Study on Crime Busting using Mathematical Model

Xiufen Wang, Fangfang Dou,Wenwen Mao, Zhihong Ma
(Tianjin Agricultural University, Tianjin 300384)

Abstract: In the paper we investigate a plot to crime. We use statistical analysis of basic principle and related constructs criminal hunt models, and statistics of each node and the number of suspected nodes correspond to . Then talking about the topic number it can be. With the help of correlation analysis and correlation coefficient, we obtain a suspect weight node and doubt topic, calculates scores of each node. After that, we select integral before 14 as a conspirator. Our model and method successfully determine priorities and scope of suspicious nodes ， for the sake of semantic network analysis and text analysis, needs further optimized. Last, an evaluation of the model developed in this paper is given, listing its advantages and limitations, and providing suggestions on measuring its performance.

Key words：Statistics, Relevance, Weight, Priority, Index

1. Introduction

ICM is investigating a plot to crime. Investigators are clearly involved in the plot, hoped to be able to determine to the other members and their leaders who were arrested. Criminals and criminal suspects are in a big company which is a comprehensive work in the office. ICM recently discovers a fraction of the information, contains 82 employees, the company believes that it can help them find the most likely unknown partner and leadership. Modeling of the goal is to determine the comprehensive office who was the most likely criminal. Column a sequence list is helpful to the supervision of the ICM, and further interrogation. Clear division of accomplice and the accomplice can also help us differentiate each group of people.

Before giving the case data, director gives her works in other city a few years ago . the solution of this kind of circumstance (EZ) investigation included. She says it is a little simple example, and will be helpful to the task.

Problem 1: Refer to the questions given in the case analysis of accomplice and new situation, the list of the accomplice, suspected of the text of the message code, names. XLS, switchable viewer.XLS, Messages. The information in the XLS. Analysis belongs to the possibility of a conspiracy to size model and algorithm, the new situation of 83 nodes (people) prioritize, and explains the model and indicators.

Problem 2: If there is new information to determine the topic 1 is associated with crime, and Chris is one of the conspirators, then the priority list will be how to change?

Problem 3: The file switchable viewer. XLS description to the topic of "dialogue", whether to use the "capabilities (semantic network, text analysis) to improve model?

Problem 4: Asked to report to include "a deeper message content network, semantics and text analysis how to model and Suggestions helpful" this discussion. The model in other aspects of the role and promotion.

2.Model Assumptions and Symbols

2.1 Model Assumptions

1） he hypothesis, the relationship between the conspirators are equivalent position no weight;

2） The assumptions are equivalent relationship between the three suspected topic, weight is the same size;

3） Note 37 and note7 represent different people;

4） Contact the suspect node number can represent the close degree and the accomplice, talking about how many suspect events can represent involving frequency size;

5） The contact frequency does not affect the suspect node degree of suspicion.

2.2 Symbols

Table 1 Symbols

mark	representation
A	The number of contact suspect nodes
$A1$	In addition to the suspect nodes of node and doubt contact the
$A2$	Doubt that node and other nodes contact number
M	Doubt the number of nodes
N	Doubt that the number of events
Z	The comprehensive weights
$VAR00009$	83 nodes (question 1)
C	Doubt node weights
D	The weight of suspected event
$VAR00010$	Doubt node (question 1)
$VAR00011$	Suspected incident (question 1)
$C1$	Doubt node weights (question 1)
$D1$	The weight of suspected event (question 1)
$VAR00018$	83 nodes (question 2)
$VAR00016$	Doubt node (question 2)
$VAR00017$	Suspected incident (question 2)
$C2$	Doubt node weights (question 2)
$D2$	The weight of suspected event (question 2)
r	The correlation coefficient
x	M or N
y	The representation of a given event 1, 0, 1

3. Modeling and Solution

3.1 Modeling and Solution to Problem I

Relationship is a no deterministic relationship, the correlation coefficient is the linear correlation between the amount of research variables. Due to the different research objects, there are several definitions of correlation coefficient.

$$r=\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\bullet\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}}$$

Simple correlation coefficient: also called the correlation coefficient and the linear correlation coefficient, generally represented by the letters P, used to measure the linear relationship between two variables.

