The Use of Data Visualization in E-Commerce: A Review

Yaman Sanobar, sanobar@cs.ubc.ca
University of British Columbia, Department of Computer Science

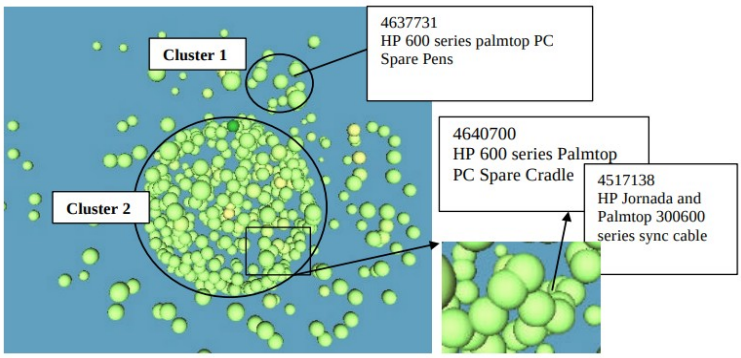

Fig. 1. An example of visualizing customers purchasing similar products, grouped into clusters "taken from [1]".

Abstract— Data in E-commerce has significant potential in improving and growing online businesses. The data collected is complex and huge, and require analysis and visualization in order to help owners and decision-makers. The paper introduces a review of using data visualization in e-commerce where Clickstream data is of the highest value. Clickstream data is the data generated from the activities of users navigating e-commerce websites. To conduct the review, we used the PRISMA checklist and the PRISMA flow diagram to follow a systematic approach for searching and picking relevant papers. We scanned and then identified relevant papers, then created a record for each paper containing key findings. From the reviewed papers, we identified central approaches and trends in Clickstream visualization research. We categorized the reviewed papers into topics and summarized the visualization methods used in each paper. We discussed the topics and we provided recommendations for designers concerned with designing applications for visualizing Clickstream data.

Index terms— Clickstream visualization, e-commerce data visualization, Clickstream analysis

1 Introduction

The amount of data of online businesses flowing each second through the internet is enormous. The importance of this data for enhancing online businesses or improving user experience can be valued when handling collected data properly. Data collection, analysis, and visualization have a great opportunity to help online stores to provide valuable insights to their customers. This will lead to the fulfillment of the overall needs of all parties involved in online shopping. For instance the visualization of data helps the owners of online stores in understanding the behavior of their customers. Clickstream data which is collected from the activities of customers navigating online stores, is stored and analyzed. Then it is organized and visualized to help owners understand the website trends and the overall status of the online store [2]. Besides, e-commerce websites provide customers with a huge amount of choices that can be overwhelming. Visualization can help customers in browsing these choices and reduce the effort of clicking and scrolling over every single product [3].

A wider look at the topic can be obtained if it is viewed from the perspective of business intelligence. Business intelligence systems combine data with data analysis tools to visualize data and then display it for planners and decision-makers [4]. The difference between data visualization for business intelligence and other types of visualization is mainly in the purpose of the visualization. Data exploration and analysis, and making decisions based on data are the main goals for visualizing data for business intelligence [5]. The importance of using data visualization in e-commerce comes from the importance of data for online stores to derive successful businesses. The success of many e-commerce websites depends on data, and how it is being used to improve operations [6]. Moreover, data mining is key when it comes to handling complex and huge amounts of data. In e-commerce, data mining is beneficial for the process of decision-making and figuring out the tendency of the development of buying and selling [7]. However, as with dealing with any huge amount of data, many challenges appear. Throughout the cycle of mining the data, many technical difficulties occur, such as collecting the right Clickstream data directly and in a timely manner [8].

1.1 Background

E-commerce is defined as "the use of computer networks to conduct business—basically the buying and selling of goods and services and information" [9]. E-commerce Clickstream data is generated from the interaction of users with online stores. It is a detailed log of the behavior of customers when browsing or doing certain tasks on e-commerce websites [10]. The analysis of these logs provides metrics of success about the activities of customers. Examples of such metrics are the increase in the number of users on the site, the reduction in the number of users leaving the site, the increase of time spent on the site, the chances for the customer to return to the site, and the increase of conversion rate [11]. Figure 2 shows the process of knowledge discovery in databases of e-commerce Clickstreams. From another perspective, "Clickstream data provide tremendous insights into how easily the site is navigated, what pages are causing the greatest confusion, and what pages are critical

in reaching a desired destination." [10]. E-commerce Clickstream data is generated from every task of every customer navigating an e-commerce website, and it is generated from different types of interactions within a website. Therefore, it is huge in size and complex in type. However, even though it is a challenging task to analyze this data, Clickstream data is particularly valuable for recognizing and understanding user behavior.

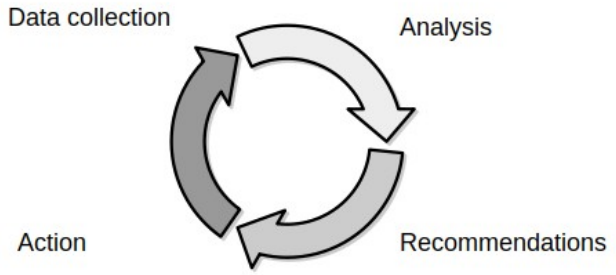

Fig. 2. Knowledge Discovery process in Databases (KDD) of e-commerce websites "adapted from [15]".

