Geospatial Discriminative Patterns with Principle Direction for Effective Crime Detection

R. Saradha[1], **S. Deepika**[2]

[1, 2] Department of Computer Science and Engineering,
Avinashilingam Institute for Home Science and Higher Education for Women .Coimbatore, India

Abstract: ***Identifying and mapping crime have a propensity to cluster geographically. This kind of grouping led to the wise practice of crime and hotspot analysis, which helps to analyze, identify and visualize crime. The process of crime detection has two different issues, which are accuracy and mapping of crimes. The proposed system provides a visual and graphical representation of crimes using machine learning and data mining approaches. The result from the accurately identified crime reports can be more beneficial to the public. Several existing system used mapping methods for hotspot. But those existing approached failed to map crime locations effectively. So the proposed work created a framework for crime detection with visual mapping using data mining approaches. The proposed system used support finding, Geospatial discriminative patterns to gain the significant difference between the normal class and crime class. The data should be perfectly matched with the class accurately, so the system uses Genetic approach for effective fitness finding of class. The system uses both real world dataset and user constructed dataset for evaluation.***

Keywords: GD pattern, hotspot analysis, prism mapping, spatial temporal mining

1. Introduction

With technological advancement information has spread with great speed. This fact has contributed to the news of violence become more frequent in the media. This crime can be considered a phenomenon that happens in space a given place and time. By involving aspects of location, it would be no better than base map science to solve situations for analysis and interpretation. The difficulty of analyzing spatial data through tables have shown that visualization is in forms of maps and more natural interpretation of information is made in full. Thus, as one of the goals of Cartography is the preparation of documents with reference maps to display information in an accessible and clear. In this context the wide usage of thematic maps will assist in analyzing the spatial distribution of crimes. In the proposed system, the approach presents a framework that uses spatial data mining concepts to map crime locations and non crime areas which helps to investigate the relationship between socio-economic and criminal variables. The proposed system has introduced a complete formal and useful mechanism to identify and map crime and proven to be very powerful in identifying the linkage between target objects and its related factors. The components of the proposed method are shown below. The proposed technique uses a spatial data mining concept, Geospatial Discriminative Patterns (GD Patterns) which helps to study the relationship between target crime hotspots and their underlying related variables. Introduce a model, Crime Visual Tool (CVT), to identify crime hotspots through their related variables and activities. Use a support calculation based method to cluster the crime related variables which has highest possibility into groups.

Envision the locations of those clusters in a rational way to assist domain scientists in further analysis, using the footsteps of GD Patterns and association analysis. Additionally the system performs Principle direction process to effectively start mapping the crime locations using the spatial datasets.

1.1 Spatial Data Mining

Spatial data mining is a technique for ‘‘extraction of implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial databases.

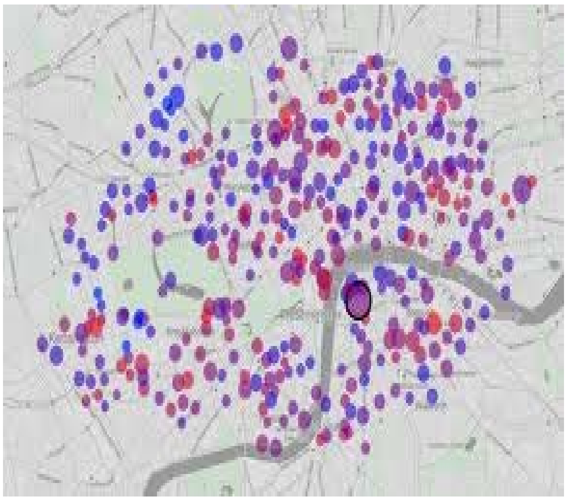


Figure1: Point mapping

2. Overview of the Related Work

The GD Pattern is an application of integrating spatial association rules with emerging data mining techniques and patterns. Applications using association rules have been developed to explore the spatial and temporal relationships among objects using census data. In the existing work association rule mining techniques have been used to explore the non-linear relationships among socioeconomic-vegetation variables. Some of the authors present a similarity measure method for summarizing large number of emerging patterns. Some authors adopts the relative risk ratio as the measure of pattern emergence and uses spatial data mining techniques in investigating vegetation remote sensing datasets.

