



City Research Online

City, University of London Institutional Repository

Citation: Nguyen, P., Turkay, C., Andrienko, G., Andrienko, N. & Thonnard, O. (2017). A Visual Analytics Approach for User Behaviour Understanding through Action Sequence Analysis. Paper presented at the EuroVis Workshop on Visual Analytics, 12-13 Jun 2017, Barcelona, Spain.

This is the published version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: <http://openaccess.city.ac.uk/17284/>

Link to published version:

Copyright and reuse: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

City Research Online:

<http://openaccess.city.ac.uk/>

publications@city.ac.uk

A Visual Analytics Approach for User Behaviour Understanding through Action Sequence Analysis

Phong H. Nguyen¹, Cagatay Turkey¹, Gennady Andrienko^{1,2}, Natalia Andrienko^{1,2} and Olivier Thonnard³

¹City, University of London, UK

²Fraunhofer IAIS, Germany

³Amadeus, France

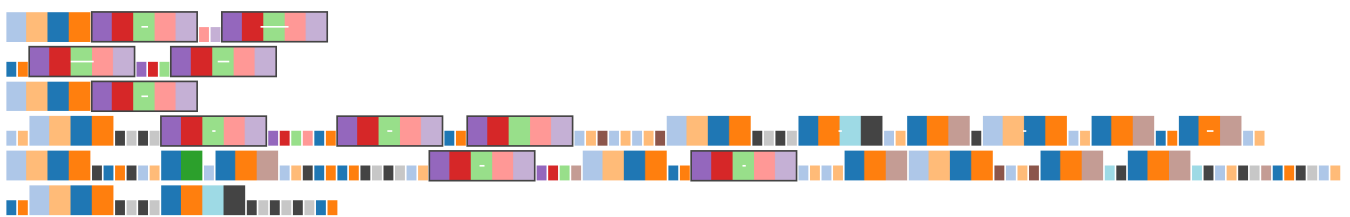


Figure 1: Comparison of user sessions using small multiples of visual summaries. Each row summarises a session, showing its actions and activities (meaningful sub-sequences of actions produced by a pattern mining algorithm) in temporal order with colour indicating the action type. Consecutively repeated activities are aggregated and the length of the white line in an activity indicates the number of occurrences in that aggregate. Frequent activities are visually highlighted, allowing analysts to quickly understand what happen in the sessions and facilitate comparison between different sessions.

Abstract

Analysis of action sequence data provides new opportunities to understand and model user behaviour. Such data are often in the form of timestamped and labelled series of atomic user actions. Cyber security is one of the domains that show the value of the analysis of these data. Elaborate and specialised models of user-behaviour are desired for effective decision making during investigation of cyber threats. However, due to their complex nature, activity sequences are not yet well-exploited within cyber security systems. In this paper, we describe the initial phases of a visual analytics approach that aims to enable a rich understanding of user behaviour through the analysis of user activity sequences. First, we discuss a motivating case study and discuss a number of high level requirements as derived from a series of workshops within an ongoing research project. We then present the components of a visual analytics approach that constitutes a novel combination of “action space” analysis, pattern mining, and the interactive visual analysis of multiple sequences to take the initial steps towards a comprehensive understanding of user behaviour.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical user interfaces (GUI)

1. Introduction

User action sequence is a data form that is widely gathered in various domains and provides valuable insight into how digital systems are being used [WPB01]. Commonly, these sequences are series of timestamped and “labelled” atomic actions that are performed by users of a system over a period of time and organised into sessions. In cyber security, action sequences are analysed to build models of user behaviour [GH11], which are eventually utilised to identify anomalous activities and threats [CBK12]. However,

such sequences are not straightforward to analyse since they contain several semantically related *patterns* (series of actions) that are driven by *user intent* which varies significantly across users and time. The inherent “noise” in these sequences (i.e., irrelevant actions performed by users) and the data volume make the analysis even more challenging. These characteristics of the domain often lead to a high level of uncertainty within the fully automated analysis of such data and thus requires a thorough understanding of the actions and the activities that take place within sessions.

