Algorithmic tools for data-oriented law enforcement

Cocx, T.K.

Chapter 9

A Distance Measure for Determining Similarity between Criminal Investigations

In comparing individual criminal investigations on similarity, we seize one of the opportunities of the information surplus to determine what crimes may or may not have been committed by the same group of individuals.

For this purpose we introduce a new distance measure that is specifically suited to the comparison between criminal investigations that differ largely in terms of available intelligence. It employs an adaptation of the probability density function of the normal distribution to constitute this distance between all possible couples of investigations.

We embed this distance measure in a four-step paradigm that extracts entities from a collection of documents and use it to transform a high-dimensional vector table into input for a police operable tool. The eventual report is a two-dimensional representation of the distances between the various investigations and will assist the police force on the job to get a clearer picture of the current situation.

9.1 Introduction

In contrast to all applications discussed so far, that dealt with a single, rather static database stored at police headquarters, data mining is also particularly suited for usage on case related data, that is gathered for a single investigative purpose. Since tools that deal with this kind of data are confronted with data of an beforehand unknown structure or quantity, an entire new set of challenges need to be addressed for such innovations to be successful.

One of the key problems in this area of policing is the sheer amount of data that is stored on computers, that are nowadays confiscated more regularly during criminal investigations. A contemporary police force should be capable of dealing with this stockpile

of data during the course of the investigation, yielding any valuable piece of information both timely and accurately. Usually, people working on these investigations have no technical background, so information should be presented to them in a comprehensive and intuitive way.

This chapter discusses new tools that deal with the extraction of logical concepts from police narrative reports and documents found on crime scenes in order to automatically establish an educated guess on what crimes may be committed by the same (group of) criminals. To this end we employ *text mining*, a *distance measure* and an *associative array clustering technique*. We discuss the difficulties in case-comparison and the specific distance measure we designed to cope with this kind of information.

The main contribution of this chapter is the discussion in Section 9.6, where the distance measure is introduced.

9.2 Project Layout

As discussed earlier, useful information exists in unstructured data like police narrative reports, intercepted emails and documents found on crime scenes. It would be desirable to employ this information for case comparison in order to supplement the forensic work done on-site and provide police officers with information about crimes that may be committed by the same perpetrators. However, this information is usually deeply hidden in the unstructured setup of such, often free-text, documents. Even after the employment of a suitable text miner, the enormous amount of extracted entities still poses many problems. As is common with police work, some cases suffer from a lack of data, while generate stockpiles of paper work. Comparing cases that differ extremely in size in terms of entities extracted from this unstructured data, is one of the challenges. Another is that of preparing the resulting data of this comparison, visualizing it and presenting it to the officers on the case. Our research aims to address both these challenges and to set up a police framework for comparing cases on the basis of (collected and created) documents.

9.3 System Architecture

Our "case-comparison" system is a multiphase process that relies on a commercial text miner, a table transformation unit, a distance calculator and a visualization tool. We therefore describe our process as a *four-step paradigm* (see Figure 9.1) and will elaborate on the individual components and their in- and output in the following sections. Black boxed, the paradigm reads in a collection of unstructured documents and provides a comparison report to the end user.

The documents we use as input for our case comparison system consist of two different types, both of which are provided by the individual regional police departments for analysis:

- Police narrative reports: one of the types contained in our document collection is that of the police written narrative reports. These reports are created to describe

