ISSN : 2231-0711

Vemula Geeta et al | International Journal of Computer Science Engineering and Technology( IJCSET) | March 2015 | Vol 5, Issue 3, 57-61

# Big Data Analytics for Detection of Frauds in Matrimonial Websites

**Vemula Geeta,**
*Asst.Professor ,*
*Dept. of CSE , SNIST,*
*Hyderabad, India*

**P. SivaJyothi,**
*Asst.Professor ,*
*Dept. of CSE ,SNIST,*
*Hyderabad, India*

**Prof.T.Venkat Narayana Rao,**
*Dept. of CSE,*
*SNIST,*
*Hyderabad*

***Abstract* -In India , online matrimonial websites plays a major role being trusted by millions of Indians globally. At the same time these act as the playground for the fraudsters and imposters duping thousands of victims by posting false and fraudulent profiles every year. Analytics is the future of big data because only transforming data into information gives them value and can turn data for social benefit. This paper investigates the benefits of Big Data technology and main methods of analysis that can be applied to the particular case of fraud detection in Indian Matrimonial Websites.**

***Keywords*—Big Data, Matrimonial Fraud Detection, Hadoop, Social Networks**

## I. INTRODUCTION

With the online matrimony getting trendy in the country, rising number of cyber duping cases has become the matter of great worry to the matrimonial web sites who appeal defenselessness beyond a point in checking the nuisance. It is a gigantic business and the websites don't want to let their consumers down by the acts some unlawful minded people who use the sites perverted. Meanwhile, the alarming growth of cyber crime pose risk to the the social order as their worldwide growth rate is supposedly four per cent per week. Internet based matrimonial service provider can't go beyond a proper limit in carrying out verifications, so it is equally crucial for the users to remain vigilant. Since fraud is on raise, the increase integrated fraud prevention is required. According to the well-known market research organization Gartner [9]. "Security and fraud risk exposure is increasing because of lack of appropriate verification procedures and are threatened at multiple points of vulnerability. Online services are reevaluating how they tackle fraudsters since a fragmented approach is constantly leaving websites at greater risk of attack. A more comprehensive approach to security ensures all layers protection function together".

The Matrimonial market has registered a boom in India as its ongoing fiscal year business estimates around 1400 million after taking a giant leap from 700-800 millions in last year. The survey, conducted by JustConsult, an online research agency, claims that matrimonial hunt is the second largest gainer among online activities as it has registered growth of 33 per cent during the last year only. It has evolved into a prime and popular service provider as twelve million users which form about half of the urban online users undertake matrimony search An online survey reveals and for each user it has to maintain all the particulars, photos, contact details, personal information, and any fact about users which creates a huge amount of data that needs to be processed and verified.

The conservative database technologies are not suitable for performing this type of exploration due to their inner limitations. The new big data technologies are yet to be understood and the advantages they provide in fighting fraud need to be explored. By implementing big data analysis, the probable imposters' profiles that are "hidden" inside huge archives of data can be identified and appropriate measures can be taken. Such profiles would go unobserved in the absence of big data technologies because the human brain is not capable to associate the huge quantities of data available in the online matrimonial sector. Already some match making organizations are using big data analytics for finding the suitable match, this paper focuses on role of big data analytics in verifying whether the profile is genuine or not.

In order to prevent matrimonial fraud, big data analytics should use the following technologies: business rules, abnormality detection, text mining, database searches and social network analysis. These technologies will be approached during the following sections.