2.1 GD Pattern

Here we give a brief introduction of *Closed Frequent Patterns* GD Patterns and related definitions.

Geospatial database object : A geospatial database object is a tuple of the form: *{x, y, V*1*, V*2*, ..., Vn, C}*,[1] where *x, y* indicate the object's spatial coordinates, *V*1*, V*2*, ..., Vn* are the categorized values of the explanatory variables, and *C* is the class label of target crime. *C* is 0 if the area is not a hotspot (normal area) and 1 if the area is a hotspot. Using *C*, objects in *D* are labeled into the class of *Dh* (hotspots) if *C* = 1, or *Dn* (normal area) if *C* = 0.

Transaction and pattern: In a geospatial database, a transaction *T* is the group of explanatory variables (*V*1*, V*2, ..., *Vn*) in an object.[3] An pattern *X* is a set of values of explanatory variables (e.g. *V*1 = 1, *V*3 = 4). For example, disregarding the class label *C*, in dataset *D* each object can be viewed as a transaction in location (*x, y*) with a fixed-number of variables.

Support*[6]* : A pattern is said to be supported by a transaction when it is a subset of the transaction. For example, given a transaction *T { V*1=1, *V*2=1, *V*3=2, *V*4=2, *V*5=3, *V*6=5*}*, patterns *X*1 *{V*1=1, *V*2=1, *V*5=3*}* and *X*2 *{V*1=1, *V*3=2, *V*4=2 are supported by *T* , though *X*3 *{V*1 = 1, *V*5=5, *V*6=3*}* is not because it is not a subset of *T* . The number of transactions that support an pattern *X* is called the support count (suppcount) of *X*. The support of *X* is the ratio of *X_s* suppcount and the total number of transactions in a geospatial database (Formula 1). *sup*(*X*) =*supp count*(*X*)*τ*(1)where *sup*(*X*) is the support of pattern *X* and *τ* is the number of transactions.

Congested frequent patterns: An pattern *X* is said to be a congested pattern when none of its immediate super-sets has exactly the same support as *X[1]*.A congested pattern can represent a set of non-congested patterns without losing any support information, because the support of non-congested patterns can be calculated directly from the congested pattern. Using congested patterns will effectively reduce the total number of patterns. Furthermore, *X* is a congested frequent pattern if the support of *X* is greater than a user-defined minimum support threshold (*ρ*). We are only interested in congested frequent patterns because infrequent patterns are likely to be insignificant and may happen by chance.

2.2 Optimization of Criminal HotSpots

The patterns we are looking for should meet two requirements: (1) to significantly represent the situation or conditions of explanatory variables in objects in *D*; (2) to significantly distinguish classes (*Dh*, *Dn*) from dataset *D*. A closed frequent pattern can satisfy the first requirement. To capture the difference of classes, the patterns should be more frequent in one class than in another.

Geospatial Discriminating Patterns*[2] (GDPattern)*: In a geospatial database, a closed frequent pattern X is also a GDPattern if the growth ratio(*δ*) of X is larger than a user defined threshold. Here, growth ratio of a pattern is defined as the ratio of its supports in different classes. *δ* = *sup*(*X*,*Dh*) *sup*(*X*,*Dn*)(2) where *δ* is the growth ratio; *sup*(*X*,*Dh*) is the supports of closed frequent pattern X in class *Dh* and *sup*(*X*,*Dn*) is supports of closed frequent pattern X in class *Dn*.

Footprint of a GDPattern: The footprint of a GDPattern X is the objects that support X in geospatial dataset *D* (Fig. 2). It is the set of cells whose correspondent objects support X in the grid map of study area. Footprints of GDPatterns provide a way to measure the spatial distribution of those patterns in studied area.