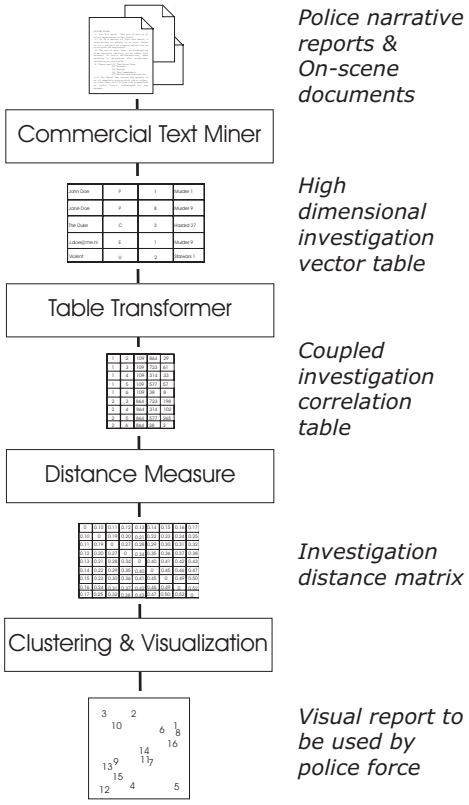


Figure 9.1: Four-step paradigm

a crime, the people involved and the Modus Operandi (MO). Protocols exist how these reports should be written, but these rules are not always strictly followed. Also, these reports suffer from an abundance in police terminology (for example, almost all reports contain the words “rep.” (report) and “serial number”) and they are likely to have typing mistakes in for example the way names are written. Some of these spelling mistakes are intentionally introduced by suspects to avoid cross referencing. Due to these effects, the police narrative reports are often reasonably polluted.

- Crime scene documents: digital documents found on crime scenes are often very rich in information. They contain valuable information like email contact lists that can give an idea of other people involved or lists of goods acquired to commit crimes. Since they are mostly created by the perpetrators themselves they are less likely to have errors or typing mistakes. Therefore, the crime scene documents are less noisy than the narrative reports, but are unfortunately also more rare.

Table 9.1: Different types recognized by entity extractor

Type	Description	Percentage
K	License plate	0.90%
P	Person	5.34%
u	URL	0.02%
O	Organization	0.48%
L	Location	0.69%
e	Email address	0.04%
D	Product	0.03%
i	IP address	0.02%
U	Unknown	92.45%

When processing these documents we first subdue them to a text miner that yields a table with concepts, the number of appearances and the investigation they belong to. This table is then transformed to a high-dimensional vector space, where each vector represents a different investigation. We then extract comparison numbers in our transformation unit, that is presented to our distance calculator. The distance matrix that results from this is now fed to the visualization and presentation tool, that can be operated by the analyst himself. The results will be presented visually and are ready to be interpreted and used by the individual police department that provided the documents.

9.4 Entity Extraction and Table Transformation

An important step in the process of getting from a collection of documents to a comparison overview is that of entity extraction or text mining. As mentioned earlier some specialized tools were created by different research programs. The INFO-NS program [28] suggests a framework for evaluation of commercial text mining tools for police usage. In practice, a lot of police departments employ one of these commercial suites for their data mining endeavors. In order to comply with this situation, we chose to employ the use of the SPSS Lexiquest text mining tool [41] as the starting point for our comparison framework. Through a simple operating system script the documents are fed into the text miner one investigation at a time. The text miner then yields the following table:

Entity	Investigation	Type	Amount

In this table Type refers to one of the types defined in Table 9.1 that also shows the percentage of these types in the dataset used for our experiments. The resulting table is

primary keyed by both entity and investigation but since the final objective is the comparison of investigations it is necessary to transform this table to an investigation based one that contains all characteristics per investigation. The table should therefore contain information about what entities are present in each investigation. The table we are creating therefore has an integer field for every entity. If the key investigation-entity is present in the original table, the corresponding field in the new table will be equal to the contents of the amount field and 0 otherwise. The number of distinct investigations we can retrieve from the table will be denoted by m. This yields the following high-dimensional vector table:

Investigation	Entity 1	Entity 2	...

where the number of dimensions, apart from the key attribute Investigation, is equal to the number of distinct entities in the previous table, which we will denote by n.

This table contains highly accurate entity usage information per investigation, for it contains all available information. We can now employ this information to get insights into the way different cases are alike in terms of used concepts.

9.5 Multi-Dimensional Transformation

The high-dimensional table that resulted from the previous step can now be used to describe distances between the various investigations in n-dimensional space. Naturally, we need to transform this table into a two-dimensional representation of the given data in order to come to a useful visualization in our tool. In order to achieve this dimensional downscaling we assume similarity can be constituted by the sizes of the investigations in terms of entities extracted and the entities they have in common.

