Journal of Computer Sciences and Applications, 2020, Vol. 8, No. 2, 40-45
Available online at http://pubs.sciepub.com/jcsa/8/2/1
Published by Science and Education Publishing
DOI:10.12691/jcsa-8-2-1

Associations Rankings Model for Cellular Surveillance Analysis

Michael M Kangethe*, Robert Oboko

School of Computing and Informatics, University of Nairobi, Nairobi, Kenya
*Corresponding author: mich01mk@gmail.com

Received June 12, 2020; Revised July 13, 2020; Accepted July 22, 2020

Abstract This is the study and implementation of an association surveillance technology framework model for GSM mobile networks. This enables the efficient and automated identification of entity associations and potential relationships between several entities and events based on a hierarchy of interactions. The approach to this problem is to develop a weighted graph network $G=(V(W),E)$ where $V=\{w(SID1),w(SID2),\ldots,w(SIDn)\}$ w represents the association sore between the ShadowID represented as a node *SID* and the Person of interest(POI) represented as the root node. This model and algorithm are developed as an automated surveillance system framework that enables the tracking of individual entities ' relationships with others based on their interaction and by their physical proximity to the entity of interest. As the future of automated surveillance will not just include the collection of geographic and visual data but also intelligence on the particular entity's interaction log information from activity patterns which can be mapped in an easy to present format to the interested parties.

Keywords: *graphs, surveillances, data mining, association scoring, GUI: - Graphical User Interface, national security, predictive analytics, Shadow ID, POI, MS (Mobile Station)*

Cite This Article: Michael M Kangethe, and Robert Oboko, "Associations Rankings Model for Cellular Surveillance Analysis." *Journal of Computer Sciences and Applications*, vol. 8, no. 2 (2020): 40-45. doi: 10.12691/jcsa-8-2-1.

1. Introduction

The need for automated reliable surveillance technologies has given rise to advance intelligence analysis algorithms using data from various sources such as surveillance reports, video feeds and Telecommunications data like as SMSs and calls. These algorithms have been used with varying successes to identify people of interest in relation to specific crimes and events and also identify patterns of interest in large volumes of data with relative ease.

Automated intelligence analysis of large data has been a vital technological necessity in the war against unwanted events (Crime, Terrorism etc.) as it has enabled law enforcement and intelligence agencies sift through large volumes of unstructured data from different sources in order to flag and identify patterns that would initially be impossible to find or laborious and time consuming to identify relationships and patterns.

GSM Technologies have over the past two decades become the most common form of communication between people and has in effect become the backbone behind almost every activity, event, association, transaction and planning. With the majority of adults having access to a cellular device in both the developing and developed countries, traditional approaches to planning and associating has changed by bringing people virtually closer to each other eliminating the old barriers and challenges of long-distance associations and the use of physical paper mediums to record schedules and messages. With this data (GSM) new avenue of lawfully collecting and analyzing intelligence emerge with the development of better data structures and corresponding algorithms from case studies and scenario hypotheses. This paper discusses and demonstrates one reliable way to solve the stated problem.

2. Background Concept

Data mining is the process of posing queries and extracting patterns, often previously unknown from large quantities of data using pattern matching or other reasoning techniques [8].

Data mining in surveillance has been used as a vital tool in the analysis of data from financial records to calculate individuals credit scores to determining the possibility of an individual's proclivity to a particular outcome. It is a very powerful tool and its utility can never be underestimated.

Predictive Analytics is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future.

Data Mining and Predictive Analysis offers a clear, practical starting point for professionals who need to use

data mining in homeland security, security analysis, and operational law enforcement settings. The use of predictive analytics in intelligence and security analysis enables the development of meaningful, information-based tactics, strategy, and policy decisions in the operational public safety and security environment [12].

Graph Theory is ultimately the study of relationships. Given a set of nodes & connections, which can abstract anything from city layouts to computer data, graph theory provides a helpful tool to quantify & simplify the many moving parts of dynamic systems. Studying graphs through a framework provides answers to many arrangements, networking, optimization, matching and operational problems [10].

