DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5274

Volume 9, No. 1, January-February 2018

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

AN INSIDER:MULTI-CLASSRECORD LINKAGEPROCESSING METHODOLOGY FORCUSTOMER 360º VIEW

Anusuya Kirubakaran
Research Scholar
Dept. of Computer science
Mother Teresa Women's University,India

M.Aramudhan
Associate Professor
Department of Information Technology
PKIET, India

Abstract:In the modern age,financial organizations&economicdevelopments are interconnected; which supports economy in terms of savings, investment, infrastructure, trade, employment, capital market, venture capital, foreign capital, regional development, electronic development,entrepreneurship development, political stability, and control of economy rapid growth.In order to support varieties of business models, financial sectors offer multiple products across different time framessince Inception. As the products increases, deployment of infrastructure varies from time to time due to advancement in technologies.Since some of the products are independent to each other and possess changing customer base with different dimensions are stored in isolated heterogeneous systems.To drive the customer centric organizations, business intelligence teamsare challenged to relate the customers between isolated systemshaving limited common set of data factors for linkages. To keep this issue in mind, we have proposed a versatile record linkage methodology which relates the customers andclassifiesas individual, household, corporate, etc. irrespective of multiple features(personal details, demographic details, etc.) versus multiple classes(E.g. Household, individual) at once to improve the time complexity. To relate multidimensional fuzzy customer data processing with minimal computational complexity, this methodology finds the source and target dataset's relativity using pairwise similarity measure and creates factor table to execute the record linkages usingdeep random forest.

Keywords: Record Linkage, Deep Random Forest, Data integration

1. INTRODUCTION

The business mantra - "Customer Is king" plays a key role on every part of the business in the competitive world as customers have a terrific supremacy in terms of choices available due to globalization on the digital era, influence in social media, ventures focusing on long term returns, etc.Henceforth the organizations started focusing on customer-centric from organization centric approach and to deliver an excellent customers experience for survival; therefore customer touchpoints are Omni presence over the phone , self-services, email, social media, mobile apps, physical customer care centers to ease out the customers support and building confidence over the organization. Still the customer facing sectors like banking, Telecom, ecommerce, etc. are struggling with customer relationship management due to poor quality of data scattered across multiple data sources and different formats in structured and unstructured data. With the help of business intelligence team, firms are trying to 'WOW' the customer experience in terms of personalized gifts, greetings and offers during memorable days, customized products to suit their requirements andquick solution to the issues through multiple channels. Hence the business intelligence teams build the customer 360 degree view by accumulating data from the various product relationship and their transactions by their family members, group of entities by household mapping. Having the 360 degree view of a customer, it would be easier for the organizations for better CRM practices, promotional activities for cross sell and up sell, collections contactablity during defaults , risk profiling and increasing revenues.

In this paper, we are trying to address the issues faced by the business intelligence team in banking and financial sector for grouping the customers as household across the banking products in savings accounts, checking accounts, fixed deposits and investments , credit cards, mortgages, personal loans, forex, etc. Challenges in household mappings are due to different ways of capturing the accounts information for different products, platforms across multiple stages. Few data discrepancy causes are given below:

- Banks would have started with savings and current accounts products during inceptions, over a period addition of new products like deposits, credit cards, mortgages on the advanced infrastructure adds the complexity of linkages.)
- Due to the appetite of business monopoly, acquisition of other players & competitor and merger results in the failure of linkages due to cross platforms and data migration activities.

Overall, the aim is to perform record linkage irrespective of different kinds of infrastructure, missing data, typos and interchange of information in the first name, middle name, last name, DOB, address, mobile numbers, email address, unique identifiers, nominees, social media id's, etc.for improvingtime complexity and accuracy measured in terms of false positives (Non linkages of related customers as household) and false negatives (Linkages of different customers as household) ratio.

2. RELATED WORK

There are substantial body of literature found for record linkages and most of the studiesfocus on blocking scheme using cluster-based blocking, locality sensitive hashing [1] which helps to select the optimal candidate selection for data linkage comparison.The blockingscheme proposed with machine learning algorithms by Matthew & Craig [2], Phan [3] which learns the blocking filter trends rather than manual intervention and reduces the data warehouse processing cost.Even though blocking scheme helps to optimize the computational complexity, data quality loss issue arisesdue to minimal data candidate's selection which is addressed by fault-tolerant duplicate data blocking methods [4] and few researchers focus on privacy preserving for multiple data sourcesintegration [5, 6, 7, 8, 9, and 10]. Since record linkage plays major role in identity management and fraud deduction, the entity extraction [11,12] required from multiple data systems like data warehouse, operational system, social media etc., which mandates probabilistic fuzzy matching dataset algorithms rather than deterministic data comparison algorithms [13,14,15]. Also record linkage studies propose clustering& event based scalable indexing [16, 17, 18] for efficient candidate selection and data set record linkage filtering enabled using hashing mechanism [19] iteratively for multiple data source information integration [20] with missing data [21] as well as for hierarchical data [22]. And few record linkage mechanism proposed by the researchers focus on data integration which helps toclean [23]the data repositories holds huge volume of data [24].