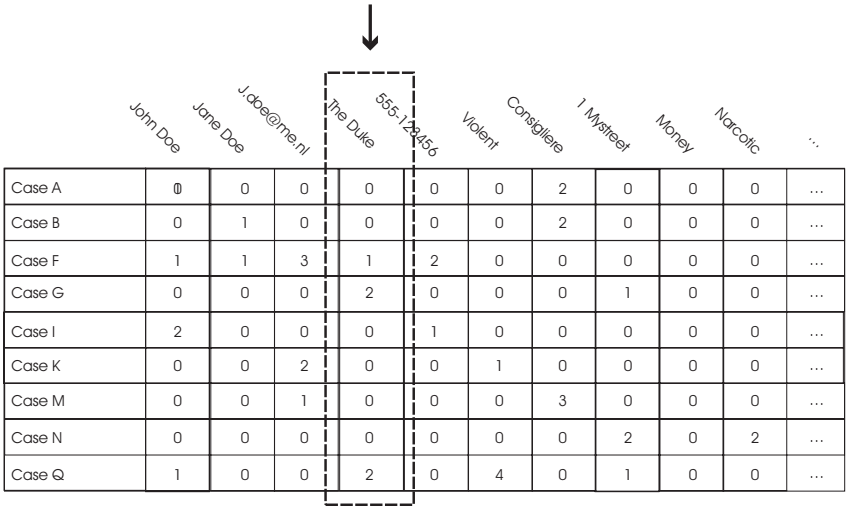
According to this assumption, we compare the investigations couple-wise on each individual entity (see Figure 9.2) and score the couples on the amount of common entities according to the following method: every time both investigations have a value larger than or equal to 1 in a column, the score for overlapping is raised by 1.

The algorithm treats every investigation as a subset of the total set of entities. The calculation of the amount of common entities in two investigations is therefore synchronous to the calculation of the amount of items in the intersection of both sets (see Figure 9.3), and goes as follows:

$$\text{Overlap} = |\text{Inv}_1 \cap \text{Inv}_2| \ ,$$

where Inv_i is the set of entities for investigation i ($i = 1, 2$). We also let Size_i denote $|\text{Inv}_i|$.

It is possible to utilize a filtering technique at this point to exclude all entities with type “U” (unknown) from the comparing method. This will probably yield highly accurate results, due to the high expressive power of the recognized entities. For example, two cases sharing the person “John Doe” are more likely to be similar than two cases sharing the word ‘money’. However, due to the high percentage of entities that are being classified as unknown, leaving them out can cause undesired shortcomings to the algorithm.



	John Doe	Jane Doe	J.doe@me.nl	The Duke	555-123456	Violent	Consigliere	1 Mystreet	Money	Narcotic	...
Case A	0	0	0	0	0	0	2	0	0	0	...
Case B	0	1	0	0	0	0	2	0	0	0	...
Case F	1	1	3	1	2	0	0	0	0	0	...
Case G	0	0	0	2	0	0	0	1	0	0	...
Case I	2	0	0	0	1	0	0	0	0	0	...
Case K	0	0	2	0	0	1	0	0	0	0	...
Case M	0	0	1	0	0	0	3	0	0	0	...
Case N	0	0	0	0	0	0	0	2	0	2	...
Case Q	1	0	0	2	0	4	0	1	0	0	...

Figure 9.2: Comparing the investigations on common entities

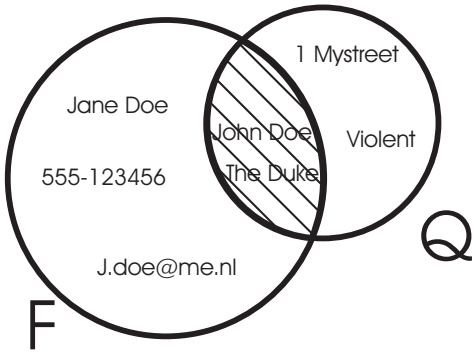


Figure 9.3: Viewing the transformation process as calculating intersections

For example: the word 'violent' may well be a keyword in comparing two individual investigations, but is still categorized under type 'U'.

The mentioned algorithm provides a table with $\frac{1}{2}m(m-1)$ rows, where each row represents one couple of investigations. Each row consists of three columns: the size of the first investigation, the size of the second and the amount of common entities:

Size_1	Size_2	Overlap

This table is comparable to the ones used in data mining on shopping baskets where the goal is to determine which customers exhibit similar shopping behavior.

9.6 Distance Measure

To constitute similarity between different investigations we introduce a distance measure, that calculates the distance between two such criminal cases. The distance measure we propose is a function over the parameters we stored in the table resulting from the previous step. The higher the function outcome and thus the distance, the less alike two investigations are. Our function yields a distance value between 0 and 1.