The notion behind graph theory id to map relationships between entities to others and assigning proximity values e.g. by its mathematical representation a graph is just an ordered pair $G= (V, E)$ comprising:

- V a set of vertices (also called nodes or points);
- $E \subseteq \{\{x, y\} \mid (x, y) \in V2 \wedge x \neq y\}$ a set of edges (also called links or lines), which are unordered pairs of vertices (i.e., an edge is associated with two distinct vertices).
- The directed graph G below is made of the nodes $\{a,b,c,d\}$ and edges $\{a,b\},\{b,c\},\{b,d\},\{a,c\}$

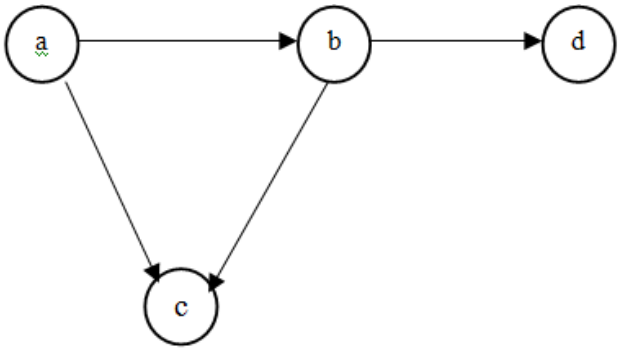

Figure 1. Example Graph

As we become a more "connected" society, a greater need exists to understand complex network structures. While many in the field of data mining analyse network data, most models of networks are straightforward - focusing on many connections of a single type. In order to better understand relationships between different types of entities and extract meaningful structure from heterogeneous data, data mining algorithms need to be developed for new models of complex graphs [11].

3. Methodology

For the applicability of the model, the factors of interest and their relationship to each other had to be quantified. This was achieved by discretely defining what a human-based event is and the Individual universal conditions satisfying the occurrence of any and all events.

Rules defined and observed during the Research and framework Development Life Cycle:

1. An event ***E*** is defined as an occurrence as a result of Human ***P*** activity over a period of time ***T*** at Location ***L*.**
2. For an Event ***E*** to occur there has to be a Location L_i or group of locations $L_i,\ldots,L_n$ where the Event ***E*** occurs.
3. For Event ***E*** to occur there must be the Time $T_i,\ldots,T_n$ to which Event ***E*** occurs where ***i*** represents the number of Times and or the duration to which ***E*** occurred.
4. For Event ***E*** to occur it must involve Entity ***P*** or Groups of Entities $P_i,\ldots,P_n$.

The system Framework will use a mathematical model to perform two main operations:

1. Generating each of the Nodes from the Root node
2. Calculating the Node weight score

Since this model will focus on metadata legally accessible data from the telecommunications service providers (I.e. Call/SMS logs and Pinged locations).

The Call logs and coverage Area active MS (Mobile Station) data, both Criteria of data can be collected independently computed independently but for the purpose of the developed framework they both will contribute in unequal measures to provide the final value which is the association Likelihood value to the collected ShadowIDs (*Any Particular Active cell phone ID that was detected at a particular point)* in relation to either a particular Entity or event of interest.

In general, the framework developed should compute the results of both the call logs and proximity logs independently and compound the result into one final ranking value for each and every entity that has been observed during the collection of the data.

3.1. The Ranking Model

After revising the model several times, we were able to arrive at a suitable formula which would consistently produce desirable results from sample data generated for the test.

This research and framework development has been greatly influenced and guided by the attributes identified in every event.

Time: The time in which the active MS was observed during the data collection period.

Location: This is the specific location in which the ShadowID was observed to be at a particular instance of time during the data collection process.

ShadowID: these are all the identified Numbers during the calculation of close associate and they make the Nodes to the graph.