Even tough, substantial study found on the literature, the big data processing complexity for multi-class potential was missing in the previous research work. So, we tried to address the issue with minimal computational complexity.

3. PROBLEM DECOMPOSITION

As we have myriad small, medium& large sized business around the globe increases year-on-year and provides the opportunity for financial services evolution. These evolutionin-built multiple products[25] deployment results in vast amount of data collection for Assets and Liabilities' transactions (E.g.: ATMs, Call Centers, Web-based & mobile sources, Industry data, Trading data, Loan, Mortgage, Regulatory data&Social media). While the rate of data grows rapidly, the quality of data decays over a period of time irrespective of best data management tools and practices. As per the study [26], a bank has 500 million data elements per $ one billion in assets products and those data are stored in isolated systems having poor data quality in terms of correctness, completeness, consistency and heterogeneous data format variations. As per theUS Postal Service estimation, 40% of the data keyed by users are either incorrect [27].To enable revenue growth, operational efficiency, risk management and customer satisfaction, the data need to be viewed in different perspective as per the linkages given below

- Unique Customer Identification
- Household – Link the related customers(E.g.: Family, Corporate)
- Campaigns - "Right Message to Right Customers"

To perform the effective and accurate above said record linkages, unique customer identification system [28] requirement arises, which is being achieved through deterministic and probabilistic quasi identifier (E.g.: first name, last name, gender, birth date, address, pin code, email address, phone number etc.,) data matching system. And the customer data single view identification system should be capable of handling [28]data quality issues, customer's Householding linkages, Flexibility of changing data with multidimensional 360 degree view.

In this paper, we propose a versatile record linkage generic methodology for customer data integration irrespective of individual or corporate customer using conglomerate of logically related quasi elements,similarity measure factor table, deep random forest record linkage prediction tree, relevant customer relativity with record linkage type classification.

4. METHODOLOGY

A. ***Domain& Linkage Nature Dependent Quasi Identifiers:***

In financial sectors, data spread across multiple data storage systems for each and every product. Even though the data has been isolated, all the system will have common set attributes which is preferredto perform record linkages. These lists of attributes are combined to link the customer information, called as quasi identifiers [29]. In this work, we have three different data linkage scenarioas given in problem definition and the set of quasi identifiers differs for each scenario.

I. Unique Customer Identification:

A single customer can deal with multiple products like savings account, current account, mortgage etc., and the data disjunction occurs due to different time period product formation, the data standard variation, data migration, bank entities incorporation and other external factors. To identify the unique customer and link the products, the list attributes areName,

Date of Birth, Gender, Address, Contact Number, Unique Identification Number (Passport, Driving License, and Social security Number), Account number and few other domain specific attributes.

II. Householding:

A group of connected customer is linked together for all product portfolios to deliver better campaign & risk management. This householding apply for a family those who banks together, and the customer who belongs corporate entity. To link the multiple accounts, the list attributes majorly used are Nominee's Name, Date of Birth, Gender, Address, Contact Number, Unique Identification Number and master account number.

III. Campaigns:

As "Right Message to Right Customers" is important for sending out the promotional offers to customers are crucial, because it would leads to economic and reputational damage of an organization. So individual and house hold email id and contact number should be captured with high accuracy.Same way for promotional activity the quasi identifiers differs. And the quasi identifier

selection varies based on business nature and nature of record linkage process.

B. Factor Based Conglomerate Of Logically-Related Multifarious Quasi Elements:

As stated earlier, record linkage process depends oncustomer's primary and demographic details captured on various stages indifferent products by business across heterogeneous systems.And each and every data source system designedwith superfluousdata elements having diversified data formats& standards. Even though the data spread across multiple systems, each data source has common data elements for relativity. Using the common elements, themultidimensional customer data integration can be implemented at once. To do so, conglomerate the logically related-multifarious common data elements (LRDE)into factor set as samples shown in Table 1 which should be capable of providing alternative feature to link the customer. Based on the domain (E.g.: Bank, E-commerce) R number of factors set are defined for record linkage usage. Among the list of factors derived, the occurrence of each or combinatorial factor decides the customer identity, householding, campaign details and other business data linkages. The factor set can have duplicate elements (E.g.: FS1, FS2, FS3) between R number factors.

