Modified Page Rank Model in Investigation of Criminal Gang

Ting LU
Research Center for Cloud Computing
North China University of Technology,
Beijing 100043, China

Qi Gao
Department of Mathematics,
Shandong Normal University,
Jinan 250002 China.

Xudong Wang
Department of Information System,
Shanghai University,
Shanghai 201800,China

Cong LIU
Department of Computer Science, Shandong
University of Science and Technology,
Qingdao 266590, China

Abstract:In this paper, we consider a criminal investigation on the collective guilt of part members in a working group. Assuming that the statistics we used are reliable, we present the Page Rank Model based on mutual information. First, we use the average mutual information between non-suspicious topics and the suspicious topics to score the topics by degree of suspicion. Second, we build the correlation matrix based on the degree of suspicion and acquire the corresponding Markov state transition matrix. Then, we set the original value for all members of the working group based on the degree of suspicion. At the last, we calculate the suspected degree of each member in the working group. In the small 10-people case, we build the improved Page Rank model. By calculating the statistics of this case, we acquire a table which indicates the ranking of the suspected degree. In contrast with the results given in this issue, we find these two results basically match each other, indicating the model we have built is feasible. In the current case, firstly, we obtain a ranking list on 15 topics in order of suspicion via Page Rank Model based on mutual information. Secondly, we acquire the stable point of Markov state transition matrix using the Markov chain. Then, we build the connection matrix based on the degree of suspicion and acquire the corresponding Markov state transition matrix. Last, we calculate the degree of 83 candidates. From the result, we can see that those suspicious are on the top of the ranking list while those innocent people are at the bottom of the list, representing that the model we have built is feasible. When suspicious topics and conspirators changed, a relatively good result can also be obtained by this model. In the current case, we have the evidence to believe that Dolores and Jerome, who are the senior managers, have significant suspicion. It is recommended that future attention should be paid to them. The Page Rank Model, based on mutual information, takes full account of the information flow in message distribution network. This model can not only deal with the statistics used in conspiracy, but also be applied to detect the infected cells in a biological network. Finally, we present the advantages and disadvantages of this model and the direction of improvements.

Keywords: Page Rank ;Markov Chain; Criminal Gangs; Biological Model; Algorithm; Markov Chain

1. INTRODUCTION

1.1 Background

With the advent of the information age, information technology development has brought new challenges and opportunities to the public security work. Because of its superiority, people are paying close attention to how to integrate and reuse to the public security data resources and how to achieve comprehensive application of global data, the automatic association of all kinds of information as well as the automatic mining function of a variety of cues, which will provide integrated applications in the information handling. Using the semantic network analysis and the relevant models plays a major role in improving the efficiency and accuracy of detection.

1.2 What is “Crime Gangs”?

Crime Gangs is a crime form, referring to the two or more persons jointly committing one or more crimes. Intelligence information (such as the collection of the relevant people’s conversations in this given case) is the technical basis of the intelligence analysis. The implementation of criminal intelligence analysis is based on intelligence gathering, mining and establishing crime patterns, analyzing and judging criminal clues as well as a variety of relationships. Building an efficient model to analyze the relevant information will surely provide some useful clues for the detection of cases and the judgment of criminal suspects.

1.3 The Mission of Our Team

In this paper, we would like to introduce the theory of information analysis to solve the problems, which is described as follows: In the previous case, it is our team’s duty to identify the guilty parties among the 10 candidates. Similarly, the other problem is to solve a relatively larger case, where there are about 83 candidates.

The analysis and process of these problems are divided into three steps. Firstly, we use the method of mutual information to score all the topics mentioned. Secondly, we can acquire the correlation matrix by calculating the score of the associated topics among the candidates. Thirdly, by solving the Markov Transfer Matrix, we get the stable point of Markov Transfer Matrix, which is the final index of 83 candidates. By analyzing this index, the 83 degree of suspicion can be determined.

