Visualization of Log Data Summaries

Anurag Kolhe, Aseem Baranwal, Debi Prasad Das and Srija Myana

Abstract—Visualization has the capability to reduce the time spent in understanding the data by a factor of a hundred. Human brains are outstanding at recognizing patterns and have a great rate of recall for the same. There are many use cases where it is necessary to quickly summarize and comprehend a large amount of data, which initially seems impossible to do. Understanding the work of analysts and their justifications for their conclusions is one such instance. We implemented the project for Log Visualizer which feeds on data provided by analysts' user interaction and generates a visual summary for their different interactions. Random Segmentation is used for easy data retrieval and to maintain a bias-free system. We have used Natural Language Processing to filter out words and list their frequencies. In order to conduct background research, we created a word cloud that demonstrated the usage and significance of highlighted and searched words. To create the User Interface, we have used Microsoft Tableau which facilitates easy data imports and shows interesting observations in a matter of seconds. Using Tableau makes our project highly scalable and versatile. Following the principles of Ben Shneiderman and Bertin, we have used visual encodings of color, shape, size, and position along with overview and detail to give a clear context of interactions performed in a segment of time along with a constructive summary of the steps taken by an analyst during an instant of time.

Index Terms—Logs, Summary, visualization, Scatter Plot, Stacked Bar, Pie Chart, Star Plot

1 Introduction

Many people have remarked on the narrative possibilities of data visualization in recent years. Renowned secret intelligence agencies such as the FBI, NSA, and MI6 often employ visualization to evaluate data and trends in order to anticipate threats even before they manifest. Experienced analysts conduct this task. Among the numerous ongoing investigations, cases are occasionally put aside to focus on other priorities, only to be reviewed later. When revisiting a case, it is not necessary to assign the same individual their duties. As a result, the next allocated analyst must start from scratch, wasting valuable time and resources. Frequently, prior work has to be presented by someone else, which involves reasoning out how the prior work led toward the conclusion. This practice compels them to restart the case from the beginning, resulting in a waste of personnel and resources. This unnecessary labor may be avoided by using data visualization to highlight the patterns adopted by the analyst. The system can gather logs of numerous interactions performed while the analyst is working. The essential records include the amount of time spent on a certain activity, the keywords searched, highlighted words, and documents accessed. Having such a relevant amount of data may subsequently be put to use and made into simple visualizations such as bar charts, pie charts, star plots, and so on.

Let's look at another significant data visualization use case. Investigating Authorities have identified a group of individuals and the possible kidnapping vehicle as the main suspects in a local kidnapping. The project can assist covert intelligence agencies in gathering all relevant data from their current insights. The suspect list may be further narrowed down using all the search histories, associated connections, and websites they recently visited, and utilizing all of that, a complete tale can be produced. However, since there is a lot of raw data to go through to get the pertinent information, accomplishing all of this can be very time-consuming. This work can be made easier by using a log summary visualizer, which can quickly display vital information like the most searched word and the place most frequently connected to suspects. It can also employ visual encodings to demonstrate the suspects' longest stays at a certain place as well as the precise tasks they carried out. Thus, it has the potential to reduce the total duration spent in catching the suspects by a factor of 1000.

• *Anurag Kolhe, Aseem Baranwal, Debi Prasad Das, and Srija Myana are graduate students in Computer Science at the University of Florida.*
E-mail: anuragkolhe@ufl.edu — a.baranwal@ufl.edu
ddas1@ufl.edu — srijamyana@ufl.edu

Visualizations that compare the time spent and the frequency of contacts in consecutive stages can reveal a lot about an analyst's approach [11]. This concept of creating visualizations from log data has the ability to minimize the duplicate effort to less than 1% of previous efforts. We present visualizations from three data sets containing logs from analysts' conclusions on Arms Dealing, Terrorist Activity, and Disappearance in this article. Furthermore, translating absolute counts to frequencies and conducting correlation analysis using pairwise correlations will be extremely beneficial in helping "users" form strong memories. We pay less attention to the reader's cognitive and emotional experience by stressing the graphical and interactive aspects of tale visualization. Thus, this will give us an impartial result from the analyst's interactions.

