An Optimal KD Model for Crime Pattern Detection Based on Semantic Link Analysis-A Data Mining Tool

C. Ram Singla, Deepak Dembla, Yogesh Chaba[1] and Kevika Singla[2]

CITM,Faridabad
[1]GJU S&T, Hissar
[2]Sapient, Gurgaon
E-mail : crslibra1010@rediffmail.com ; deepak_dembla@yahoo.com ;
yogeshchaba@yahoo.com ; kvk@rediffmail.com

ABSTRACT

Knowledge discovery in databases is a rapidly growing field, whose development is driven by strong research interests as well as urgent practical, social, and economical needs. In this paper we look at the use of link analysis in data mining to help detecting the crime patterns and speed up the process of solving crimes. In our study, we have seen that how this data mining tool in knowledge discovery process learns the typical behavior ('profile') of serial killers by applying link analysis, and helps in detecting the offender's movements. We looked at the use of data mining for identifying crime patterns using semantic link association algorithm. Our contribution here was to formulate crime pattern detection as machine learning task and to thereby use data mining to support police detectives in solving crimes. We identified the significant attributes and developed the scheme for weighting the significant attributes. Our modeling technique was able to identify the crime patterns from a large number of crimes making the job for crime detectives easier.

Keywords: *Knowledge discovery in databases, data mining, Crime-patterns, link analysis, visualization*

1. INTRODUCTION

Be it a satellite orbiting our planet, a medical imaging device, a credit-card transaction verification system, or a supermarket's checkout system, the human at the other end of the data gathering and storage machinery is faced with the same problem: What to do with all this data? The value of raw data is typically predicated on the ability to extract higher level information: information useful for decision support, for exploration, and for better understanding of the phenomena generating the data. Each year more operations are being computerized, all accumulate data on operations, activities and performance and all these data hold valuable information, e.g., trends and patterns, which could be used to improve business decisions and optimize success. Therefore, humans need assistance in their analysis capacity. This need for automated extraction of useful knowledge from huge amounts of data is widely recognized now, and leads to a rapidly developing market of automated analysis and discovery tools. Knowledge discovery and data mining are techniques to discover strategic information hidden in very large databases. Automated discovery tools have the capability to analyze the raw Analysts data and present the extracted high level information to the analyst and decision-maker, rather than the analyst finding himself or herself and thus acts a sort of amplifier of basic human analysis capabilities.

Data mining is a process concerned with uncovering patterns, associations, anomalies, and statistically significant structures and events in data[1].It is defined as the process of discovering meaningful new correlations, patterns and trends, often previously unknown, by sifting through large amounts of data, using pattern recognition, statistical and mathematical techniques. It can not only help us in knowledge discovery, that is, the identification of new phenomena, but it is also useful in enhancing our understanding of known phenomena. One of the key steps in data mining is pattern recognition, namely, the discovery and characterization of patterns in image and other high-dimensional data. A pattern is defined as an arrangement or an ordering in which some organization of underlying structure can be said to exist. Patterns in data are identified using measurable features or attributes that have been extracted from the data. Data mining is an interactive and iterative process involving data pre-processing, search for patterns, knowledge evaluation, and possible refinement of the process based on input from domain experts or feedback from one of the steps [5].

Knowledge discovery is defined as "the non-trivial extraction of valid, novel, potentially useful, unknown, and ultimately understandable patterns in data" .A clear distinction between data mining and knowledge discovery is drawn. The knowledge discovery process takes the raw results from data mining (the process of extracting trends or patterns from data) and carefully and accurately transforms them into useful and understandable information. Data mining is just single step in larger process known as Knowledge Discovery in databases(KDD)[3].Other steps in KDD process, in progressively order, include data cleaning, data integration, data selection, data transformation,(data mining),pattern evaluation, and knowledge presentation[6].

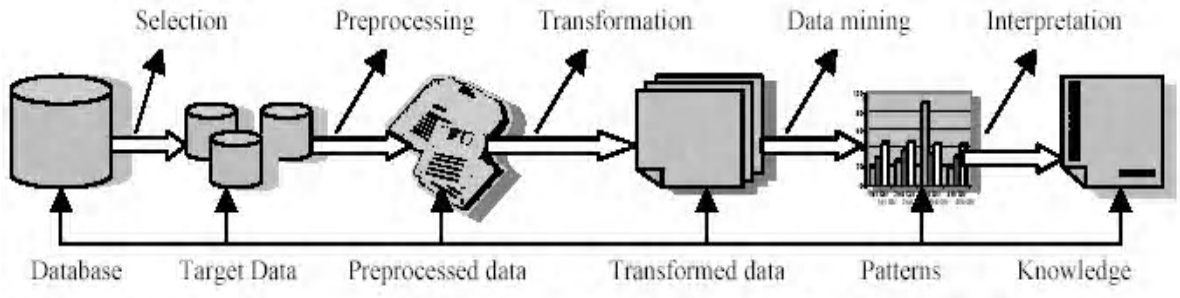
Data mining uses a variety of mathematical algorithms to analyze historical data. The results of this analysis are then used to build models based on real world behavior, which are in turn used to analyze incoming data and make predictions about future behavior.

