A Visual Analytics Approach for Crime Signature Generation and Exploration

Wolfgang Jentner, Geoffrey Ellis, Florian Stoffel, Dominik Sacha and Daniel Keim

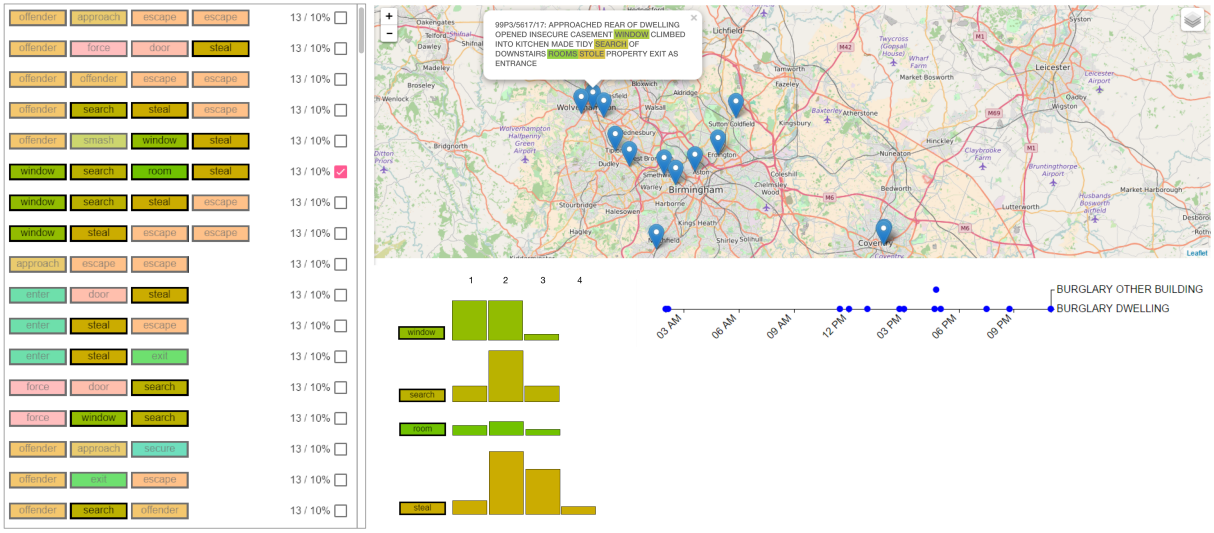

Fig. 1. Screenshot of the current system. Crime fingerprints satisfying a minimum support of 10% are displayed on the left. Each concept is colored differently. Crimes containing the selected fingerprint *window search room steal* are displayed on the map and on a timeline below the map. The crimes are displayed according to their respective crime types. The Modus Operandi text of one of the crime reports is displayed as a popup on the map; concepts of the selected pattern are highlighted. The bar charts show the number of occurrences of each of the selected concepts within the patterns and where each one occurs in the sequence.

Abstract—The exploration of volumes of crime reports is a tedious task in crime intelligence analysis, given the largely unstructured nature of the crime descriptions. This paper describes a Visual Analytics approach for crime signature exploration that tightly integrates automated event sequence extraction and signature mining with interactive visualization. We describe the major components of our analysis pipeline — crime concept/event extraction, crime sequence mining, and interactive visualization. We illustrate its applicability with a real world use case. Finally, we discuss current problems, future plans, and open challenges in our development of a solution that incorporates automated event pattern mining with human expert feedback.

Index Terms—Event extraction, sequence mining, event analysis, interactive visualization, visual analytics, crime analysis

1 Introduction

When analyzing crime reports, it is important to find similar crimes, series of crimes, and potential regions or times of significant crime activity. However, it is impractical to perform such tasks manually given an excess of one million crime reports over 3 years, even when limited to burglary cases. Apart from basic data such as the date, time, place, officers involved, the most detailed information on a crime is contained in a block of text called the *Modus Operandi*. Unfortunately, the freely written text is unstructured, does not follow any strict grammar rules, is written in uppercase and has little punctuation and hence poses a challenge for automatic processing.

Our approach is to extract semantically important concepts from this text. The set of concepts and their order of occurrence represent a *crime signature*. This serves as an ordered list of events which can be mined for sequential patterns with suitable mining algorithms. The resulting data and associated metadata are fed into our Visual Analytics (VA) tool which, with the use of multiple visualizations, assists the analyst in making sense of the sequential patterns and, hence, gaining a better understanding of the criminal activity.

