

Crime Incident Association with Consideration of Narrative Information

Xiaofeng Wang, Donald E. Brown, James H. Conklin

Abstract—Law enforcement agencies need to discover underlying patterns of crimes in a short time. One of the key steps is associating related criminal incidents with each other. This paper describes a new methodology to associate crime incidents automatically, as well as, accurately by using narrative information. Specifically this paper shows a new method for measuring the similarity between crime incident narratives. Evaluations show that our method could measure similarity between narrative information quickly and accurately. In addition, a toolkit based on this methodology was developed to perform crime incident association.

I. INTRODUCTION

LAW enforcement agencies in the United States collect detailed data on criminal incidents in their jurisdictions. These data include location of incidents, description of offenders and narrative information about incidents.

Crime analysts search these reports to compare incidents with each other and associate incidents with similar fields. These similar incidents possibly indicate multiple incidents committed by the same person or group of persons. Once the crime analysts have these associated incidents, they could discover underlying patterns of crimes and use these patterns to assist in apprehension. This task of associating crime incidents according to the similarity between incidents is called crime incident association. [1]

However, performing crime incident association manually is time consuming. As noted by Brown and Hagen [1] performing pairwise comparisons on just 500 cases would require more than 1 million hours. Therefore, the objective of crime incident association research is to develop a method to associate crime incidents automatically, as well as, accurately and thereby to significantly reduce the time required by manual methods.

Brown and Hagen [1] developed methods for automating criminal incident data association by scoring the similarities between incidents. Only categorical and numerical variables were considered. Prats [2] developed methods to perform crime incident association by considering narrative data. Prats integrated narrative information into criminal data association

Xiaofeng Wang is with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: xw4u@virginia.edu).

Donald E. Brown is with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: brown@virginia.edu).

James H. Conklin is with the Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22904 USA (e-mail: conklin@virginia.edu).

using a Term Frequency – Inverse Document Frequency (TF-IDF) similarity function. The TF-IDF is used in information retrieval. However, this method cannot measure the similarity between narrative information accurately and quickly.

Text mining, also called text data mining, is a process of deriving useful information from text. Text mining firstly structure the input text by parsing narrative data, then derive patterns within the structured data, and finally evaluate and temperate the output. Typically, text mining tasks include information extraction, text categorization, summarization, and clustering [3].

Text mining has been applied to many disciplines to process narrative information. Yetisgen-Yildiz and Pratt [4] developed a literature-based discovery system called LitLinker to mine the biomedical literature for new, potentially connection between biomedical terms. Instead of applying natural language processing techniques, they used Medical Subject Headings (MeSH), keywords assigned to the document, to capture the content of the documents. This system can process narrative information quickly.

This paper describes an improved methodology to perform crime incident association with consideration of narrative information. This methodology not only enhances the accuracy of association, but also reduces the time to associate incidents.

II. METHODOLOGY

Although Prats measured the similarity between narrative information, and integrated it into crime incident association methodology developed by Brown and Hagen, there are still some problems with his approach. The first problem is the computation time. Since there are usually more than ten thousand words appearing in the collection of narrative information, comparing a few number of narrative information will take a long time. The second problem is the accuracy of the method. Since all the words in the narrative information will be compared, some useless information may influence the accuracy of association. Therefore, we developed a new methodology to measure the similarity between narrative information and integrated this method into crime incident association methodology developed by Brown and Hagen.

A. Methodology for Crime Incident Association with Consideration of Narrative Information

Instead of using Information Retrieval technology to measure the similarity between narrative information with all

words in it, the new method generates a word dictionary from all the narrative information, uses this dictionary to identify the high information content words in the narrative information, and then measures the similarity between the high information content words of narrative information to calculate a similarity score between narrative information. With the similarity score between narrative information, we follow the methodology developed by Brown and Hagen to associate crime incidents. There are four steps in the new methodology.