This open-to-interpretation and multi-faceted (i.e., time, users, action patterns) nature of action sequences make this a problem domain that calls for methods involving a human analyst within the analysis process, and visual approaches have already shown great potential in tackling such challenges [SFK13, Leg15]. This paper describes a visual analytics approach that aims to provide a comprehensive understanding of user behaviour through the visual analysis of action sequences. We are primarily motivated by a case where action sequences are analysed to build models of user behaviour within cyber systems. Here, we firstly present the motivating case study and discuss the initial results from a series of user workshops that we have carried out to elicit the requirements. We then introduce the components of a visual analytics approach that constitutes of a suite of computational and visual methods. In order to analyse the similarities within the actions and to visualise sessions within the context of actions, we compute *action spaces* and demonstrate how they can help to distinguish interesting sessions. We then discuss the derivation of semantically relevant functional units through an activity mining approach which results in activities that can be utilised as “analysis objects” within further visual investigation. We then demonstrate how a single session can be summarised and how multiple sessions can be compared to confidently understand and evaluate the nature of particular “suspicious” sessions.

2. Domain Characterisation

In this section, we firstly describe a case study where the analysis of action sequence data is an essential task. The presented case study has been identified within an ongoing collaborative research project as an example where a visual analytics approach can highly benefit. In the next section, we analyse a dataset that comes from this case study and comprises of 17000 sessions performed by around 1400 distinct users.

2.1. Motivating Case Study

A company is interested in detecting misuse and fraudulent activities that its employees may carry out whilst using one of its applications. To enable this, the company captures user actions for further analysis and modelling. The log data is organised into *sessions* – identified by a unique id assigned automatically at the beginning of a user session. Each session is performed by a single user and contains an ordered list of *actions* with two types of information recorded: the *time* when an action was performed and the *type* of that action (such as `SearchUser` and `DisplayOneUser`). The type of actions are determined automatically by developers of the application and logged accordingly. Once all such data is logged, a probabilistic model is built to automatically compute an anomaly score for each session. If the computed score is high, an investigation into that session needs to be conducted to validate the score and search for an explanation. Currently, such an investigation is highly manual and time consuming. The analyst needs to examine sessions in a spreadsheet-like format (data tables) with the help of pie charts showing summary statistics of action types appeared in selected sessions (such as top 10 most common actions). Both the large number of sessions required to analyse and their lengths worsen the manual investigation problem.

2.2. User Requirements

Based on the previous case study and our observation of several investigation sessions performed by end users, we set the following requirements for supporting such an anomaly investigation.

R1 – Single session exploration: Help analysts understand what happened in a single session.

R2 – Multiple sessions comparison: Help analysts identify similarities and differences among multiple sessions.

3. Analysis and Design

3.1. Action Space

There is often an underlying functional and semantic similarity between actions and thus actions often occur jointly in many sessions. We hypothesise that an analysis approach that focuses on identification and externalisation of these similarities and differences has a role in understanding the “normality” of sessions. Hence, we devise an approach to analyse sessions within the context of *action space*. Action spaces are 2D mappings where each action is represented and positioned according to its *relation* to other actions. To construct an action space, we compute pairwise “distances” between actions based on the median distances between the occurrences of the action names in the action sequences.

Our approach applies a simplified text mining technique on lexical co-occurrence [LB96], considering sessions as “sentences” and actions as “words”. The distance between two words (w) occurrences in a text is then the number of other words between them plus 1 (hence, if w_2 immediately follows w_1 , the distance from w_1 to w_2 is 1). We also set a parameter N_{max} as the maximal allowed number of words between two words. For each instance where w_1 and w_2 (to re-iterate – w_1 and w_2 are unique action *types*) co-occur, we find the distance between them. We then average distances from all the instances (i.e., sessions) where these two co-occur and set the median distance to be the eventual distance. Once the distance matrix is computed for all pairs, we apply a projection algorithm (Sammon’s mapping [Sam69]) to obtain a 2D arrangement of the set of actions according to their pairwise distances. The actions that often occur together in a session are expected to appear close in the resulting action space.

This action space then gives us a new medium where we can visualise sessions as trajectories where actions are “visited” in a sequence. An example action space can be seen in Figure 2 (left) with all the sessions in the dataset are rendered to indicate which actions are often performed jointly. A “typical” session in this action space is expected to span a number of actions that are close (i.e., often co-occur) and an “interesting” session flows through a number of actions that are in a distant location in the space.