The above are the three priority factors that have been taken into consideration when building the framework model. Due to their independence and differing Weights contribution toward the final outcome, some of the variables have been combined to form new secondary variables that have been used in the designed model to compute the final outcome. The development of the new variables is a combination of two or more Primary set Variables as detailed below:

Duration: This is the time range $\{\Delta t = t_2 - t_1 \mid t_2 \geq t_1\}$ that will be selected by the party to determine the duration of the event, although it isn't directly associated to the framework model, it is used as a filtering mechanism that will narrow down the focus scope in the data.

Run_Level: This is the Depth of computation to which is used to reveal ShadowIDs that might be associated with the POI but were not detected during the First Pass. In this case, the Run level is computed as the iteration of the computation loop, where if the primary POI is SID1 and after the first pass the collected ShadowID list contains SID2, SID3….SIDN, The Run_Level will be increased by one after each time the set POI moves from SID1 to SID2 to SID3 and so forth. The run level in effect will determine the depth of the Graph from the root node.

Frequency: This is the number of times the $ShadowID_X$ has been identified during each calculation, it is usually as a result of adding all the number of times a particular shadow ID has been identified when each run is being computed, although it is independent to the *Run_Level* , it does affect the final Ranking outcome.

All the above have been used in varying weights (effect to the model's accuracy) to compute the final Ranking outcomes as their relationship to the Final Ranking which is represented as $\boldsymbol{R\_SID_i}$ where $\boldsymbol{R}$ is the Final Rank Value and $\boldsymbol{SID_i}$ is the particular individual $ShadowID_i$ instance.

Due to the nature of how people communicate. We were able to identify one crucial factor, *People know each other tend to communicate longer than those who don't.* Thus, the longer the call duration the closer the association. From the above analysis we were able to derive the model to reflect our analysis as below.

$$R_{SID_1} = \sum_{i=1}^{\infty}\left[\left(\frac{1}{2^{n_{D1}}}\right)+\left(\frac{1}{2^{n_{c1}}}+D\right)\right] \qquad (1)$$

Where: from the Overall Ranking Equation (1) above

- $\boldsymbol{R\_SID_i}$ is the overall Rank Value for $ShadowID_i$. *i* is the instance to which that particular Process is at a point in time, in this case, the number of appearances is also represented as ***i***.
- ***n*** *is* the *Run_Level* for Proximity Rankings ***pi*** and for the Call Logs Rankings.
- $\mathbf{D_i}$ is a function of the Call Duration Interval and Rankings represented by Equation (2) below.

$$D_i = \left[\frac{(d+c_i)}{\left(\frac{d+c_i}{c_i}\right)}\right]. \qquad (2)$$

The Calls Duration Ranking for each identified ShadowID is represented in the equation (2) above and is used to calculate the possible association between the POI and all other ShadowIDs in the collected data.

From the above formula the generated graph can look as below when the final computation is done for two Run levels.

The Diagram shows the output from the model which has generated four Nodes from a selected POI as the root node. It also shows each node score with respect to its association with the root node. The graph has a depth of 2 thus indicating the run level limit was 2 assuming the start point is 0.

