UDC 004.912

Vatter C., Mozgovoy M.

Data mining in forensics: a text mining approach to profiling criminals

1. Introduction. Data mining is the process of processing and analyzing large amounts of data and summarizing it into useful information[1]. Information gathered from data mining can help individuals and organizations make better decisions or draw conclusions about a particular subject. Data mining can also reveal trends in the data, which can be advantageous in predicting future outcomes.

In the field of forensic linguistics—the practice of analyzing and using speech and writings in the court of law as evidence for criminal cases—discovering trends and idiosyncrasies becomes imperative in creating a solid case either for or against conviction. In particular, if a certain text matches certain patterns—unique to a certain individual or criminal profile—it can serve to strengthen a case for conviction during an ongoing criminal investigation.

The purpose of this study is to attempt to use data mining in order to discover common trends in certain criminal texts. The question becomes whether or not there are clusters created based on document similarity that reflects either authorship or type of crime.

A corpus of texts was collected through a series of web searches and categorized according to type of criminal with which it was accredited. At the time of writing this paper, the corpus includes a total of 157 files. This corpus is comprised of transcripts, speeches, journals, letters, poems, songs, manifestos, and other miscellaneous items. Great care was taken to make sure that the way in which the original pieces were presented is preserved; meaning that all spelling errors, capitalization choices and slang were left unaltered when transferred into pure text format. The entire corpus consists of 8 categories, which include: bombers and terrorists, cult leaders, dictators, ransom notes and kidnappers, school shooters, serial killers, and spree killers, as well as a category

Vatter, Corey – Undergraduate, Student, Rose-Hulman Institute of Technology; e-mail: vattercw@rose-hulman.edu

Mozgovoy, Maxim – PhD, Associate Professor, University of Aizu; e-mail: mozgovoy@u-aizu.ac.jp

for other notable criminals that do not fit any of the other particular categories.

In this study, the categories that will be focused on for analysis are school shooters, spree killers, and serial killers, as they seem to have the most in common as with relation to type of crimes and usually having emotionally charged motives.

2. Method and Data. The software that was used for text mining and analysis in this study was QDA Miner 4 with WordStat 7. The focus of this software is for directly dealing with text analysis, and is suited well for analyzing similarities and trends between documents.

Using QDA Miner, the texts—referred to as 'cases' in the program—were manually sorted into the following categories:

- School Shooter (35 cases)
- Serial Killer (44 cases)
- Spree Killer (6 cases)

When working with text, in order to find accurate and meaningful similarities, many parameters need to be established before beginning analysis. These parameters may change depending on the type of results desired. In this study, we decided to remove any common English articles and stopwords (i.e. 'a', 'the', 'and', 'I', etc.) and ignore capitalization to allow the analysis to focus on the important content of the texts being processed. We also decided not to include a stemming algorithm (i.e. Porter[2]) in our peliminary study, so as to differentiate between different tenses of a word and to not ignore spelling errors that could have been made by any one author—as these errors and tenses add to the uniqueness of a particular text.

The process for performing hierarchical clustering includes generating a distance matrix based on the cosine coefficients that are deterimined by the frequency of particular keywords present in the corpus. These cosine coefficients—which can also be referred to as a similarity index—determine the relative distance between texts, with the most similar texts having less distance between them in visual representations. The similarity index can be any value between 0.00 and 1.00—1.00 indicating that two texts are completely identical.

WordStat, in particular, includes a few different visual representations of the data, including dendrograms and similarity proximity maps. The similarity proximity plot takes into account a set of cases and their similarity index to calculate their distance from one another on the map as well as the strength of link between the two cases as illustrated with a line which becomes thicker the larger the similarity. In cases where the similarity between documents is at or below the insignificance threshold, a link is not visible on the map. Dendrograms are tree diagrams used to visualize an arrangement of clusters based on hierarchical clustering[3].

The method by which the keywords were clustered in this study included same case co-occurence identification, cosine theta indexing, and second order co-occurence grouping [4]. The same case co-occurence identifies and records everytime two words are used in the same text. Cosine theta indexing measures the cosine of the angle between two vectors of values in the vector space model—which ranges from -1.00 to +1.00—and takes into account both the occurence and frequency of a word within the given text. Second order co-occurence allows for the consideration of two words that are used in similar environments—such as synonyms or alternate spellings and forms—to be grouped together even if they don't occur near each other. These factors were combined to determine the similarity indicies between each text in the corpus during the clustering process.