The multiple correlation coefficient : also called multiple correlation coefficient. The multiple correlation is the correlation between variables and multi variables. For example, some kind of commodities between seasonal quantity demanded and the price level, the income level of workers and other phenomena of complex correlation.

The canonical correlation coefficient: is the first original each variable principal component analysis, comprehensive index of the new linear relation, the linear correlation coefficient between comprehensive index to study the correlation between the original variables between groups.

Theorem: Necessary and sufficient conditions | ρ XY | = 1, there exists a constant a, b, such that P{Y=a+bx}=1

Correlation coefficient ρ XY value between -1 and 1, ρ XY = 0, X, Y are not related;

| ρ XY | = 1, X, Y complete correlation, at this time, X, has a linear relationship between Y

| ρ XY | < 1, X changes caused some changes in the absolute value of Y, ρ XY is bigger, X changes caused by change of Y is greater, | ρ XY | > 0.8 called highly relevant, when, that is | ρ XY| < 0.3, called the low correlation, the other was moderately correlation.

In view of the text messages which are qualitative material, so we need to the rightness quantization process. In order to determine the size of the possibility of each node (people), we count each node (people) contact frequency with the criminal and various nodes (people) the frequency of a suspected conversation. At the same time, according to the following formula:

$$\mathbf{A2}=\frac{A^{*}M}{M-1}$$

Got 83 nodes of the corresponding data in annex 1

For different node (person), I with 1, 0, 1 on behalf of the accomplice, unknown, the accomplice， Then use SPSS software to calculate the correlation coefficient, as a comprehensive weight[1]. The following table 2:

Table 2 relevance weights

Control variables			VAR00011	VAR00010
VAR00009	VAR00011	correlation coefficient	1.000	.680
		Significant (both) sides)	.000	.000
		df	.0	.80
	VAR00010	correlation coefficient	.680	1.000
		Significant (both) sides)	.000	.
		df	80	.0

According to $Z = M * C + N * D$ so as to obtain comprehensive parameters of the two variables and sorted according to their priorities. The following table 3:

Table 3 83 nodes sorting and comprehensive weights

note	Weighted score	note	Weighted score	note	Weighted score	note	Weighted score
21	9.3956	15	1.96	60	1	78	0.32
67	9.3956	28	1.96	69	1	79	0.32
7	7.96	37	1.7556	29	0.96	82	0.32
54	7.32	24	1.68	32	0.96	26	0
43	7	45	1.68	41	0.96	52	0
18	5.5644	19	1.64	25	0.68	53	0
49	5.2444	38	1.64	66	0.68	55	0
81	4	50	1.64	5	0.64	58	0
48	3.68	3	1.6	8	0.64	59	0
40	3.36	1	1.36	9	0.64	61	0
20	3	6	1.32	11	0.64	62	0
10	2.96	30	1.32	42	0.64	63	0
34	2.64	31	1.32	80	0.64	64	0
13	2.28	33	1.32	12	0.32	68	0
16	2.28	36	1.32	23	0.32	70	0
17	2.28	46	1.32	39	0.32	71	0
0	2.04	65	1.32	51	0.32	73	0
27	2	4	1.28	56	0.32	74	0
35	2	22	1.28	57	0.32	76	0
44	2	47	1.28	72	0.32	77	0
2	1.96	14	1	75	0.32		

After get the order, in order to delimit the suspect node scope for all nodes in the cluster analysis, we divide it into three categories as comprehensive weight above doubt node scope, not doubt node of other scope. Using the SPSS cluster analysis results in annex 2[2], Specific classification as shown in figure 1:

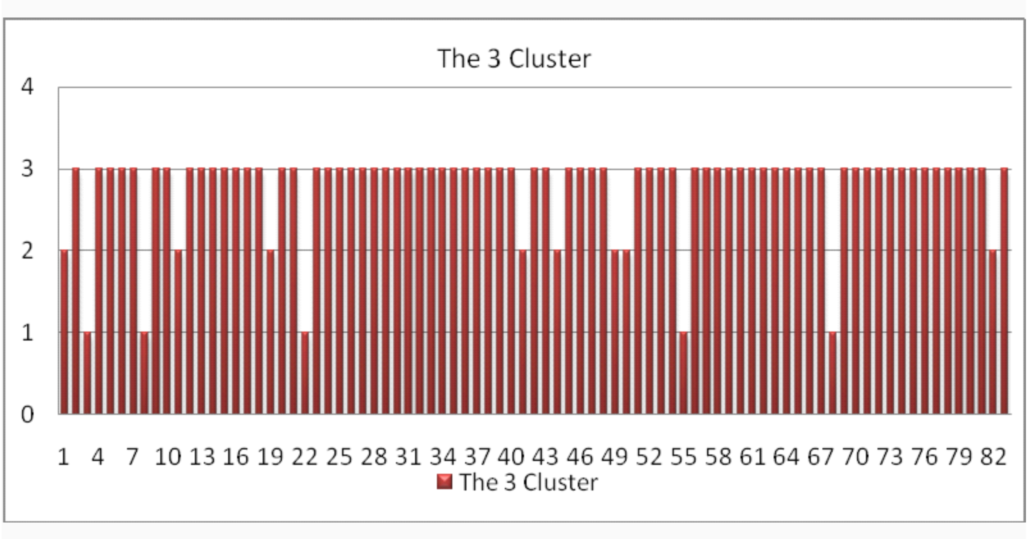

Figure 1 The classification of the accomplice

This article selects the nodes and the known criminal relationship between the frequency of the frequencies. Get a suspected conversation as a important index to identify accomplices. At the same time in order to determine the two different weights of the indexes in the identification process, we use SPSS to the correlation between the data processing, and using the correlation coefficient as the weight. Using excel software to calculate various nodes (people) scores, in the cluster analysis of accomplice.

3.2 Modeling and Solution to Problem II

On the premise of problem one adds a questionable subject and a conspiracy, but these will not lead to the change of the model and algorithm.

In question 1 on the basis of ontology is just on the conditions changed, the same model is established in this paper. First of all , the various nodes with doubt and suspicion node number associated statistics, will doubt the node at the same time. The measured nodes and the suspicious nodes with 1, 0, 1, was in the application of SPSS software to calculate the suspect events and suspicion nodes. The correlation coefficient is the weight of these two variables, specific see appendix 3, the correlation coefficient to see the following table 4:

Table 4 correlation

Control variables			VAR00017	VAR00016
VAR00018	VAR00017	correlation	1.000	.695
		Significant (both)	.	.000
		df	0	80
	VAR00016	correlation	.695	1.000
		Significant (both)	.000	.
		df	80	0

According to $Z = M * C + N * D$ can get 83 comprehensive weighting of each node, question 2 is actually in the consideration question 1 model is in line with the change in conditions change, after proper affected by variables rather than the result of the constant has identified. Then the comprehensive weight of each node according to the 83 order to get the following table 5:

Table 5 83 nodes sorting and comprehensive weights

note	Weighted score	note	Weighted score	note	Weighted score	note	Weighted score
21	15.1412	14	2.305	46	1.305	55	0.305
67	10.44	27	2.305	65	1.305	57	0.305
54	8.61	44	2.305	69	1.305	62	0.305
7	8.22	24	2	41	1.22	63	0.305
2	7.475	35	2	60	1	64	0.305
43	7.2077	45	2	66	1	71	0.305
18	6.2927	28	1.915	68	1	72	0.305
49	5.4934	31	1.915	8	0.915	75	0.305
40	4.39	50	1.915	9	0.915	78	0.305
48	4.305	22	1.83	11	0.915	79	0.305
81	4.305	37	1.7073	29	0.915	52	0
0	3.6038	1	1.695	5	0.61	53	0
10	3.525	6	1.61	12	0.61	58	0
20	3.305	19	1.61	42	0.61	59	0
34	2.915	38	1.61	56	0.61	61	0
17	2.83	4	1.525	80	0.61	70	0
13	2.525	47	1.525	82	0.61	73	0
15	2.525	25	1.305	23	0.305	74	0
16	2.525	30	1.305	26	0.305	76	0
32	2.525	33	1.305	39	0.305	77	0
3	2.44	36	1.305	51	0.305		

According to the clustering analysis in problem 1, 83 nodes can be obtained before the sort of 13 nodes is divided into doubt range, in order to compare problem 1 and 2 conditions changed, the change of priorities can have a clear comparison in the table below.

