ISSN 2320-9984

Volume 4 Issue 2 | June 2016

ISSN: 2320-9984 (Online)

International Journal of Modern Engineering & Management Research

Website: www.ijmemr.org

A Novel Approach to Identify Terrorist Using Combination of K-Mean & Farthest First Algorithm

Amit Kumar
Research Scholar
Department of Computer Science & Engineering
Vindhya Institute of Technology & Sciences
Jabalpur (M.P.), [INDIA]

Sanjay Gupta
Head of Department,
Department of Computer Science and Engineering,
Vindhya Institute of Technology & Sciences
Jabalpur (M.P.), [INDIA]
Email: sanjiit@rediffmail.com

Saurabh Kumar Singh
Assistant Professor
Department of Computer Science & Engineering,
Jabalpur Engineering College
Jabalpur (M.P.), [INDIA]
Email: ssingh@jecjabalpur.ac.in

Abstract—*Nowadays identification of suspicious person (terrorists) is challenging task for researchers. Studies are going on this topic. Terrorist data mining is introduced a new term among other data mining flavours. This paper describes a novel work to identify terrorist or suspicious person with the use of k -mean & farthest first cluster algorithm.*

Keywords:— *Clusters, real time mining, web mining.*

1. INTRODUCTION

The data mining is a task of data processing which is used various kinds of algorithms and other techniques which are employed for recovering the meaningful information from the raw content. To accomplish this research data mining techniques is applied over the student skill database and may be in dataset which content the details of hundred peoples which are having the visited links and logging time duration of the system and server and using the no of sims.

1.1 Data mining:-

Data mining is the process of extracting useful patterns or trends often previously unknown from large amounts of data using various techniques such as those from pattern recognition and machine learning. Data volumes and the lack of sophisticated and sensitive network tools and techniques to utilize the data effectively and very efficiently, there have been several developments in data mining and the technology is being used for a wide variety of applications. In other word we can say that the data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cut costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categories it, and summarize the relationships identified.

Data mining finds these patterns and permits businesses to make predictions about how buyers in that same area code will behave in the future.

How Data Mining Works: Data mining uses a variety of mathematical algorithms to analyze historical data. The results of this analysis are then used to build models based on real world behaviour, which are in turn used to analyze incoming data and make predictions about future behaviour.

1.2 Terrorist Mining:-

The discussion about the global security came into limelight after 9/11 attacks. A prime challenge faced by law enforcement agencies is the large crime RAW data volumes and lack of sophisticated tools and techniques utilized the data effectively and efficiently. Likewise the web traffic generates vast amount of data from which only a small portion is critical to the intelligent terrorist network mining has emerged as novel field of research applied to investigation of organized crime. Relationship among criminals/terrorist organization which can be viewed as a network where nodes represents terrorist and links represent relationships or association between terrorists.

Traditionally analysis of terrorist network was a manual process consuming much time and effort due to information overload and thus failed to generate valuable knowledge on time hence effective techniques are of essence to amend the information overload problem. This paper describe the techniques that generate patterns distinguishing between legitimate and threat groups and helps law enforcement agencies to decide which networks to put under scrutiny. Data mining and automated data-analysis techniques are not complete solution. They are only tools, but they can be powerful tools for this new intelligence requirement. These techniques can assist analysis and investigators by automating some low-level functions that they would otherwise have to perform manually. These techniques can help prioritize attention and provide clues about where to focus, thereby freeing analysts and investigators to engage in the analysis that requires human judgment. In addition data mining and related techniques for useful tools for some early analysis and sorting task that would be impossible for human analysis. They can find link, pattern and anomalies in masses of data that human could never detect without this assistance. These can form the basis for further human inquiry and analysis.

1.3 Research Challenges: -

We discuss a few of challenges in this section. Data mining technologies are now being applied for many applications the matter is, are they ready for identifier and/or preventing terrorist activities. For example- can we complete eliminate false positive and false negatives. False positive might be difficult for various individuals. False negative might increase terrorist activities. The object is to find the needle in the haystack. We need knowledge oriented data mining to eliminate false positive and false negative as much as possible.

Second challenge is mining data in real time. Currently we tools to detect credit or debit card violation. These tool functions in real time, but how one build models in real time model building can is a challenge among the research community.