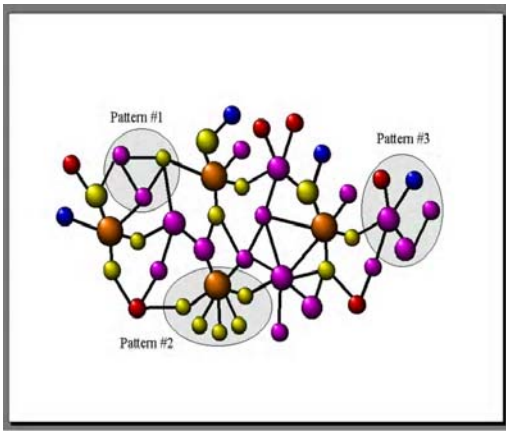
2. DATA MINING TASKS

At the core of KDD process are the data mining methods for extracting patterns from data. The methods can have different goals, dependent on the intended outcome of overall KDD process. Various data mining methods are *Data Processing, Prediction, Regression, Classification, Clustering, Link Analysis (Associations), Model Visualization, Exploratory Data Analysis (EDA), Time Series and Sequence Discovery[7].*

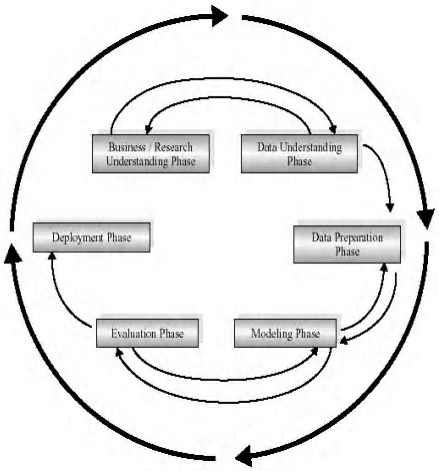
2.1 Link Analysis (Associations)

Given a set of data items, identify relationships between attributes and items such as the presence of one pattern implies the presence of another pattern. These relations may be associations between attributes within the same data item ('Out of the shoppers who bought milk, 64% also purchased bread') or associations between different data items ('Every time a certain stock drops 5%, a certain other stock raises 13% between 2 and 6 weeks later'). The investigation of relationships between items over a period of time is also often referred to as 'sequential pattern analysis' [10].





Can Find Unusual Patterns in the Network Structure



3. AN OPTIMAL KD MODEL FOR CRIME PATTERN DETECTION BASED ON SEMANTIC LINK ANALYSIS

Data Mining aims at discovering hidden instances of patterns of interest, such as patterns indicating an organized crime activity. An important problem targeted by this case study is identification of criminal networks, based on available intelligence. We present an approach using semantic link analysis and visualization of findings. The system finds the critical path of the serial killers who are striking over again and again and determines links between their crime activities locations occurred in the past, travel record, and background history, etc. These findings increase the chance of finding these repeat offenders.

The links as external memory aids, which are crucial tools for arriving an unbiased conclusion in the face of uncertain information. For example, if we have information that the offender selects Chiquita as the point of inception of crime and then commuting his criminal activities through Healdsburg and then through Windsor through to the Stony Point (as shown in the figure 1), then this Information aids in visualizing the connections for investigating the current and next movements of serial killers across the global networks[4].Criminal intelligence analysis therefore requires the ability to integrate information from multiple crime incidents or even multiple sources and discover regular patterns about the structure, organization, operation, and information flow in criminal networks[11].

Table 1: Sample Crime Example

Crime Type	Location	Victim Age	Victim Sex	Victim Profession	Weight Assigned
Murder	A	30	F	Contractor	60
Murder	B	27	M	Contractor	20
Murder	C	31	F	Architect	55
Murder	D	28	F	Architect	57
Murder	E	29	M	Engineer	30

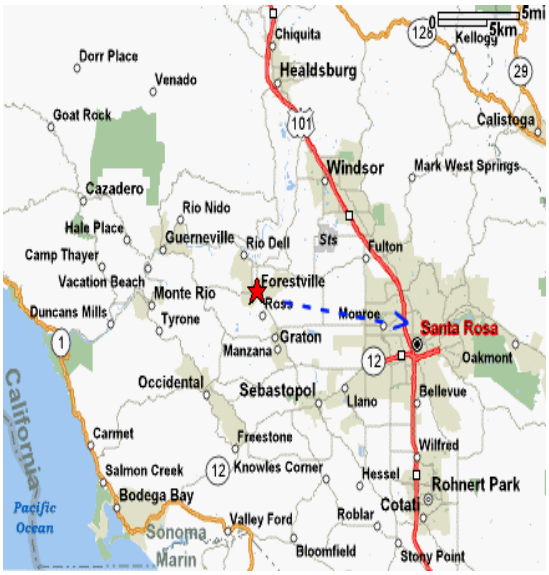


Fig. 1: Map

Based on the criminal records in the various locations and its association with different attributes as mentioned in above table, we calculated the similarity between each pair of location to form a similarity matrix. Then, this matrix was fed to a multidimensional scaling (MDS) algorithm [2] which generated a two dimensional graph of the location link structure. The proximity of nodes (locations) in the graph reflects the similarity level. Figure 2 shows the visualization of the criminal activity link structure. Interestingly, domain experts recognized the existence of six patterns representing linked locations in the crime network. On the left side of the network resides the pattern, Crime network 'E'. It signifies the most likely killer area.