1) Generating Word Dictionary

The word dictionary is generated from the collection of all the narrative information. Two fields are included in this word dictionary: Words and IDF. For the i^{th} entry in the dictionary, $word_i$ represents a certain word appeared in the narrative information and IDF_i [2] represents the importance of $word_i$ in distinguishing narrative information. IDF_i is computed from (1).

$$IDF_i = \log_2 \left(\frac{N}{n_i} \right) \quad (1)$$

Where N is the total number of reports of narrative information available for association and n_i is the number of reports that contain the i^{th} word. The greater the number of reports containing the i^{th} word, the less distinctively the i^{th} word describes any of those reports [5].

2) Identifying the High Information Content Words

In a report of narrative information, not all the words are important. High information content words are words that are useful in distinguishing and representing narratives. For example, suppose that in a report it records “A man carried a Japanese Sword. Several people were wounded by this man carrying the Japanese Sword”. Although both the words “man” and “Japanese Sword” appear twice in the narrative, only the word “Japanese Sword” is the high information content word, because it rarely appears in reports while the word “man” appears much more frequently in reports. The word “Japanese Sword” gives us more information about the incident than the word “man”. Therefore, the high information content words are more useful to identify the similar incidents. We hope to extract the group of high information content words to represent the whole report of narrative information.

To measure the importance of a single $word_i$ in the j^{th} report of narrative information, we use the method developed by Prats [2]. We firstly compute the term frequency tf_{ij} , which counts the number of times the i^{th} word occurs in the j^{th} report. The more times a word occurs in a report, the more likely the report is about this word [3]. Then, we compute the importance of $word_i$ in the j^{th} report by (2).

$$w_{ij} = tf_{ij} \times IDF_i \quad (2)$$

Where w_{ij} denotes the importance of $word_i$ in the j^{th} report of narrative information, tf_{ij} denotes the term frequency, and IDF_i is from the word dictionary. w_{ij} represents the trade-off between the high term frequency within narrative information and high distinctiveness of term frequency within the entire narrative information collection.

Also, we need to choose the number of high information content words to represent the narrative information. We randomly choose 20 reports of narrative information from the crime incident report database to compute the importance score for each words in the narrative information. The result is shown in Fig. 1. From the figure we know, before the top 20th word the importance scores are high, and after the top 20th word the importance scores decrease slowly. Therefore, we suggest choosing top 20 words with the highest importance scores as the high information content words to represent the narrative information.

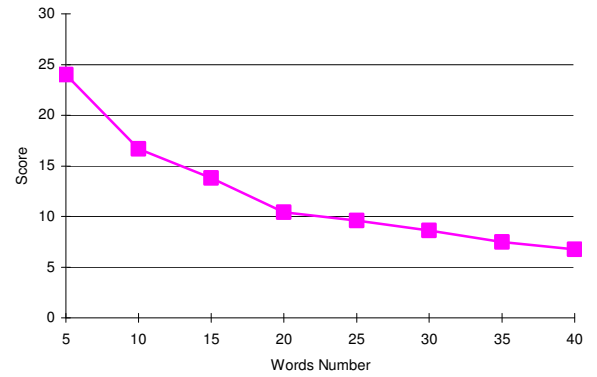


Fig. 1. Importance score vs. number of words

3) Measure the Similarity between the High Information Content Words

We measure the similarity between the high information content words of the narrative information and use this similarity to represent the similarity between the narrative information. We use function (3) to measure the similarity between two sets of high information content words.

$$S_{ij} = \frac{N_{ij}}{(n_i + n_j) / 2} \quad (3)$$

Where S_{ij} denotes the similarity between the i^{th} set of high information content words and the j^{th} set of high information content words, N_{ij} denotes the number of the same words in both sets of high information content words, n_i denotes the number of words in the i^{th} set and n_j denotes the number of words in the j^{th} set. By definition, $S_{ij} \in [0, 1]$. When $S_{ij} = 0$, the i^{th} and j^{th} narrative

information are very dissimilar, while $S_{ij} = 1$ they are very similar.

4) Associate Crime Incidents

With the similarity between narrative information, we use Total Similarity Measure (TSM) developed by Brown and Hagen [1] to compute the similarities between any pairs of incidents. Given a target incident, we could get a TSM score for each other incident. The higher the TSM score is, the more likely the incident should be associated with the target incident.