Privacy is a challenge with respect to data mining for counter terrorism the challenge is to extract useful information from data mining but at the same time maintain privacy. Several efforts are under processing for privacy preserving data mining.

Web mining is challenge for detecting unusual patterns in a way web mining encompasses data mining as one has to mine all the data on the way in addition we need tools to mine the structure of the web as well as usage patterns. A Third challenge is multimedia in data mining. Mining unstructured database still a challenge. Do we fetch structure from unstructured database and then mine structure data. Or do we apply mining tool directly on unstructured data. Furthermore, while there is progress on text mining, we need work on audio and video as well as on image mining. Other directions include graph and pattern mining. All the dots, essentially one builds a graph structure based on the information he or she has. If multiple agencies are working on the problem, then each agency will have its own graph. The challenges are to be able to make inference about missing nodes and links in the graph. Finally finding the data to test the ideas is still a major challenge. How can we get unclassified data? How can we find large data set consisting of multimedia data type? Is it possible to develop a test-structure where one can apply the various data mining tools to determine their efficiency?

The above are some of the challenges for data mining against terrorism discussed at worldwide. That is while data mining could become a use full tool for counter terrorism; there are many challenges that need to be addressed. They include mining multimedia data graph mining, building models in real time, knowledge directed data mining to eliminate false positive and false negative, web mining and privacy sensitive data mining. Research is progressing in the right direction how ever there is still much to be done.

A good concept to learning is the best idea of so-called concept learning. Suppose we want to learn the concept of a kangaroo, so that on subsequent observation it can be recognized. Somebody shows us whole host of animals and negative examples of the concept of a kangaroo. Now learning algorithms would try to reduce a definition of the concept of a kangaroo from the positive and negative examples. This definition can take various forms depending on which qualities of the animals we investigate.

There is a variety of different techniques to enable computers to learn concepts. A very important quality of good learning algorithms is that they learn consistent and complete definitions. A definition of a concept is complete if it recognizes all the instances of a concept; in our case, this means that it does not classify good examples of kangaroos as non-kangaroos. A definition of a concept is consistent if it does not classify any negative examples as falling under the concept: in our case, this would mean that it would not recognize fish or birds as being kangaroos. An incomplete definition of the concept of a kangaroo would be too narrow and would fail to recognize all the kangaroos. An inconsistent definition of the concept of a kangaroo would be too broad, that is, it would classify some non-kangaroos as being kangaroos. Note that a definition can be both inconsistent and incomplete at the same time, although this would obviously be a very bad definition of a concept. A very important element i machine learning is the language in which we express the hypothesis describing the concept. This language could be a specialized computer language like Prolog or lisp, or a special form of knowledge representation using database tables. In the case of the kangaroo hypothesis, it might well be a set of attributes that could be stored in a database. The issues of the expressive power and the structure of the language in which we formulate our hypothesis are very important in machine learning.

2. PROBLEM STATEMENT

The system required must focus on finding the suspected terrorists out of many with a given set of information related to their online activity. The proposed system works on databases, i.e. extracted and collected database. Both the database relations have the following attributes:

1. Name
2. Links (addresses visited)
3. Time
4. No. of Sim Cards (number used)

Our system works in a way such that it matches the various attributes of the database and forms clusters with similar attribute values. The database contains the previously collected database and the currently extracted database under supervision. If the data under supervision is in a cluster with terrorist's previously collected data, then it may be considered "suspicious".

3. PROPOSED WORK

Process:

The need is to track the terrorist activity on the internet using smart measures. For this purpose a set of database and a tool known as "WEKA" is used.

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

In our work we use 2 databases:

1. The collected database: The pre-known verified data of terrorists
2. The extracted database: The data of users under surveillance

Both these databases are merged to form a single excel sheet. (In our case we have taken 100 records in total)

This excel sheet should be converted into the ".csv" format before being used in "WEKA" tool.

This gives us the data in a comma separated format.

For the next step we need to divide the entire data into a number of clusters with each cluster having a similarity in the all the elements in it. In this way we can find out a suspected user if it is clustered with a terrorist data i.e. some of its elements are similar to it using the "Farthest First Clustering" or the "K-means clustering" algorithms.

For this, we take three attributes for each user:

1. Links (addresses visited)
2. Time
3. No. of SimCards (number used)

We assume that these three attributes of the user under surveillance can be extracted using some tools and we have these attributes of the terrorist's data.