## II. APPLICATION OF BIG DATA ANALYTICS FOR FRAUD DETECTION IN MATRIMONIAL SERVICES

The area of big data is of major interest to the technical society as the volume of the databases has been rising beyond the limits of present technologies. There is ongoing research of big data technologies with relevance in different sectors. There are currently many research initiatives for developing big data technologies. Some of these initiatives are funded by private companies such as IBM ORACLE, SAS and Microsoft. Other initiatives are funded by public bodies and/or the open source community. A notable technology open source is Hadoop which is often integrated with commercial technologies. Although massive data amounts were produced during the last two years, the term "big data" was present in the research literature starting with 1970s, but it has seen an explosion of publications since 2008 [3]. Big data is defined as "a collection of composite data sets so large and complex that it becomes complicated to process using on-hand database management tools or conventional data processing applications". The industry standard definition of Big Data projects it along five dimensions: volume, velocity, variety, veracity and value as shown in Fig.1

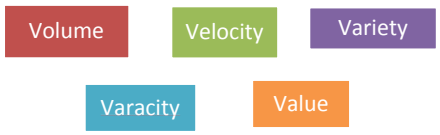


**Fig 1. The five Vs of Big Data**

*Volume*
Refers to the huge amounts of data generated every second. Now we are talking about Zettabytes or Brontobytes. If we take all the data generated in the world between the commencement of time and 2008, the same amount of data will soon be generated every minute. This makes most data sets too big to store and analyse using traditional database technology. New big data tools use distributed systems so that we can accumulate and analyse data across databases that are scattered around anywhere in the world.
*Variety*
Refers to the diverse types of data we can now use. In the past we only focused on ordered data that neatly fitted into tables or relational databases, such as financial data. In fact, 80% of the world's data is unOrdered (script, text, images, video, audio, etc.) With big data technology we can now scrutinize and bring together data of diverse types such as mails, messages, social media conversations, images, sensor data, video or audio recordings.
*Velocity*
Refers to the swiftness at which new data is generated and the speed at which data moves around. Just think of social media messages circulating in seconds. Expertise allows us now to analyse the data while it is being generated (referred to as in-memory analytics), without ever putting it into databases[6][8].
*Value*
One more V is value. Having access to big data is no good unless we can turn it into value. Organizations are starting to generate astonishing value from their big data.

More than that, when working with big data, the meaning of each event can be interpreted only in relationship with preceding events. So, we heave streams of data that must be analyzed all together, like a sequence, so the traditional analytic methods work poorly on these cases. Traditional analytic tools approach data at entity level, as each entity provides useful information. The shift to detailed stream data changes the needs and requires for complex ETL tools. First used by Internet giants like Yahoo, Ebay or Facebook, Apache Hadoop is the most popular big data platform. Hadoop is an open source platform for processing big data that uses distributed processing across clusters of servers. It has become the "de facto" standard for monitoring, processing and analyzing vast amounts of data. Hadoop is a Java based framework and uses simple parallel programming models using clusters of economical servers that locally store and process huge volumes of data. The result is a fundamental decrease of data storage cost. The analysts are free to write code in almost any modern language using the streaming APIs viable in Hadoop[8]. The platform offers a high level of scalability as processing requirements are distributed on thousands of machines and its software is designed to detect and solve failures at application level. This way, its clusters are very flexible. Core Hadoop has two main systems:

*a).Hadoop: Distributed File System(HDFS)*: Self-healing high-bandwidth clustered storage.
*b)MapReduce*: Distributed fault-tolerant resource management and scheduling coupled with a scalable data programming abstraction.

In the beginning (2008), Hadoop had significantly less capabilities than relational databases and had limited supporting tools. But now, it has more strong SQL capabilities and access to all SQL-based applications. Cloudera was the first that introduced commercial support for Hadoop in 2008, followed by MapR and Hortonworks. IBM and EMC have each its own Hadoop distribution. Microsoft and Teradata offer balancing software Hortonworks' platform. Oracle resells and supports Cloudera, while HP, SAP work with multiple Hadoop software providers [4].
Classic business intelligence tools use relational databases for storage and query execution. In order to use the traditional analysis methods and techniques, there have been many efforts to develop SQLlike languages for big data access, query and manipulation: BigSQL, HiveQL, CassandraQL, JAQL, Sparql, Shark etc, each of them associated with a specific big data platform [6].
Many users concluded that no type of big data is optimal for all their requirements. For example, combining Hadoop for unstructured data staging with in-memory business intelligence tools for query acceleration, with stream computing for continuous data provision and with massively parallel processing RDBMS for data warehousing and data management.