It is not just the amount of common entities that constitutes the distance between two investigations; the different sizes of the investigations should be taken into account as well. It is common practice in for example the analysis of the earlier mentioned shopping baskets, to let a difference in size have a negative effect on similarity. If we take a look at two shopping baskets, and we observe that one basket contains a newspaper and a bottle of wine and another basket contains the same paper and wine but also a hundred other items, no analyst would claim similar shopping behavior of the two customers, although 100% of one of the customer's acquisitions is also in the other one's basket. Therefore, distance measures like the symmetrical distance measure [26]:

$$\frac{(\text{Size}_1 - \text{Overlap}) + (\text{Size}_2 - \text{Overlap})}{\text{Size}_1 + \text{Size}_2 + 1}$$

that also incorporates size differences, are often employed in this area. However, this does not hold for the comparison of investigations. Although the size in terms of entities extracted may well be an indication of difference between two cases (many or few computers found on scene) it is, as mentioned earlier, not uncommon for law enforcement cases to differ largely in size while they still involve the same people (the police was at the scene very quickly vs. the criminals had time to destroy evidence).

As a consequence, the symmetrical distance measure mentioned above is not applicable in the area of case comparison. Naturally, the sole use of common entities is not applicable either. We therefore introduce a new distance measure specifically suited for the comparison of criminal investigations based upon entities extracted.

We propose a distance measure based upon the random amount of common entities two investigations would have if they were drawn randomly from the entire set of entities. The deviation between the size of the randomly selected intersection and the actual amount of common entities then constitutes distance. The size of the entire collection of entities, the universe of entities, will be denoted by A. In calculating this value we only count each distinct entity once instead of using each occurrence of a single entity in the table. This will more accurately represent the probability space for each individual investigation subset. We will denote the average size of the intersection of two randomly drawn subsets having sizes X and Y as E, which can be calculated as follows:

$$\frac{X}{A} \cdot \frac{Y}{A} = \frac{E}{A} \quad \Longleftrightarrow \quad E = \frac{X \cdot Y}{A^2} \cdot A = \frac{X \cdot Y}{A} \ .$$

We can now easily calculate the difference (Differ) between the actual value Z and the expected value E as follows:

$$\text{Differ}(Z) = Z - E \ .$$

As is clear from the calculation of E, the expected value depends on the three variables X, Y and A. As a consequence, a very large universe A can lead to very low values of E and thus to very large differences between E and Z. This variation can be considered to be disruptive to the process in the following two cases:

- Some very large investigations without any relation to the other investigations, for example, two large Finnish investigations among a series of English investigations, are included in the list to be analyzed. The large number of unique entities these Finnish investigations would contribute to the universe A would implicate that all other investigations would have very low expected values and therefore very high differences. This can put all those investigations at a distance from each other that is far less than it should intrinsically be.
- When a lot of different investigations are to be compared, they all contribute a number of unique entities to the universe A. This means that, while the actual chance of two investigations having a certain overlap does not change, the calculated E would decrease approximately linearly to the amount of investigations included in the analysis. This can, as was mentioned above, lead to too small distances between the investigations.

As a measure for countering these effects we propose to calculate A from the actual overlapping values instead of just using the total amount of entities. We have implemented this method and compared it to the standard method described above.

We will base the alternative calculation of A upon the actual overlapping entities in all the possible couples, meaning that we calculate A out of X, Y and Z, instead of calculating E out of X, Y and A. Our method will then average all the individually calculated A's and use this number for A instead of the total amount of entities. Calculating A will go as follows:

$$A = \frac{\sum_{i=1}^{m} \sum_{j=i+1}^{m} \frac{X_i \cdot Y_j}{Z_{ij}}}{\frac{1}{2}m(m-1)} \tag{9.1}$$

In this summation we omit the pairs (i, j) with $Z_{ij} = 0$. Having obtained the differences we can calculate the distance between two investigations.