Table 6 The different of problem 1 and problem 2

problem 1	21	67	7	54	43	18	49	81	48	40	20	10	34
problem 2	21	67	54	7	2	43	18	49	40	48	81	0	10

It can be seen from the above table priority listing problems 1 and 2 suspicion nodes, great changes have taken place in the model can be drawn from the adaptability. It is better, it is not a dust is changeless, and it will change according to the condition of random change.

3.3 Solution to Problem Ⅲ

Semantic network is a directed graph, the concept of vertex said, while the side said the semantics of the relationship between these concepts. Semantic network is used to express complex concepts and their mutual relations, thus forming a semantic network composed of nodes and arcs describe figure.

Text analysis is refers to the representation of a text and its feature selection; Text analysis is a fundamental problem of text mining, information retrieval, it abstracted from text to represent text to quantify of key information. Then from the structure of the original text into a structured computer. we can identify the processing of information, namely to scientific abstraction of text, establish its mathematical model, a term used to describe and replace text. It allows the computer to realize the recognition of the text, based on this model calculation and operation [3].

The file switchable viewer. XLS description to the topic of "dialogue", as in the title has explicitly given suspected event specific code, directly in contact with these topics as the quantitative standard, does not use these "capabilities (semantic network, text analysis) to improve the model.

3.4 Solution to ProblemⅣ

This model does not reflect to the message content in the network, semantics and text analysis how to model and suggestions to help, but if you can in a text message for message content and graphics analysis of network, semantics and text analysis will strengthen the accuracy of model. Because by directed line segment of a graphic can deduce some suspicious nodes. Some are Wu Duan nodes, through the analysis of the key words in a text message and statistical can deduce more suspect number of events, so that we can improve the accuracy of the models.

After the establishment of the model, we aim to quantify the text information to find the main influence factors in statistics. And determine the correlation between the invariant and variable. Get the weight to calculate the comprehensive weight sorting of each node. Designated limits suspect, the application of clustering analysis method to the solution of the objective scope of the draw[4].

The entire model establishment of logical thinking and able to solve some objective factors of network database is used to identify (sure), prioritize problems. After the clustering analysis of classification of similar node, more accurate classification is not influenced by artificial factors. This model can be generalized to other scheduling problems, such as: when there is a [shows that infection node and have identified some infections] various images or chemical data, this model can be combined to calculate the relevant factors in the comprehensive weights and sorting method of accurate found in biological networks of infected or diseased cells.

4. conclusion

Strengths

1) The model adaptability is better, it can adjust the results according to the change of variables;
2) This model is powerful generalization, it can use in many ways;
3) The model text messages can be quantified from qualitative

quantitative change, and problem solving is convenient, simple;

4) The same kind of quantitative expressed in the same Numbers can get to calculate the correlation coefficient, and increase the accuracy.

Weaknesses

1) Without taking into account of all the information in the topic;
2) Select variables is a bit subjective;
3) Not using the subject of the semantic network analysis and text analysis;
4) The model is not the results of inspection.

References

[1].ShaoKai Ni.Seven kinds of determine review refers to the right than the weight method. South China J Prew Med,dec2002.

[2].Xiao Mei Zou,Chun Bo Xiu.Research on Factors Related to Crime Rate Based on Cluster Analysis, Forum on contemporary law,2010.

[3].LI Ming-Qin, LI Juan-Zi, Wang Zuo-Ying, Lu Da-Jin. Semantic Analysis and Structured Language Models, Journal of Software, 2005.

[4].Fang Pan, Zi li Zhang.Criminal Network Analysis Based on Fuzzy Hierarchical Clustering,Journal of Southwest China Normal University (Natural Science Edition), Jun 2009.