Automated Data Collection for Credit Score Calculation based on Financial Transactions and Social Media

Jay Lohokare[1], Reshul Dani[2*], Sumedh Sontakke[3]*
[*]Department of Computer and IT, [^]Department of Electrical Engineering
College of Engineering, Pune
[1]lohokarejs13.comp@coep.ac.in, [2]reshulsd13.comp@coep.ac.in, [3]sontakkesa15.elec@coep.ac.in

***Abstract*— Financial credibility of a person is a cardinal factor in the approval of loans or permitting credit transactions. Today, this credibility is based on the 'credit score' of the person which is calculated from the person's past performance on debt obligations. This paper provides a unique alternative solution to collect this data. Harnessing the fact that almost everyone has smartphones today, there can be a smartphone application that collects all such data and submits it to the official body. Financial transactions are not the only parameters that can determine credibility of a person. This paper proposes accessing social media data to get insights into general social status of a person. Today, all transactions are conveyed by banks and other institutes to the users via SMS. Hence, having access to SMS will result to getting data related to all such transactions. The proposed solution will have smartphone application that will capture bank transaction data and data related to online purchases through SMS. Use of artificial neural networks will enable calculating the final credibility score based on the various data parameters collected. The primary contribution of this paper is that it provides a system to automatically collect all data required to calculate the credit score. What is unique in this approach is the involvement of data from social media. This approach is better than the existing solutions as it will collect data other than just transactional data thus enabling calculation of a more effective credit score.**

***Keywords*— Credit Score; bank transactions; loan approval; CIBIL; SMS; social media; R**

I. INTRODUCTION

Today, credit score of a person is a factor considered essential by all financial institutes for all kinds of approvals. This score tells the banks whether or not a person taking the loan will be able to repay it within the decided time. Higher the credit score, better the loan and credit prospects for the person. A credit score generally ranges from 300 to 950 [2]. Loan providers will prefer credit scores greater than 750. Traditionally, the credit score is calculated only on basis of the data available with banks. The body calculating the credit score takes this data from the banks on monthly or yearly basis. This data is usually strictly related to financial transactions. Hence, the score generated is based just on the transaction and loan history of a person with banks. This though accurate to some extent, lacks consideration of non-financial factors that have a strong influence on the financial credibility of a person. This paper gives a solution for automatic collection of data which will be obtained from every user instead of banks. Thus, the credit score calculating body will not have to rely on various banks to provide the information. The solution proposed also describes new sources of data that contribute to the credit score. The paper will describe the technology stack and the benefits of including social media based results for credit score calculation. In India, any loan or credit card approval needs a high CIBIL TransUnion Score. This score is calculated by CIBIL based on data it collects from banks. The solution given in this paper tries to create an autonomous system that will collect, analyse the data and then calculate the credit score by itself.

Fig. 1 shows the process of collection of data proposed by us. This process is based on collecting data from smartphones of all users through an application.

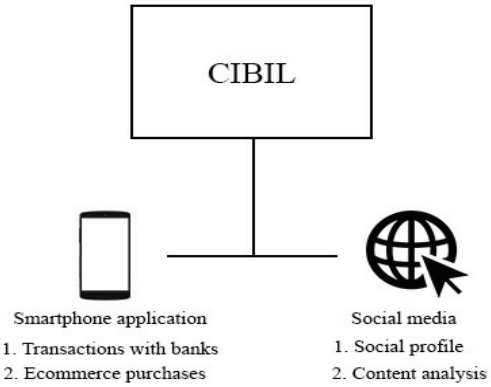


Fig. 1. Proposed method of collecting data

The basic concept is that every user gets SMS from banks for all sorts of transactions. Same is true for all purchases made on any online stores. These SMS contain data like amount for transaction, account number and corresponding account balance. The solution also collects data from social media platforms and considers it as an important parameter contributing to the credit score. This data when collected will automatically generate entire summary of a user's financial transactions.

The source of all transaction data for a particular user will be the same, irrespective of the number of bank accounts the user has. This will remove the unnecessary trouble of getting

978-1-5090-3404-8/17/$31.00 ©2017 IEEE

data from multiple banks. The solution facilitates availability of all data of a user from single source.

Thus, the paper presents a solution to fetch data needed to calculate the credit score automatically, and to make the credit score more accurate by including other factors which contribute to the score.