Table 1. Sample Factor Set

Factor Set Name	Element List
FS 1	First Name, Gender, DOB
FS 2	Middle Name, Gender, DOB
FS 3	Last Name, Gender, DOB
FS 4	Email
FS 5	Address
FS 6	SSN
FS 7	Driving License Number
FS 8	Nominee First Name, Gender, DOB
FS 9	Nominee Middle Name, Gender, DOB
FS 10	Nominee Last Name, Gender, DOB
FS 11	Mobile Number
FS 12	Home Contact Number
FS 13	Latest Call Center Number
.. FS R	Element r_1 r_2 r_3 … r_n

Table 2. Sample for Factor Decision

INPUT: FS1			OUTPUT
First Name (FN)	Gender (G)	DOB (G)	FN Λ G ΛG
Unmatch	Unmatch	Unmatch	0
Unmatch	Match	Match	0
Match	Unmatch	Unmatch	0
Match	Match	Match	1

C. Factor's Boolean Matrix:

After grouping the data elements, source and target data element's pairwise similarity measure and the element level result collaboratively derives the decision factor match as true or false to relate the customer record as given in Table.3. Based on the data quality and complexity, similarity measure algorithms are preferred as distance based or token based or phonetic based or hybrid algorithms. To implement this step, select the source dataset S and target dataset T of size Sfrom the nearest neighbor records.A data source S is a collection of records X={r_1, r2…r_n} where each records has finite number attribute or elements i.e. r= {a_1, a_2….a_k}. Same way target data source T is a collection of record Y= {r_1, r_2…r_m} where each records has finite number attribute r= {a1, a2….ax). For each and everyelementlogically grouped to factors, find the pairwise similarity [30]matchand return the result as 1 for match and 0 for un-matched pair.

Table 3. Factor Boolean matrix

Factor	Elements	Source	Target	Similarity Measure	Factor Decision
FS1					
	First Name	BARANI	PRAVESH	False (Un-Match)	0
	DOB	24/12/2000	24/12/2000	True(Match)	
	Gender	M	M	True(Match)	
FS2					
	Last Name	KIRUBA	KIRUBAKARAN	True(Match)	1
	DOB	24/12/2000	24/12/2000	True(Match)	
	Gender	M	M	True(Match)	

After element level pairwise similarity measure between S & T records set, factorBoolean matrixFfor the predefined factors needs to be generated as given in section B. The each factor values for this table are derived from its predefined element's pairwisesimilarity measure results. When the factor defined with single element then the direct pairwise similarity measure result will be taken. On the other hand, when the factor has more than one element, the elements results works as AND gate. In precise, if each andevery element belongsto each factor has similarity measure as "Match" then the factor decision becomes 1 otherwise factor decision becomes 0 as samples shown in Table.2.The concept here is, each factor has mandatory elements and results needs to be true for all elements. When any deviation occurs, factor values will leads to false positive and false negative record linkages. At end of this step, the factor table F is created with S number of records and each record will have R number of factors. And each factor have (0 or 1)values which is the outcome of factor decision derived using similarity measure.

D. Record Linkage-Deep Random Forest:

In the data driven world, data grows tremendous with wide diversity due the usage of hand held devices and electronically driven businesses. And in traditional record linkage methods, the customer data integration (Unique customer, householding) executed separately foreach element in large scale data becomesa time-consuming process and involves redundant data processing. In order to overcome this process, the proposed methodology uses the deep random forest decision tree machine learning algorithm toimplement severalcategories of record linkages at once for the factor matrix derived above, which has multiple features & multiple classes. The selected deep random forest or random decision forest algorithm generates subtree predictors for factor matrix record classifications with optimal combination of features (factors) and provides high accurate data classification. The prediction trees are generated as given below.