The normal distribution is the most widely used distribution in statistics and many statistical tests are based on the assumption of normality. One of the most used representations of this distribution is the probability density function (see Figure 9.4), which shows how likely each value of the random variable x is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

where σ denotes the standard variation and μ denotes the mean. Since we want to give a quantization of how notable the Differ function outcome is, we can take this function

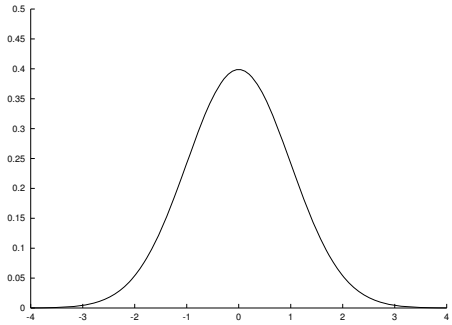


Figure 9.4: Normal probability density function; $\sigma = 1, \mu = 0$

as basis for our distance function. We will thus employ an adaptation of that function to calculate the distance of two investigations by using the above mentioned function Differ. First, because we want to normalize our outcome between 0 and 1 we need to top of our function at $\frac{1}{2}$ by changing the factor before the exponential part of the function into $\frac{1}{2}$. Then we take the minimal value of X and Y as our standard deviation, since it is logical for the intersection of two large subsets to deviate more from the expected value than the intersection of two smaller subsets. We can flip the function part left of the mean to represent a positive deviation as a smaller distance between the two investigations. If we denote $\min(X, Y)$ as the minimum value of X and Y, our final distance function will look like this:

$$\text{Dist}(Z) = \begin{cases} \frac{1}{2} \exp\left(\frac{-\text{Differ}(Z)^2}{\frac{1}{2}\min(X,Y)}\right) & \text{if Differ}(Z) \geq 0 \\ 1 - \frac{1}{2} \exp\left(\frac{-\text{Differ}(Z)^2}{\frac{1}{2}\min(X,Y)}\right) & \text{otherwise} \end{cases}$$

This function calculates distance with respect to any size difference that may occur between two investigations while not incorporating any negatives effect for that difference. The proposed measure is symmetrical (X and Y can be exchanged). If two investigations are very much alike, their distance will approximately be 0; if they are very different their distance approaches 1.

As is clearly illustrated in Figure 9.5, the form of the graph of the distance function differs significantly between different sized investigations. This enables us to compare different sized subsets, since similarity is constituted by the probability space rather than integer values. For example, three overlapping entities in two small subsets can now judge the two cases to be just as similar as 10 overlapping entities in a large and a small subset, or 100 overlaps in two very large cases.

If we apply this distance measure to all the rows in the last table we are able to create a distance matrix M, where for each $1 \leq i \leq M$ and $1 \leq j \leq M$ element M_{ij} represents the distance between investigations i and j. Due to the fact that the distance between i and j is the same as between j and i our distance matrix is symmetrical. Having calculated all the distances we can display the investigations in our visualization tool.

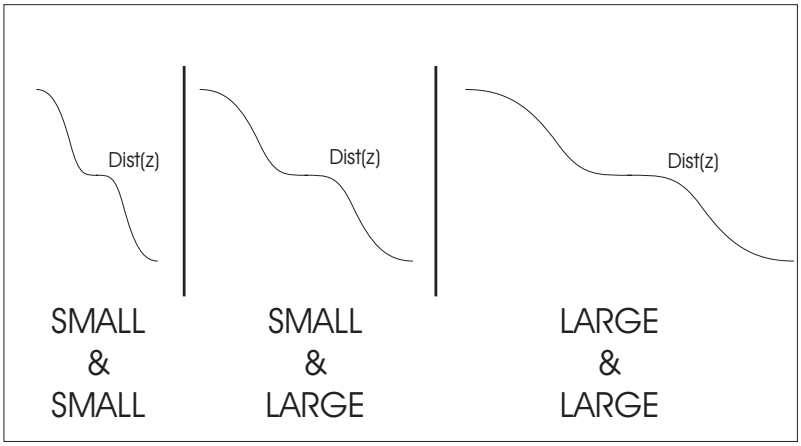


Figure 9.5: Distance function of different sized couples of investigations

9.7 Visualization

It is desirable for police personnel to be able to view the total picture in one glance. In most cases it is not possible to display a high-dimensional situation, such as our initial vector table, perfectly in a two-dimensional plane, especially after the amount of transformations our data went through. We therefore employed the associative array clustering technique [27] to display (an approximation) of all the distances between the different investigations in one two-dimensional image. This technique can be viewed as Multi-Dimensional Scaling (MDS, see [15]), and is especially suited for larger arrays. The image we now have can be fed back to the officers on the case to enhance their understanding of the situation.