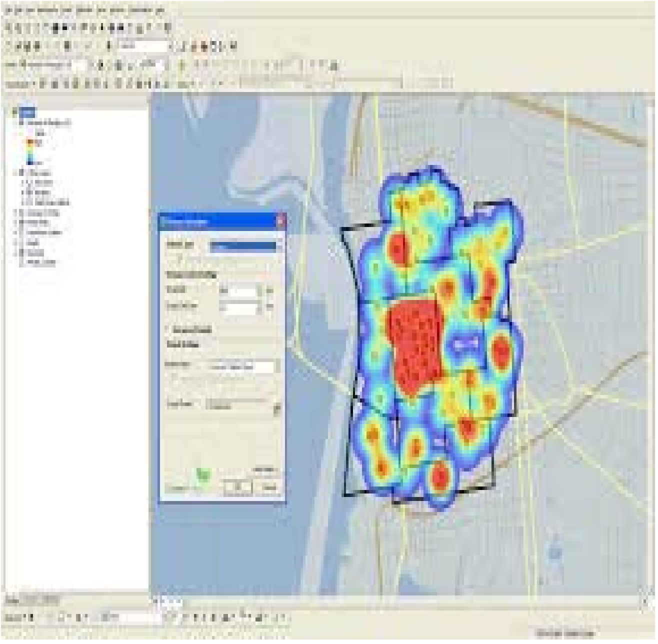


Figure2: GD pattern

An example map of GD Patterns Footprints, by selecting Residential Burglary (RB) data as the target crime, nine other variables are used as explanatory variables from the experiment dataset and 1,500 GD Patterns are mined with a growth ratio larger than twenty. The red area is RB hotspots with a user defined threshold and hallow squares with slash lines are footprints of the 1,500 GD Patterns. Hence, with a rational threshold of growth ratio the GD Patterns mined from *D* are significantly different between classes and are capable of digging out the meaningful information underlying the spatial distribution of target crime hotspots.

3. Problem Formulation

The existing system analysis section briefly presents some literatures related to criminology, spatial data mining, and hotspot mapping techniques. Additionally, this gives a concise introduction to the existing crime mapping application.

Existing hotspot mapping methods can be essentially divided into three main categories: point mapping, choropleth mapping, and kernel density estimation (KDE) usually, these methods aggregate the density of a target crime, which results in a net loss of information. For example, in prism mapping, incident-level data is first aggregated into arbitrary administrative or political boundary areas. During this step, spatial details within and across the thematic areas are lost. Second, when hotspots are generated based on aggregated data, there is a further decline of precision in the resulting map. Because traditional methods mainly rely on target crime density, particular areas with relatively less crime may be left out of hotspots, even though crime related variables indicate they are under similar risks as those hotspots.

- The Hotspot Analysis (HSA) tool implemented by Esri ArcGIS (ESRI, 2011). This study shows the impact of variable and its co occurrences. The Occurrence of crime has been linked to a number of different variables.
- Some other papers explain the need of adjacent areas of crime hotspots which are at higher risk.
- The probability of crimes or the social penalty for committing crime may be lower in crime hotspots than in other neighborhoods, which leads to the ‘‘contagion’’ of criminal activity in crime hotspots.
- Recent work done by Short,Bertozzi, and Brantingham (2010) also discusses how an area is affected by the activity scope of offenders.

4. Proposed Work

In the proposed work GD Patterns are used as a tool to discover the statically significant difference between target crime hotspots and normal areas spatially, with respect to the underlying related variables. The proposed system uses a special kind of tool for Spatial and Temporal Analysis of Crime. The system presents a spatial data mining framework to study the spatial distribution of crimes through their related variables and support. The system also uses some advanced classification techniques for accurate hot spot mapping. Finally the system implements in quantum GIS as tool with the production of graphical output of crime spot mapping. The system also formulates the query framework for effective crime enquiry and hotspot areas. Use of genetic and principle direction concepts the system will able to find the associate locations on the specified area.

5. Methodology

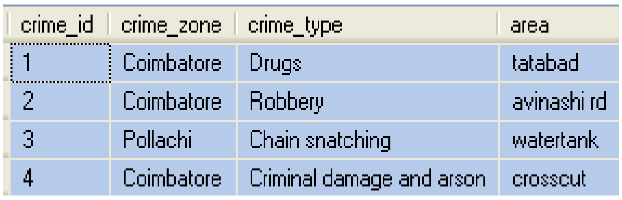
5.1 Dataset Collection

The system used both synthesis and real word datasets. The collection of dataset is the major portion of the proposed system. The system used following datasets for experiments.
UC Repository dataset:

Attribute Information: (122 predictive, 5 non-predictive, 1 goal)

Synthesis dataset:
Attributes:
Crime_id
Crime type
Age
Area
Zone
Locations, etc.,
Instances: 100 and more
Class: 2
Classes: Normal, Crime zone.
Crime type: 15
Areas included: 5 locations

The system implements with the dynamic dataset’s, which are categorized into two types. The types are predefined and user defined values.



crime_id	crime_zone	crime_type	area
1	Coimbatore	Drugs	tatabad
2	Coimbatore	Robbery	avinashi rd
3	Pollachi	Chain snatching	watertank
4	Coimbatore	Criminal damage and arson	crosscut

5.2 Crime Mapping

Crime mapping helps the police department to protect the people from the crime more effectively. An understanding of where and why crimes occur will help to fight against the crime. Simple map shows where the crimes have been occurred.

A. Visualizing the Crime Location

Digital maps visualize the crime scenario in quicker manner. At which place the crimes have occurred that will be visualized, Inspite of searching from the list of events, mapping is easy to visualize the crime hot spot.

B. Integrate the community characteristics

Community characteristics mean the most possible places for occurrence of the crime activities. For example slums, institutions, theme parks, alcohol permit location etc.

C. Producing the maps

At any geographical level the maps can be produced. Where the crimes have occurred that particular place will be shaded darker. The number of crime incidents percentage change will be displayed by shading the area’s location.

5.3 Crime Analysis

Crime analysis is a set of systematic and analytical process for providing the information regarding crime patterns. Crime investigation is an important activity for identifying the crime hotspot. This carries the number of department functions that includes spatial monitoring and deployment, patrol exploitation, special operations, strategic units, investigation, development and explore, crime avoidance and organizational services. Crime analysis is classified into three categories, these are following as,

- Tactical is an analytical process for providing the information to assist operations personnel (patrol and investigative officers) for identifying the crime trends, patterns, series and hotspot. It includes at which time crime is occurred and associating the criminal activities by crime method.
- It includes the preparation of crime statistical reports, resource acquisition and its allocation studies.
- It focuses on provisioning on economic, geographic or social information to administration.

6. Classification of Clusters

The system identifies the areas that contain the more number of clusters (crime areas). The similar type of crime activities will be grouped together. Based on the clusters' result, which cluster contains the more number of criminal activities that will be called as crime hotspot for the particular crime.

6.1 Comparison of location of crime hotspot

The crime hotspot that have been identified over several months.

6.2 Comparison of hotspot with different crime types

The identified crime hotspot will be compared with the other type of crime hotspot. For example burglary type of crime hotspot will be compared with the murder type of crime hotspot.

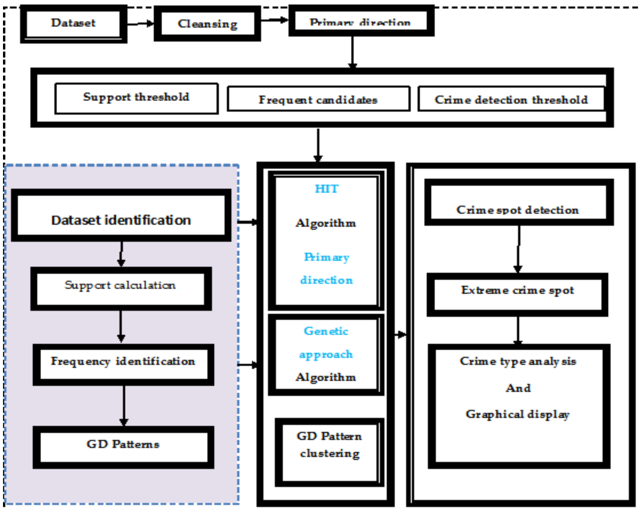


Figure3: System Architecture

6.3 Proposed Algorithm

```
Input: database dt
Candidate threshold (ct)
Support and closed frequent threshold(s)
Output: set of crime spots
Steps:
1. Read the transactional database Dt;
2. Count=1 [find unique crime type CRt]
3. Count++[find the transaction count of CRt]
4. Match with the candidate threshold
5. G=mine GDPatterns (gp) using Ct ct,s.
6. Find max(G) among other clusters.
7. Map the G value in the Dt.
8. Visualize the results.
```