E. Training dataset preparation:

To classify the record linkage type, prepare the training dataset D which has 1,2,3…C number of classes for record linkage classification with N Number of record instance and each record instance has 1,2,3….F number of features as same as shown in Table.4

Table 4.Sample Training Dataset

Class 1- Individual Customer Class 2- Household Class 3- Individual Customer & Campaign Class 4- Household & Campaign									
Example Number	Factor 1 (First Name, Gender, DoB)	Factor 2 (Middle Name, Gender, DoB)	Factor 3 (Last Name, Gender, DoB)	Factor 4 (Email)	Factor 5 (Address)	Factor 6 Nominee	……	Factor R	Class
1	1	0	0	0	0	0		0	Class 1
2	0	1	0	0	0	0		0	Class 1

3	0	0	1	0	0	0		0	Class 1
4	0	0	0	1	0	0		0	Class 4
5	0	0	0	0	1	0		1	Class 2
6	0	0	0	0	1	1		0	Class 2
7	1	1	1	1	0	0		0	Class 3
8	1	1	1	1	1	0		0	Class 3
9	0	0	0	0	0	1		1	Class 1
10	1	1	1	0	0	0		0	Class 1
11	1	1	0	0	0	0		1	Class 1
12	0	0	0	1	1	1		1	Class 4
14	1	0	1	0	0	0		0	Class 1
15	0	1	1	0	0	0		1	Class 1
..									

F. Training Deep Random Forest:

Using the training data D, blend of random trees[31]created with multiple layers K (1) , . . . , K (n) of classifiers, where each layer K (`) consists of a forest. The output of each individual tree is a vector of class probabilities, as determined by the distribution of classes present in the leaf node into which the sample is sorted. Specifically, given any decision tree, each leaf of the tree is assigned a vector of class probabilities, p = (p_1, . . . ,p_r), corresponding to the proportion of training data assigned by the tree to the leaf in each class. This is done for all of the training data, hence transforming the data to be of dimension R × m`, where K is the number of classes for the training dataset and m` is the number of trees in the current layer. The outputs of each layer become the inputs to the next, until the data have been mapped through the final layer K (n). The final class prediction is made by averaging all the class probability output vectors from the mn decision trees in K (n), and predicting the class with the highest probability.

G. *360º View Classification:*

To classify the records belongs to multi-class for 360 º view with minimal computational complexity, the records belongs tofactor table F will be searched indeep random forest's mn decision trees ofK(n) layers. Since the deep random forest's decision trees trained to classify the records as Class 1 (Individual Customer), Class 2 (Household), Class 3(Individual Customer & Campaign), and Class 4 (Household& Campaign), The factor table vs deep random forest search returns classification value which helps to perform the data linkages on isolated data source.Overall, the detailedalgorithm for the above process is given below.

Algorithm: Record Linkage

```
Input
-Source dataset S
-Target dataset T
-Factor set F
-Record Linkage Training dataset D
-Factor table matrix M

    For each record x ∈S
        For each record y∈ T
            For each element i∈x,      j∈ y
                Find similarity measure (i,j)
            End
            For all factor f∈ F
                For all element in e∈ f
                    M (f) = true if ∀ e = true
                End
            End
        End
End

    Create Deep Random ForestpredictorR (D)
```

```
For each record r∈M
        Link_result=Search (r, R)
        Link record(Link_result)
End
```

5. RESULTS

The time complexity has been evaluated for varieties of classes for varying number of source and target records and the results are shown in Fig.1. For the proposed approach, the time complexity varies minimally based on the total number of classes to be linked. On the other hand, the existing approach process the source and data as many times directly proportional to the total number classes. The processing methodology increases the computational complexity as shown in Fig.2.