We work over the entire process in 2 stages:

1. Level 1 Suspicious Activity
2. Level 2 Suspicious Activity

If the links attribute of the data match with a terrorist link data i.e. they are in the same cluster, we term it as LEVEL 1 SUSPICIOUS ACTIVITY and call it eligible to go through LEVEL 2.

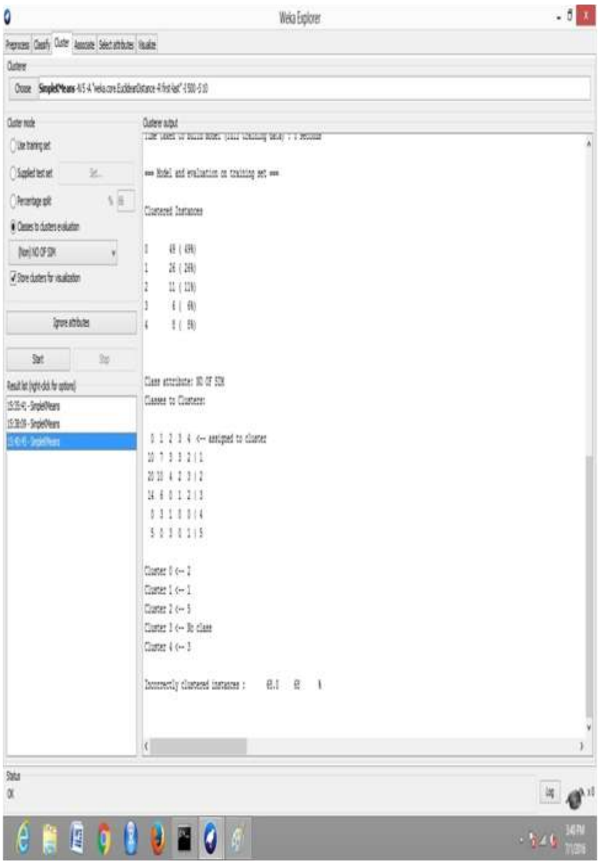
For LEVEL 2, we use the surfing time on a suspicious website and the no. of sim cards used by that particular user. If the "elapsed time" and the "no. of sim cards s" exceed the threshold limit for both we may say that the particular user falls under LEVEL 2 SUSPICIOUS ACTIVITY.

For "links" clustering we use the "Farthest first clustering algorithm"

For the "time" and "no. of SimCards" attributes clustering purpose we use the "K-means" algorithm with k=5.

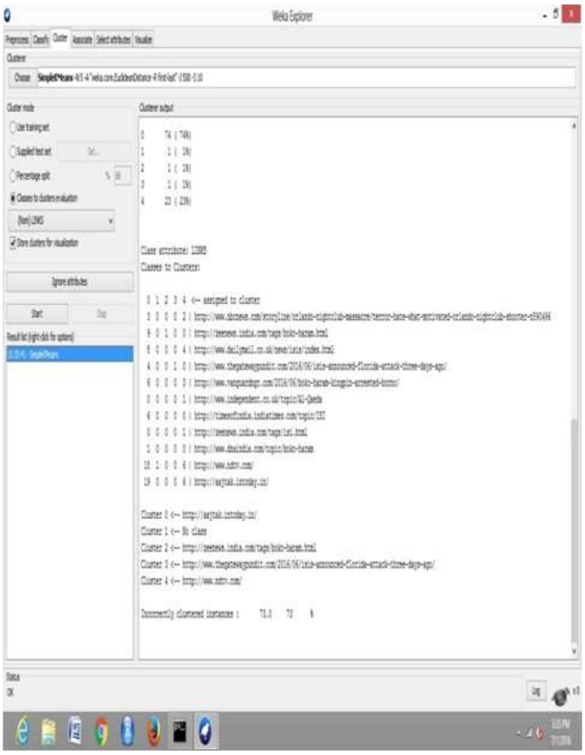
4. RESULTS AND SIMULATION

WEKA-Weka stands for Waikato environment for knowledge analysis. It consists of machine learning algorithms for performing data mining task. These algorithms use data set which has graphical user interface or it can be imported from your own java code by using weka java library.