The report describes the project's tactics for making the visualizations neutral and informative while retaining a high rate of recall. This approach's use in the analysis verification and the fast summarization of a series of events can have game-changing results. Future research can aid in evaluating numerous user techniques, and an efficient strategy can be determined using various metrics such as time, the number of searches, opened documents, and so on, depending on the effective cost of conducting each interaction.

2 Related Work

Visual expression and storytelling are fundamental components of human civilization; storytelling has even been called "the world's second-oldest profession." Information retrieval and storytelling using visualization are quicker, unbiased, and language-independent. The paper outlines a few fundamental ideas guiding log summary visualization without describing millennia of accomplishment.

2.1 Data Quantification

The paper on Case Study Using Visualization Interaction Logs by Hua et al.(2016) [16] involves an experiment that used logs of interactions to comprehend and illustrate the insights of diverse participants using quantitative techniques. The contributions aided in analyzing the design decisions made by participants and their practical application by allowing for the usage of interaction logs to address assessment goals. Last but not least, by discovering a connection between interactions and ideas from our case study, the experiment made it simple to make design suggestions. The interactional qualities were compared to insight characteristics, and the findings established the parameters for the visual analytics program intended to promote insights development. The experiment examined each component individually (for instance, it counted the number of analysts who could generate an actual fact and how they did it).

2.2 Data Preprocessing

Another class of researchers aims at applying a variety of data preprocessing techniques, such as data transformation, data cleaning, data reduction, and data integration on the raw data to make it suitable for mining. The authors of Suad et al.(2017) [18] make review various data preprocessing techniques in Data Mining. Large data sets are sorted through data mining in order to uncover useful patterns, models, and correlations that may be used in data analysis to solve problems. In data mining, the quality of the data is crucial. Due to their enormous size, raw data are insufficient, inconsistent, noisy, and missing. Data mining results may be impacted if this data is not processed. Following preprocessing, a number of mining techniques, including clustering, classification, and regression, are used to extract precise patterns from the data. Data reduction lowers the volume needed to keep data, followed by Data cleaning, which eliminates noisy, incomplete, and superfluous data, Data integration, which merges data from many sources, and Data transformation, which changes the data format. The experiment examined large raw datasets which contained some noise, inconsistency, and missing information and used the Data Preprocessing techniques on these datasets to process the data, analyze and prepare it for Data Mining.

Fig. 1. Word Cloud for Highlighted Words

2.3 User Session Identification

Aaron et al.(2015) [12] proposed a technique to identify user sessions based on the user interaction timestamped log data which approach is most frequently employed in behavioral analysis and web analytics. The study used indicators such as work activities and inactive time and included datasets from several disciplines. Although it was indicated that using one hour as the inactivity threshold's boundary would be adequate for the majority of study types, it gives us the insight that there isn't a universal inactivity threshold. Additionally, Activity Theory (AT), which establishes a link between actions and operations, has the potential to significantly alter system design, enabling users to work more productively. Using timestamp-based log data analysis, the authors examined the user's website activity.

2.4 Improved K-means clustering algorithm

Padmaja et al.(2016) [8] used an improved K-means clustering algorithm to analyze preprocessed data and find user activity trends in web log files. Web log data must be transformed and interpreted in order to be used for web data analysis, which also includes knowledge discovery. Consideration of a few factors allows for an analysis of the algorithm's effectiveness. Date, time, CS_method, S_id, User agent, C_IP, and time was taken are the parameters. The study was conducted using real data that was gathered over a two-year period from the web servers of two separate groups of academic institutions. This dataset offers a more accurate log data analysis for identifying internet user behavior.