The next section provides a short overview of related work in sequential pattern mining algorithms and event visualization techniques. Section 3 gives details of the crime data and analysis tasks and in Section 4 the main components of the visual analytic system are described. A use case in Section 5 illustrates the application of our approach and in Section 6 we review our plans to add further functionality to the system. Finally, our research contributions are discussed in Section 7.

2 Related Work

The formal definition of sequential pattern mining was established by Agrawal et al. [3]. Numerous algorithms followed to seek efficient solutions to that problem [7, 13]. Extensions are able to mine closed sequential patterns [11] and maximal sequential patterns [8]. Additional constraints allow the search for more specific patterns [16]. The application area is very broad and includes web log mining, click streams, and bio data [13]. Typically, these applications are dealing with many

• *All authors are with the University of Konstanz, Germany.*
Data Analysis and Visualization Group
E-mail: forename.lastname@uni-konstanz.de

Proceedings of the IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis. Available online at: http://eventevent.github.io

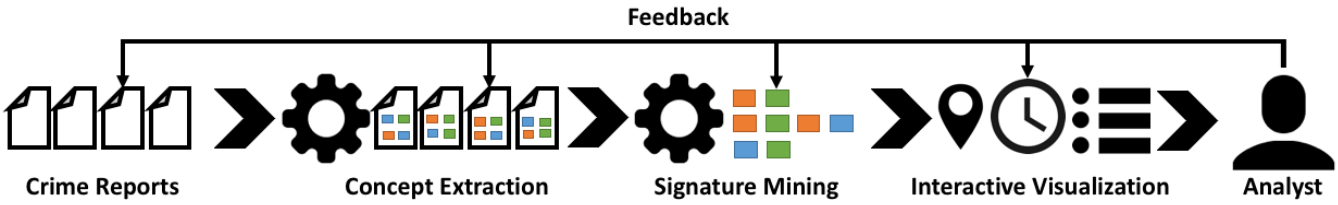

Fig. 2. The VA pipeline combines concept extraction, signature mining with interactive visualization. The analyst is in the loop and can feedback to each component, which includes filtering on crime reports, steering the context extraction and sequential pattern mining, and interact with the visualizations.

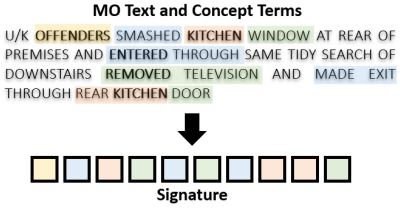

Fig. 3. Concept and Fingerprint Extraction Example. Text taken from [2]. Concept words are indicated bold, the color indicates different detected contexts.

but short sequences.

Event visualization techniques in text analysis are usually coupled to topic modeling algorithms. LeadLine [6] detects events in news and microblogging data. If geo locations are provided they can be additionally displayed on a map. ThemeRivers [12] are often used to display such events in streaming data. Spatio-temporal events can be also displayed in a space-time cube [10] or by animating events on a 2D-map [5]. Activitree [18] uses a tree layout to identify significant sequences in temporal event data.

3 Crime Report Data and Analysis Tasks

Police analysts have to deal with a very large set of crime reports. These reports are related to available metadata such as time, geographic information, or other categorical information such as the type of a crime (e.g. burglaries). Importantly, the provided crime report contains a sequential textual description of the crime. The analysts are interested in specific word occurrences within this sequence, e.g., "how the offender entered a house" or "what has been done within" and "how they left it". We call such a sequence of interesting terms in such a textual description a *crime signature*. Note, that such crime signatures are relevant for two different events and sequential analysis: (1) The sequence of terms has to be analyzed to extract a set a meaningful crime signatures, (2) for each crime signature we can further analyze to check if there are also temporal or geographic relations. E.g., if we treat a specific crime signature as an event in time and place, it is possible to uncover patterns, either algorithmically or visually, in both space and time. Consequently, analysis tasks and questions are to (1) analyze frequent crime signatures and their co-occurrences. Note that due to the nature of sequential pattern mining, frequent patterns can be contained in other frequent patterns. This sequence containment can be exploited to draw a context in between patterns. (2) Furthermore, the crime signatures have to be related to the temporal (days, day, night, periods) and geographic dimensions (specific regions, districts).