3. Results. After a series of trial and error, it was determined that setting the number of clusters to 36 gave the best results in splitting the texts into reasonable groups with highest form of cohesive similarity, as well as removing the outlier texts present in the corpus (see Figure 1).

When interpreting the dendrogram, each color represents a different cluster. After setting the analysis to 36 clusters, 19 clusters of size greater than 1 text were generated and 17 cases that have low similarity indices were removed. It is important to note that the shorter the distance is between two documents along the line that connects them, the closer they are in similarity.

In Figure 2, we present the general shape of the similarity proximity map. The numbers on the links between nodes correspond to the similarity index. The proximity map reveals various similarity clusters.

Figures 3 and 4 show examples of such clusters. Figure 3 depicts the main cluster of school shooter cases with strong similiarity indices. Figure 4 shows an example of authorship similarity, using the Zodiac killer as the example, as the Zodiac killer had a very unique style of writing.

4. Discussion. An interesting finding was that most of the texts written or spoken by a particular individual had higher similarity indices. For example, the Zodiac killer's letters to police and newspapers always seemed to follow a unique format using similar words and sentence structures. Three documents in particular, which were sent to 3 different newspapers on the same day, had similarity indices of over 0.750 (Figure 4).

It was also intriguing to see that the similarity proximity map displayed clusters of cases that also reflected types of crime. As seen in Figure 3, there was a central cluster of school shooter texts written by different people from different times and locations. What is most remarkable is that the central documents in this particular cluster were written by Eric Harris, one of the perpetrators of the Columbine shooting in 1999-an event that inspired future school shooters. These writings can be seen to have links with other writings of successive school shooters' writings, meaning that there is a linguistic link of influence or at the very least a similar linguistic style.

The results of our study reflect the notion in forensic linguistics that these three criminal groups are quite similar, as some of these clusters are ambiguous and contained a large amount of outliers. However, with computational analysis of linguistics using text mining, we were able to discover small sub-clusters of cases that reflect the type of crime committed or the unique writing style of a particular individual.

References

1. Frand J. "Data Mining: What is Data Mining?"[Internet resource]: URL:http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm.

2. Porter M. F. "An algorithm for suffix stripping"// Program. 1980. Vol. 14, No 3. P. 130–137.

3. "Dendrograms and Clustering"[Internet resource]: URL: https://docs.tibco.com/pub/spotfire/5.5.0-march-2013/UsersGuide/heat/heat_dendrograms_and_clustering.htm (date: 03.13).