1.2 Scope

The goal of this paper is to provide an overview of the usage of data visualization in e-commerce. This is done by surveying the topic of visualizing e-commerce Clickstream data. Much work has been found on Clickstream data analysis, but not many are concerned directly and broadly with the visualization of e-commerce Clickstream data. An example of such work is a novel visual interface for analyzing Clickstream data [11]. Through data and task abstraction, the interface proposes an iterative loop of viewing, refining, and recording Clickstreams. The loop refines collections of action sequences to create meaningful segments from Clickstream data. Another example is a visual analytics system to analyze and monitor the order processing of e-commerce warehouses [12]. The focus of the system is on the real-time analysis and visualization of streaming data generated from e-commerce warehouses. As for survey papers, the only relevant survey we found discusses the topic of combining the technologies of e-commerce with business intelligence [13]. The survey provides insights, suggests new research directions, and proposes an architecture for combining business intelligence with e-commerce.

The scope of this paper is the current research on the systems and tools of visualizing Clickstream data and the relation of that with data visualization in e-commerce. We followed a systematic way of searching and picking relevant papers, then identified key tools and trends in Clickstream visualization and analysis research. Through reviewing the papers, we found multiple tools that have been developed to tackle the difficulty of analyzing data of such characteristics. The earlier tools are more concerned with the navigation path of users browsing websites. It uses directed-graph visuals to help understand user behavior. Data clustering tools gave the ability to handle the data size problem of Clickstream data and provided a higher-level view of its structure. A few systems combined several techniques to propose a solution for analyzing and visualizing Clickstream data. We categorized the reviewed papers into four main topics. We summarized the reviewed papers and analyzed their methods of visualization. Then we discussed the methods used and the identified trends. We showed how having low-level and high-level views is essential in order to optimally analyze Clickstream data. Lastly we provided recommendations for designers concerned with visualizing Clickstream data.

2 Method

The method used for identifying and selecting relevant papers is PRISMA, an evidence-based reporting items method for systematic reviews and meta-analysis [14]. We selected certain elements from the PRISMA 2020 Checklist to ensure the review has a systematic approach for surveying the topic. We used the PRISMA 2020 flow diagram (Figure 3) which illustrates the process of identifying and screening records. The inclusion and exclusion criteria are 1. articles of conferences or journals, 2. full text must be in English, 3. articles published after 2000. The reason we excluded articles published before 2000 is that we noticed the loose of significance when compared to newer studies. The database used for search is Google Scholar. The keywords used for the search are:

- "e-commerce Clickstream visualization"
- "e-commerce Clickstream visualization survey"
- "visualizing e-commerce Clickstream data"

We used each query sentence once, then the first 20 records of the search results were navigated. We noticed the search results after the first 20 records are out of scope. As a preliminary selection, Each paper was skimmed and scanned, then it is selected if it identifies sufficient information about either Clickstream data visualization or Clickstream data analysis. Selected papers were recorded in a table where each record contains the paper title, link to the paper, paper keywords relative to the review, search sentence used, and important findings. In the preliminary selection, a total of 60 records were scanned. 26 out of 60 were duplicates, which made the total number of unique records skimmed and scanned 34 records. The number of records that didn't comply with the inclusion and exclusion criteria is 3; 2 records are not articles, and the full text of 1 record is not in English. As a result, the total number of papers selected in the preliminary selection is 9. Figure 3 illustrates the identification process and the total number of seed papers identified, screened, and selected.

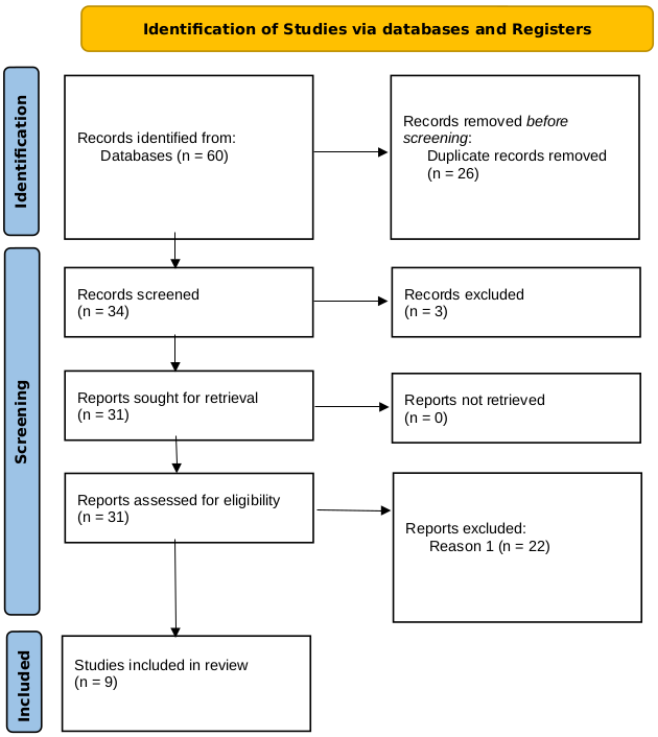

Fig. 3. PRISMA flow diagram for selection of papers.

While reading seed papers, a second round of identification was done through backward citation. 4 Out of 9 papers were marked for backward citation search. In the preliminary selection round, we identified the overall scope of the survey. Then in the second round of identification and based on the scope identified, we selected relevant papers to scan from the backward citation search. In the backward citation search, a total of 18 records were scanned. 6 records were excluded because they didn't comply with the inclusion and exclusion criteria; 4 records are published before the year 2000 and 2 are not articles. As a result, 3 records were selected from the backward citation search, which makes the overall number of records selected 12. Figure 4 illustrates the whole filtering process and the total number of records selected. As for forward citation search, we made few and modest attempts but not rigorous and deep ones, which is one weakness towards this paper. Table 1 shows the papers selected for review. Papers are categorized into the main topics or trends that we discovered or identified through the surveying process. The first 9 papers in the table are the seed papers while the last 3 are the papers from the backward citation search.