7. Crime Classification

To classify the crime incidents based on the similarity between the crime objects stored in the class, structure crime classification is used. Classification is the hierarchy of these attributes. These attributes are represented by classification in three ways,

1) Classification of crime place
2) Classification crime types
3) Classification of crime time

The structured crime classification algorithm is used to identify the more similar objects in the data sets. Algorithm, to find the hotspot and normal from the dataset.

```
Input: Database DB
Output: crime area
1. Assign S=DB 2. Apply purification attribute Ai by Cn
3. Repeat
   a. Find the similarity of crime attribute objects(C.Ai,
      C.Ai+1)
   b. Find the probability of particular crime classification =
      Probability (C.Ai, Classification)
   c. Threshold T=(Cluster Area-Sparse Area)
   d. Find F(C)=classification U Pi(Ci)
   e. If F(C)>positive description
Produce a hot spot
Else
Produce a cold spot 4. Go to step 3
```

Let S denotes a set of crime incidents. Ai be an attribute of crime incidents and Ci be a classification of each crime attribute Ai. For two elements x1, x2 in the tree of Ci, if there is a path from x1 to x2 is called the parent of x2. Furthermore, x1 is a generalization of x2. In the structure crime classification algorithm, the national dissipation between the events is similar and the events are more similar. Choose the crime attribute Ai in the crime class C. Find the similarity of each crime attribute of crime objects if both objects have the same similarity, join these two objects have the same crime attribute incident and put into the same class C. And finally find the F(C) based on the probability of crime incident occurring in the particular class to which it is merged. If F(C) is greater than the positive description, it produces a crime hot spot, and otherwise it produces a crime cold spot.

8. Crime Clustering

Clustering is data mining technique for grouping the similar type of crimes will be grouped together. In this paper the burglary crime will be clustered, based on the clusters' result the crime hotspot will be identified.

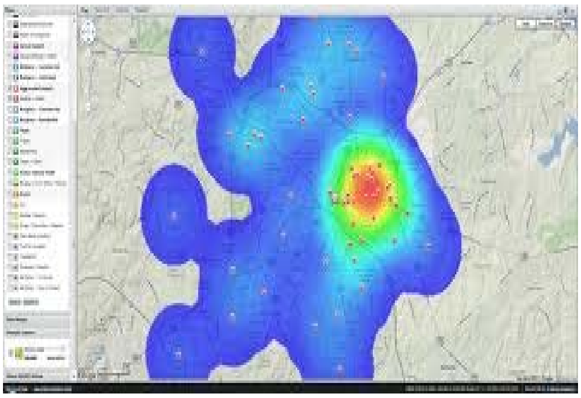


Figure 4: Crime cluster

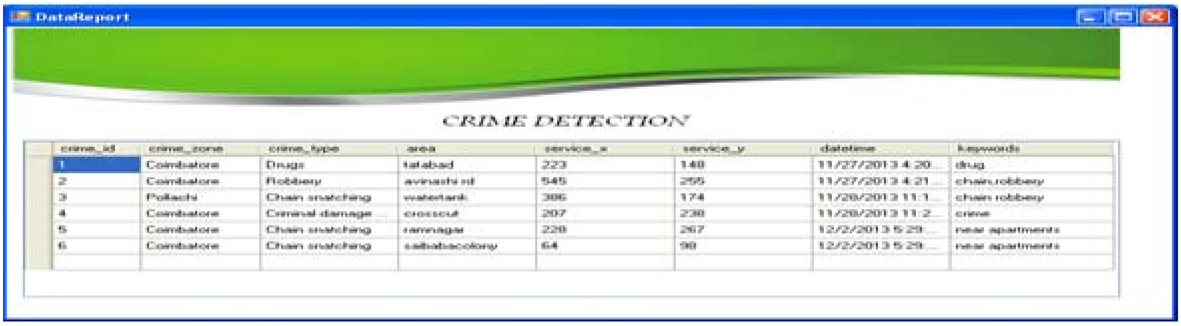
9. Results and Analysis

The implementation of the proposed system used visual studio environment with C#.net language along with QGIS tool .This chapter presents implementation and experiment which conducted by using the ant colony optimization and