3.2. Activity Mining

A session contains an ordered list of timestamped and labelled actions, with labels determined by the developers of the application. Even though each action is associated with a meaningful label indicating its purpose (such as `SearchUser` and `DisplayOneUser`), it is still challenging to understand the nature of a ses-

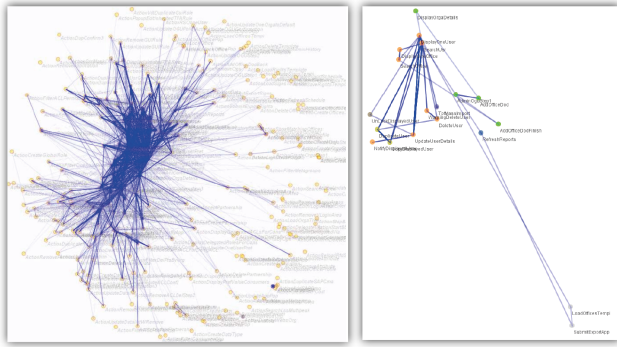


Figure 2: An “action space” where similar actions are mapped close to each other. Sessions overlaid as trajectories that “visit” actions (left). A single session where some “distant” actions have been carried, making this a less usual session, thus worthy of further investigation (right).

sion due to the large number of actions (many sessions containing more than 100 actions). Moreover, early investigations reveal that actions do not appear randomly. They often appear together as short “patterns” where a higher level *activity* is carried out, such as `SearchUser` → `DisplayOneUser` to retrieve the details of a user. Figure 3 shows an example of a session having a frequently recurring pattern.

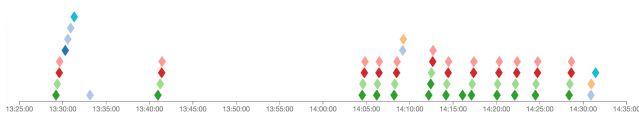


Figure 3: Actions in a session can follow a recurring pattern. Each diamond is an action and colour-coded by action type. Sequence “red → light red → green → light green” appears 12 times.

In order to both simplify the action space and represent the session data with a higher semantic level (addressing **R1**), we mine the *activities* from raw user actions. More specifically, given an ordered list of actions in a user session, we split the list into contiguous and disjoint sequences so that each sequence (an ordered sub-list of actions) represents a meaningful activity that the user performed (We refer to these artefacts as *activity* from now on in the text). It is reasonable to assume that those activities are a subset of frequent action sequences because small activities are supposed to be repeated many times in many different tasks, which are carried out in many different sessions (as similarly evidenced in other analysis settings [GGZL16]). Extracting frequent action sequences can be implemented using classic sequential patterns mining algorithms such as AprioriAll [AS95] and GSP [SA96]. However, the number of sequences produced by those algorithms can be very high and the majority of them may not represent meaningful activities. To exclude non-activity sequences, we apply several constraints such as the maximum time gap between two adjacent actions in a sequence.

We visualise the sequences produced by the mining process to


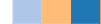
communicate the frequent activities performed as shown in Figure 4. The visualisation consists of multiple rows, where each represents an activity and is split into two parts: the right part visualising the actions in an activity and the left section listing statistics on these actions. Each activity is represented as a contiguous sequence of colour-coded squares, where each square represents an action. To characterise the “frequency” of an activity, three statistics are visualised in nested bars: the number of times the activity appears (biggest bar), the number of sessions having that activity (medium bar), and the number of users performing it (smallest bar). For instance, compare activity  (second top) and activity  (second bottom). The former repeats many times more than the latter, but taking place in a few sessions by a few users, whereas the latter is spread more evenly across several sessions and users.



Figure 4: Frequent activities. Each row represents an activity. The right part indicates its sequence of actions with colour-coded squares representing action types. The left part shows three statistic values of the activity: the number of times it appears, the number of sessions it appears in, and the number of users performing it.