We will use the above algorithms and techniques for implementing our project. In the next section, we present an overview of our tool which provides details of various user activities such as opening a document, searching for a keyword, highlighting, dragging, etc over different segments of time which in turn helps to judge the user qualities or any specific details about the user to any person.

3 Design

The project is implemented using an incremental model defined by Hans et al.(2015) [2]. We started with the idea of the project to be implemented in multiple modules to be implemented in parallel. Implementing such a model helps in achieving a great speed of project development while also maintaining that the code remains error-free. Since Log Visualizer involved a significant amount of data preprocessing and data manipulation, we decided to create the modules containing the data values (termed as the back-end) in the beginning. The next steps included the manipulation of this data to make a user-friendly interface that satisfies Schneideman's Mantra defined by Shneiderman et al.(2002) [1]. According to his mantra, a good visualization project involves the project overview at first which is showcased by the different bar charts and star plots. Additionally, each segment can be clicked to reveal a pie chart that showcases the individual time spent by the analyst to complete the user interaction which satisfies his second rule of zoom and filter among the dataset. Further, the project can be interacted with to showcase the individual duration of different interactions performed, satisfying his third concept of details on demand.

The model is implemented on the dataset of user interactions by an analyst. These interactions involve actions like searching and highlighting words and opening, reading, and dragging documents. The whole time frame of user interactions is divided into multiple segments which gives a better visualization of analysts' state of mind. Segments are created at random and don't signify any special event. This adds randomness to our dataset and prevents our project from getting biased, thus remaining true to our goal. Further, for each segment, we determine the values of different interactions performed. We undertake this action while also keeping track of the number of those specific interactions performed and the total time spent in performing those interactions. The different possible interactions include highlighting and searching for words and dragging, opening, and reading documents.

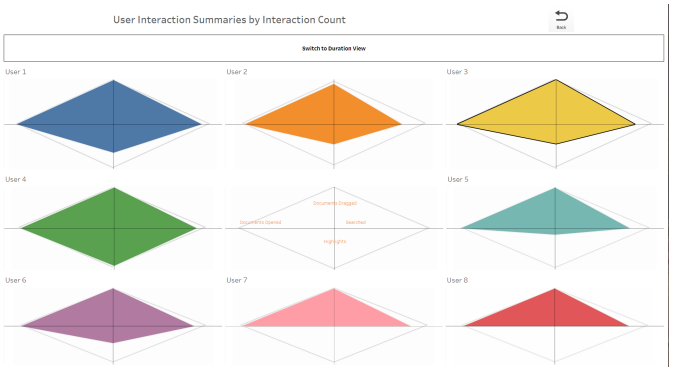

Fig. 2. Duration Overview of different users presenting user summary

A star plot is a perfect tool to visualize this dataset because star plots are very effective at displaying comparisons of high-quality data.

The size and form of the polygons show any general variances, and several properties may be easily compared to one another along their own axes. Selecting the star plots for overview gives an idea of how the user behavior changes over the period of data analysis. On clicking the star plot, the different segments are viewed as bar charts to facilitate easy comparison. Then they are integrated with the pie charts to show the duration spent on the data.

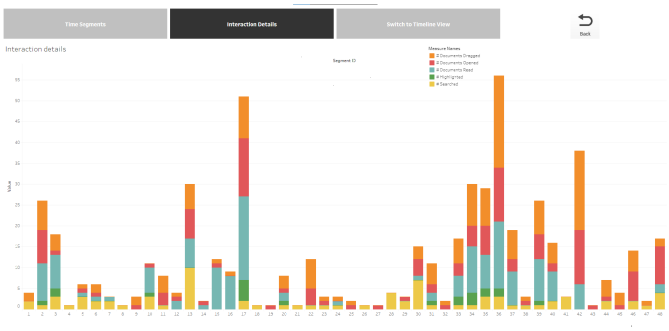

Fig. 3. Interaction Overview of a user presenting segment summary

Interaction data fed by the analysts is used to visualize each segment. For each segment, this data is visualized using the stacked bar chart where each bar represents the count of interactions performed and each stack represents the activity performed. The initial bubble chart shows the different segmented durations and the total time spent. On clicking any bubble, a new pie chart opens up that showcases the total number of interactions of a particular type that were performed by a user. Further, the tool also allows the user to go into even more detail using the timeline view which can help the user to narrow down the interactions performed to a mere second. Thus, this action makes our implementation modern and user-friendly and provides detail on demand. Having such a type of design doesn't lose details but also maintains ease of use for novice users.