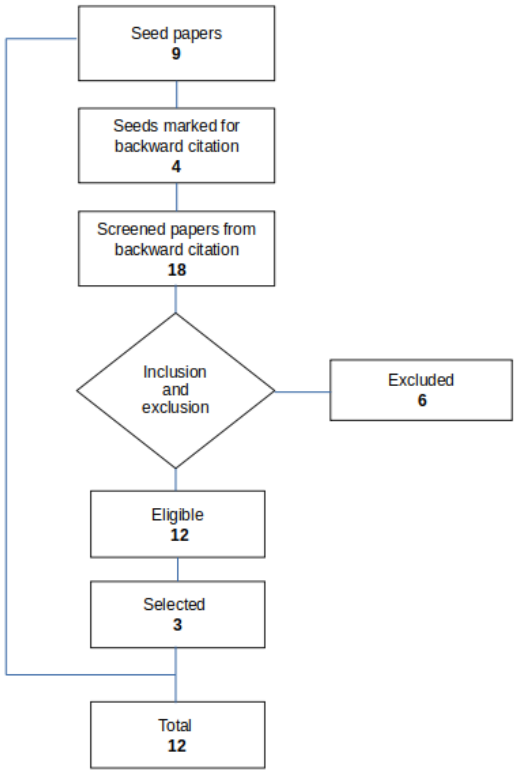

Fig. 4. The overall filtering process for selecting review papers.

Table 1. The twelve identified papers for review

Paper	year	Authors	Category
[15]	2001	Brainerd & Becker	Path Navigation
[16]	2002	Waterson, Hong, Sohn, Landay, Heer, & Matthews	Path Navigation
[17]	2004	Ting, Kimble, & Kudenko,	Path Navigation
[18]	2012	Wei, Shen, Sundaresan & Ma	Data Clustering
[19]	2014	Kateja, Rohith, Kumar & Sinha	Visualization Approach
[20]	2016	Wang, Zhang, Tang, Zheng & Zhao	Data Clustering
[21]	2017	Hanamanthrao & Thejaswini	Analysis Architecture
[22]	2017	Frhan	Visualization Approach
[23]	2018	Anand, Vamsi & Kumar	Analysis Architecture
[24]	2001	Hong & Landay	Path navigation
[25]	2015	Zhao, Liu, Dontcheva, Hertzmann & Wilson	Visualization Approach
[26]	2015	Shi, Fu, Chen & Qu	Visualization Approach

3 Review

Before digging into the reviewed papers, we traced the work of Lee and Podlaseck to give an example of the work that this paper is focusing on. Then we give an example of a relevant visualization interface and the way we introduced visualization interfaces in this paper. First, Lee and Podlaseck presented a visualization system that provides analysts with abilities to explore Clickstream data using parallel coordinates [27]. Then the authors presented a visualization system for Clickstream data based on Starfield display [28]. Lastly, the authors developed a system to utilize the two previous visualization techniques to visualize Clickstream data and measure the effectiveness of an online store; parallel coordinates to visualize sessions of users and Starfield graphs to visualize the performance of products [29]. Additionally, Lee and others proposed an interface to navigate catalogs of online stores using parallel coordinates [30]. Figure 5 shows the interface where the catalog contains 10 products from 3 different categories and 5 attributes. In visualization terms, marks are the basic graphical elements of visualization, and channels are the ways to control the appearance of marks [31]. Hence, for Figure 5 the mark used is a vertical line and the channel used is color. Each product is represented as a vertical line, each color denotes a different category, and each row represents an attribute. The user can hover over any product to see the full description of it by knowing the corresponding attributes. Table 3 shows the what-why-how analysis table that summarizes an analysis of a visualization [31]. In this paper, we used this table after certain visualizations to easily understand figures and to provide a quick analytical view of those visualizations.

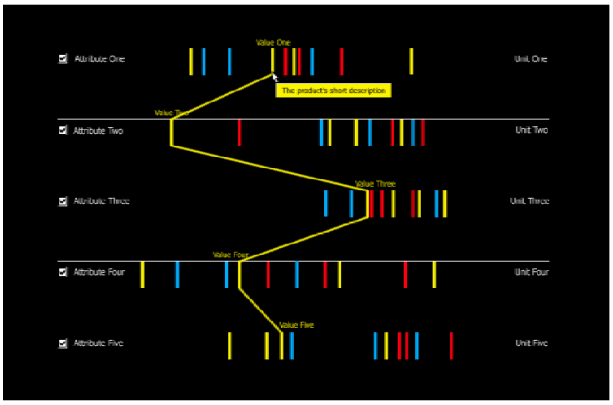

Fig. 5. Visualizing catalog of products using parallel coordinates "taken from [30]".

Table. 2 Analysis summary for Figure 5	
Idiom	Parallel Coordinates
What: data	Table
How: encode	Vertical lines express products, horizontal positions express attributes value, Color hues express category
Why: task	Hover over a product to see its full description
Scale	Attributes: 5, Items: 10

Throughout the surveying process, we noticed the interconnection between Clickstream analysis and visualization. Besides, Clickstream data share the same noticeable properties of big data which is volume and velocity [19], [23]. The data is collected from every user navigating an e-commerce website, and the activities of users that produce Clickstream data are generated at any time interval. Furthermore, Clickstream analysis is a typical process for understanding user behavior [21], which is a very important process but a challenging one for any e-commerce website [18], [20]. Several systems and tools proposed a solution for both analyzing and visualizing Clickstream data. In this review, we categorized the reviewed papers into main topics. Through that, we identified prominent trends and tools in the research of Clickstream visualization and analysis.