4 Visual Analytics Approach

The VA pipeline consists of three major components (1) crime signature extraction from crime report data, (2) sequential crime signature pattern mining, and (3) interactive visualization of crime signatures including related temporal and geospatial data.

4.1 Concept Extraction

We analyze a dataset containing numerous crimes, that are described by metadata such as date, time, crime category, as well as a text field commonly referred to as *Modus Operandi* (MO, the description of procedure). Our work is based on the MO data, because it contains the most detailed information about the incident such as the procedure of entering a building in a burglary case, which is not found in any metadata. Higher level information extraction to identify phrases or chunks of text require that the analyzed text contains a syntactic structure [1]. Unfortunately, the MO only contains fragments of phrases and does not provide a useful overall syntactic structure, as illustrated by an example in Figure 3. This collection of unconnected chunks, e.g. `ENTRY GAINED UNTIDY SEARCH`, present a problem for state-of-the-art natural language processing toolkits [14, 15]. In addition, the entity detectors from both, OpenNLP and Stanford CoreNLP libraries, are not able to identify names or places in the MO text at all. Given these problems with current NLP methods, we adopt low-level techniques in order to extract information from the MO text. This gives the highest degree of flexibility and the most control over the processing and extraction methods, which is particularly useful in tailoring the information extraction pipeline to this specific kind of text data. In contrast to keywords, we would like to have a phrase like structure from the document, that makes it possible to understand the context of a word. To do so, we propose to extract *concept n-grams*, a type of word level n-grams containing at least one concept word and one or more other terms. Multi-word terms are known to be useful descriptors of the document contents [9], so this is a suitable way to represent the contents of an MO. A concept word is considered to stem from a manually curated list of words that have application relevance in the processed type of data. In this respect, we employ a set of ten different concept classification, ranging from clothing brands, parts of a building, words describing movements and behavior of persons. After the MO analysis, a crime report is represented by the set of extracted concept n-grams, which is the input for the signature mining process described next.

4.2 Crime Signature Mining

The algorithm we use to find frequent crime signatures is related to the SPAM approach [4] but is adapted to mine for closed and maximal sequential patterns if required. Due to the nature of sequential pattern mining algorithms, patterns can be contained in patterns consisting of more items. The set of maximal sequential patterns only consists of patterns that are not contained in any other pattern. However, this may lead to a loss of information as a contained pattern might have a higher support meaning that it supports more sequences. A pattern which is not contained in any other more complex pattern or has a higher support is called a closed sequential pattern. Note, that even the set of maximal sequential patterns can be large, thus requiring further filtering for interesting or targeted patterns. As shown in Figure 3, concepts are extracted and represent events occurring in order and with a "timestamp" respective to their position within the text. A frequent crime signature is a subset of the extracted concepts whereas the events have to occur in the same order but may have gaps. The number of gaps is one of the algorithm parameters, which can be set to unlimited to find all possible patterns. Patterns are called frequent if they satisfy a given minimum support which can also be set by the user. By default, the algorithm returns the set of maximal sequential patterns; ones which are not contained in any longer pattern. This helps considerably to reduce the size of the result set. As the user modifies the minimum support, the result set of frequent patterns changes. Higher minimum