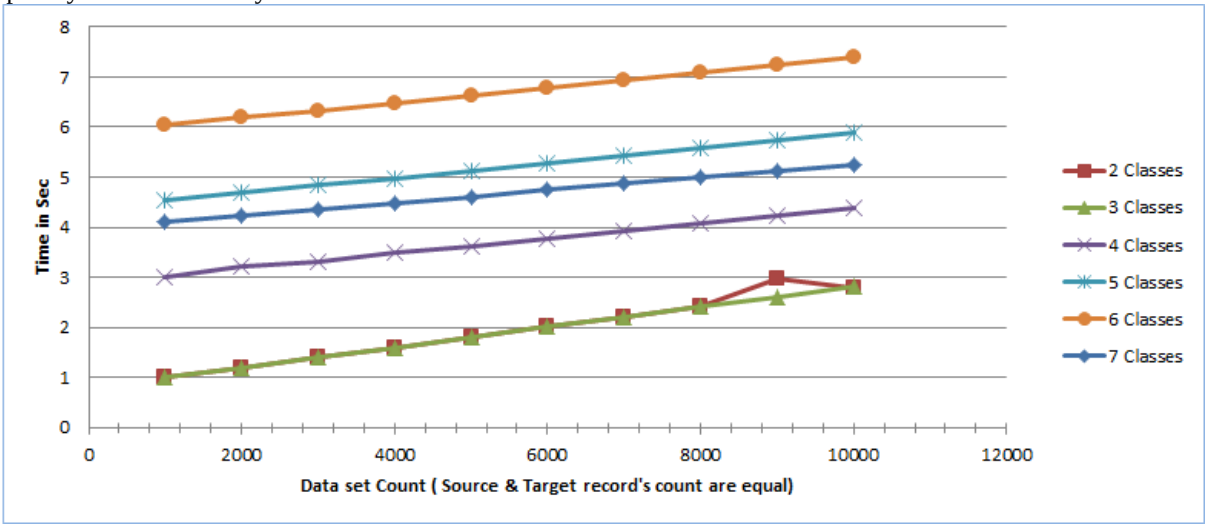

Fig. 1Time complexity

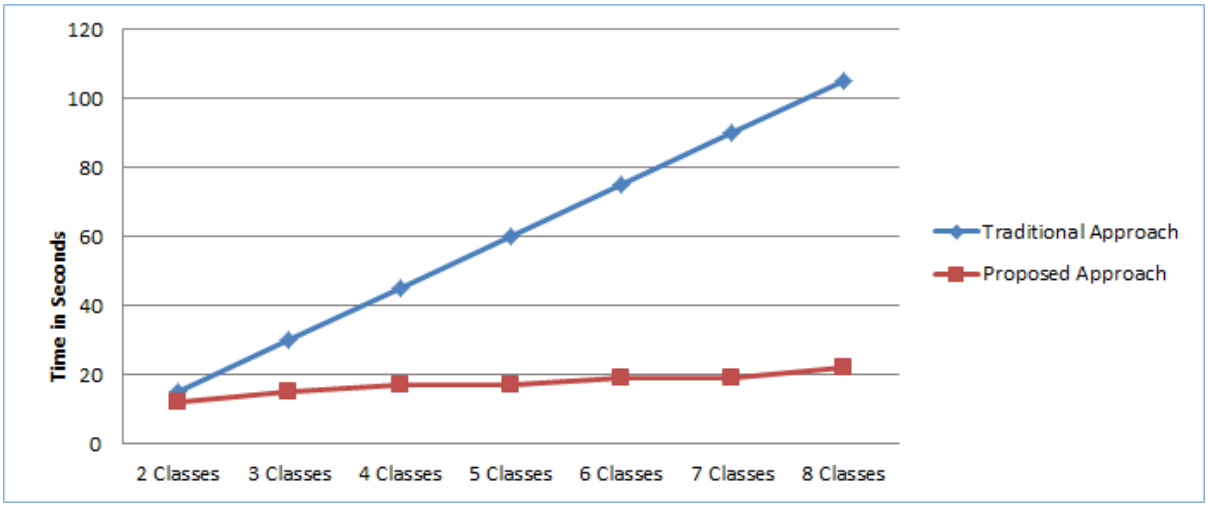

Fig. 2Computational complexity

6. CONCLUSION

In recent days, customer facing organizations are fighting for its survival and retention of the customers, hence trying to please the customers with multiple attractive products & offers, leveraging the accessing medium to the latest mode of communication (like mobile, internet, etc.).. As discussed earlier, even though data spread across multiple isolatedlocations, 360 degree data view becomes basic necessity of the customers and organizations. Failure of the same leads to customer dissatisfaction and impacts the goodwill, reputation of the business. In order to support this process of customer identifications, in this paper we proposed theversatile learning record linkage methodology which identifies relevant customer record and classifies the record as individual, household, corporate and so on. This proposed approach can be extended to any data intensive domains like ecommerce, insurance, Income tax, National security agencies, health care, etc.

