K-means clustering

In this section we will introduce a personalized and parallelized K-means clustering algorithm as an example of the iC3i analytic models. The classic k-means clustering [20], which is the technique used to divide n data points into k clusters with related points to minimize the total distance between the points to the cluster centers. The algorithm involves selecting cluster center that is k data points, and then involves associate each data point to cluster center. Update the center of cluster, this is usually carried out by average values of each dimension for all points in the cluster.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

For obtaining initial insight and to reduce complication, it is efficiently and widely used in much application. For example in banking it is used to segment customer into many clusters based on the profile and transaction information, and then the bank can provide its services and marketing message to each cluster. There are several limitations while analyzing large amount of data. The quality of data is low. The cluster is sensitive to error, makes this hard to interpret business value of each cluster if the data is incomplete. Identifying the cluster with closely related data in customer data analytics is valuable. To segment the result which will be more appropriate for banking and to improve the strength of the k-means, a customized algorithm has been developed in iC3i solution with few steps as follow:

1. Select K data points as the cluster centers.

a) Choose first data point as the first cluster center.

b) Compute the minimum distance between a data point and each defined cluster center. The Manhattan distance is used here. For two D-dimensional data points x and y, the metric is defined as

$$D(x,y)=\sum_{i=1}^{D}|x^i - y^j|,$$

where x^i is the coordinate of x in the ith dimension.

c) Next select the new cluster center from the data point with the largest minimum distance from the defined cluster centers.

d) Repeat Steps b) and c) until K cluster centers have been obtained.

2. Using standard k-means algorithm assign each data point to the closest cluster with distance metric shown in formula (1).

3. Amend the cluster centers in the following way: suppose there are J_k data points $x_k,1, \ldots, x_k;J_k$ in the kth cluster where k= 1, . . . , K, and the current cluster center $C_{k,old}$, then the updated cluster center is given as

$$Ck;new= \sum_{j=1}^{J_k} w_{k,j}x_{k,j},$$

where

$$w_{k,j}= 1/ d(x_{k,I},C_{k,old}) / \sum_{j=1}^{J_k} 1/ d(x_{k,j},C_{k,old})$$

The weighted average of all data points is new cluster center, and the associated weight is inversely proportional to the distance between the old cluster node and the point.

4. Reconstruct the data points to their closest cluster and drop any data point xj, such that the point is far away from any cluster center, that is

$$\min_{1\le k\le K} d(x_j,C_k)> \tau_1$$

where τ_1 is a predefined distance threshold.

5. Apply (2) to update cluster centers using the remaining data points in each cluster.

6. Repeat Steps 4 and 5 until

$$\max_{1\le k\le K} d(C_{k,old},C_{k,new})< \tau_2$$

where τ_2 is a predefined tolerance threshold.

This algorithm is further proceed with Mapreduce model: the data points are divided into different groups and in each group the Manhattan distance between data points and cluster center are calculated. The partial weighted sum of data points in each group can be computed in parallel and to obtain new center the reducer will add the partial sums. The Mapper is used to calculate distances between data points and cluster centers in each group, the data points which located far away for the point is dropped. The next cycle also follow the same

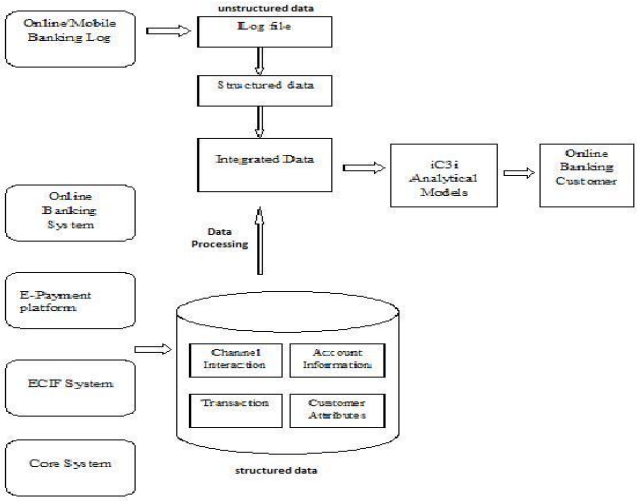

Figure 2
The detailed view of iC3i framework

method for remaining data points. This process continues until merging.