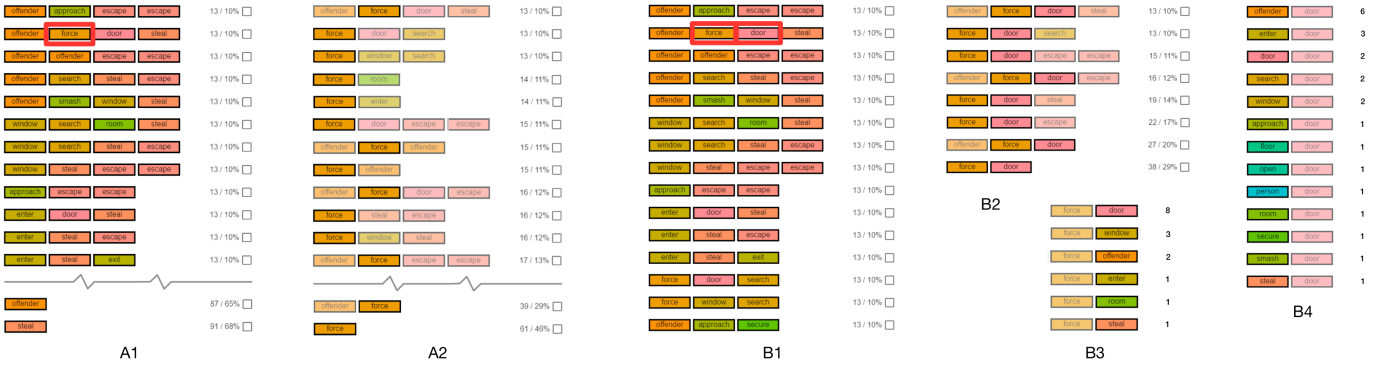

Fig. 4. Interaction examples with the filtering and recommendation of patterns: The user selects a concept (A1) resulting in a filtered list of patterns containing that concept (A2). In (B1), the user opts for two concepts which results in three grouped lists. (B2) contains patterns with both selected concepts and (B3) and (B4) with either selected concept. B3 and B4 are provided automatically to the user.

support results in shorter patterns as long item-sequences become more unlikely and will, therefore, be filtered out.

4.3 Visualization and Interaction

Figure 1 is a screenshot of the system prototype, consisting of four main components. The sidebar on the left displays the extracted frequent crime signatures ordered, in this example, by increasing support. The analyst can select one or more of the sequences to explore further (e.g. *window search room steal* as shown) and the crime reports metadata pinpoints the location of each crime on the zoomable map and also shows when they occurred on the timeline below. The crime classification is used to differentiate the types of crimes (e.g. different classes of burglary). The user can also view the Modus Operandi text of any crime report, with concepts colored as in the crime signature list on the left. To allow the analyst to relate sequential, temporal, and geographic patterns, additional information is presented automatically to assist in making sense of the often complex data. For example, in Figure 1, the small bar charts show the number of occurrences of each of the selected concepts within the crime signature list. Because the position of a concept within the sequence is relevant to understanding its meaning, we also show an aggregate where each one occurs in the sequences.
In addition to providing the user with basic filtering based on space and time, as well as being able to browse and select particular crime signatures, the system allows the user to explore individual and sequences of concepts. Fig 4 shows two examples of this. In diagram A1, the user is interested in and hence has selected the concept *force*. The system subsequently shows the user a list of the crime signatures containing this concept, as in A2. Note that *force* is the last entry in the list with a support of 46%. In the second example, B1, the user has selected the sequence *force door* and similarly, is presented with all the sequences which contain this pair (see B2). Following, the principle of automatically providing additional information, the system shows all other concepts which follow *force* (see B3) as well as all the other concepts which proceed *door* (see B4). The user is also informed of the number of each sequence which is in the crime signatures list. The user can interact with these supplementary lists to explore any of these sequence pairs as they wish. Due to the importance of the position of a concept within a sequence, it is also possible to filter on a concept at a particular position in a crime signatures by clicking on the appropriate bar in the small bar charts shown in Figure 1.

The system additionally lets the analyst interact with computational components of the pipeline. For example, the crime signature mining algorithm can be steered by removing specific concepts that are not of interest, as well as tuning parameters, such as the minimal length and the minimal support of a pattern.

5 Example Case

An analyst is considering burglaries on New Year's Eve. Searching for frequent patterns in the crime reports on that day gives a list as shown in Figure 1. They are interested in the way the suspects entered the building. The list contains two patterns which only consist of one concept: *door* and *window*. Searching for these two concepts filters the list and only patterns containing either remain. The first one is mentioned in 75 crime reports and the second concept occurs in 59 crime reports. By selecting both patterns the analyst can see that most of these crimes were reported between 10am and 9pm. By selecting them individually they can see that more burglaries containing *window* happened in the city of Coventry, UK. By reading some of the crime reports the investigator learns that in some of the cases the door was used by the criminal to exit the building. An example fragment of the Modus Operandi is "EXIT VIA REAR DOOR" that occurs at the end. More complex patterns, containing more items, are contained in the result list. These are *window steal* and *door steal*, both contained in 43 crime reports. The patterns can be interpreted as follows: Firstly, either *window* or *door* has to be mentioned in the Modus Operandi, later on, the word *steal* must occur. This is likely to remove crimes where the suspect exited via a door or a window and thus increases the precision. Figure 5 shows a comparison between the two patterns. The crime reports are plotted onto the map and the timeline on the bottom. In general, the southeast of Birmingham, UK is less affected. Burglaries, where the suspect entered through a window, took place twice as much in the city of Coventry as when doors were used as a method of entry.