3.3. Visual Summaries of Sessions

3.3.1. Single Session

To help analysts gain understanding of a single session (addressing **R1**), we devise a summary visualisation. Since both the timestamp and the action type are of importance, they are both considered in the visualisation. We start with a standard timeline representation [NXWW14, NXWW16] due to its simplicity and familiarity. Actions are shown as diamond glyphs along a horizontal time axis at when they happen and are colour-coded based on their types (Figure 3). Only the most common 20 actions performed by the session user are colour-coded (using the colour set provided in the D3 library [BOH11]) and the rest of the actions are indicated with grey. Our approach here can be compared to the LifeFlow [WGGP*11] and EventFlow [MLL*13] methods that aim to explore and summarise a large number of action sequences. These methods often limit themselves to known subsequences and here we investigate how we can incorporate derived actions and multiple aggregation levels within the visual summaries.

Because of the large number of actions and the limited meaning they carry, we replace sequence of actions with activities whenever possible. An activity is represented by an “extended” diamond

covering the duration from the first to the last action in that activity. The glyph is also horizontally divided into equal segments where each is colour-coded to represent an action. Figure 5a shows the same session as in Figure 3 but with an emphasis on frequent activities. To emphasise activities and simplify the representation even further, consecutive activities are aggregated as in Figure 5b. A little white dot indicates when an activity instance happens to subtly preserve the temporal information. These three representations (actions → activities → aggregated activities) allow analysts to investigate a session at different levels of details. At the highest level, the analyst observe what activities are performed and further examine individual activities to dive into raw actions as necessary.

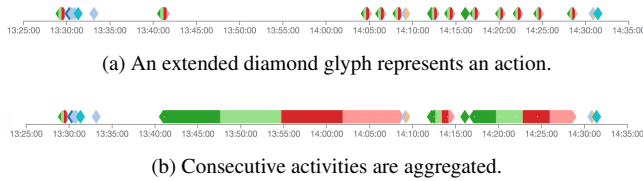


Figure 5: Visual summary of a session using activities.

With this representation, when activities are highly frequent or the session is long, the activity glyphs become narrow and harder to read. We address this issue by sacrificing the absolute temporal information, and only preserving the relative order between activities. Another scenario where this representation is effective when the focus of the analysis is on the type of activities themselves rather their time. Figure 6 shows the same session as in Figure 5 but only preserve temporal order between activities instead of their absolute timestamps. The activity and action glyphs are replaced with rectangles and positioned in the order they occurred. Action glyphs have half width and half height of activity glyphs to make activities more noticeable. A horizontal white line is shown in an activity with its length indicating the number of times the activity consecutively repeats.



Figure 6: Visual summary of a session based on relative temporal order.

3.3.2. Multiple Sessions


Very often, an analyst queries all sessions by a user to reconstruct what that user typically does, and to compare a given session against the other sessions of that user (addressing R2). We support this through small multiples of visual summaries of sessions with each session shown in a separate row (Figure 1).

When an analyst selects a user to investigate, all of his or her sessions are shown in the timeline using small multiples. Sessions are ordered by the time they begin, allowing the analyst to quickly understand the user activities over time. The analyst can highlight the session that needs to be investigated and visually compare it with other sessions. When an activity or an action is hovered, all other occurrences are highlighted (the ones with black borders in

Figure 1), allowing the analyst to quickly see the similarities and differences.

The timeline and the activity visualisation in Figure 4 are linked together to allow a quick activity lookup to understand its statistics. Also, a user can click on an activity in Figure 4 to show all sessions that have the selected activity in the timeline, providing further contextualisation of activities.

4. Application Example

Here we discuss a brief example of how different components in our approach can be used to gain an understanding into user behaviour. Observing an overview of frequent activities in Figure 4, we identify an interesting recurring activity  in the second row containing five actions: DuplicateToExistingUser → FilterUser → DuplicateToExistingUser1 → DuplicateToUserConfirmation → DupConfirm1. Clicking on that activity makes all sessions containing it display in the timeline for further investigation. This “user duplication” activity is dominant in all of this user’s sessions. We then notice that all of those sessions performed by the same user (the user list is not described in the paper due to limited length), and select that user to examine all of his sessions, which are shown in Figure 1. It turns out that the user performs only one more session (the bottom one), and this session may involve a task different from what he usually does. Depicting several session of the user in a small multiple setting and the aggregation of actions into activities enables analysts to gain an overview of the common tasks done by a user and eventually reconstruct a comprehensive understanding of behaviour.