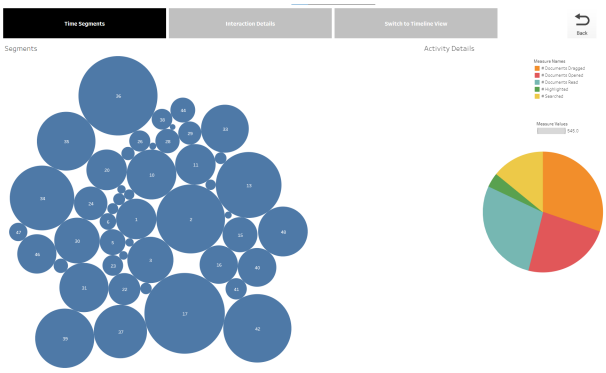

Fig. 4. Time Segmented View of a user presenting segment summary

Following Ben Shneiderman's footsteps, we have given the functionality to the user to zoom and filter and find interactions performed by the analyst at any particular instant. This gives the ability to the advanced users to filter out unnecessary details and get elaborated details at any particular instance of time. A scatter plot is essentially the best visualization that can be used to achieve this herculean task. Not only, does this give extrapolated details about the dataset, but also has great sensemaking abilities as it employs the position encodings to do the process of sensemaking.

We have used different visual encodings in our project including length, size, area, angle, and color which makes the overall visualization easy to understand. Using these encodings ensures a high degree of recall which is very important in this project as it contains a lot of data to process. This also helps the end user to form a conclusion easily

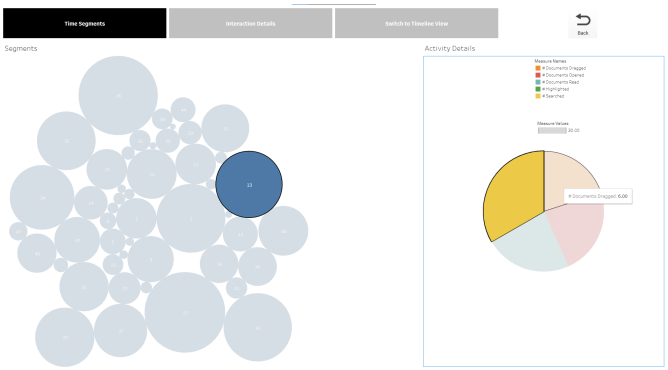

Fig. 5. Hovering over Time Segmented View gives more details about a segment

without being too distracted by the overwhelming dataset.

4 Implementation

The dashboard is the first screen that provides a brief introduction to the concept of log data visualization and how it can be significantly advantageous over the conventional log database. It has two tracks, duration study view, and interaction study view. To compare various users who completed the same task, both views are used. As described above in design, small multiples of star plots can do wonders for providing an overview of user interactions and a good notion of how other analysts have tackled the topic.