2. PROBLEM ANALYSIS

2.1 Topics’ Suspicious Degree Analysis

First, we analyze the suspicious degree of talks between each other, using the mutual information and entropy in information theory, defined as M(A,B) = I(A) + I(B) - I(A,B), to obtain the scores of the 12 undetermined topics and the scores of suspicious ones(Obviously, the scores of the suspicious are supposed to be higher than the others). Further illustration is put in models.

2.2 The Transmission of Suspicious Information in the Working Group Analysis

The work group is divided into five parts, namely, the confirmed suspects, the people with some suspicion, the undetermined people and the innocent (listed by the degree of suspicion). Because we refer to Google's PangeRank algorithm, we need to obtain the correlation matrix of this group of people first. Then we can use this algorithm to get the Markov Transfer Matrix and later get the stable points using the nature of Markov. The acquisition of the correlation matrix is based on the message topics among the 83 people. We can use g_{ij} to represent the total scores of all the topics, thus getting an 83*83 correlation matrix.

2.3 The Initial Vector Distribution of the Working Group Analysis

To solve the Markov stable point, we need to use the initial vector distribution. We define the scores of the suspicious as α (The total number of this group of people is n); the scores of the innocent is 0; the scores of the undecided is β (The total number of this group of people is m). They satisfy the equation nα + mβ =1, and then we can decide the values of α and β through the search of α with step size of 0.1

3. BASIC ASSUMPTIONS OF OUR MODEL

3.1. All data is true and reliable.
3.2. There is valid evidence to prove the people who are suspicious.
3.3. There is valid evidence to prove the people who are innocent.
3.4. The suspicious topics have higher degree of doubt than that of the undermined ones.
3.5. An objective and subjective attitude is kept in the process of handling the cases.
3.6. These topics give a valid summary of the messages.

4. MODEL AND SOLUTION

Our whole model is divided into 3 parts, including 3.1, 3.2 and 3.3. In the 3.4 part, we use the whole model to analyze a simple example. And in 3.5 and 3.6, we solve the real problem presented in question 1, and its changed version in question 2.

4.1 Mutual Information Algorithm

Let me put a simple example to better illustrate this algorithm.
"What a good weather today! I would like to go fishing." A says to B. The words, such as "today", "weather", "good", "I", "fishing" involved in this conversation, can be seen as key words, and we can use $a_1, a_2, \cdots, a_n$ to represent them. When B says to A:"Yeah, really good weather! I will go climbing." Also, the words such as "weather", "hiking", "good", "I", can be extracted as key words and we can use $b_1, b_2, \cdots, b_m$ to represent them. Here, we list the relationship between the words of these two people.

Table 1: the language relationship table

A	today	good	fishing	weather	I		
	1	1	1	1	1	0	0
B				weather	I	good	climbing
	0	0	0	1	1	1	1

Notes: "1" in the second row and second column of the table rep resents the key word "today" said by A, but "0" in the fourth ro w and second column represents that B didn't say the key word "1"indicates the value of mentioned topic in the chat, "0"indicates the value of topic which is not mentioned in the table.

We can use the formula of mutual information M(X,Y) = I(X) + I(Y) - I(X,Y)to get the value of the mutual information, defined as M(X,Y), where I(X) represents the entropy from A, I(Y) represents the entropy from B.

$$I(X) = -P_0 \log P_0 - P_1 \log P_1 \qquad (4.1\text{-}1)$$

$$I(Y) = -P_0 \log P_0 - P_1 \log P_1 \qquad (4.1\text{-}2)$$

$$I(X,Y) = -P_{10} \log P_{10} - P_{11} \log P_{11} - P_{01} \log P_{01} \qquad (4.1\text{-}3)$$

P_0 and P_1 denoting the probability of 0 and 1 respectively; P_{10}, P_{11} and P_{01} denoting the probability of (1, 0), (1, 1) and (0, 1) respectively.

4.2 Markov Chain Model

In order to get the Markov state transition matrix, we first need to get the 83*83 Markov correlation matrix(rows represent the person's words, columns represent the words that others said to him /her). According to each topic, sum up the weight of the words involved in the conversation between one person and the others so as to get the total weight of one person, in the same way, we can come to the other people's weight, thus getting an 83*83 Markov correlation matrix.