II. Existing Solutions

Today, the bodies like CIBIL depend on banks as source of data. The banks have to send the transaction data of every user to CIBIL from time to time (with certain frequency). Currently, the frequency of data collection by CIBIL is around 1 month [1]. Banks and financial institutes partner with CIBIL and send this data to one central portal. As the banks have data related to the financial transactions, credit card usage and loans, the data collected is purely based on how the person has managed his transactions with the financial institutions.

Fig. 2 shows how CIBIL collects data of a user from various financial institutions. The user makes transactions with these financial institutes. The transaction data is saved by the institutes and later shared with CIBIL.

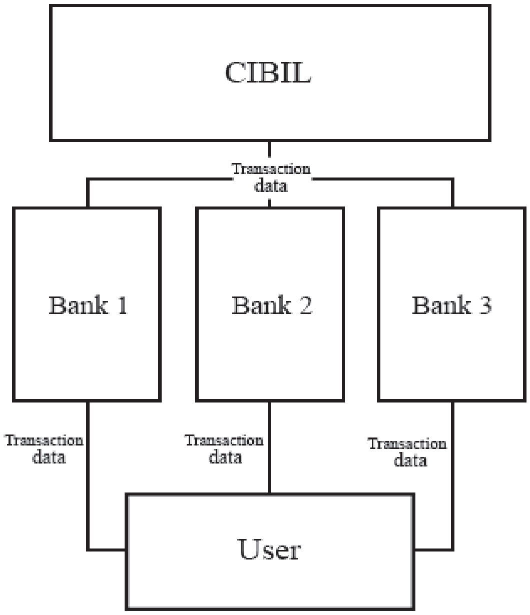


Fig. 2. CIBIL getting transactions data of a user from 3 banks

The process of getting data from financial institutes is tedious as it involves collecting data from multiple sources for every user. CIBIL has to search for data related to a user in entire pool of data it gets from various financial institutions. This collected data lacks any level of understanding a person's non-financial background. Those people who have no access to formal channels of credit are financially excluded due to a low or nil credit score. People from low and middle income countries are not on file in world public credit registries. The non-financial parameters which are actually very important don't reflect in the credit score of a person. This method hence needs modifications and improvements.

Currently, there is no data being fetched from social media platforms by CIBIL. There is no attempt made to understanding the background of the person. Factors like educational status, credit scores of relatives, job details, social status can highly affect a person's financial credibility. These factors should be considered while calculating the credit score

III. Need for automated system to calculate credit score based on transcational data and social media content

The current systems need CIBIL to depend on the partner financial institutions and banks to obtain the transactional data of a person. This results to a system with multiple sources of data, making it arduous to map this data to individual users.

The system proposed in this paper makes data collection easy as the source of all data related to a user is now one single application. There will be no need to depend on any institutions or banks to get access to this data. CIBIL will automatically get all important data related to a user from the smartphone application as seen from Fig. 3.

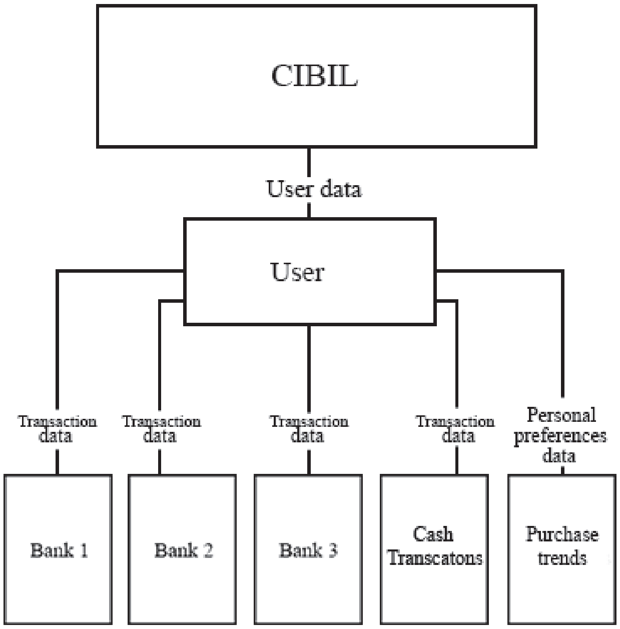


Fig. 3. Collection of user data from smartphone applications

Due to no access to data other than transactions data, CIBIL is currently unable to consider social factors while calculating the credit score. For example, for a person 'A' who is just out of college and has just started earning money, there won't be high amount transactions in his account. Using the traditional systems, this will result to a low CIBIL credit score. Now, if the system proposed by this paper is used, data obtained from social media will be considered in calculating the credit score. Suppose 'A' has a parents or spouse with high CIBIL credit score, this should ideally drastically improve the credibility of

'A'. This factor will be taken into consideration by the proposed system there by resulting to a higher score.