REFERENCES

1. Rebecca C. Steorts, Samuel L. Ventura, Mauricio Sadinle, and Stephen E. Fienberg.A Comparison of Blocking Methods for

Record Linkage. Springer International Publishing Switzerland 2014.
2. Matthew Michelson and Craig A. Knoblock. Learning Blocking Schemes for Record Linkage. American Association for Artificial Intelli- gence, 2006.
3. Phan H. Giang. A machine learning approach to create blocking criteria for record linkage. Springer Science &Business Media New York 2014
4. Alexandros Karakasidis , Georgia Koloniari , Vassilios S. Verykios. Scalable Blocking for Privacy Preserving Record Linkage. ACM, 2015.
5. DinushaVatsalan, Peter Christen, Vassilios S. Verykios. A taxonomy of privacy-preserving record linkage techniques. Elevier 2012.
6. DinushaVatsalan and Peter Christen. Scalable Privacy-Preserving Record Linkage for Multiple Databases. ACM, 2014.
7. Mohamed Yakout, Mikhail J. Atallah, Ahmed Elmagarmid. Efficient Private Record Linkage. IEEE International Conference on Data Engineering, 2009.
8. Alexandros Karakasidis and Vassilios S. Verykios. A Simulator for Privacy Preserving Record Linkage: Engineering application of neural networks. Communications in computer and Information science Springer 2013. Page 164.
9. Chris Skinner.Assessing Disclosure Risk for Record Linkage: Privacy in Statistical Databases. UNESCO Chair in Data Privacy International Conference, PSD 2008 Istanbul, Turkey, September 24-26, 2008 Proceedings. Page 166-172.
10. Rahul Shukla and Le Gruenwald. Research Issues in Privacy-Preserving Record Linkage. IEEE, 2010.Mehmet Kuzu, Murat Kantarcioglu, Elizabeth Durham and Bradley Malin. A Constraint Satisfaction Cryptanalysis of Bloom Filters in Private Record Linkage.
11. Flavio Villanustre. Large-scale Entity Extraction and Probabilistic Record Linkage. IEEE 2014.
12. KedarBellare, Suresh Iyengar, Aditya Parameswaran,VibhorRastogi. Active Sampling for Entity Matching. ACM.
13. ArnMigowski. Accuracy of probabilistic record linkage in the assessment of high-complexity cardiology procedures.scielo,2010.
14. Josep Domingo-Ferrer, Vicen, Torra. Distance-based and probabilistic record linkage for re-identification of records with categorical variables.
15. Adrian Sayers, Yoav Ben-Shlomo, Ashley W. Blom and Fiona Steele. Probabilistic record linkage. International Journal of Epidemiology, 2015, 1–11.
16. ThilinaRanbaduge, DinushaVatsalan, and Peter Christen. Clustering-Based Scalable Indexing for Multi-party Privacy-Preserving Record Linkage. Springer International Publishing Switzerland, 2015.
17. Abdulla,Mamun, Robert Aseltine, SanguthevarRajasekaran. Efficient Record Linkage Algorithms Using Complete Linkage Clustering. Plos one, 2016.
18. Timo Reuter, Philipp Cimiano, Lucas Drumond, KrisztianBuza, Lars Schmidt-Thieme. Scalable Event-Based Clustering of Social Media Via Record Linkage Techniques. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 2011.
19. Hung-sik Kim, Dongwon Lee. HARRA: Fast Iterative Hashed Record Linkage for Large-Scale Data Collections. ACM, 2010.
20. Akiko Aizawa, KeizoOyama. A Fast Linkage Detection Scheme for Multi-Source Information Integration
21. Toan C. Ong, Michael V. Mannino, Lisa M. Schilling, Michael G. Kahn. Improving record linkage performance in the presence of missing linkage data. Elevier 2014.
22. Steven N. Minton, Claude Nanjo Steven, N. Minton and Claude Nanjo. A Heterogeneous Field Matching Method for Record Linkage.
23. Indraji Bhattacharya, List Getoor. Iterative record linkage for cleaning and integration. ACM 2004.
24. Anja Gruenheid, Xin Luna Dong, Divesh Srivastava. Incremental Record Linkage. 40th International Conference on Very Large Data Bases, September 1st - 5th 2014, Hangzhou, China. Proceedings of the VLDB Endowment, Vol. 7, No. 9.
25. OracleEnterprise Architecture White Paper. Big Data in Financial Services and Banking: Architect's Guide and Reference Architecture Introduction. FEBRUARY 2015.
26. EcoSystems. Data Quality and Integration In Banking. 2011.
27. Aureus Systems. Customer 360 Degee Banking. 2013.
28. Trillium Software. Create A Single Customer View:Solution Guide.
29. Quasi Identifier. https://en.wikipedia.org/wiki/Quasi-identifier. Date Accessed:12/09/16.
30. Similarity measure. https://en.wikipedia.org/wiki/Similarity_measure. Date accessed 9/29/16
31. Miller, K., Hettinger, C., Humpherys, J., Jarvis, T., &Kartchner, D. (2017). Forward Thinking: Building Deep Random Forests. arXiv preprint arXiv:1705.07366.