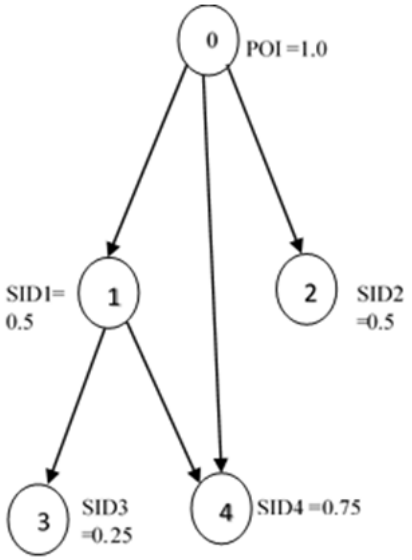

Figure 2. Rank Model Computation Graph Result

3.2. The Pseudo algorithm

3.2.1. Proximity Ranking Pseudo procedure

- Select and set POI.
- Select the Ranking Criteria.
- If Event-Based Rankings user selects Location and/or Date and/or Time-stamp specific to the Event E.
- Set the Association maximum Depth level N.
  //N is the set Maximum Depth Level distance of associations the user can set
- Create a new List LSID of Associated ShadowIDs.
- Add POI to list at i+1 if it doesn't exist.
  //i is the index of the last ShadowID item on the list.
- Create location and Time list TL.
- Add unique Locations and their appearance date and time-stamp where ShadowID=POI in list TL with each item in l+1.
  //l is the index number of the List item.
- From Proximity_log Record 0 - x.
  //x is the last Record
- if $Record\_Location_i$ = $POI\_Location_L$ and $Record\_Time_i = Record\_Time_L$
- Add Unique $ShadowID_L$ to list LSID if doesn't exist Proximity_Rank =
  //n is the Depth level.
- Set AL = n.
  //AL is the Appearance Level which is the initial depth from the original POI in which the ShadowIDi 1st appears.
- if ShadowID exist in list LSID compute new Rank to Proximity_Rank+
- if set n > 1 Set new POI = $ShadowID_{L+1}$
- Repeat from Steps 6 until the end of the list LSID.
- Set n+1
- for every $ShadowID_L$ where AL = new n
- Repeat Step 6 until n = N.Repeat Step 6 until n = N.

3.2.2. Call Rankings Steps

- For each ShadowIDi in list LSID
- Set POI = ShadowIDi

- if CallLog_POIc or ContactIDc = ShadowIDi and CallLog_POI*c* or ContactID*c*=POI
- *//c is the i^{th} position in the callLogs table.*
- Set *n* = ShadowID*i_AL.*
- Set ShadowID*i*_CallRank = ShadowID*i*_CallRank*i* + (+ D_i.)
- *// D_i. Is the CallLog_Duration*
  - *Final Rankings:*
- Set ShadowID*i*_OveralRank = Proximity_Rank*i* + CallRank*i.*

3.3. Model Testing and Evaluation

The test was conducted based on three cases in two criteria.

1. Entity Based Rankings-Rankings based on a specified Person of Interest (POI) and Event.
   - This is where the Association Rankings was purely based on a particular Entity and had no Regard to any Event in particular.
2. Event-Based Rankings: Rankings based on a specified Person of Interest (POI) only conditions such as the time and location.
   - This is where the Association Rankings were Based on a particular Event in which a particular Entity (POI) is Associated.

Out of this Criterion, they were further subdivided into two sub-criteria as the Models Framework Validity and performance accuracy was measured on two criterions

a) General Associated ShadowID Listing:

This is the listing of all the ShadowIDs that are associated with a particular Entity (POI) irrespective of their order of Association.

b) The overall Top ShadowID Listing:

This is the ordered Listing of the ShadowIDs in relation to their Level of Association, the main aim was to List in the first 10 all the Related ShadowIDs above all other ShadowIDs that may have been detected.

c) Accuracy Testing Results Based on the display of all the actual ShadowIDs in the Listing by the system.

The conditions set in the model test and evaluation was based on increasing data size. In this case, we set all three events to contain 10 individuals with varying degrees of associations to each other. Each generated event contained multiple locations and time of appearance within those locations in which some overlap with other events.

3.3.1. Performance Results

As shown in the table Figure 1: above the general listing accuracy remained optimal as all the ShadowIDs were listed during the Event Simulation but Ordered listing Accuracy Decreased down to 80% of the possible listings. This affected its overall Model Accuracy to 90%. This means that only 90% of all the Associated ShadowIDs were to be listed in the first slots during the model simulation process. This was due to the presence of Noise in the data. The above event was generally best suited for Entity Based Rankings.