3.1 Path Navigation

Earlier tools focused on the browsing path of several users completing a single task on a website. Visualizing the navigation path is used to improve usability and detect browsing patterns.

WebQuilt is a visual analytics tool that logs and visualizes web data for web usability testing [16], [24]. It uses a proxy server to log Clickstream data [17], and it uses directed graphs to visualize the browsing paths of users [25]. The visualization interface of WebQuilt allows analysts to compare alternative browsing paths taken by users with an optimal path for a single task to improve usability [15]. The tool consists of five independent components: a proxy logger, an action inferencer, a graph merger, a graph layout, and a visualization interface [24]. The visualization interface is an interactive one, in which it visualizes a single task for multiple users in the shape of nodes and arrows. It is supported with filtering capabilities through checkboxes. It has semantic zooming that allows analysts to navigate the full browsing path and zoom in to view pages if required. Figure 6 shows the visualization interface of WebQuilt where each node represents a visited web page by the user, and arrows represent the transition between pages. The interface is color-coded, where entry pages are in green and exit pages are in cyan. The thickness of the arrows represents traffic and the color of the arrows represents the time spent on a page before transitioning. The zoom slider is on the left of the screen and the filtering check-boxes can be found at the bottom. Through this interface, analysts can detect several usability issues for a single task. Analysts have the ability to know which pages the users spent a significant amount of time on, pages where users left the track of the task on, and exit pages [16].

Another tool that proposed path navigation is a Clickstream visualizer called ClickViz [15]. Similar to WebQuilt, the system illustrates direct graphs of nodes and arrows to analyze the behavior of users navigating a website. However, Clickviz has the ability to segment Clickstream data based on the attributes of users. As a result, ClickViz allows analysts to illustrate paths based on the types or the demographics of their users, and enable them to gain a deeper understanding of user behavior. Figure 7 shows two paths that are drawn using ClickViz, where each node is a web page and each arrow is a transition between two pages. Colors of arrows represent an attribute value that could be selected from a list of attributes. ClickViz has an advantage in understanding Clickstream data when compared to WebQuilt because of its ability to segment arrows. Arrows between nodes are drawn based on attribute type in addition to user count, which allows analysts to compare browsing paths based on attributes. The tool has a hierarchical layout to visualize the paths as shown in Figure 7 in addition to a circular layout. Figure 7 displays an example of the browsing paths of a checkout task. It is used to compare user behavior between purchasers and non-purchasers. Users with the attribute "purchase" are in blue and users with the attribute "non-purchase" are in red. The purchasers took a direct route through the checkout process, while the non-purchasers route is random and shows early abandonment.

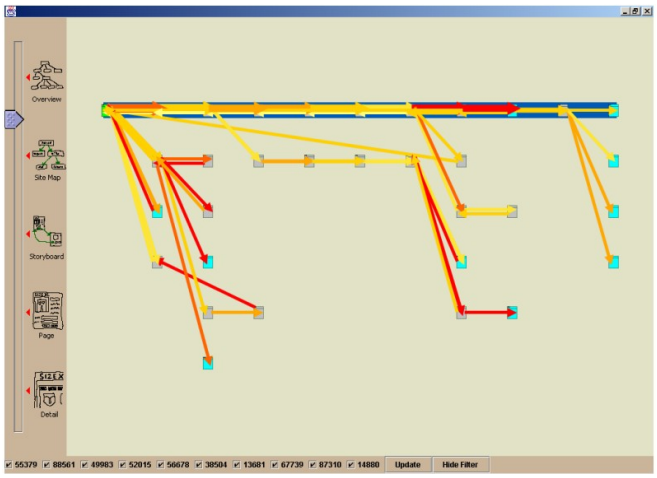

Fig. 6. The visualization interface of WebQuilt "taken from [16]"

Table. 3 Analysis summary for Figure 6	
Idiom	Browsing Paths
What: data	Multidimensional Table
How: encode	Nodes express pages, colors of nodes express type, arrows express transition, thickness of arrows expresses traffic, and colors of arrows color express time
Why: task	Compare alternative paths with optimal path
Scale	Single task, dozen of nodes, multiple users

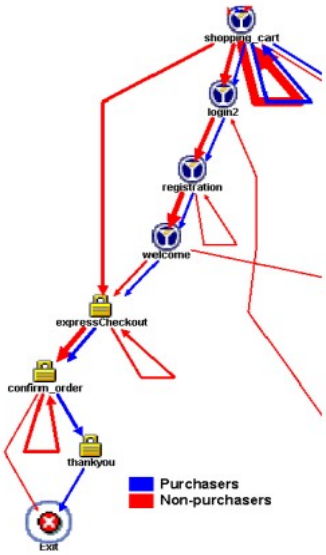

Fig. 7. A hierarchical browsing path visualized using ClickViz "taken from [15]"

Footstep Graph is a Clickstream visualization technique that has an advantage over the often-used technique for path navigation which is footstep maps [17]. Figure 8 displays a footstep map which is a graph of nodes and arrows. Each node represents a web page visited and the arrows represent the transition between nodes in a task. The numbers without parenthesis represent the sequence of the clicks and the numbers with parenthesis are the time spent before transitioning from one node to the next. The downside with footstep graphs is it becomes complex and unreadable when the navigation path is dense and long. Furthermore, the information obtained from footstep maps is hard to be interpreted directly into insights. It requires analysts to manually deal with data in order to generate valuable recommendations. To overcome this difficulty, the authors proposed a better technique called Footstep Graphs that normalizes the navigation path based on time and sequence (Figure 9). The graph is a 2D plot where the x-axis represents time and the y-axis represents the visited nodes on the user's navigation path. Horizontal distances are the time spent between two nodes and vertical distances represent the transition between nodes.