The proposed solution can be divided into 3 major parts. The first part consists of smartphone application for mining data related to transactions through SMS. The second part consists of server mining data from social media using R language [3]. The third part consists of ANN (Artificial neural networks) to calculate the final credit score.

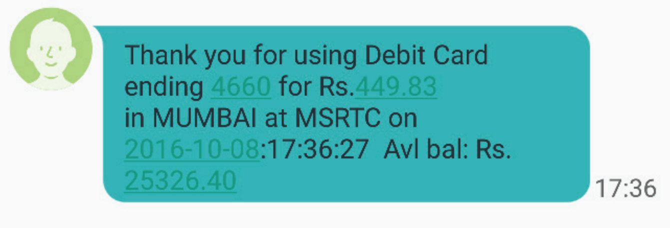
IV. Collection of Data Through Smartphones

Most of the bank accounts today have a mobile number linked to them. An SMS is sent to the user's mobile whenever a transaction made with the bank. The solution implemented in this paper harnesses this data available in SMS to capture every transaction.

Fig. 4 shows a typical message received when a user does some transaction related to a bank account. This data contains amount credited/debited, bank name, account number, date, available balance and purpose of the transaction (MSRTC in case of Fig. 4). Similar SMSs are received from banks for transactions, loans details and credit details.

These messages can be accessed by smartphones programmatically. For Android, SMS Broadcast listeners can be used to access SMS content. SMS content once received will be mapped to pre-set templates of data to find the necessary parameters. For example, in Fig. 1, the template that will be mapped on the SMS string will be: "Thank you for using 'A' card ending in 'B' for Rs. 'C' in 'D' at 'E' on 'F' Avl bal: Rs. 'G' ".

This mapping will be done for all SMSs received from HDFC bank. The listener will know which bank the SMS is from by checking the SMS sender information and comparing it with a list of such verified senders.



Thank you for using Debit Card ending 4660 for Rs.449.83 in MUMBAI at MSRTC on 2016-10-08:17:36:27 Avl bal: Rs. 25326.40

17:36

Fig. 4. Typical SMS received on smartphone for transactions with a bank (SMS from HDFC bank)

The data captured by such a mapping will be:

Amount = 449.83

Card type = Debit

Account number ending = 4460

Timestamp = 2016-10-08

Balance = 25326.40

Vendor = MSRTC

Bank = HDFC

SMS can also be a source of data related to bills, policies, investments and other such forms of transactions which may not be captured by banks. Timely payment of bills, payment of income tax, and investment in insurance policies are few of the factors that reflect the financial credibility of a person and hence should be considered in credit score calculation. The proposed solution has access to SMS and hence can extract the mentioned data thereby letting CIBIL access such parameters.

Thus, such SMSs will provide all transaction data required by CIBIL to calculate the credit score. At the same time, CIBIL will also know what the transaction was for, what the person purchased, where the user spends more, thus helping to understand the person in a better way. Thus, all that will need to be done is for users to install a single smartphone application which sends all this data. The smartphone application will be a zero User Interface application. It will maintain the extracted data locally in a SQLite database. The app will periodically send the new data to servers whenever internet is made available. This is shown in Fig. 5. The data transfer to servers can be implemented using REST APIs or light weight protocols like Message Queue Telemetry Transport (MQTT).

The solution implemented uses PHP based REST API to send the data from SMS to servers. An android application captures transactions and sends them to the database. The database used is MongoDB [4]. There are three collections per user in MongoDB. The first collection stores all raw data from SMS and social network platforms. Second collection stores daily and monthly summary. Third collection stores the overall summary obtained from the first two collections as well as the final credit score calculated.

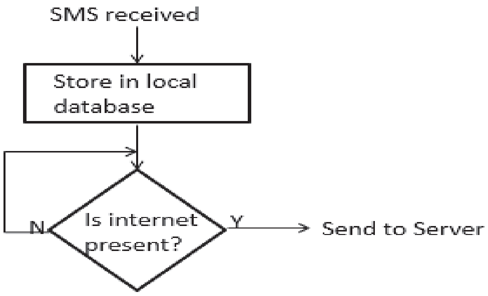


Fig. 5. Sending data from the smarphone application to the server

V. Fetching Data From Social Media

Social networking platforms provide important insights into nature and status of a person. Many essential trends that can't be captured from transactions can be found through social media. This paper proposes capture of the following data from social networking platforms:

1. Educational details
2. Family members
3. Followers count
4. Sentiment of content
5. Locations checked in at
6. Professional background

These parameters affect to a great scale the financial credibility of a person. Higher educational qualifications indicate better job prospects and thus better income. Recently completed education can foretell a possible change (increase) in credibility. Family members and their credit score highly influence a person's financial status. A person with spouse having high credit score will tend to be better able to repay loans. The number of followers on social media depicts the social status of a person, trust and fame in the society. Professional background of a person – the jobs, job profile, responsibilities also give an idea of the financial credibility. Sentiment analysis of social media content can shed light on general nature of a person.