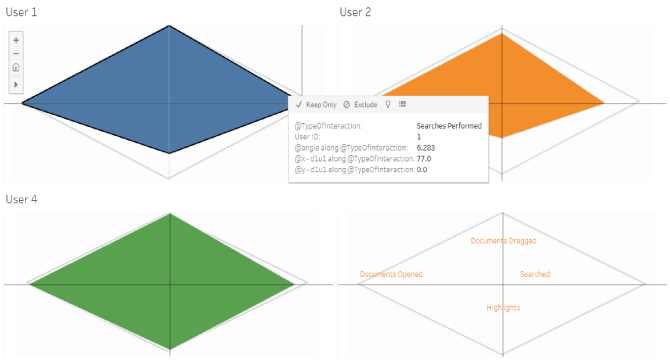

Fig. 6. Hovering on duration Overview screen presenting user summary and more details

Fig. 7. Hovering on interaction Overview screen presenting user summary and more details

In the next view, after selecting a specific analyst interaction, the user can drill down into our large dataset in the duration view or the

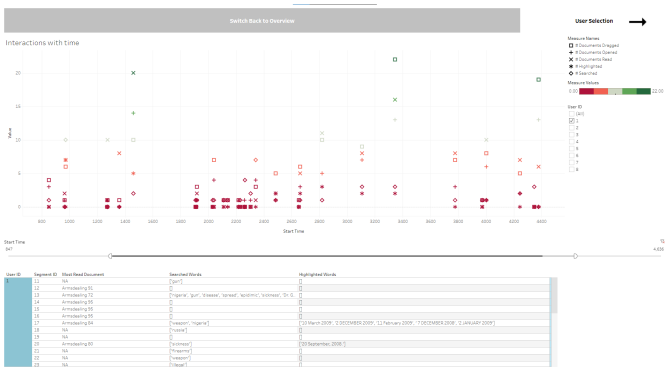

Fig. 8. Timeline view gives a concrete idea and a drill down of a particular instant

Fig. 10. Hovering on scatter plot and narrowing down slider width to drill down into the data

interaction view. For the duration view, we've utilized a bubble chart that shows which interaction the user spent the most time on using the area and size encodings. We also have a pie chart that shows the different types of interactions the analyst performed over that period in more detail on the same page. Additionally, to highlight various areas for the interaction view, we used a stacked bar chart. The size encoding is used to compare different encodings within a segment and to compare the number of interactions made in one segment to that of some other segment. Different color encodings are implemented to differentiate between various interactions.

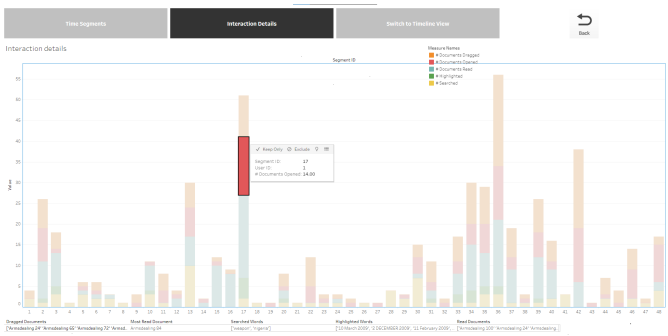

Fig. 9. Hovering on interaction Overview screen presenting user summary and more details

The last view is the final drill-down that is built to be the closest to the actual log data and can be used after the user/analyst has concluded to explore more into the log data because there is so much a visual summary can tell a user. Based on a slider, the user can go through the timeline and figure out what interaction was performed and what specific activity was carried out in the interaction. For eg. – it specifies what has an analyst searched during the instant of time and how many of those searches were performed. Different types of figures and colors are used to clearly specify what the user has done at some particular instant and in what amount. Further, the y-coordinate of a particular figure gives the total count of the interactions performed at some particular time stamp. In addition to this, selecting a hue for depicting the number improves sense-making. The table at the bottom of the view gives a clear drill down of the different types of interactions performed and what was the actual interaction that was done by an analyst or a user.

Following that, a user might begin with a summary of the many interactions conducted from the star plot overview. In the following phase, he/she might begin to search for the overall time spent and the most time spent on a certain activity or interaction. He/she can dig down into the interaction display segmented bar charts as a drill down. Finally, if the user has identified a suspect or is interested in any interaction performed by the analyst, they may navigate to the timeline view to precisely depict the work performed by an analyst (Eg. - The exact search query of the analyst and the exact documents read by the user).