Although we use Google's Page Rank model，some people are not the information receivers in the case, so it is unavoidable that the sum of a few columns in the association matrix is 0. Taking this into account, when computing the Markov Transfer Matrix, we do some special handling on the correlation matrix to make the sum of all the columns of the correlation matrix non-zero.

Define the Markov Transfer Matrix as $A = (a_{ij})$, then

$$c_j = \sum_i g'_{ij} \qquad (4.2\text{-}1)$$

$$r_i = \sum_j g'_{ij} \qquad (4.2\text{-}2)$$

C_j represents the sum of columns in matrix $G'(g_{ij})$, r_i represents the sum of rows in matrix $G'(g_{ij})$, g_{ij} represents the elements in the correlation matrix $G'(g_{ij})$.

The formula used to solve the elements in transfer matrix A is

$$a_{ij} = \frac{(1-d)}{n} + d * \frac{g'_{ij}}{c_j} \qquad (4.2\text{-}3)$$

d is a model parameter, we choose to define d = 0.85 based on experience. a_{ij} represents the summed scores of the message topics involved in the conversation from person j to person i. From the basic nature of the Markov chain, there is a stationary distribution in the regular Markov chain $x = (x_1, x_2, \cdots, x_N)^T$,then we can get the sorting in accordance with the conspiracy possibility.

4.3 Improved Page Rank Model

As we have known，there are 7 conspirators among the 83 people in the second case and each person's scores can be defined as $P(i), i = 1, 2, \cdots, 7$. And the scores of each innocent person can be defined as P(j), j = 1,2,...,8. The model is as follows:

$$\max \quad M(\lambda, s) = \sum_{i=1}^{7} P(i) - \sum_{j=1}^{8} P(j) \qquad （4.3\text{-}1）$$

$s.t.$

$$Ax = x \qquad (4.3\text{-}2)$$

$$\sum_{i=1}^{n} x_i = 1 \qquad (4.3\text{-}3)$$

The objective function indicates the maximum probability gap between the known conspirators and the innocent (All the conspirators are put in the upper part of the list, and all innocence in the lower part of the list, leaving the undecided people located in the middle of the list). We use k to express the maximum of mutual information in the 12 topics. λ represents the steps from the upper bound to 20. Through the debugging and analysis, we find that when λ is given a value of 0.5, the results will be much more in line with the subject requirements. s is the initial vector distribution, the percentage of each conspirator is 0.1, and the percentage of each innocent person equals to 0, so the undetermined equally allocate the remaining percentage. We define that s ranges from 0.5 to 0.9 is distributed in steps of 0.1, and we can get an initial vector distribution on the subject on the whole.

4.4 Solution of The 10-people Case

In practice, we conducted some exploratory analysis on the previous case, and get 4 ranking list of the probability of all the people when defining k =1，10，12，15，17，20，respectively. Through comparative analysis, we find that when k=17, the ranking we get will be more in line with the true outcome of the case than other values, that is, people like Inez don't get off, people like Carol are not falsely accused, and people like Bob do not have the opportunity to get reduced sentences. When k=17, we can obtain a ranking list of suspicious degree, listed as follows:

Table 2：The Sorting Table of the degree of Suspicious (k=17)

Ranking	Number	Name
1	7	George
2	2	Bob
3	9	Inez
4	4	Dave
5	5	Ellen
6	1	Anne
7	8	Harry
8	10	Jaye
9	3	Carol
10	6	Fred

Conclusion Analysis:

(1)George and Dave are the given co-conspirators, ranking fourth in the list. Meanwhile, Bob ranks second, illustrating he is one of the co-conspirators, so he cannot get off. Inez ranks third, meaning that we can also catch him.

(2)Jaye and Anne are both ranked relatively rearward, which is in accordance with the given information. And based on Carol ranking No.9, we can basically determine her innocence. Overall, the solution derived from our models can meet the actual requirements and have some promotional value.