The capture of such data from social media can be facilitated by using the APIs of the respective platforms. Twitter and Facebook provide APIs to access public data of a user. These APIs can be called from R Studio to fetch, extract data and store results in databases. Accessing the data involves following simple steps:

1. Register for developers program of the platform. For example, twitter developers can register twitter developers site
2. Enable API access and generate a access key
3. Call the APIs from R using this key

The APIs allow querying data with multiple filters and parameters. In the solution implemented, data needs to be fetched by user name. The data obtained in R studio is in form of JSONs and can then be then extracted into databases.

In the solution implemented, database gets all transaction data from Android application through REST APIs. This data is stored in a MongoDB collection. An R script fetches data from social media and stores it in the same collection. Another R script creates monthly summary of data from this collection and stores in a 'monthly summary' collection. Credit score calculation takes as input data from 'monthly summary' collection and store the result in a summary collection. The whole architecture is shown in Fig. 6.

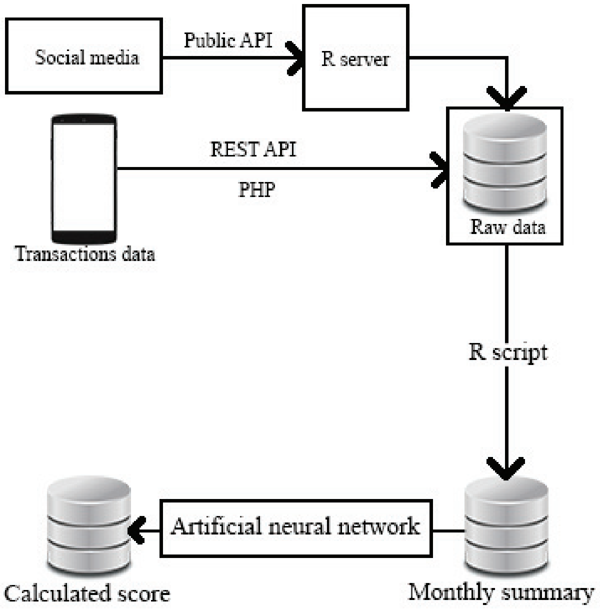


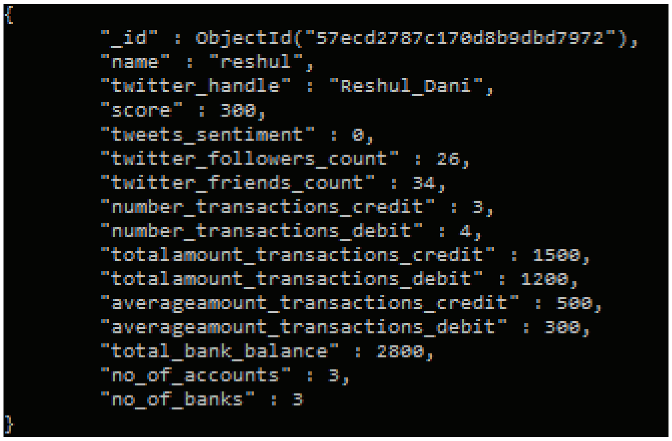
Fig. 6. Architecture of proposed solution

VI. Calculation of Credit Score

The data collected will thoroughly represent the financial status of a person. There are many methods to calculate the credit score from the data [5]. In the most basic way, the credit score is calculated by assigning weights to the various parameters [6].Credit scoring is a data mining classification problem. One classification technique that out performs the rest is Artificial Neural networks [7]. The ultimate aim is to solve a regression problem that outputs a credit score after it is trained with past ground truths. For now the credit score has been calculated by assigning weights out of 100% to the parameters collected. For example, twitter sentiment will have lesser weight compared to credit score of your family.