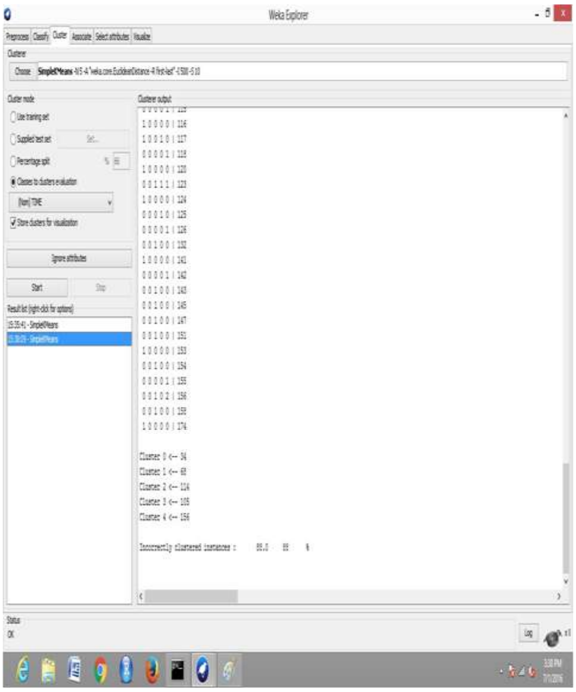
```
SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

=== Model and evaluation on training set ===

Clustered Instances

0       49 ( 49%)
1       26 ( 26%)
2       11 ( 11%)
3        6 (  6%)
4        8 (  8%)


Class attribute: NO OF SIM
Classes to Clusters:

  0  1  2  3  4  <-- assigned to cluster
 10  7  3  3  2 | 1
 20 11  4  2  3 | 2
 14  6  0  1  2 | 3
  0  3  1  0  0 | 4
  5  0  3  0  1 | 5

Cluster 0 <-- 2
Cluster 1 <-- 1
Cluster 2 <-- 5
Cluster 3 <-- No class
Cluster 4 <-- 3

Incorrectly clustered instances :	65.0	 65	 %
```

Figure 1: Using Simple K-mean algorithm with respect to the time clusters evaluation.



```
SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

0       74 ( 74%)
1        1 (  1%)
2        1 (  1%)
3        1 (  1%)
4       23 ( 23%)


Class attribute: [illegible]
Classes to Clusters:

  0  1  2  3  4  <-- assigned to cluster
  3  0  0  0  2 | http://www.cbsnews.com/storyline/orlando-nightclub-massacre/terror-hate-what-motivated-orlando-nightclub-shooter-c590496
  9  0  1  0  0 | http://zeenews.india.com/tags/boko-haram.html
  8  0  0  0  4 | http://www.dailymail.co.uk/news/isis/index.html
  4  0  0  1  0 | http://www.thegatewaypundit.com/2016/06/isis-announced-florida-attack-three-days-ago/
  6  0  0  0  3 | http://www.vanguardngr.com/2016/06/boko-haram-kingpin-arrested-kano/
  0  0  0  0  1 | http://www.independent.co.uk/topic/Al-Qaeda
  4  0  0  0  0 | http://timesofindia.indiatimes.com/topic/ISI
  0  0  0  0  1 | http://zeenews.india.com/tags/isi.html
  1  0  0  0  0 | http://www.dnaindia.com/topic/boko-haram
 18  1  0  0  6 | http://www.ndtv.com/
 19  0  0  0  6 | http://aajtak.intoday.in/

Cluster 0 <-- http://aajtak.intoday.in/
Cluster 1 <-- No class
Cluster 2 <-- http://zeenews.india.com/tags/boko-haram.html
Cluster 3 <-- http://www.thegatewaypundit.com/2016/06/isis-announced-florida-attack-three-days-ago/
Cluster 4 <-- http://www.ndtv.com/

Incorrectly clustered instances :	73.0	 73	 %
```

Figure 2: Using Simple K-Mean algorithm with respect to the No of Sim Cards clusters evaluation



```
SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

 1 0 0 0 0 | 116
 1 0 0 1 0 | 117
 0 0 0 0 1 | 118
 1 0 0 0 0 | 120
 0 0 1 1 1 | 123
 1 0 0 0 0 | 124
 0 0 0 1 0 | 125
 0 0 0 0 1 | 126
 0 0 1 0 0 | 132
 1 0 0 0 0 | 141
 0 0 0 0 1 | 142
 0 0 1 0 0 | 143
 0 0 1 0 0 | 145
 0 0 1 0 0 | 147
 0 0 1 0 0 | 151
 1 0 0 0 0 | 153
 0 0 1 0 0 | 154
 0 0 0 0 1 | 155
 0 0 1 0 2 | 156
 0 0 1 0 0 | 159
 1 0 0 0 0 | 174

Cluster 0 <-- 94
Cluster 1 <-- 68
Cluster 2 <-- 114
Cluster 3 <-- 105
Cluster 4 <-- 156

Incorrectly clustered instances :	88.0	 88	 %
```

Figure 3: Using Simple K-Mean algorithm with respect to the links clusters evaluation.