## II. FRAUD IN ONLINE MATRIMONY SERVICES

In India, the business of matchmaking is fast growing. The conventional practice which was once firmly managed through family relations and word of mouth, is fast developing to include mobile and social technology, becoming the medium of preference for meeting likely partners.The main activities involved in matrimonial services are as given in Fig.2.

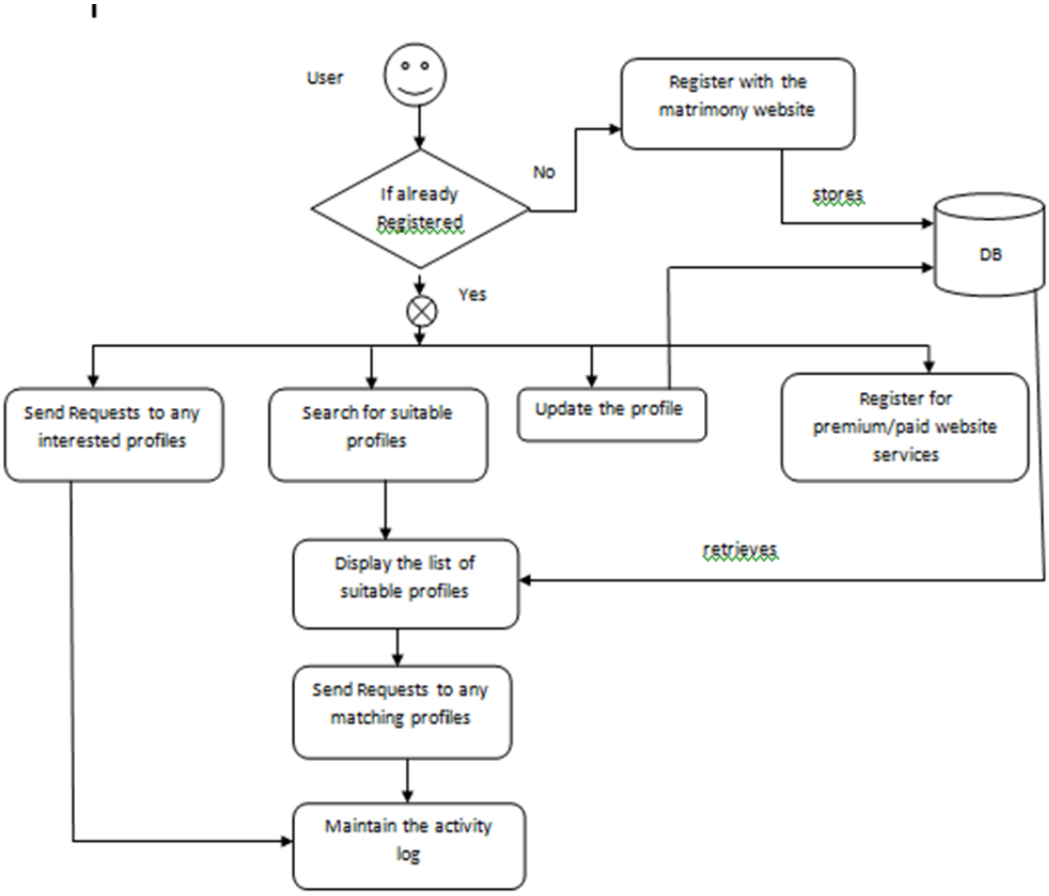


Fig.2. Activities involved in general matrimonial websites

As there is increasing misuse of matrimonial web sites, it became mandatory for matrimonial services to undertake some additional preventive measures as verifying the phone numbers and residence addresses of each user and maintaining harmony of family background. Cyber crime is a threat for the worldwide society as it jumps the geographical limits promptly and its perpetrators are almost unseen as there are rare chances to snatch them [7].