The associative array visualization technique strives for the creation of a flat image of all considered elements where the physical distance between them is linearly related to the distance in the matrix, while minimizing the error in that distance, sometimes referred to as "stress". It is an iterative process that starts off at a random situation and through a specified number of iterations tries to improve that situation until it reaches a more or less stable state. This is a state where the error made in the placement of the elements is at a (local) minimum.

The algorithm works as follows: starting at the earlier mentioned random position, where all the elements are in an arbitrary position, the algorithm investigates a random couple of elements and when the distance in the image is relatively larger than the requested distance, the pull operation is executed. If, on the contrary the current distance is smaller than the distance in the matrix the push operation will push the elements away from each other. As can be seen in Figure 9.6 the push and pull operations move the elements in the target image away from or towards each other on the line defined by the two elements.

In every iteration all couples of investigations will be evaluated and their respective

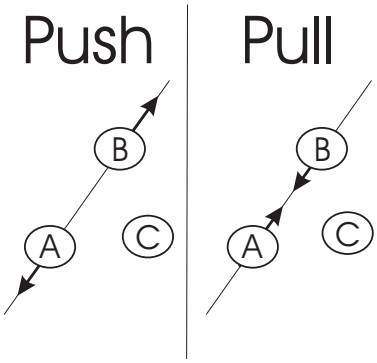


Figure 9.6: Push and pull operations

distances corrected. Since the image usually can not be displayed entirely correct in the two-dimensional plane, the image might differ a bit depending on the random starting point, but is consistent enough to give a good overview of the similarity between investigations. Also note that rotations and reflections may be applied without affecting the outcome.

It is imperative for the tool that performs this task to be operable by the police officers that request the similarity analysis. The tool to be designed therefore needs to be extended by a graphical user interface (GUI). Developing a GUI not only serves the purpose of making the application usable by police personnel, but also gives insights in the formation of the image and enables us to detect problematic situations and improve the algorithm.

The tool we developed allows a number of settings and has a main screen where the user can see the image unfold in different speeds. The user can then output the images in PDF format for usage in a document. The user can customize the screen to display investigation labels of numbers if the titles overlap too much.

As a simple demonstration of the algorithm's possibilities, we tried to regain the image of four points forming the corners of a square and a fifth point in the center, by means of its distance matrix. The result depends on the random starting position of the five points and if done correctly would represent the original image reasonably accurately in compliance with rotation and mirror symmetry. Figure 9.7 is one of the possible results.

9.8 Experimental Results

One of the major tasks of the police is the dismantlement of synthetical drugs laboratories. Several of these have recently been located and rendered out of order. Investigation of these crime scenes has led to the acquirement of digital documents, interception of email traffic and the compiling of numerous narrative reports by officers and forensic experts assigned to these investigations. Given the nature of the different laboratory sites, case detectives suspect common criminals to be involved in exploiting some of these

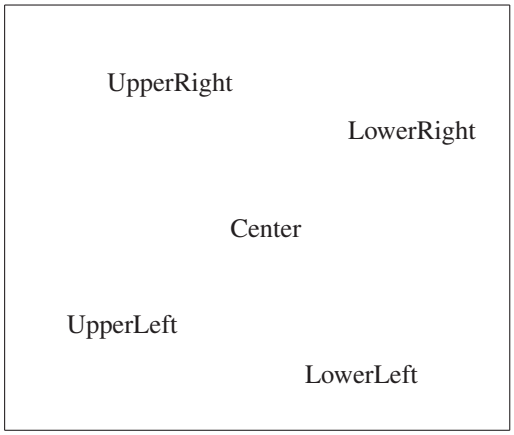


Figure 9.7: Visualization of a centerpointed square based upon its distance matrix; visualized using labels

locations for the creation of synthetical drugs. Employment of a clustering technique should provide answers to the questions about common perpetrators in these and future cases. Research on the collected data should therefore focus on the following:

- Producing a comprehensive report on the similarity of current investigations into the construction and employment of synthetical drugs laboratories.
- Using the data to produce a tool that enables the police to perform similar tasks in similar future situations.

As was mentioned earlier, the data in such documents are often polluted by the inclusion of enormous amounts of police terminology and the large number of typing mistakes. Also, since the commercial text miner is not specifically trained on police documents, a lot of entities were labeled as unknown or as the wrong type. Incorporating this knowledge into our scripts we decided to use all types of entities instead of the inherently more powerful classified entities alone. We present some results for this analytical task containing $n = 28$ police investigations, together having $m = 152{,}820$ distinct entities. Here, A is either m, using just the amount of entities, or is computed according to Formula (9.1).