5 Experimental Results

This section presents the initial performance of our proposed model used for project implementation. Presently when any end user uses our tool, he/she can able to get information about various analysts' activities in different time segments. The various activities include the number of documents opened, highlighted words, searched words, their frequency, and the time spent by the analyst on any document. The Word Cloud used in the back end of our project shows a clear picture to any end-user about the most frequent words or documents searched by analysts based on their activities. The generated word clouds for searched and highlighted words are depicted in Figures 5 and 6 respectively. The Visualizer tool gives details about any analyst based on their activities in a given time frame. The analyst in the beginning starts searching for a word in a document. The time spent on that document increases till he doesn't find that word in the document. Then he searches for another document. Once he gets the word the time spent by him on that document goes on decreasing.

The tool was reviewed by Professor Eric Ragan and other peers. A group of peers used our tool and wanted to draw inferences about various users based on their activities in a given time segment. Some analysts were searching for words that were repetitive and some users were just reading the documents and highlighting. Based on their activities they were able to visualize users' qualities.

In the future, we will incorporate the changes suggested by Professor and finish designing the user interface.

6 Future Milestones

The current project doesn't support dynamic databases and has been programmed to work only on a suitable type of data set. Working towards extracting data dynamically and fitting that according to our model will make the usability of the project even better. The second future work includes the usage of the current project for the comparison of approaches taken by different analysts. Using this, another use case to predict the personality type of the users can be done and they can be strategically put onto cases that can bring out the best from a personality type. Though we evaluated our tool now with fewer participants in the future, we will evaluate it by including participants from various diverse on a large scale. We hope that in the future our tool will assist renowned secret intelligence agencies in evaluating the data and drawing trends about any user's activity and averting any possible threats.

7 Conclusion

We can conclude that the power of NLP methodologies combined with various visualization techniques can generate meaningful results. Their advancement has given both individuals and companies access to reliable, potent, and scalable visualization tools. They are made to aid in the analysis and visualization of data using different analytical

techniques. In this work, we looked at ways to enhance the basic word cloud visualization with more details and interactive elements to make a more potent Log Visualizer Tool. With the help of this technology, we can readily identify behavioral and activity trends over time about any user by correlating logs, metrics, and their interactions performed to gain a more comprehensive picture of them.

We put this research work into practice for the Log Visualizer tool, which uses data from analyst-user interactions to produce a visual overview of those interactions. To ensure a heuristic system and make data retrieval simple, random segmentation is used. Initially, NLP techniques such as word clouds are used to generate and tokenize the data and get meaningful information from the given text. We have implemented different visual encodings to generate a visual summary of the user's activity session. We have used Microsoft Tableau to build the user interface since it makes it simple to import data and quickly displays insightful findings. Our research work is incredibly flexible and adaptable because of Tableau. We have used various visual encodings of color, shape, size, and position, along with overview and detail, in accordance with the principles of Ben Shneiderman and Bertin, to provide a clear context for interactions that took place during a specific period of time and a helpful summary of the actions an analyst took at a particular point in time.

An analyst is provided with various visualizations which could be helpful to make sense of a user's action and derive some verdict from it about the user. These types of visualizations are helpful to derive some meaning from unstructured data. A total of 12 participants were given to use the tool and draw conclusions about User-1's activity. They were asked the following questions including where did the analyst spend most of the time and what activities he/she performed, the amount of time spent on reading a document, the interaction trends, and the info about the most read-document and most searched keywords in the analysis task. All of the participants were successfully able to draw correct conclusions about User-1's activity.

ACKNOWLEDGMENTS

We want to thank Professor Eric Ragan and his peers for their support of our project.

REFERENCES

[1] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. *Proceedings 1996 IEEE Symposium on Visual Languages*, August 2002.

[2] Hans-Jörg Schulz, Marco Angelini, Giuseppe Santucci, Heidrun Schumann. An Enhanced Visualization Process Model for Incremental Visualization. *IEEE Transactions on Visualization and Computer Graphics*, July 2015.