III. PERFORMANCE EVALUATION

In this section, we describe the performance of the new method and compare it with the method developed by Prats. Two criteria are important to evaluate the performance: response time and accuracy.

A. Evaluation of Response Time

1) Criterion

Response time is the time between the moment the user submits a request and the moment the system returns the result. The shorter the response time is, the better the method is.

In this evaluation, we measure the response time to get the similarity scores for different sizes of narrative information sets of different methods. Since the first and the second steps in the new method will be available before the user submits requests to the actual system, we do not count the time to generate the dictionary and the lists of high information content words as part of response time.

2) Test Set

The test set is from the crime incident reports of New Kent County, VA (ORI: VA0630000). There are 6350 reports of narrative information describing the crime incidents that happened between 1998 and 2006. In total, 311098 words appear in the narrative information (excluding article words or pronoun words like “the”, “a”, “an”, “that”, etc.). There are 15351 different words in this set of narrative information. On average, each report includes 49 words.

3) Result

Given a target incident from this set, we measure the response time to get the similarity scores for reports of narrative information. The results are shown in Table I.

TABLE I
RESULT FOR EVALUATION OF RESPONSE TIME

METHOD	TIME FOR 10 REPORTS	TIME FOR 500 REPORTS	TIME FOR 6350 REPORTS
NEW METHOD	<1s	1s~2s	2s~3s
PREVIOUS METHOD	5s~6s	98s~100s	>1080s

New method refers to the method developed by this paper; previous method refers to the method developed by Prats.

From these results, we observe that the new method measures the similarity between narrative information much faster than the previous method. Therefore, the new method is a more efficient approach to perform crime incident association.

B. Evaluation of Accuracy

In this section, we measure the accuracy of the new method without comparing it with the previous method.

1) Criteria

a) Distribution of Similarity Scores

Associated reports are reports about criminal incidents which have the same offender. Non-associated reports describe crimes which have different offenders. A good measure of similarity should distinguish associated reports from non-associated reports. Ideally, the similarity between associated reports should be large while the similarity between non-associated reports should be small. Therefore, the distribution of similarity scores of associated reports should be different from the distribution of similarity scores of non-associated reports.

In this evaluation, we evaluate the similarity scores from narrative information. With each similarity score, we know whether the pair of incidents should be associated. Then there are two groups of similarity scores: one group of similarity scores that measure the narrative information between the associated reports (we call it as Group A) and another group of similarity scores that measure the narrative information between the non-associated reports (we call it as Group B). If the distributions of these two groups are the same or similar, the new measure is not useful in distinguishing incidents or in associating incidents; if the distributions of these two groups are different, the new measure is useful.

b) Receiver-Operator Characteristic Curve (ROC curve)

To measure the accuracy of whether the similarity scores could classify the incidents rightly, we could measure the true positive rate and the false positive rate of the classification. True positive means when the similarity score indicates the incidents should be associated and these two incidents are actually associated, while false positive means the similarity score indicates the incidents should be associated but these two incidents are actually not associated. Ideally, we hope the true positive rate is 1 while the false positive rate is 0.

ROC curve is a plot to capture the trade-off between the true positive rate and the false positive rate. The more the curve close to left up corner, the better the classification is.

2) Test Set

The test set used in this evaluation is the same test set used in the last section. Besides the incident number and narrative information, we also have the suspect person id associated with the incident. According to the goal of crime incident association [1], we consider two incidents are actually associated if these two incidents have the same suspect person id.

We measure the similarities between the narrative information of 35 incidents happened in 2005 and the narrative information of all the 6350 incidents in the test set. Therefore, there are 35×6349 similarity scores (we do not compare similarity between narrative information from the

same incident). Each similarity score is with a label marketing whether the incidents are actually associated.

3) Result

The box plot of two groups of similarity scores are shown in Fig. 2. From the box plot we observe that the distributions of similarity scores are different for different groups. The median of similarity scores from Group A is greater than the median of similarity scores from Group B. 50% of similarity scores in Group B are less than 0.054 while only 25% of similarity scores in Group A are less than 0.059.