5. Conclusion

This paper presents initial results from an ongoing project where a comprehensive understanding of user behaviour is required for the robust modelling of users and identification of anomalies. Through a motivating case study, we list a number of high level requirements and describe how a visual analytics approach might be instrumental in addressing these. We observe that the multi-faceted nature of action sequences requires one to investigate sessions thoroughly from multiple perspectives and through comparisons – making visualisation a suitable approach. We aim to continue this work with further iterations of the designed solutions and with evaluation examples where direct impacts on analysts’ decisions can be observed. Further lines of research are to investigate ways to cluster and simplify the action space (in particular to reduce the number of hues used in the sessions), and devise ways to infer higher levels of semantics from users’ activities, such as high level tasks or roles as evidenced in the literature [GGZL16].

6. Acknowledgement

This work is supported by the European Commission through the H2020 programme under grant agreement 700692 (DiSIEM).

References

- [AS95] AGRAWAL R., SRIKANT R.: Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on (1995)*, IEEE, pp. 3–14. 3
- [BOH11] BOSTOCK M., OGIEVETSKY V., HEER J.: D³ data-driven documents. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2301–2309. 3
- [CBK12] CHANDOLA V., BANERJEE A., KUMAR V.: Anomaly detection for discrete sequences: A survey. *IEEE Transactions on Knowledge and Data Engineering* 24, 5 (2012), 823–839. 1
- [GGZL16] GUO H., GOMEZ S. R., ZIEMKIEWICZ C., LAIDLAW D. H.: A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. *IEEE transactions on visualization and computer graphics* 22, 1 (2016), 51–60. 3, 4
- [GH11] GREITZER F. L., HOHIMER R. E.: Modeling human behavior to anticipate insider attacks. *Journal of Strategic Security* 4, 2 (2011), 25. 1
- [LB96] LUND K., BURGESS C.: Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers* 28, 2 (1996), 203–208. 2
- [Leg15] LEGG P. A.: Visualizing the insider threat: challenges and tools for identifying malicious user activity. In *2015 IEEE Symposium on Visualization for Cyber Security (VizSec)* (Oct 2015), pp. 1–7. doi: 10.1109/VIZSEC.2015.7312772. 2
- [MLL*13] MONROE M., LAN R., LEE H., PLAISANT C., SHNEIDERMAN B.: Temporal event sequence simplification. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2227–2236. 3
- [NXWW14] NGUYEN P. H., XU K., WALKER R., WONG B. W.: Schemaline: timeline visualization for sensemaking. In *Information Visualisation (IV), 2014 18th International Conference on (2014)*, IEEE, pp. 225–233. 3
- [NXWW16] NGUYEN P. H., XU K., WALKER R., WONG B. W.: Time-sets: Timeline visualization with set relations. *Information Visualization* 15, 3 (2016), 253–269. 3
- [SA96] SRIKANT R., AGRAWAL R.: Mining sequential patterns: Generalizations and performance improvements. In *International Conference on Extending Database Technology (1996)*, Springer, pp. 1–17. 3
- [Sam69] SAMMON J. W.: A nonlinear mapping for data structure analysis. *IEEE Transactions on computers* 100, 5 (1969), 401–409. 2
- [SFK13] STOFFEL F., FISCHER F., KEIM D. A.: Finding anomalies in time-series using visual correlation for interactive root cause analysis. In *Proceedings of the Tenth Workshop on Visualization for Cyber Security (New York, NY, USA, 2013)*, VizSec '13, ACM, pp. 65–72. URL: <http://doi.acm.org/10.1145/2517957.2517966>, doi:10.1145/2517957.2517966. 2
- [WGGP*11] WONGSUPHASAWAT K., GUERRA GÓMEZ J. A., PLAISANT C., WANG T. D., TAIEB-MAIMON M., SHNEIDERMAN B.: LifeFlow: visualizing an overview of event sequences. In *Proceedings of the SIGCHI conference on human factors in computing systems (2011)*, ACM, pp. 1747–1756. 3
- [WPB01] WEBB G. I., PAZZANI M. J., BILLSUS D.: Machine learning for user modeling. *User modeling and user-adapted interaction* 11, 1-2 (2001), 19–29. 1