HIT (Hotspot Identification Tool) algorithm. The implementations are represents as follows. The First step of the implementation is the consideration with some real data sets, including both numerical and categorical domains, in order to assess the capability of the approach in mining interesting knowledge. The implementation uses the initial direction which used initial principle analysis with the use of HIT algorithm. Then the system produces the graphical results which compute the frequent candidates values and crime spot among the dataset. In order to point out differences and to show that the approach this presents a new and dynamic technique which is more powerful in characterizing groups of hotspots effectively. This section also provides experimental results on real numerical and predefined datasets. This facilitates the implementation with the characterization and crime type analysis of the data by using the logical scenario.

10. Conclusion

The key insight behind the proposed methods is identifying hotspots by searching, utilizing, and presenting patterns in geographic space. By preprocessing the crime related data sets into a transaction based geospatial dataset, the system develops a model, called HIT using primary direction with the impact of genetic approach, to map crime hotspots through the related variables. Then this introduces a similarity method to summarize the identified GDPatterns into clusters. Based on these clusters, a relevant report of crime hotspots and related variables is visually presented for domain experts. The proposed system presents a spatial data mining framework to study the spatial distribution of crimes through their related variables and provides visual alerts to the user. The system utilizes some new approaches regarding the crime detection, which use related variables in crime hot spot mapping with the effective of primary direction. In the framework the system address an iterative and inductive learning process to study the spatial properties of crime. In some cases the system uses self training process to identify the crime spot. Experiment results show that the proposed HIT model outperforms in precisely identifying crime hotspots in and out with datasets. Additionally, by using a similarity measure method, this demonstrates the characteristics and its accuracy can be well shown using probable efficiency algorithm.



CRIME DETECTION

crime_id	crime_zone	crime_type	area	service_x	service_y	datetime	keywords
1	Coimbatore	Drugs	tatabad	223	148	11/27/2013 4:20 ...	drug
2	Coimbatore	Robbery	avinashi rd	545	255	11/27/2013 4:21 ...	chain,robbery
3	Pollachi	Chain snatching	watertank	386	174	11/28/2013 11:1...	chain robbery
4	Coimbatore	Criminal damage ...	crosscut	207	238	11/28/2013 11:2...	crime
5	Coimbatore	Chain snatching	ramnagar	228	267	12/2/2013 5:29:...	near apartments
6	Coimbatore	Chain snatching	saibabacolony	64	98	12/2/2013 5:29:...	near apartments

Figure 5: Crime Detection

References

[1] Bifet, A. and Kirkby, R. 2009. Massive Online Analysis, Technical Manual, University of Waikato.

[2] Bolton, R. and Hand, D. 2001. Unsupervised Profiling Methods for Fraud Detection, Proc. of CSCC01.

[3] Brockett, P., Derrig, R., Golden, L., Levine, A. and Alpert, M.2002. Fraud Classification using Principal Component Analysis of RIDITs, The Journal of Risk and Insurance 69(3): pp. 341-371. DOI:10.1111/1539-6975.00027.

[4] Caruana, R. and Niculescu-Mizil, A. 2004. Data Mining in MetricSpace: An Empirical Analysis of Supervised Learning Performance Criteria, Proc. of SIGKDD04.DOI: 10.1145/1014052.1014063.

[5] Christen P. and Goiser, K. 2007. Quality and Complexity Measures for Data Linkage and Deduplication, in F. Guillet and H. Hamilton (eds), Quality Measures in Data Mining, Vol. 43, Springer, United States. DOI: 10.1007/978-3-540-44918-8.

[6] Cortes, C., Pregibon, D. and Volinsky, C. 2003. Computational methods for dynamic graphs, Journal of Computational and Graphical Statistics 12(4): pp. 950-970. DOI: 10.1198/1061860032742.

[7] Experian. 2008. Experian Detect: Application Fraud Prevention System. Whitepaper, http://www.experian.com/products/pdf/experian detect.pdf.

[8] Fawcett, T. 2006. An Introduction to ROC Analysis, Pattern Recognition Letters 27: pp. 861-874. DOI: 10.1016/j.patrec.2005.10.010.