By comparing the cities of Coventry and Birmingham for three years, the support of patterns found in both regions can be pairwise compared. Although no statistical tests for significance are performed, some interesting observations can be made. For the majority of the patterns, the support only differs by less than one percent. This is especially true for patterns with fewer items. Longer patterns show more variation. Note, that the overall number of crimes in the cities is different. While more than 50,000 crimes (burglaries) were reported in Birmingham, less than 10,000 crimes were reported in Coventry which is due to the different sizes of both cities.

6 Next Steps

The prototype system described in this paper is still in development and in this section, we highlight some of our planned next steps. An important issue is the filtering and highlighting of relevant or potentially interesting patterns to the user. This will include a recommender system which is trained on what patterns are selected by the user. An early example is already provided in Figure 4, here we can see how the user selects a concept out of one pattern (A1) which leads to a filtering of patterns containing that specific concept (A2). In (B1), two concepts are selected by the user. The system recommends patterns containing both selected concepts or either and automatically groups them accordingly. The concept extraction algorithm allows the extraction of different variants of concepts. Currently, we use single terms concepts, however, it is possible to gather n-grams as well as high-level categories. This may lead to a better semantic understanding of the patterns but requires different layout techniques. Furthermore, it should be possible to steer the concept extraction interactively, which includes

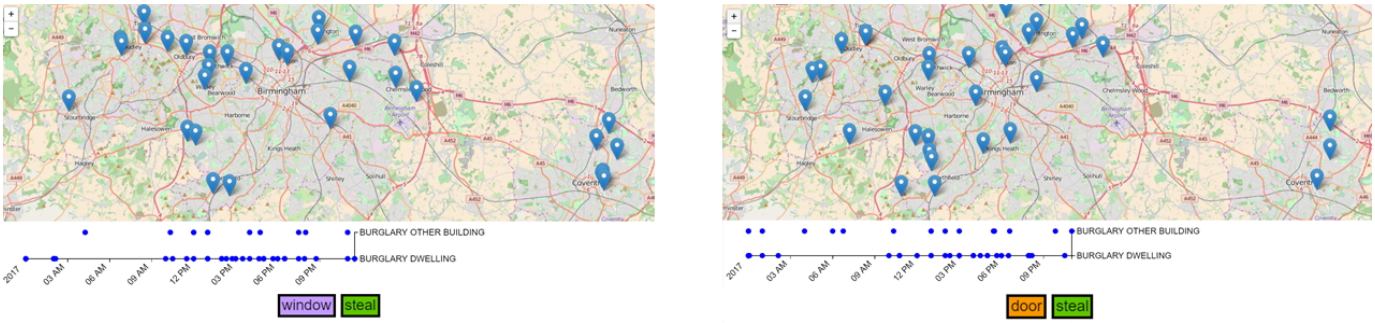

Fig. 5. A comparison of the patterns “window steal” and “door steal” shows that he crimes were reported around the same time but at different locations.

the addition of new concepts which are deemed important to the user. The processing pipeline can be re-executed with the new input. Based on the recommender system and enhanced filtering capabilities, we wish to provide further flexibility for the user to summarize patterns into logical interesting groups. For instance, this may include patterns of the same length, similar terms, or similar positions of terms. To help the analyst understand if specific sequential patterns belong to multiple crimes or if they are disjoint, we would like to investigate the use of Euler Diagrams [17]. Each pattern will represent a set, with events (crime reports) belonging to one or more sets based on the support of the pattern. As briefly introduced in the previous section, it should be possible to compare different sets of patterns based on the user’s spatial or temporal selections. This could be achieved with side by side displays, suitably aligned, accompanied by statistical results if valuable to the user. The crimes serving as temporal events and classified by their crime signature can be analyzed again for temporal event sequences. This can be useful to (semi-)automatically find crime sequences. As mentioned in the introduction, there are a large number of crime reports and hence the tool as well as the visualizations should be capable of handling this quantity of data. From the related work, the Modus Operandi could be visualized with a stream visualization, the temporal and spatial events of crimes with a space-time cube or suitable animation. Due to the support of the VALCRI project[1] we have access to crime report data, as well as analysts and police officers which we intend to engage in further design studies and evaluation.