4. Provalis Research. "Hierarchical Clustering and Multidimensional Scaling"// WordStat v7.0.

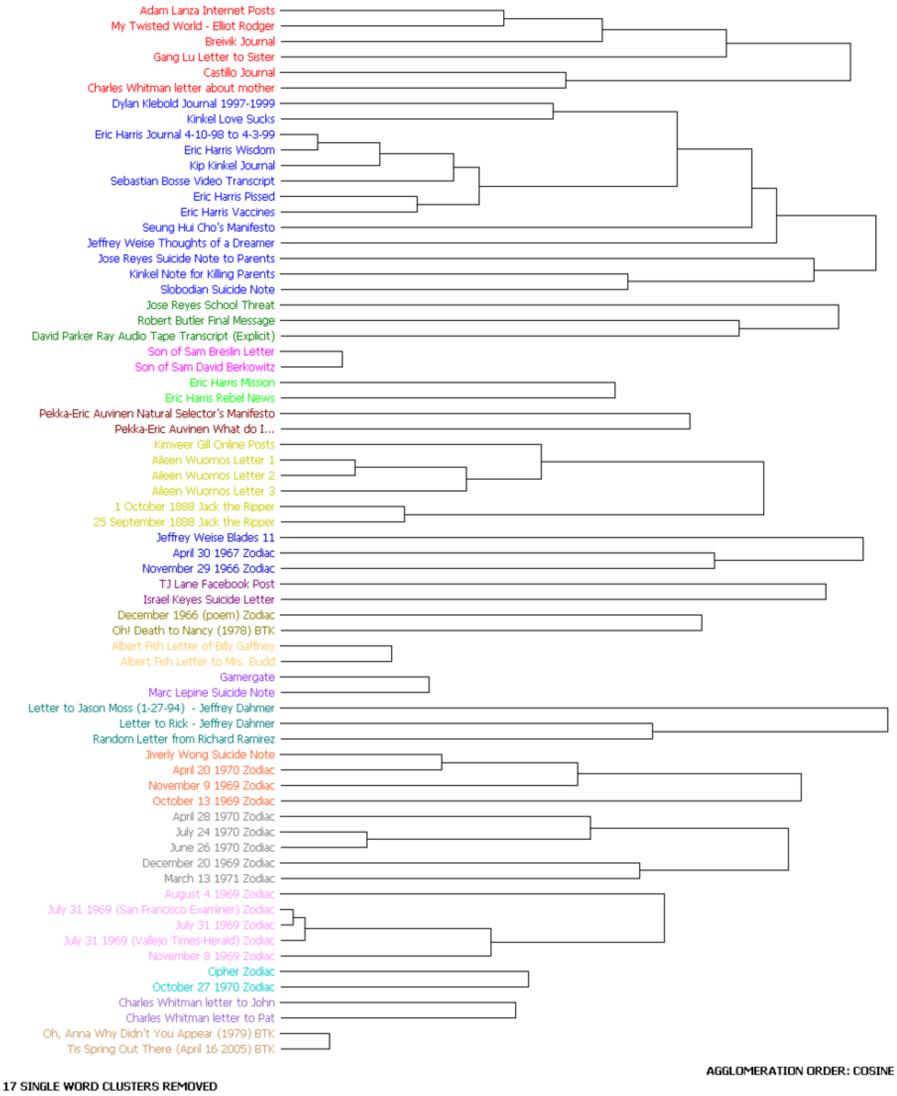

Fig. 1. Agglomeration Dendrogram for Text Similarity

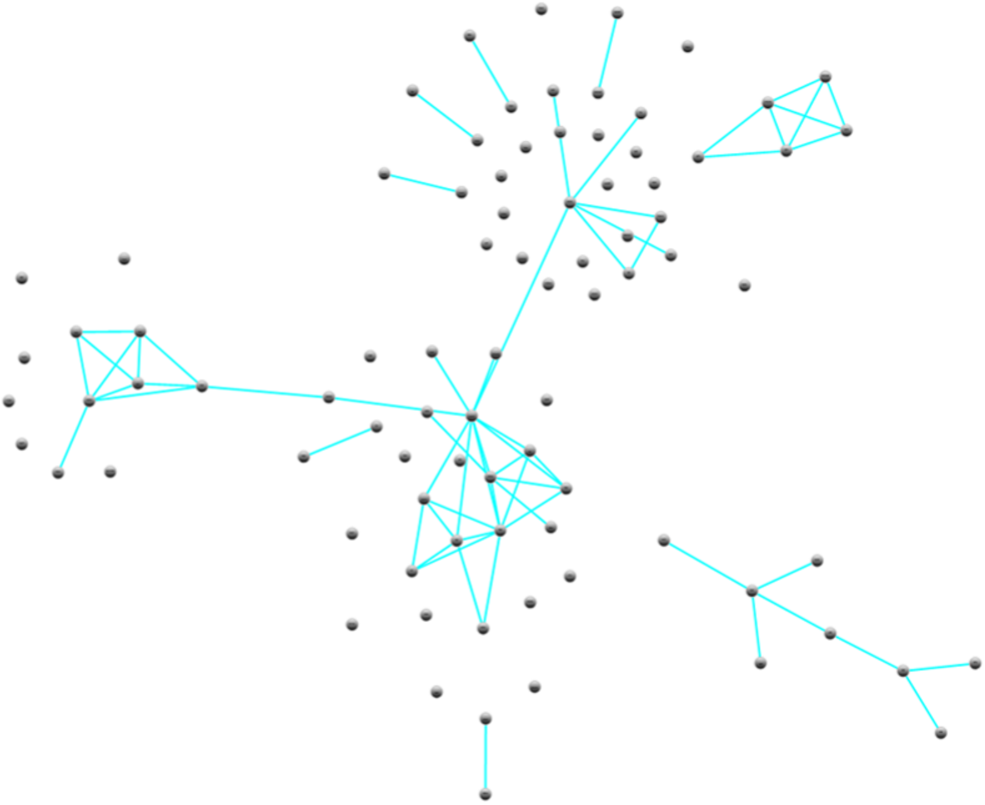

Fig. 2. Entire similarity proximity map with relevant similarity

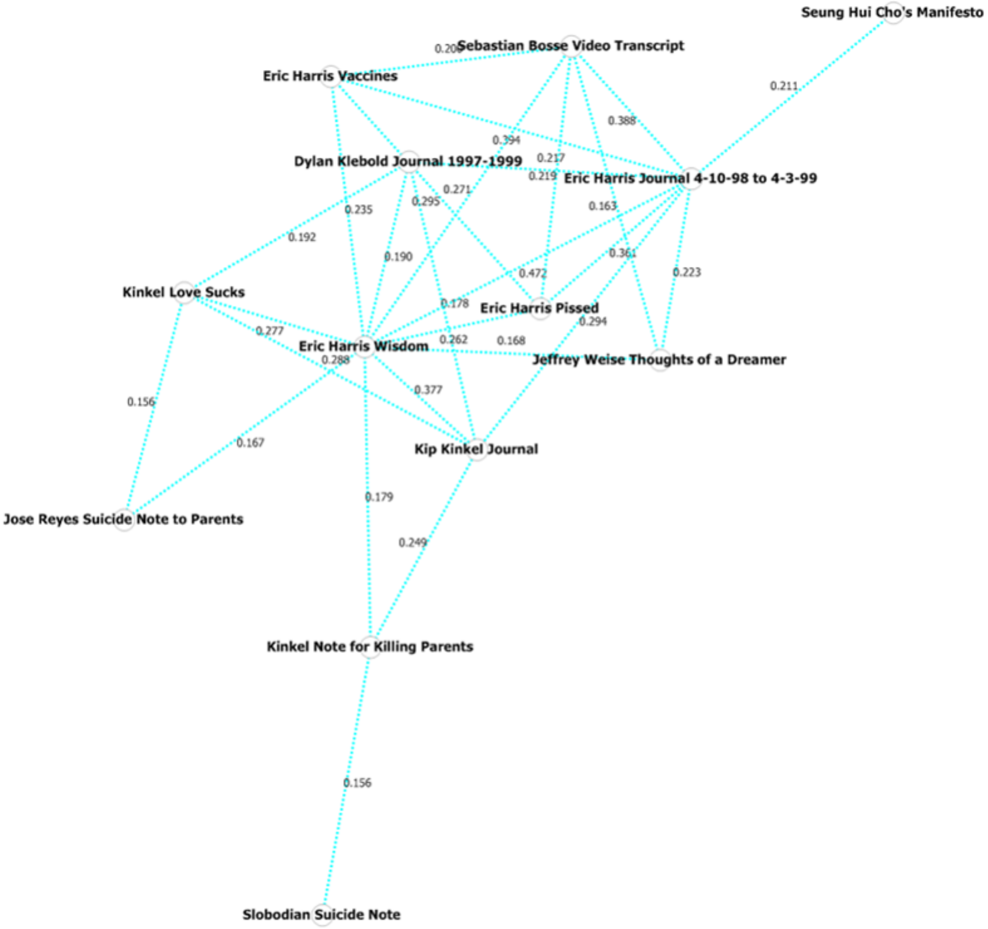