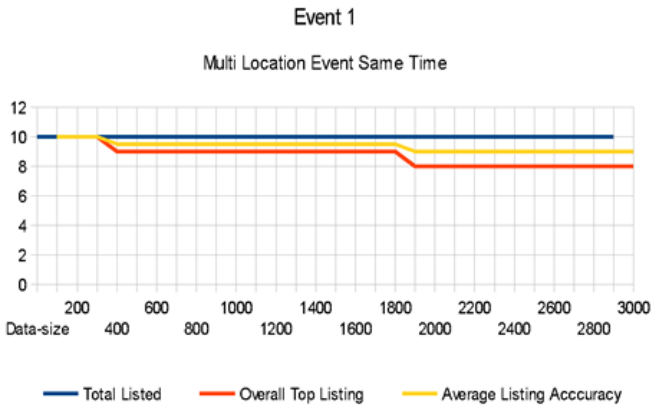

Figure 3. Event 1 Framework Performance Results

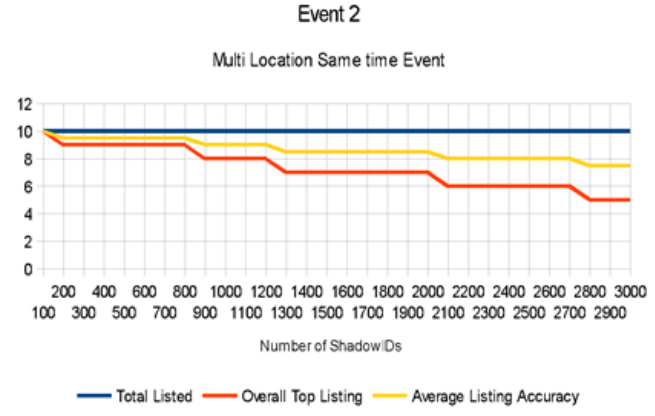

Figure 4. Event 2 Framework Performance Results

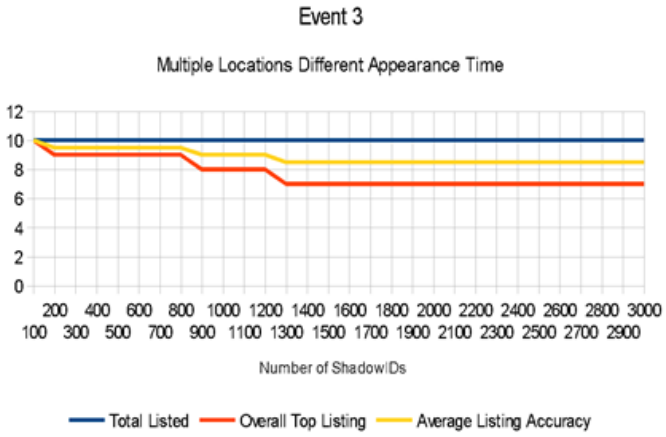

Figure 5. Event 3 Framework Performance Results

As shown in Figure 2: above the general listing accuracy remained optimal as all the ShadowIDs were listed during the Event Simulation but Ordered listing Accuracy Decreased down to 50% of the possible listings. This affected its overall Model Accuracy to 75%. This means that only 75% of all the Associated ShadowIDs were to be listed in the first slots during the model simulation process. This was due to the presence of Noise in the data. The above event was generally best suited for Single Location-Based Events Rankings like Daily or periodic Meetings.

As shown in table Figure 3 above the general listing accuracy remained optimal as all the ShadowIDs were listed during the Event Simulation but Ordered listing Accuracy Decreased down to 70% of the possible listings. This affected its overall Model Accuracy to 85%. This means that only 85% of all the Associated ShadowIDs were to be listed in the first slots during the model simulation process. This was due to the presence of Noise in the data. The above event was generally best suited for Single Location-Based Events Rankings like the DeadDrop Event where one person would place an Item in a hidden location for the other party to come and collect later on several occasions.

From the above-generated events we injected each individual event into a separate database table of 100 records, we tested the general listing accuracy, and ordered listing accuracy of the model by increasing the record counts both the call logs and the proximity logs by the order of 100 from the first 100 records(including the event-specific records) to the final 3000 records.

Event 2 showed a lesser Ordered listing accuracy due to the observed fact that there are people who are usually within the same location as a result of either residence or location of occupation. The above-average entity listing order accuracy faults were however overcome to a great extent by the included computation association depth levels which are the computing of association levels based on both the ShadowID and all the listed associations in a tree format where the parent was the POI and the child was the ShadowID. This was done by giving the ShadowID unique fixed values-per-level. This resulted in higher rankings for highly associated ShadowIDs and the coincidental ShadowIDs that are not really associated with the event remained with their ShadowID values reducing their overall rankings.