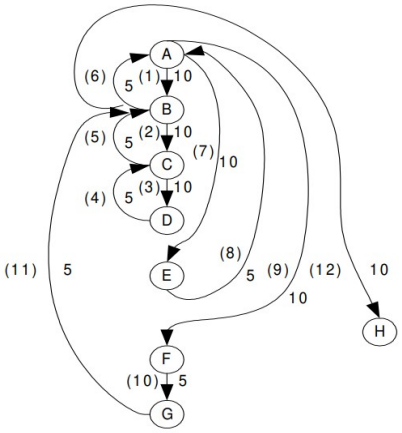

Fig.8. Footstep map "Taken from [17]"

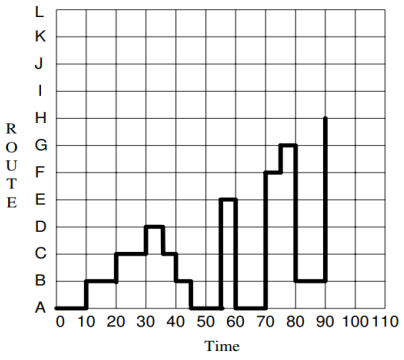

Fig. 9. Footstep graph "Taken from [17]"

Table. 7 Analysis summary for Figure 9

Idiom	Footstep Graph
What: data	Multidimensional Table
How: encode	X-axis express visited nodes, y-axis express time spent between nodes, line represent navigation path
Why: task	Interpret patterns to identify user beahvior
Scale	Dozen of nodes, single user, single task

Interpreting user behavior from the footstep graph is much easier than Interpreting it from the footstep map. The authors identified many path patterns, like the upstairs pattern when the user moves forward only in the website, and the downstairs pattern when the user moves backward only. The Mountain pattern is when the user goes upstairs and then downstairs. This denotes an unsuccessful pattern because the user spent a lot of time browsing but they end up returning to the same node without finishing the task. A valley pattern happens when the user goes back to a page they visited earlier and then moves forward to a new page. The valley pattern could denote a successful pattern if it is the last valley on the graph, which indicates that the user finished the task. It could indicate an unsuccessful pattern if the valley pattern is followed by a downstairs, fingers, or another valley pattern. It denotes that the user made a repeated mistake. The last pattern the authors mentioned is the fingers pattern which has the shape of fingers. It denotes that the user is doing a browsing loop because of confusion or ambiguity of the website. Finally, the authors proposed several examples of how to extract recommendations and insights from path patterns.

3.2 Data Clustering

Data clustering tools provided significant capabilities for understanding Clickstream data compared to path navigation. It gave the ability to gain a higher level view and a holistic understanding of e-commerce Clickstream data.

One tool proposed visual analysis for Clickstream data by exploring a map of clusters of browsing patterns of users [18]. The tool utilizes Self-Organizing Maps with a Markov chain neural network model to map and cluster Clickstream data. The data is visualized on an interactive interface which enables analysts to explore the clustered patterns and know the demographics of their users. Before developing their tool, the authors interviewed analysts at eBay. They found three questions that are frequently asked by analysts and derived tasks from the questions:

- "What are the most frequent user behavior patterns?".
- "What are the demographics of the users who follow a specific behavior pattern?".
- "How do the behavior patterns correlate with the performance of the online service?".

Their approach to handling Clickstream data is to map Clickstreams in a 2D plot, then visualize samples of representative data. Before visualizing the data, they made adjustments to the placement of data to remove overlapping and reduce visual clutter. Figure 10 displays a comparison of processed data before and after the adjustment is applied. We can notice how clusters of patterns are clearly formed after the adjustment.

Fig. 10. Visualizing Clickstream data before and after adjusting the Clickstream 2D layout "Taken from [18]"

Throughout the visualization interface, analysts can explore and cluster different patterns to gain insights about user behavior, and make correlations between demographics and clusters. Figure 11 shows the full visualization interface. The mark used is a rectangle and the channels are color hue and area size. Figure 11(a) illustrates the 2D representation of Clickstream patterns, in which each rectangle represents an action done by the user. The type of action is mapped in color hue and the size of each pattern encodes the frequency at which the pattern is occurring in the dataset. Figure 11(I) shows a single pattern selected which is a horizontal line of actions. Figure 11(b) shows the action legend, and Figure 11(c) shows the currently selected pattern which changes according to the selection of the analyst. Figure 11(d) shows the demographics and statistics of users of the current selection. It uses histograms to display the segment of buyers, the gender of users, the years of being an eBay user, and the cart size of users from the corresponding pattern. Lastly, Figure 11(G) shows that the analyst has the ability to select multiple Clickstream patterns to form clusters. This way, analysts have the ability to understand Clickstream data from several aspects at different scales.

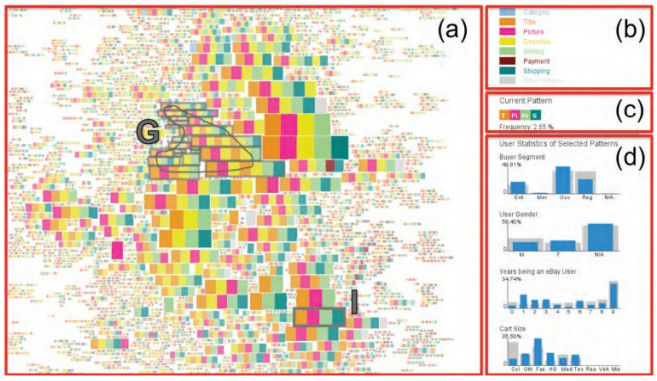

Fig. 11. The interface of a visual cluster exploration tool, (a) Clickstream patterns; (b) action legend; (c) the selected Clickstream; (d) user statistics of selected patterns; (I) single pattern; (G) cluster of similar patterns "Taken from [18]"