The ROC curve for the test set is shown in Fig. 3. The curve is apart from the diagonal. For example, at the false positive rate of 0.2, we could get a true positive rate of 0.62. Therefore, the new method is useful for classification. Also, from this plot we could decide a threshold for association given a false positive rate.

From this evaluation, we conclude the similarity scores from the new method are useful in distinguishing associated reports from non-associated reports and thereby helpful to associate incidents.

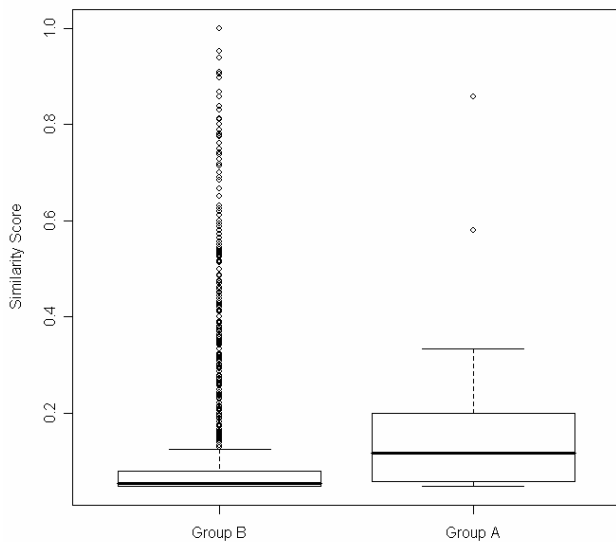


Fig. 2. Box plot of similarity scores

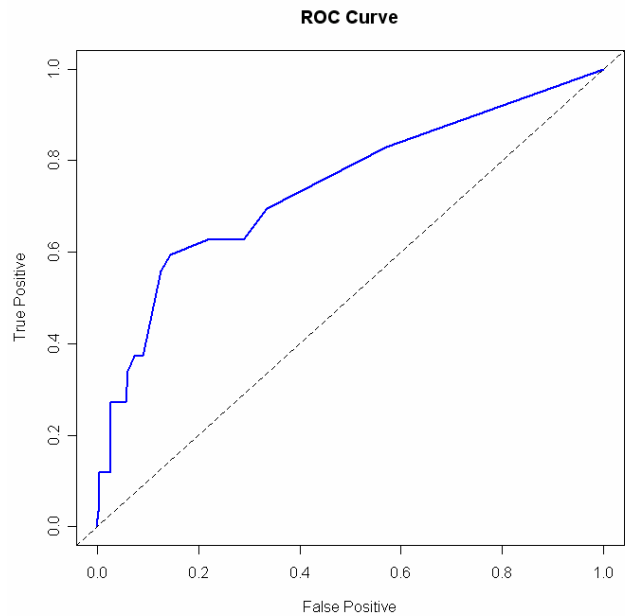


Fig. 3. ROC Curve

IV. CONCLUSIONS

This paper describes an approach to automatically associate narrative reports about crime incidents. This method is faster than previous techniques and provides accurate results. This technology can be used by police to identify associated crimes.

For future research, advanced text mining technologies will be tested in measuring the similarity between narrative information. For example, we could consider the meaning of the words instead of simply matching the words.

REFERENCES

- [1] Brown, D.E., Hagen, S., "Data association method with applications to law enforcement", *Decision Support Systems*, vol. 34, pp. 369-78, Feb. 2002.
- [2] Prats, F. G., "Textual analysis and linking of narratives (TALON): Utilizing TF-IDF for the incorporation of narrative information into incident data association", M.S. Thesis, Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, 2005.
- [3] Konchady, M., *Text Mining Application Programming*. Charles River Media, 2006.
- [4] Yetisgen-Yildiz, M., Pratt, W., "Using statistical and knowledge-based approaches for literature-based discovery", *Journal of Biomedical Informatics*, vol. 39, pp. 600-11, Jan. 2006.
- [5] Salton, G., *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Compute*. Addison-Wesley, Massachusetts, 1989.
- [6] Prats, F. G., "Textual Analysis and Linking of Narratives (TALON)", in *Proc. 2005 IEEE Systems and Information Engineering Design Symposium*, Charlottesville, VA, 2005, pp. 177-82.