From the data available in databases and activity logs we can identify some unusual patterns and can be categorized as types of frauds discussed below[2].

*a)Case 1*: Admin could observe that two or more profiles with same mobile number, email id, location etc but with different username, qualifications, education, preferences etc exists. This clearly shows that the individual is imposter and trying to lure victims with different profiles. Such profiles must be tracked and scrutinized before recommending the profile to other users.

*b)Case 2*: Admin can observe some profile with lots of negativities with respect to the behavior, family background, education, job profile etc. In such cases it is highly possible that fraudsters try to defame or tarnish the social image of anybody out of enmity by posting all false/ negative data about that person or his family background.

*c)Case 3*: Admin can observe that particular user is sending enormous number of requests to different profiles with or without matching preferences. Here there is high possibility that the profile user could be imposter. Such profiles should be identified and need to investigate thoroughly.

## III. MATRIMONIAL PROFILE VERIFICATION METHODS FOR FRAUD DETECTION

Today when marriages are generally arranged via matrimonial websites and newspaper-ads, one can never be sure of the background of the individual. Hence, we help in making detailed investigations so that the consumers do not land up with the immoral person.

At present most of the matrimonial websites don't include users' profiles verification measures. In such scenarios it will be the responsibility of user to do the background check before proceeding for the proposal. This can be done personally or by employing private investigator.

These investigations involve investigation pertaining to all major aspects of individuals' past as well as present life. From family background to salary package they give detailed report and recommendation. A routine investigation may take upto 15 days or so. But the process is tedious and expensive for any individual and also not necessary that the results are 100% accurate.

Here in our paper focuses on performing big data analysis on the suspicious profiles which we obtain by observing the home database and activity logs. The website must inculcate big data analysis for verification processes and

mark the profile with appropriate indicators so that other users will be attentive before blindly believing the request proposal.

With the available resources like databases and activity logs , analyst can identify the suspicious profiles using following methods[1].

*a)Anomaly Detection:*Anomaly detection algorithms are very simple to set and functions without human intervention. Some profile parameters are taken into account and then thresholds are set. If a threshold is exceeded, then the profile/event is signaled for further investigation. The effectiveness of this method is influenced by the choice of parameters to be monitored, of the analysis period, and of the threshold value settings [5].

*b)Business Rules:*If fraud patterns are known, one can resort to checking every activity by applying business rules. Based on an aggregate score or exceeding a set threshold, the profile can be marked as suspicious, and then carefully investigated. This technique is very simple to apply, once the system was originally set. Its weaknesses are two: setting initial parameters can lead to many false alarms that require further investigation, and the system is flexible to adapt to new methods to defraud the system, new business rules. You can add new business rules only if they meet the new method of fraud.

For example, the Fig.3 shows a business rule in matrimonial services

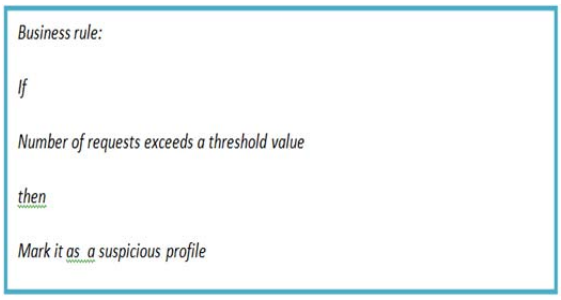


Fig.3. Sample Business Rule

*c)Database Searches:*If by observing each database any profile found to be suspicious, then the data can be compared with other geographically distributed database sources.

*d)Predictive modeling:*Predictive modeling is very successful in detecting fraud. By applying data mining tools, fraud propensity scores can be calculated. Then, using predictive models, they can automatically tell the probability that profile data is fraudulent and it must be subjected to detailed analysis. To preserve accuracy, models must be constantly updated to include new types of illegal events.