There are big advantages of applying this personalized and parallelized K-means clustering algorithm in banking data analytics. Normally Euclidean distance is used instead here we used Manhattan distance which makes the algorithm more robust because the distance in each axis is not squared which may result in some error. This is used especially in big data analytics where there may be many noise in data. Leaving the data point which is not close enough to any cluster reduces noise of non-target customers for product or service promotion

This leads to beneficial in term of robust and computing time. Lastly using the parallelized algorithm for massive amount of data following map reduce model speed up the implementation of analytics model.

Business applications:

The main aim of customer analytics is to create a more understanding between the customers and their behavior to maximize their longstanding value to the company. Customer analytics can be applied to many applications, like customer marketing, credit scoring and approval, profitable credit card customer identification, high-risk loan applicant identification, payment default prediction,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

fraud detection, etc. The following are five examples that have been built in the current iC3i solution.

1. *Client distribution and inclination analyses:* This module produces fine-grained customer segmentations in which customers share similar preference for different sub-branches or market regions. Based on these results, banks can get deeper insights in their customer characteristics and preferences, so as to improve customer satisfaction and achieve precision marketing by personalizing banking products and services, as well as marketing messages.
2. *Future client classifying:* This module helps banks identify hidden high-revenue or faithful customers who are likely to become profitable to the bank, By this method, banks can get a more complete and accurate target customer for high-value customers, which can improve marketing efficiency and bring huge profits to the banks.
3. *Client chain analysis:* By the idea of the customer and product attraction through study of social media networks, customer network study can improve customer holding, and recall.
4. *Advertise latent analysis:* Using economic, analytical and physical data, this module generates the structural distribution for both present customers and hidden customers. With the commercial capability distribution map, banks can get a clear overview of the target customer's locations, and identify the customer lacking areas for investing, which will support the banks' customer marketing and research.
5. *Medium portion and process breakthrough:* Based on the banks' strategy and structural distribution of client resource, this module optimizes the configuration(i.e., location, type) and operations of service channels (i.e., retail branch or automated teller machine) Maximizing gain, client satisfaction, and reach against costs can improve client retention and attract new customers.

Due to its ability to integrate multiple data sources quickly with parallel computing, iC3i provides a flexible framework that can be applied to many other applications. In the next section, a specific case study and its results are examined.

CONCLUSION

In this paper, the iC3i framework was introduced as a method to handle huge amount of data regularly and analyze client behavior for retail banks. It goes a head the limitation of traditional customer analytics that has been done in the banking sector, using unstructured data that has not been used before. Further, the results can be understood as business rules that help with decision-making that can generate real business value for a bank. Thus the result demonstrates that the iC3i framework provides a useful way to handle complex banking data analytics.

However, this paper just outlines basic work on the iC3i framework; there are many possible delay of the method. It can be extended for other data analytics applications, not limited to customer relationship management or the banking industry. Finally, since the iC3i framework is scalable by adding more parallel analytical models, this lays the foundation for even bigger projects using more data.

REFERENCES:

[1] J. Mervis, BU.S. Science policy: Agencies rally to tackle big data,[Science, vol. 336, no. 6077, p. 22, 2012.
[2] A. Labrinidis and H. Jagadish, BChallenges and opportunities with big data,[in Proc. VLDB Endowment, 2012, vol. 5, no. 12, pp. 2032– 2033.
[3] IBM Corporation, Big Data. [Online]. Available: http://www.ibm. com/big-data/us/en/
[4] IBM Corporation, What is Big Data: Bring Big Data to the Enterprise, 2012. [Online]. Available: http://www-01.ibm.com/software/data/bigdata/
[5] M. Petersen, BInformation: Hard and soft,[Northwestern University, Evanston, IL, USA, Jul. 2004, Working Paper. [Online]. Available: http://www.kellogg.northwestern.edu/faculty/petersen/htm/papers/softhard.pdf.
[6] B. W. Agarwal, S. Ambrose, S. Chomsisengphet, and C. Liu, BThe role of soft information in a dynamic contract setting: Evidence from the home equity credit market,[J. Money Credit Banking, vol. 43, no. 4, pp. 633–655, Jun. 2011.
[7] M. Lin, N. R. Prabhala, and S. Viswanathan, BJudgin borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending,[Manag. Sci., vol. 59, no. 1, pp. 17–35, Jan. 2013.
[8] S. Agarwal, S. Chomsisengphet, C. Liu, and N. S. SoulelesBBenefits of relationship banking: Evidence from consumer credit markets,[Social Sci. Res. Netw., Rochester, NY, USA, May 2009, Working Paper. [Online]. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1647019.
[9] M. Puri, J. Rocholl, and S. Steffen, B On the importance of prior relationships in bank loans to retail customers,[Eur. Central Bank, Frankfurt am Main, Germany, Nov. 2011. [Online]. Available: http://www.ecb.europa.eu/pub/pdf/scpwps/ecbw p1395. pdf.
[10] R. S. Swift, Accelerating Customer Relationships: Using CRM and Relationship Technologies. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.
[11] A. Parvatiyar and J. N. Sheth, BCustomer relationship Management: Emerging practice, process, discipline,[J. Econom. Soc. Res., vol. 3, no. 2, pp. 1–34, Jul. 2001.
[12] A. H. Kracklauer, D. Q. Mills, and D. Seifert, BCustomer management as the origin of collaborative customer relationship management,[Collaborative Customer Relationship ManagementVTaking CRM to the Next Level. Berlin, Germany: Springer-Verlag, 2004, pp. 3–6.
[13] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, BApplication of data mining techniques in customer relationship management: A literature review and classification,[Exp. Syst. Appl., vol. 36, no. 2, pp. 2592–2602, Mar. 2009.
[14] S. Y. Kim, T. S. Jung, E. H. Suh, and H. S. Hwang, BCustomer segmentation and strategy development based on customer lifetime value:
[15] case study,[Exp. Syst. Appl., vol. 31, no. 1, pp. 101–107, Jul. 2006.
[16] N. C. Hsieh and K. C. Chu, BEnhancing consumer behavior analysis by data mining techniques,[Int. J. Inf. Manag. Sci., vol. 20, pp. 39–53, 2009.
[17] U. D. Prasad and S. Madhavi, BPrediction of churn behavior of bank customers using data mining tools,[Business Intell. J., vol. 5, no. 1, pp. 96–101, Jan. 2012.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 03, March-2017

[18] McKinsey Global Inst., Big Data: The Next Frontier for Innovation, Competition, Productivity, New York, NY, USA, May 2011. [Online]. Available: http://www.mckinsey.com/insights/business_tec hnology/big_data_the_next_frontier_for_ innovation.

[19] C. T. Chu, S. K. Kim, Y. A. Lin, Y. Yu, G. R. Bradski, A. Y. Ng, and K. Olukotun, BMapreduce for machine learning on multicore,[in Proc. 20th Annu. Conf. NIPS, 2006, pp. 281–288.

[20] J. B. MacQueen, BSome methods for classification and analysis of multivariate observations,[in Proc. 5th Berkeley Symp. Math. Statist. Probab., 1967, vol. 1, pp. 281–297.

[21] Parallel Processing Systems forBig Data: A Survey: By Yunquan Zhang, Member IEEE, Ting Cao, Member IEEE, Shigang Li, Xinhui Tian, Liang Yuan, Haipeng Jia, and Athanasios V. Vasilakos, Senior Member IEEE