VII. Results

The proposed architecture was successfully implemented. The transactions data obtained through SMS on Android were sent it to MongoDB using PHP based REST API. Social networking platforms like twitter were crawled to obtain the proposed parameters and data. Fig.7 shows the document from the summary collection. This collection contains overview of data of a user along with calculated score.



```
{
        "_id" : ObjectId("57ecd2787c170d8b9dbd7972"),
        "name" : "reshul",
        "twitter_handle" : "Reshul_Dani",
        "score" : 300,
        "tweets_sentiment" : 0,
        "twitter_followers_count" : 26,
        "twitter_friends_count" : 34,
        "number_transactions_credit" : 3,
        "number_transactions_debit" : 4,
        "totalamount_transactions_credit" : 1500,
        "totalamount_transactions_debit" : 1200,
        "averageamount_transactions_credit" : 500,
        "averageamount_transactions_debit" : 300,
        "total_bank_balance" : 2800,
        "no_of_accounts" : 3,
        "no_of_banks" : 3
}
```

Fig. 7. Summary collection in mongodb for a particular user

VIII. CONCLUSION

The aim of this solution was to provide an alternative and unique way to collect the data required for calculating the credit score. It has been shown how data from social media which is a crucial factor in determination of credit score can be collected. The solution proposed in this paper can change the way the credit systems work in a major way due to inclusion of non-transactional factors in credit score calculation. Adding qualitative measures like depth and breadth of a person's social media connections and alternate payment data refine algorithms that are used to derive credit scores. Thus reputation of users and their connections in society will also determine the credit score. Because of the smartphone application, users will have access to their credit score and this will encourage financial discipline in a person's life. The app will not just show the credit score at a particular time but also how the score has improved over a period thereby encouraging the user to be aware of how the score can be improved.

IX. FUTURE WORK

Implementing the actual neural network would have required a dataset to train the network. However, the amount of data required to train the neural network is not available. So this can be done after large amounts of transactional and social media data is gathered and made available using the proposed solution. For example, there is data of default of credit card clients' data set available from the University of California, Irvine Machine Learning Repository [9]. The Back Propagation algorithm was used to classify the entries into whether they defaulted or not. The data set was partitioned-25000 entries were used to train the 5 hidden layer neural network, and the remaining 5000 were used as test data. The classifier gave an accuracy of 72.1% in classifying the test set. This exercise was done on a relatively small data set. However as a predictive model, this neural net is very versatile. The model parameters can be varied to predict just about any kind of numerical data. Thus given the data input parameters and architecture of this neural network, the problem of calculating credit scores can be easily solved.

Another problem that needs to be addressed is data security. The data accessed by the proposed solution will be extremely confidential and needs to be encrypted in order to avoid misuse of the data. There are many existing algorithms to encrypt the data communicated from Android like [8]. Users may be reluctant to let some organization have direct access to their data. Access to SMS and direct streaming SMS data to servers may be treated as breach of privacy. It also raises many questions related to security of data shared. A simple solution over this is to fetch all including transactions and social media data on the smartphone itself. The application will fetch the data and calculate the credit score locally. It will send just this calculated score to the server. However, crawling social media data through smartphone application is the problem that needs to be solved here.

Another issue that needs to be solved is lack of internet. Users may turn off internet leading to servers not getting any transactional data at all. There needs to be a system to assure transmission of transaction data to servers irrespective of internet availability.

REFERENCES

[1] CIBIL, "About Us", [Online] Available from: https://www.cibil.com/about-us [Accessed 2/10/16].

[2] Harshala Chandorkar, "How is CIBIL your score calculated ?", 2015 [Online] Available from: https://www.cibil.com/how-your-cibil-score-calculated-%E2%80%93-harshala-chandorkar [Accessed 5/10/16].

[3] R, https://www.r-project.org/

[4] MongoDB, https://docs.mongodb.com/

[5] Wang Cong and Ning Huicong, "The study of big data based on complex network—with the example of credit reference," *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*, Harbin, 2015, pp. 614-617.

[6] FICO, "What's in my FICO scores?", [Online] Available from: http://www.myfico.com/crediteducation/whatsinyourscore.aspx

[7] E. Kambal, I. Osman, M. Taha, N. Mohammed and S. Mohammed, "Credit scoring using data mining techniques with particular reference to Sudanese banks," *Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on*, Khartoum, 2013, pp. 378-383.

[8] S. Verma, S. K. Pal and S. K. Muttoo, "A new tool for lightweight encryption on android," *Advance Computing Conference (IACC), 2014 IEEE International*, Gurgaon, 2014, pp. 306-311.

[9] UCI Machine Learning Repository, Default of credit cardsclients data set. [Online] Available from: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients [Accessed 9/10/16]