Once the suspicious profiles are detected from the above methods these can be subjected to intense investigation using big data analysis as discussed in the Fig.4

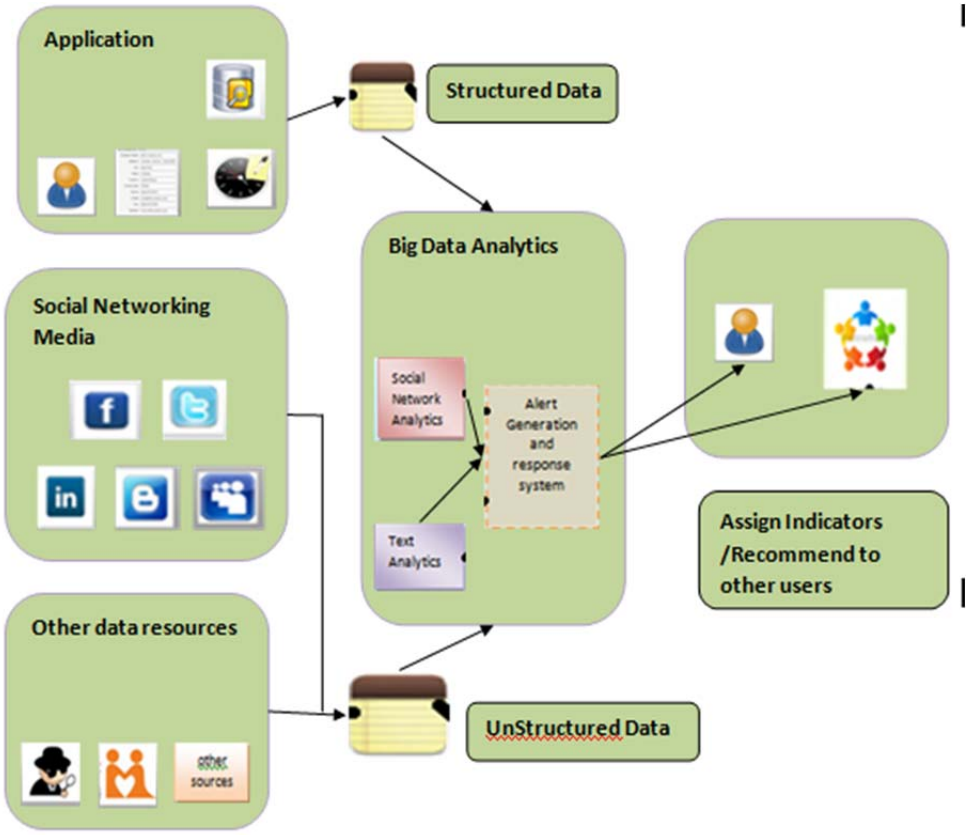


Fig.4. Big Data Analytics integrated in conventional matrimonial websites

The below is the procedure that can be followed by matrimonial services to detect fraudulent profiles:

*Step 1:* The user registers in matrimonial website by giving his interests and preferences. The users profile will be stored into home database, subsequently the activities of the users will be saved into the activity log. Using the basic fraud prevention techniques that are stated above, the suspicious profiles are detected. This data we refer as structured data.

*Step 2:* Social networks are linked with fraud detection because frauds are performed almost always of networks of persons rather than single individuals. Once an individual has been identified as suspicious, the entire social network linked to him could be analyzed for searching fraud schemes. Most fraud schemes are hidden beyond the huge collections of data. If a controller would know where to look the fraud schemes would be relatively easy revealed as they are pretty simple to identify.

*Step 3:*Based on those suspicious profiles, their data from other matrimonial services and other sources like banking and telecom are obtained and sent for big data analysis.Data collected in step 2 and step 3 form the unstructured data.

*Step 4*:Using Big data tools like Hadoop, the users profiles are validated .

*Step 5:*By detecting the fraudulent profiles, the other users are alerted.