Fig. 2: Instance of a selected region

4. RESULTS OF DETECTING CRIME ACTIVITY PATTERN

The model finds the links between various locations of criminal activities and is capable of determining the path occurred in the past, their affiliations with the offender, victim's profession, victim's age etc. Visualizing patterns can lead to a better understanding of the underlying criminal network areas [8]. In addition, the visualization serves as a tool for showing the relationships between various linked areas which can be killer, non-killer or unknown as shown in the figures 3 below. Furthermore, it helps foretell likely attributes of the victims which are likely to be affected by the killer in future in the specific area (based on the identified pattern).This identified pattern is the most likely targeted area by the killer for a particular group of people.

The crime data is investigated and the link analysis data mining technique in the knowledge discovery uncovers hidden instances of patterns of interest such as pattern indicating crime activity. These patters can also obtain non-killer areas data set and unknown data set which after further analysis and visualization conclude the most likely killer area and most likely a non-killer area**.**

5. CONCLUSIONS AND FUTURE DIRECTIONS

KDD is a rapidly expanding field with promise for great applicability. Knowledge discovery purports

Fig. 3: Visualization Scenario

to be the new database technology for the coming years. The need for automated discovery tools had caused an explosion in the number and type of tools available commercially and in the public domain. It is anticipated that commercial database systems of the future will include KDD capabilities in the form of intelligent database interfaces [9]. Some types of information retrieval may benefit from the use of KDD techniques. Due to the potential applicability of knowledge discovery in so many diverse areas there are growing research opportunities in this field.

We looked at the use of data mining for identifying crime patterns using semantic link association algorithm. Our contribution here is to formulate crime pattern detection as machine learning task and to thereby use data mining to support police detectives in solving crimes. We identified the significant attributes and their association and developed the scheme for weighting the significant attributes. Our link based modeling technique is able to identify the crime location patterns from a large number of crime locations making the job for crime detectives easier. Visualizing the various patterns, facilitates the analysis of killer locations and paves the way for more refined microscopic content analysis of these areas. In this way, we have created a model for predicting the crime hot-spots that will help in the deployment of police at most likely places of crime for any given window of time, to allow most effective utilization of police resources.

Apart from countering the crime locations, future work will explore fraud detection studies and detecting the frauds like credit application fraud detection by using this model of data mining in knowledge discovery process.

REFERENCES

1. Han J. et Kamber M. (2002), Data Mining: Concepts and Techniques,Morgan Kaufmann Publishers, Canada, 2002.
2. Agrawal R., Gupta A., Sarawagi A., "*Modeling Multidimensional Database*s", ICDE'97 pages 232-243. IEEE Press, 1997.
3. Fayyad U.M., Piatetsky-Shapiro G., Smyth P. (1996), « From Data Mining to KDD : an overview », AAAI/MIT Press, 1996.
4. Bedard, Y., 2002, Geospatial Data Warehousing, Datamart and SOLAP for Geographic Knowledge Discovery, Université de Muenster,Germany, 2002.
5. Wang, F. Pan, D. Ren, Y. Cui, D. Ding, et W. Perrizo (2003), Efficient olap operations for spatial data using peano trees, 8th ACM SIGMOD Workshop on Research Issues in Data Mining and KnowledgeDiscovery, 2003.
6. Fayyad, U. Data Mining and Knowledge Discovery: Making Sense Out of Data. IEEE Expert, v. 11, no. 5, pp. 20-25,October 1996.
7. Goebel, M., and Gruenwald, L. A Survey of Knowledge Discovery and Data Mining Tools. Technical Report, University of Oklahoma, School of Computer Science,Norman, OK, February 1998.
8. Brachman, R. and T. Anand. The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In *Advances in Knowledge Discovery and Data Mining*,ed. U.. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI/MIT Press, 1996.
9. Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery in Databases (a survey), AI Magazine, 17(3): Fall 1996, 37-54 .
10. Jain, A.K., Murty, M.N., Flynn, P.J. (1999) Data Clustering: A Review, *ACM Computing Surveys*, 31, 3:264-323.
11. Lemos, R. (2002) What are the real risks of cyberterrorism?,*ZDNet*, URL:http://zdnet.com.com/2100-1105-955293.html.