5. CONCLUSION:

In this paper suspicious person are identified by data mining approach using k-means and farthest first cluster algorithm. Accuracy can be improved using any other algorithms any filtration technique can carrying a big change in result. So future role is open for researcher to optimize this model & method.

REFERENCES:

[1] Wilson Clay. CRS Report for Congress Botnet, Cybercrime, and Cyber terrorism: Vulnerabilities and Policy Issues for Congress. 2008.

[2] CEHv6.1: cyber warfare and terrorism module 43, page 30. ECCOUNCIL [Course note]-

[3] Kohonen Teuvo, Self-Organizing Maps, 3rd Ed. Verlag-Berlin Heidelberg:Springer. 2001

[4] P.N. Tan, M.Steinbach, V.Kumar, Introduction to Data mining, AddisonWes

[5] Arnold, J.J., Tsai, M.-C., Halpern, P., et al.: 'Mass-casualty, terrorist bombings: epidemiological outcomes, resource utilization, and time course of emergency needs (Part 1)', Prehosp. Disaster Med., 2003, 18, pp. 220–234.

[6] Stewart, M.G.: 'Cost effectiveness of risk mitigation strategies for protection of buildings against terrorist attack', J. Perform. Constructed Facil., 2008, 22, pp. 115 –120

[7] Ganor, B.: 'Chapter 2: trends in modern international terrorism', in Weisburd, D., Feucht, T., Hakimi, I., et al.(Eds.): 'To protect and to serve: policing in an age of terrorism' (Springer, 2011)

[8] Powell, R.: 'Defending against terrorist attacks with limited resources', Am. Political Sci. Rev., 2007, 101, pp. 527–541

[9] Coaffee, J., Moore, C., Fletcher, D., et al.: 'Resilient design for community safety and terror-resistant cities', Proc. ICE, Municipal Eng., 2008, 161, pp. 103–110

[10] Kappia, J.G., Fletcher, D., Bosher, L., et al.: 'The acceptability of counter-terrorism measures on urban mass transit in the UK', WIT Trans. Built Environ., 2009, 107, pp. 627–636

[11] Turley, C., Stone, V.: 'Security screening trial at Heathrow Express' (Department for Transport, 2006)

[12] Turley, C., Stone, V.: 'Security screening trial at Canary Wharf station' (Department for Transport, 2006)

[13] Turley, C., Stone, V.: 'Sniffer dogs trials (London and Brighton)' (Department for Transport, 2006)

[14] Potoglou, D., Robinson, N., Kim, C.W., et al.: 'Quantifying individual's trade-offs between privacy, liberty and security: the case of rail travel in UK', Transp. Res. A, 2010, 44, (3), pp. 169–181

[15] Viscusi, W.K., Zeckhauser, R.J.: 'Sacrificing civil liberties to reduce terrorism risks', J. Risk Uncertain., 2003, 26, pp. 99–120

[16] Harper, S.: 'Impact of predicted demographic change on Great Britain's Railways: employers and service providers' (Rail Safety and Standards Board (RSSB), 2007)

[17] Donabie, A.: 'Social trends 41: transport' (Office for National Statistics, 2011)

[18] Ian McCulloh. Detecting change in longitudinal social networks. Journal of Social Structure, Vol. 12 (2011).

[19] Kathleen M. Carley, Dave Columbus, Matt DeReno, Jeff Reminga and Il-Chul Moon. ORA User's Guide 2008. Institute for Software Research School of Computer Science Carnegie Mellon University (2008).

[20] Ronald D. Fricker, Jr. Gmae Theory in an Age of Terrorism: How Can Statisticians Contribute? Statistical Methods in Counterterrorism. New York: Springer (2006).

[21] Fudenberg, D. and Tirole, J. Game Theory. Cambridge: MIT Press (1991).