## IV. CONCLUSION

Big Data technology and distributed processing power of big data bring fraud detection in matrimony services to another level. Not long ago, matrimonial fraud detection was not considered cost-effective because the cost and duration of the investigations were too high. Applying Big Data analytics methods can lead to rapid detection of fraudulent profiles which may not be possible with the help of conventional techniques. The article briefly presented the main types of fraud that are encountered in matrimonial services. An analysis of Big Data technology demonstrates its huge potential, but it shows that native tools for data analysis are still immature. The analytics methods applied in the field of matrimony were briefly described, each of them being effective for a particular type of fraud or a particular stage of the fraud detection process. All this leads to the conclusion that the best solution for detecting fraud in the matrimonial services is, at present, a integrated solution, both in terms of technologies and in terms of models of analysis.

## REFERENCES

[1] Ana-Ramona BOLOGA, Razvan BOLOGA, Alexandra FLOREA,"Big Data and Specific Analysis Methods for Insurance Fraud Detection", Database Systems Journal vol. I, no. 1/2010

[2] Richard A.Derrig,"Insurance Fraud", The Journal of Risk and Insurance",2002,Vol.69,No.3,271-287

[3] Big Data Meets Big Data Analytics, http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/big-data-meets-big-data-analytics-105777.pdf

[4] Artis, M., Ayuso M. & Guillen M. (1999). "Modelling Different Types of Automobile Insurance Fraud Behaviour in the Spanish Market". Insurance Mathematics and Economics 24: 67-81.

[5] Abbott, D., Matkovsky, P. & Elder, J. (1998)." An Evaluation of High-End Data Mining Tools for Fraud Detection." Proc. of IEEE SMC98.

[6] Barse, E., Kvarnstrom, H. & Jonsson, E. (2003). "Synthesizing Test Data for Fraud Detection Systems". Proc. of the 19th Annual Computer Security Applications Conference, 384-395

[7] Bell, T. & Carcello, J. (2000). "A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting. Auditing": A Journal of Practice and Theory 10(1): 271-309.

[8] Bolton, R. & Hand, D. (2001). "Unsupervised Profiling Methods for Fraud Detection". Credit Scoring and Credit Control VII.

[9] Rutrell, Y., "Analytics platform helps agencies fight cyber crime, government computer news", Jul 12, 2012, http://gcn.com/articles/2012/07/12/sassecurity-intelligence-latfromanalytics.aspx;

## BIOGRAPHIES:

Vemula Geeta, M.Tech Software Engineering from JNTUH, Hyderabad .She possess 8 years of experience in Academic has guided many UG students. Currently she is working as Asst. Professor at Sreenidhi Institute of Science and Technology, Hyderabad. Her areas of interest include Web technologies, Database Management Systems, Object oriented Analysis & Design, Information Security, Big Data Analytics.

P. SivaJyothi , M.Tech Computer Science & Engineering from JNTUH, Hyderabad. She possess 6 years of experience in Academic has guided many UG students. Currently she is working as Asst Prof at Sreenidhi Institute of Science and Technology, Hyderabad. Her areas of interest include Operating System, Object oriented Analysis & Design, Information Security, Big Data Analytics.

Professor T.Venkat Narayana Rao, received B.E in Computer Technology and Engineering from Nagpur University, Nagpur,India, M.B.A (Systems), holds a M.Tech in Computer Science from Jawaharlal Nehru Technological University, Hyderabad, A.P., India and a Research Scholar in JNTUK. He has 23 years of vastexperience in Computer Science and Engineering areas pertaining to academics and industry related I.T issues. He is presently working as Professor, Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology , Ghatkesar , R.R.Dist.,T.S, INDIA. He is nominated as an Editor and Reviewer to 45 International journals relating to Computer Science and Information Technology and has published 78 papers in international journals. He is currently working on research areas, which include Digital Image Processing, Digital Watermarking, Data Mining, Network Security and other emerging areas of Information Technology. He can be reached at tvnrbobby@yahoo.com