7 Conclusions

In this paper, we have described the development of a Visual Analytic system to assist crime analysts in making sense of their very large collection of crime reports. The unstructured and peculiar nature of the descriptive Modus Operandi text means that standard NLP processing is generally ineffective in extracting meaning. Modified NLP techniques are used to extract concepts or terms from the raw text, however, to put these in a context we have made use of sequence mining techniques, normally associated with temporal data, treating concepts as an event with a positional timestamp. A use case highlights the importance of the order of the concepts in the extracted crime signature for increased semantic understanding. The number of sequential patterns returned by the algorithm can be very large, so the user is able to interactively adjust the size of the result set and the length of the mined patterns and we envisage that this could be developed into a semi-automatic progress process. Flexibility to explore and analyze the patterns or crime signatures is very important to our users, so the system provides filtering at the signature level and also at the lower concept level, with the option to investigate sequences of concept terms, including adding constraints on the position of concepts within a pattern. To enhance the analysis, we automatically provide additional and related information to the user.

Acknowledgments

This work is supported by the EU project VALCRI (FP7-SEC-2013-608142).

[1] http://valcri.org/

References

[1] S. Abney. Parsing by chunks. In R. Berwick, S. Abney, and C. Tenny, eds., *Principle-Based Parsing*, vol. 44 of *Studies in Linguistics and Philosophy*, pp. 257–278. Springer Netherlands, 1992.

[2] R. Adderley. Exploring the differences between the cross industry process for data mining and the national intelligence model using a self organising map case study. In *Business intelligence and performance management*, pp. 91–105. Springer, 2013.

[3] R. Agrawal and R. Srikant. Mining sequential patterns. In *Data Engineering, 1995. Proc. 11th Int. Conf. on*, pp. 3–14. IEEE, 1995.

[4] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu. Sequential pattern mining using a bitmap representation. In *Proc. 8th ACM SIGKDD Int. conf. on Knowledge discovery and data mining*, pp. 429–435. ACM, 2002.

[5] B. Demissie. *Geo-visualization of movements: moving objects in static maps, animation and the space-time cube*. 2010.

[6] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou. Leadline: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conf. on*, pp. 93–102. IEEE, 2012.

[7] P. Fournier-Viger, A. Gomariz, M. Campos, and R. Thomas. Fast vertical mining of sequential patterns using co-occurrence information. In *Adv. in Knowledge Discovery and Data Mining*, pp. 40–52. Springer, 2014.

[8] P. Fournier-Viger, C.-W. Wu, A. Gomariz, and V. S. Tseng. Vmsp: Efficient vertical mining of maximal sequential patterns. In *Advances in Artificial Intelligence*, pp. 83–94. Springer, 2014.

[9] K. T. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: the c-value/nc-value method. *Int. J. on Digital Libraries*, 3(2):115–130, 2000.

[10] P. Gatalsky, N. Andrienko, and G. Andrienko. Interactive analysis of event data using space-time cube. In *Information Visualisation, 2004. Proc. 8th Int. Conf. on*, pp. 145–152. IEEE, 2004.

[11] A. Gomariz, M. Campos, R. Marin, and B. Goethals. Clasp: An efficient algorithm for mining frequent closed sequences. In *Adv. in Knowledge Discovery and Data Mining*, pp. 50–61. Springer, 2013.

[12] S. Havre, B. Hetzler, and L. Nowell. Themeriver: Visualizing theme changes over time. In *Information Visualization, 2000. IEEE Symposium on*, pp. 115–123. IEEE, 2000.

[13] N. R. Mabroukeh and C. I. Ezeife. A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys (CSUR)*, 43(1):3, 2010.

[14] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *Proc. 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*, pp. 55–60, 2014.

[15] T. Morton, J. Kottmann, J. Baldridge, and G. Bierner. Opennlp: A java-based nlp toolkit, 2005.

[16] R. Srikant and R. Agrawal. *Mining sequential patterns: Generalizations and performance improvements*. Springer, 1996.

[17] G. Stapleton, P. Rodgers, J. Howse, and L. Zhang. Inductively generating euler diagrams. *IEEE Transactions on Visualization and Computer Graphics*, 17(1):88–100, 2011.

[18] K. Vrotsou, J. Johansson, and M. Cooper. Activitree: Interactive visual exploration of sequences in event-based data using graph similarity. *IEEE Trans. on Visualization and Computer Graphics*, 15(6):945–952, 2009.