Table. 4 Analysis summary for Figure 11

Idiom	Cluster Map
What: data	Multidimensional Table
How: encode	Rectangles express actions, colors of rectangles express type, horizontal lines of rectangles express patterns, sizes of lines express traffic
Why: task	Cluster and explore different browsing patterns with the corresponding demographics
Scale	Dozen of action types, thousands of navigation patterns

Another system proposed is an unsupervised clustering system that analyzes and visualizes user behavior through hierarchical clustering [20]. The system identifies clusters and uses similarity metrics between users to create similarity graphs. Through these graphs, analysts can recognize the browsing patterns of user behavior in a nested manner. The system proposed is a sophisticated tool that enables analysts to answer key questions about user behavior easily and deeply. The authors identified three requirements before developing their system. First, it should function well on large and noisy Clickstream datasets. Second, it should be able to capture old user behavior patterns that lack certain attributes or categories. Third, the system should be interactive, intuitive, and present clear and understandable results. The proposed system builds Clickstream similarity graphs that are partitioned according to user behavior. The system uses a hierarchical clustering algorithm to detect and illustrates user behavior. Clusters represent user groups with similar behavior, and the edges of clusters capture the similarity distance between behaviors. Higher level clusters represent more general behaviors while lower level clusters represent less common behaviors. A structure of nested clusters shows user groups with similar general behaviors and how these groups differ in behavior internally.

The authors used two different online social network datasets to present their tool. Figure 12 displays the visualization interface of the system displaying a dataset collected from Whisper, a social messaging application. The dataset contains 136 million click events from 99,990 users over a period of 45 days. The authors categorized events into 6 main categories: browsing, account, posting, chatting, notification, and spam. The resemblance of events between this dataset and an e-commerce Clickstream dataset would be in browsing, account, and notification. The tool is built using D3.js,, and Packed Circle [32] is used to nest child clusters with parent clusters. The interface has semantic zooming and it is interactive where the user can click on any cluster to gain detailed information of the selected cluster through a pop-up box. In Figure 12, we can notice multiple nested clusters where the wider the cluster is the more general the user behavior is. In the pop-up box, each row represents one pattern of actions of events ranked according to a Chi score. The higher the score is the brighter the cluster is. The frequency column shows the frequency of the pattern in the dataset. The green bars indicate the frequency of the pattern outside the selected cluster, while the red bars indicate the frequency inside the cluster. The authors made a study afterward to test how intuitive and understandable the interface is to interpret user behavior. Additionally, the authors implemented several visualizations to display the hierarchical clusters. Figure 13 shows the different methods by which clusters are visualized through.

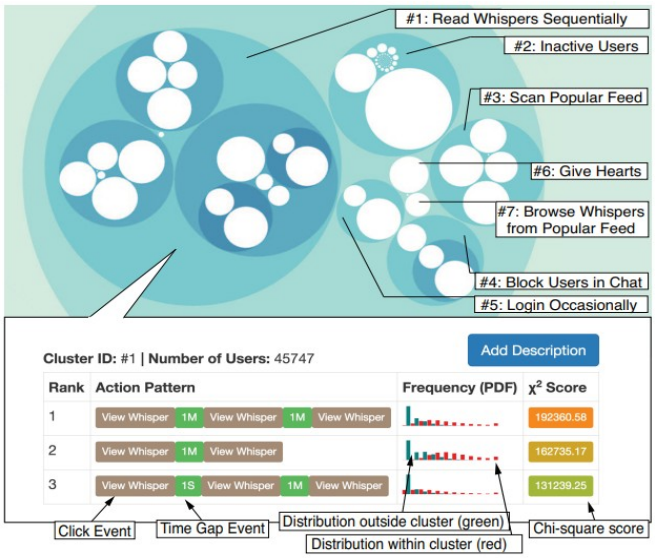

Fig. 12. The interface of Clickstream hierarchical clustering system "Taken from [20]"

Table. 5 Analysis summary for Figure 12

Idiom	Hierarchical clusters
What: data	Multidimensional Table
How: encode	Clusters express behaviors of users, sizes of clusters express frequency, luminance of color express chi-score
Why: task	Explore connection between different behaviors of users
Scale	50 clusters, thousands of users, millions of clicks

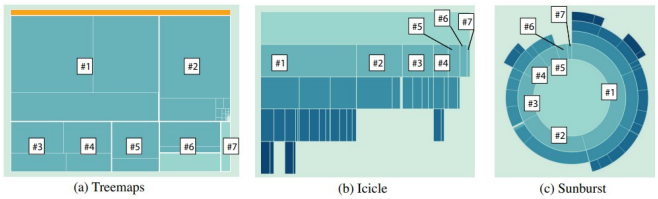

Fig. 13. The different visualization methods that hierarchical clusters could be displayed through, (a) Treemaps; (b) Icicle; (c) Sunburst "Taken from [20]"

3.3 Visualization Approaches

As Clickstream data is big and complex, researchers proposed a variety of methods and techniques to approach, handle, and visualize Clickstream data.