[9] Goldenberg, A., Shmueli, G. and Caruana, R. 2002. Using Grocery Sales Data for the Detection of Bio-Terrorist Attacks, Statistical Medicine.

[10] Gordon, G., Rebovich, D., Choo, K. and Gordon, J. 2007. Identity Fraud Trends and Patterns: Building a Data-Based Foundation for Proactive Enforcement, Center for Identity Management and Information Protection, Utica College.

[11] Hand, D. 2006. Classifier Technology and the Illusion of Progress, Statistical Science 21(1): pp. 1-15. DOI: 10.1214/088342306000000060.

[12] Head, B. 2006. Biometrics Gets in the Picture, Information Age August-September: pp. 10-11.

[13] Hutwagner, L., Thompson, W., Seeman, G., Treadwell, T. 2006.The Bioterrorism Preparedness and Response Early Aberration Reporting System (EARS), Journal of Urban Health 80: pp. 89-96. PMID: 12791783.

[14] IDAnalytics. 2008. ID Score-Risk: Gain Greater Visibility into Individual Identity Risk. Unpublished.

[15] Jackson, M., Baer, A., Painter, I. and Duchin, J. 2007. A Simulation Study Comparing Aberration Detection Algorithms for Syndromic Surveillance, BMC Medical

Informatics and Decision Making 7(6).DOI: 10.1186/1472-6947-7-6.

[16] Jonas, J. 2006. Non-Obvious Relationship Awareness (NORA), Proc. of Identity Mashup.

[17] Jost, A. 2004. Identity Fraud Detection and Prevention. Unpublished.

[18] Kantarcioglu, M., Jiang, W. and Malin, B. 2008. A Privacy-Preserving Framework for Integrating Person-Specific Databases,

[19] Privacy in Statistical Databases, Lecture Notes in Computer Science, 5262/2008: pp. 298-314. DOI: 10.1007/978-3-540-87471-325.

[20] Kleinberg, J. 2005. Temporal Dynamics of On-Line Information Streams, in M. Garofalakis, J. Gehrke and R. Rastogi (eds),Data Stream Management: Processing High-Speed Data Streams,

[21] Springer, United States. ISBN: 978-3-540-28607-3.

[22] Kursun, O., Koufakou, A., Chen, B., Georgiopoulos, M., Reynolds,K. and Eaglin, R. 2006. A Dictionary-Based Approach to Fast and Accurate Name Matching in Large Law Enforcement Databases,Proc. of ISI06. DOI: 10.1007/11760146.

[23] Neville, J., Simsek, O., Jensen, D., Komoroske, J., Palmer,K. and Goldberg, H. 2005. Using Relational Knowledge Discovery to Prevent Securities Fraud, Proc. of SIGKDD05.

[24] Oscherwitz, T. 2005. Synthetic Identity Fraud: Unseen Identity Challenge, Bank Security News 3: p.7.

[25] Roberts, S. 1959. Control-Charts-Tests based on Geometric Moving Averages, Technometrics 1: pp. 239-250.

[26] Romanosky, S., Sharp, R. and Acquisti, A. 2010. Data Breaches and Identity Theft: When is Mandatory Disclosure Optimal?, Proc. of WEIS10 Workshop, Harvard University.

[27] Schneier, B. 2003. Beyond Fear: Thinking Sensibly about Security in an Uncertain World, Copernicus, New York. ISBN-10: 0387026207. Schneier, B. 2008. Schneier on Security, Wiley, Indiana. ISBN-10: 0470395354.

[28] Sweeney, L. 2002. *K-Anonymity*: A Model for Protecting Privacy, International Journal of Uncertainty, Fuzziness Knowledge-Based Systems: 10(5): pp. 557-570.

[29] Veda Advantage. 2006. Zero-Interest Credit Cards Cause Record Growth In Card Applications. Unpublished.

Author Profile

Saradha. R received her B.E in Computer Science and Engineering from Sengunthar Engineering College Tiruchengode, India in 2012 and is currently pursuing her ME in Computer Science and Engineering in Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India.

Deepika. S received her B. Tech in Information Technology from Kalasalingam University, Srivilliputhur, India in 2012 and is currently pursuing her M.E in Computer Science and Engineering in Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India.