[3] Zeqian Shen; Jishang Wei; Neel Sundaresan; Kwan-Liu Ma. Visual Analysis of Massive Web Session Data. *IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, December 2012.

[4] Suad A. Alasadi, Wesam S. Bhaya. Review of Data Preprocessing Techniques in Data Mining. *Journal of Engineering and Applied Sciences*, September 2017.

[5] Florian Heimerl, Steffen Lohmann, Simon Lange, Thomas Ertl. Word Cloud Explorer: Text Analytics based on Word Clouds. *47th Hawaii International Conference on System Sciences*, March 2014.

[6] H. Lam, D. Russell, D. Tang and T. Munzner. Session Viewer: Visual Exploratory Analysis of Web Session Logs. *IEEE Symposium on Visual Analytics Science and Technology, 2007, pp. 147-154, doi: 10.1109/VAST.2007.4389008.*, November 2007.

[7] P. H. Nguyen, C. Turkay, G. Andrienko, N. Andrienko, and O. Thonnard. A Visual Analytics Approach for User Behaviour Understanding through Action Sequence Analysis. *EuroVis Workshop on Visual Analytics (EuroVA). The Eurographics Association, 2017, doi: 10.2312/EUROVA.20171122.*

[8] Padmaja, S. and Dr. Ananthi Sheshasaayee. Clustering of User Behaviour based on Web Log data using Improved K-Means Clustering Algorithm. *IEEE Transactions on Visualization and Computer Graphics*, February 2016.

[9] Joshi, Ashish P. and Biraj V. Patel. Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process. *Oriental journal of computer science and technology (2021): n. pag*, January 2021.

[10] K. R. Suneetha, Dr. R. Krishnamoorthi. Identifying User Behavior by Analyzing Web Server Access Log File. *IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4*, April 2009.

[11] Sina Mohseni, Alyssa Pena, Eric D. Ragan. ProvThreads: Analytic Provenance Visualization and Segmentation. *IEEE*, 2017.

[12] Aaron Halfaker, Oliver Keyes, Daniel Kulver, Jacob Thebault-Spieker, Tien Nguyen, Kenneth. User Session Identification Based on Strong Regularities in Inter-activity Time. *International World Wide Web Conference Committee (IW3C2)* , May 2015.

[13] Vishal Murgai. Log Analytics with NLP. *Stanford CS224N Natural Language Processing with Deep Learning* , December 2020.

[14] Yordan Kalmukov. Using Word clouds for fast identification of Papers' Subject domain and Reviewers' competencies. *PROCEEDINGS OF UNIVERSITY OF RUSE - 2021, volume 60, book 3.2, pp. 114-119*, 2021.

[15] C. Bertero, M. Roy, C. Sauvanaud and G. Tredan. Log Mining Using Natural Language Processing and Application to Anomaly Detection. *IEEE (Institute of Electrical and Electronics Engineers) 28th International Symposium on Software Reliability Engineering (ISSRE), 2017, pp. 351-360, DOI: 10.1109/ISSRE.2017.43.*, 2017.

[16] Hua Guo, Steven R. Gomez, Caroline Ziemkiewicz, David H. Laidlaw. A Case Study Using Visualization Interaction Logs and Insight Metrics to Understand How Analysts Arrive at Insights. *IEEE Transactions on Visualization and Computer Graphics (Volume: 22, Issue: 1)*, January 2016.

[17] Jock D. Mackinlay, Pat Hanrahan, and Chris Stolte. Show Me: Automatic Presentation for Visual Analysis. *IEEE Transactions on Visualization and Computer Graphics (Volume: 13, Issue: 6, Nov.-Dec. 2007).*

[18] M. Gray, A. Badre, M. Guzdial. Visualizing Usability Log Data. *Proceedings IEEE Symposium on Information Visualization)*, 1996.