Fig. 3. Cluster of school shooter cases with similarity indices

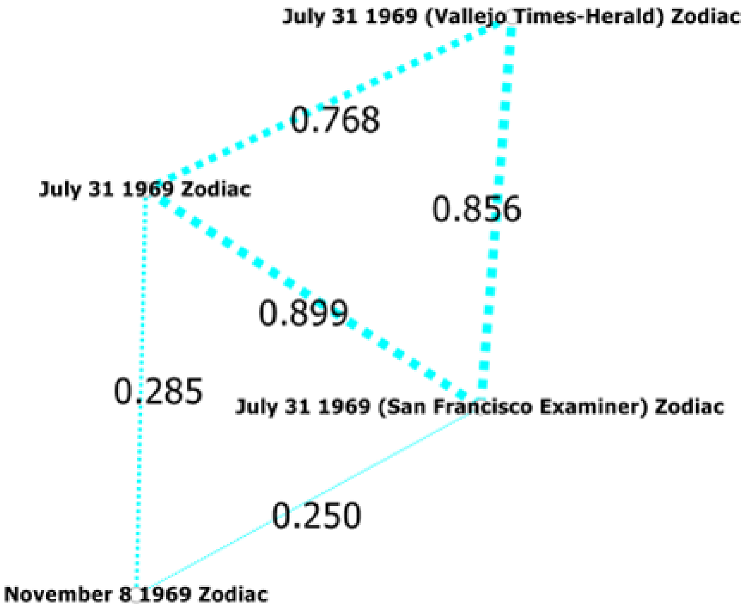
July 31 1969 (Vallejo Times-Herald) Zodiac
0.768
0.856
July 31 1969 Zodiac
0.899
0.285
July 31 1969 (San Francisco Examiner) Zodiac
0.250
November 8 1969 Zodiac

Fig. 4. Zodiac killer case cluster with similarity indices