From the previous analysis, we know that the 10-people case is a microcosm of the current case, and the results of small case subject has been given to us, so understanding and solving the smaller case can be seen as a test of our models, which may help us to modify them.

4.5 The Solution of the Current case for Problem 1

When conducting an exploratory analysis on the current case, we used the same way as the first one to analyze. We found that when k=17 or k=20, we can get two basically the same probability rankings of all the people, much more in line with the requirements listed in the subject.

First, we can obtain the ranking list of the 15 topics using the Mutual Information Method.

Table 3：Score Ranking List of 15 Topics

Number	Score
1	0.883016
2	0.857865
3	0.898072
4	0.87294
5	0.872976
6	0.857882
1	1
8	0.888021
9	0.883014
10	0.893037
11	1
12	0.888021
13	1
14	0.888034
15	0.883008

Notes: The numbers of suspicious topics are 7, 11, 13.

The suspicious degree ranking of the suspected, innocent and the senior managers based on k=17.

Table4: The Sorting Table Based on the Suspicious Degree of the Conspirators, Non-conspirators and Senior Managers (K=17)

Ranking	Name	Number
1	Yao	68
2	Alex	22
4	Paul	44
6	Harvey	50
7	Dolores	11
8	Ulf	55
17	Jerome	17
19	Jean	19
27	Paige	3
28	Elsie	38
31	Darlene	49
46	Chris	1
57	Tran	65
58	Ellin	69
60	Gretchen	5

Notes: The senior managers are Dolores, Jerome and Gretchen.
Conclusion Analysis:

(1) By analyzing the result, we can find that two senior managers, namely Dolores and Jerome are both conspirators.

(2) The already known conspirators are shown in the forefront of the ranking list, we can basically draw that they are the conspirators.

(3)The already known non-conspirators are all at the rear of the ranking list, except for Paige who ranks 27, the others can be determined to be innocent. Although there are a few minor differences, this result has been very much in line with the actual situation, which means the feasibility and stability of the model is very well.

4.6 The Solution of the Current Case for Problem 2

As to the second requirement, we can also use the Improved Page Rank Model to analyze data. First obtain a modified ranking list of the 15 topics, listed as follows:

Table 5：Score Ranking List of the 15 Modified Topics

number	score
1	1
2	1.177737
3	1.20412
4	1.193429
5	1.172717
6	1.167063
7	1
8	1.199102
9	1.193429
10	1.20412
11	1
12	1.20412
13	1
14	1.173339
15	1.193429

Then we can get the ranking list about the suspicion degree of the suspected, the innocent and the senior managers.

Table 6：The Sorting Table Based on the Suspicious Degree of the Conspirators, Non-conspirators and Senior Managers

Number	Name
1	Yao
2	Alex
4	Paul
6	Harvey
7	Dolores
8	Ulf
17	Jerome
19	Jean
27	Paige
28	Elsie
30	Darlene
45	Chris
57	Ellin
58	Tran
60	Gretchen

Notes: The senior managers are Dolores, Jerome and Gretchen, the 7 suspected people are Yao, Alex, Paul, Harvey, Ulf, Elsie.
Conclusion Analysis

The newly added conspirator Chris does not make much difference in results compared with the ones in the first requirement. According to analysis, this may be due to Chris's inactive involvement in the conversation, so even if he is one of the conspirators, his influence on others is very weak, resulting in the basically similar results.

4.7 The Solution of the Current Case for Problem 3

In this section, we introduce the conception of Semantic network and apply the theory of Similarity to the calculation of topic scores.

As we have obtained the original messages, which is a much bigger corpus than that we used in the first two questions, so we can apply the similarity analyze into our model. To describe it in detail:

First: Calculate the similarities of among the 3 suspected messages(A) and the 12 unsuspected messages (B),

A includes A1，A2，A3，B includes，

$$Sim(A_i, B_j) = \frac{\alpha}{d + \alpha} \qquad (4.7\text{-}1)$$

d : the distance between the two corpus.