Usage of our new distance measure on this data yielded distance matrices that indeed showed results that could indicate similarity between some of the individual investigations. The distance matrices showed some significant difference in distance values between the individual investigations. Application of our clustering approach to this newly generated matrix for both different calculation methods for A showed a clustering image (Figure 9.8 left and right) that indeed demonstrated that certain investigations are

closer and therefore more similar to each other than others. We infer from these images that there is some relevant similarity between certain investigations and submitted the reports, outputted by our application to the investigation teams. We are currently in discussion with the domain experts about the validity of the outcome of both methods employed by our system.

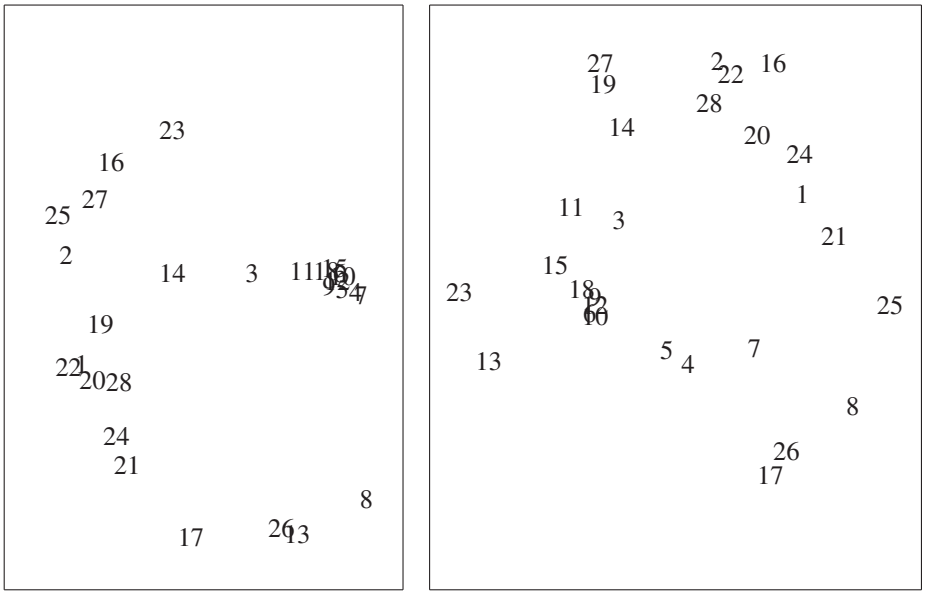


Figure 9.8: Visualization of the database of investigations using $A = n$, numbered 1 to 28 (left) and visualization of the database of investigations using Formula (9.1) for estimating A, numbered 1 to 28 (right)

In the comparison of both possible methods of calculating A, it is noteworthy that the distances represented in both images do not vary a lot between the different methods but show minor differences, as is, for example, evident in the placement of 13 in Figure 9.8.

9.9 Conclusion and Future Directions

Data mining is a suitable solution for many problems and opportunities arising from the information explosion. In this chapter we demonstrated the applicability of data mining in the comparison of individual criminal investigations to establish a quantity for similarity between them. We used a four-step paradigm to transform a set of documents into a clustering image that gives a full overview of the similarity between all investigations and is ready to be used by police experts. The new distance measure we introduced was

specifically designed for this purpose. It incorporates the differences in information size between investigations while still maintaining a realistic comparison standard.

Future research will aim at getting a clearer picture about the computation method for A. Assigning n to A describes the situation with a true to reality universe of entities while using Formula (9.1) probably delivers better end-results for largely different or a large number of investigations. Both methods of assigning a value to A therefore have their own merits and more testing on different data sets is a prerequisite in deciding between them.

The commercial text miner used in this project was a source of problematic entries in our initial table. Incorporation of domain specific text miners such as used in the COPLINK project [12] would probably lead to a significant improvement of the total system. This is one of the subjects where future research should focus on.

Extending our research in this area by creating a linguistic model that deals with the large amount of typing mistakes would be a great advantage to our entire approach and would probably lead to an even more realistic end-report and a clearer picture that the police force has of the perpetrators it pursues.