After we saw the low-level view for a single task provided by path navigation, and the high-level view of multiple actions provided by data clustering, a system proposed a solution to combine positives from both methods. MatrixWave is a system that compares two event-sequence datasets [25]; Clickstream datasets that are captured from different users at different times from the same website. A modern approach to path navigation is to aggregate paths and then visualize them using Sankey diagrams. However, these visualizations are readable for a dozen of pages only, and they become unreadable when visualizing large amounts of transitions between large amounts of visited web pages. Data clustering can handle the volume issue of Clickstream data, but it does not provide the ability to track browsing paths for a single task. MatrixWave visualizes Clickstream data using a technique called "transition matrices". Figure 14 shows a comparison between a Sankey diagram (Figure 14(a)) and a transition matrix (Figure 14(b)). In the Sankey diagram, Clickstream paths are divided into a sequence of ordered steps where each node represents a visited web page. Paths are aggregated and then visualized and analysts can compare the most frequent path from the thickness of the link between nodes. In the transition matrix, each cell in the matrix represents the volume of traffic between two visited web pages. With that, the visualization would still be readable for a huge amount of transitions between two nodes, but it loses the ability to track the full navigation path. Therefore, the system aggregates transition matrices between every two nodes along the path in a zig-zag layout (Figure 14(c)) in order for analysts to trace down navigation paths.

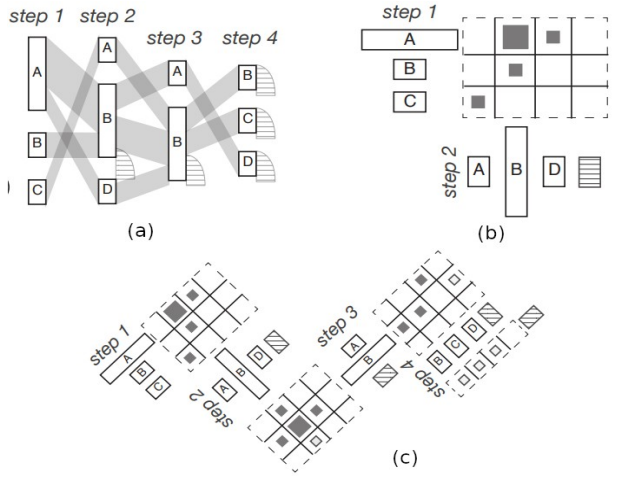

Fig.14. Comparison between different techniques to visualize navigation paths, (a) Sankey diagram; (b) transition matrix; (c) zig-zag of transition matrices "adapted from [25]"

The authors interviewed analysts in large IT companies to understand the tasks required to visualize Clickstream data. They identified four main tasks. First, to assess the number of visitors to a website and to detect pages with the highest number of visitors. Second, to know the entry points of the traffic to the website, which helps in understanding the effect of marketing campaigns. Third, to know how and when users leave the website. Fourth, to know the browsing path of users. The design of the interface shows traffic statistics on each page and between every two consecutive pages along multiple navigation paths. It gave analysts the ability to track multiple navigation paths through the zig-zag layout of multiple consecutive transition matrices.

The event-sequence datasets here are ordered Clickstream datasets. To compare two datasets, the system assumes that the structures of the datasets are the same with the exception of a small number of pages. Nodes, which represent visited pages, of the same name and step number from the two datasets are matched. Links that have the same source node and target node are matched. If a node or a link appears in one dataset only, then its equivalent is set to zero on the other dataset. The authors first considered four approaches to handle visual comparison: juxtaposition, superposition, explicit encoding, and animation. After testing the four approaches and discovering the positives and negatives of each, they used explicit encoding to visualize nodes and links. Figure 15 displays the chosen visual encoding for comparison. The visual mark is a 2D area and the channels used are a diverging color map and area size. Figure 15(a) shows how the traffic volume on the same node from the two datasets is encoded. The difference in traffic between the two event-sequence datasets is calculated and displayed. Then, differences are mapped to a diverging color scheme; purple to orange. Purple denotes traffic volume is greater in dataset one, orange denotes traffic volume is greater in dataset two, and white denotes traffic volume is equal in both datasets. The size of the node represents the average volume of traffic. A smaller size denotes a smaller average of traffic volume, which indicates the pages visited the least. A larger size denotes a larger average of traffic volume, which indicates the pages visited the most. With the same approach, Figure 15(b) shows how the traffic volume on links between two consecutive nodes is encoded along the navigation path. Rows represent nodes from the first step and columns represent nodes from the second step. Purple denotes the traffic volume in the corresponding link is greater in dataset one, and orange denotes the traffic volume in the corresponding link is greater in dataset two. Smaller squares denote a smaller average of traffic volume on that link, which shows paths taken the least. Larger squares denote a larger average of traffic volume on that link, which shows the paths taken the most. Empty cells denote no transition available between the nodes at that step, and white nodes denote traffic volume on that link is equal between the datasets.

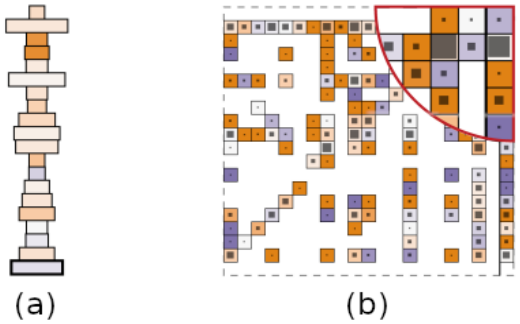

Fig.15. Explicit encoding used in MatrixView, (a) traffic on nodes; (b) traffic between two nodes along navigation paths "adapted from [25]"

Table. 6 Analysis summary for Figure 15	
Idiom	Transitions Matrices
What: data	Multidimensional Table: comparison of two event-sequence datasets
How: encode	(a) nodes express pages; (b) cells of matrices express links between two nodes, size express difference in traffic volume on nodes, color- divergence express difference in traffic volume on transition links
Why: task	Compare two event-sequence datasets, get insights for the overall traffic volume of pages, and track transition paths
Scale	Hundreds of nodes and links, millions of clicks, thousands of users

This approach is used to visualize all the nodes and links along the navigation path in a zig-zag manner, as we saw in Figure 14(c). The interface of MatrixWave is interactive with the ability to hover over any node or path to gain extra information through highlighting and pop-up boxes. Figure 16 displays the output on the interface of MatrixWave, where each transition matrix represents a step, and represents the difference in traffic volume values between the corresponding nodes and the next ones.