α : an changeable parameter.(usually between 0 and 1)

Second: give the index rank of B1，B2，……，B12

$$B_j = \sum_{j=1}^{12} Sim(A_i, B_j), (i = 1, 2, 3) \qquad (4.7\text{-}2)$$

Third: set the value of A1，A2，A3.

$$A_i = Max(B_j), (i = 1, 2, 3\ j = 1....12) \qquad (4.7\text{-}3)$$

So far, we successfully obtain the rank of the 15 messages, comparing with the rank based on the index of mutual information; this model theoretically provides a more credible result.

5. ADVANTAGES AND DISADVANTAGES

5.1 Advantages:

We introduced the concept of mutual information when computing topics, and build the relationship between the suspicious topics and the undetermined ones.

5.2 Disadvantages:

●As it is mentioned in requirement 3 that due to the limitation of the message traffic, we can just roughly determine the doubt degree of the topics by semantic network analysis.

● Though the computing of the mutual information, we can just determine the values of the 12 undetermined topics, however, the values of the 3 higher suspected ones are artificially assigned.

●Without much experience for reference in the model, the accuracy of the results may be undermined.

6. PROMOTION OF THE MODEL

Through the analysis, we found that our model is considerable promotional, such as being used as a method to find the infected or diseased cells in a biological network. Briefly speaking, if we can get the meaningful characteristics from all the cells can calculate the mutual information between these characteristics. Later, we can try to obtain the mutual information between the undetermined cells and the infected ones so as to get the probability ranking list of the undetermined. Though the detecting process, we may us the Improved Search-based Page Rank Model and the Markov Transfer Matrix to simplify our operation as well as increase the validity of our research.

We have introduced the mutual information theory in analyzing this problem in the first two parts, with a relatively satisfactory result. As far as we know, semantic network analysis which is a widely used in artificial intelligence and computational linguistics. It provides a structure and process for reasoning about knowledge or language. And another useful computational linguistics capability in natural language processing is text analysis.

7. CONCLUSION

In this paper, we introduce the model of improved Page Rank algorithm based on mutual information and successfully solved the case problems. The consequences are listed as follows: In the previous case, George and Dave are the given co-conspirators. Bob and Inez are proved other co-conspirators. Carol is innocent. Overall, the solution derived from our models can meet the actual requirements and have some promotional value. Then in the next case, by analyzing the result, we can find that two senior managers, namely Dolores and Jerome are both conspirators. This will help the investigation of the cases. The other version of the second case, the newly added suspect Chris does not make much difference in results compared with the ones in the first requirement. According to analysis, this may be due to Chris's inactive involvement in the conversation, so even if he is one of the conspirators, his influence on others is very weak, resulting in the basically similar results.

All in all, as we only use the topics which are derived from original messages in mutual information analysis, so our result will be more valid if we have a larger corpus. That is to say, we can introduce semantic network and use similarity-computing to acquire a more precise result, in that way, our judgment will be more reasonable.

8. ACKNOWLEDGMENTS

This work was supported in part by the Graduate Innovation Foundation Project of Shan-dong University of Science and Technology (YC140106). Also, we would like to thank those anonymous reviewers for their constructive comments to help us improve the paper.

9. REFERENCES

[1]Haveliwala, T. (1999) Efficient Computation of PageRank. Technical Report. Stanford.

[2]Battiti, R.(1994)Using mutual information for selecting features in supervised neural net learning.

[3]Terry. (1999) The Page Rank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab.

[4]Jackendoff, Ray S.(1972) Semantic Interpretation in Generative Grammar.

[5]Xing, W., Ghorbani A. (2004) Weighted Page Rank algorithm, Communication Networks and Services Research. Proceedings.

[6]Catherine Benincasa (2006) Page Rank Algorithm.

[7]John F. Sowa (1991) Principles of Semantic Networks.

Alexander Kraskov, Harald Stogbauer, and Peter Grassberger (2004) Estimating mutual information.