For the purpose of this research, Three Events were generated for simulation and Model Validation purposes. The Events generated were based on Actual Events but due to the Sensitive nature of some of the Events, the Locations and ShadowIDs names were changed.

The Table below Describes the Three separates Events of interest and their unique quantifiable properties.

From these criterions divided it into two tables, the Proximity logs where we used the POI or ShadowID, the Location to which the particular ShadowID appeared, the Time and Date of appearance. As the key features of interest. And the Call logs where the POI or ShadowID, Contact, Date and Time of contact, Duration. Were the key features of interest? They were computed independently based on the ranking formula.

4. Key Outputs

Some key issues have become apparent during the research based on surveillance and below are noting.

- Individuals who are socially related (like friends, family business colleagues) always do exhibit patterns in which involves appearing at the same location at the same time.
- Events are a result of active human activity over a period of time.
- It is easier to identify events when there are multiple parties involved than when created by a single individual. The more the parties involved the easier it is to identify the event.

5. Key Achievements

The research has resulted in some noted achievements which include:

- The development of a quantifiable mathematical Model that can be used with desirable confidence to computing the possible social hierarchy of relationships between individuals from observing single person movement activities only.
- A computer system framework has been developed that can externalize the human computation process in using the mathematical model to compute the rankings.

- Events based on human activities have now been defined based on symbolic relations between factors of interest.

6. Key Challenges

This research like any other has not been without its own challenges, both technical and resource-wise as they include.

- Unavailability of actual real-world data to compute real-world events thus limiting the model's accuracy.
- Limited Research time as this research has proved to be wide and has a lot of factors that still need both interpretation and analysis.

7. Assumptions and Limitations of Scope

The completion of this research and development of this framework will present some challenges which are stated below.

- The unavailability of the actual mobile subscriber data from the Telecommunication networks due to subscriber privacy and Business ethics
- The accuracy of the framework model cannot be guaranteed since it has not been tested with real data.
- Data collected has been generated from the previous collection of records from the mobile BTS and filtered based on a GSM tracking system that is installed on the Telecommunication networks

8. Conclusion and Recommendations

The development of this framework has and will enable the immediate and automated analysis of data sets. The use of multidimensional data from different sources could enhance the analysis capabilities using this model including social media data and metadata.

References

[1] Bruno Agard, Catherine Morency, Martin Trépanier. 2007. Mining Public transport user behavior from smart card data.

[2] Vikas Grover, Richard Adderley, Max Bramer,. Review of Current Crime Prediction Techniques. 2009.

[3] Nassar, Khaled, Data-Mining of State Transportation Agencies Projects Databases. 12. 2007.

[4] Tom M. Mitchell 1997. Machine Learning.

[5] Van der Veer, H.T. Roos, A. van der Zanden, Data mining for intelligence led policing. 2009.

[6] Dale Dzemydiene, Raimundas Vaitkevicius, Ignas Dzemyda, Pattern recognition based on statistics and structural equation models in multi-dimensional data warehouses of social behavioral data, pg 4 -10. 2010.

[7] Han J, Kamber, Data mining: concepts and techniques. 2006.

[8] Thuraisingham, Bhavani, Data Mining, National Security, Privacy and Civil Liberties. SIGKDD Explorations. 4. 1-5. (2002).

[9] Lee BS, Snapp RR, Musick R, Critchlow T, Metadata models for ad hoc queries on terabyte-scale scientific simulations. 2002

[10] Najera, J., 2020. *Graph Theory — History & Overview.* [online] Medium. Available at: <https://towardsdatascience.com/graph-theory-history-overview-f89a3efc0478> [Accessed 15 May 2020].

[11] Singh, L., 2020. *Exploring Graph Mining Approaches for Dynamic Heterogeneous Networks*. Georgetown University.

[12] McCue, C, Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis: Second Edition. Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis: Second Edition. 1-393, (2015).