Fig.16. The interface of MatrixView showing a zig-zag of visually encoded transition matrices "taken from [25]"

Another tool called VizClick proposes a systematic approach to visualize Clickstream data [19]. The framework is built to solve three main concerns that the authors considered essential in understanding e-commerce Clickstream data. First, to measure association and interdependence between features of Clickstream data. Second, to analyze geospatial aspects of Clickstream data. Third, to get a deeper understanding of the topography of the website and how users are interacting with it. The authors used data collected from www.adobe.com for three days, then analyzed the data using Adobe Analytics. VizClick filters raw Clickstream data, then sort it into sessions to create transition matrices and visualize the website topography. The sorted sessions are summarized and then processed to discover and visualize the association between features, and to visualize Clickstream data at the geospatial level. The framework is packaged as a website that provides three types of visualizations with three different roles; heatmap to display the association between attributes, cartogram to display the geographical aspects of data, and tree and bubble graph to display the topography of the website. The authors used entropy-based association between two variables to measure association. They selected certain categorical variables from the data then they used a heatmap matrix to visualize the association. Figure 17 shows the heatmap matrix where we can notice clusters of different colors. Cross-cluster cells are in black and each color hue encodes an association. In the center, we can notice the biggest cluster in red which represents the sections of the website that are visited together the most. The higher the color saturation is the higher the association is between attributes. Hence we can notice the diagonal of the matrix has a high saturation as each attribute is associated with itself. An example of a clear association could be obtained from the purple cluster between "os", "browser", and "resolution".

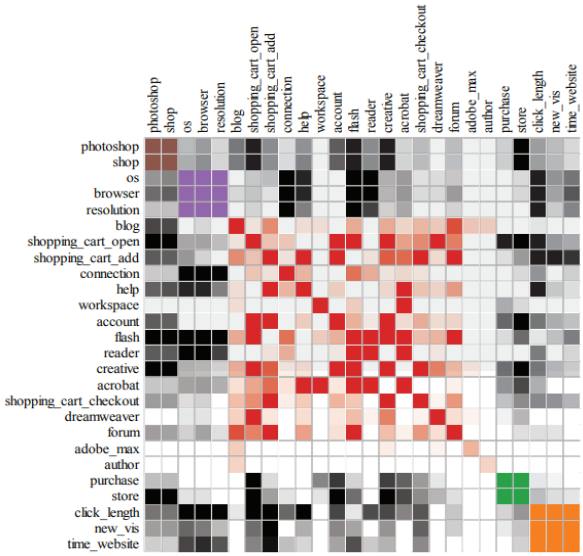

Fig. 17. Heatmap matrix for association between Clickstream data attributes "taken from [19]".

Table. 8 Analysis summary for Figure 17	
Idiom	Heatmap
What: data	Table
How: encode	Rows and columns express Clickstream data attributes, color hue express association, color saturation express degree of association
Why: task	Discover association between attributes to better understand the connection between them
Scale	Dozens of attributes, millions of clicks

To visualize geographical data, the authors used a cartogram. The mark used is area and the channels are size and color. Figure 18 shows an example of visualizing the number of new visitors (Figure 18(b)) and the number of purchases (Figure 18(c)) from each state of the US. The tool gives the analysts the ability to choose and visualize many various other geospatial metrics. The authors provided a way to normalize data before visualizing it in order to detect and remove any deviation that could be caused by the cartogram. The result would be a map similar to Figure 18(a) but it is color coded only without distortion. As for the topography of the website, the tool presents two types of visualization. First a bubble graph of clustered web pages based on modularity (Figure 19). The algorithm used in visualization produces a hierarchical tree that is visualized as a bubble graph. The parent node is the website root, the internal nodes represent clusters of pages, and the leaf nodes represent single pages from the website. The size of the bubble encodes traffic on that page. The visualization is interactive where if the analyst clicks on any node then a zoom-in to that node occurs. The second type of visualization proposed for geospatial data is a horizontal tree for the different web pages of the website that illustrates the website structure. It starts with the root URL of the website and expands into different nodes where each node is a different URL from the website. Hovering over any node would highlight its neighbors. With these visualizations in hand, an analyst

would understand Clickstream data deeply by examining the association between attributes. They would obtain a deeper understanding of user demographics, and a holistic view of the topography of the website.

There is a system with an improved Markov model called WebClickViz that compared the accuracy of the visualizations of WebClickViz with the visualizations of VizClick [22]. We noticed the mention of the Markov chain model in several papers when it comes to Clickstream analysis in general. However, we didn't include these papers as including models would be out of the scope of this paper. WebClickViz uses Apache Flume tools and provides several visualizations that are similar to VizClick's visualizations as a result. The authors made a comparison of accuracy between WebClickViz and VizClick and demonstrated how WebClickViz has better accuracy than VizClick. The accuracy of visualizations isn't much discussed in the papers we reviewed. Nonetheless, VizClick visualizations are interactive, clear, and more deeply thought of. While the focus of WebClickViz was on the model used and the analysis accuracy, not on the interface of the visualizations.

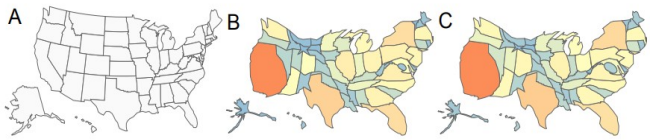

Fig. 18. Visualizing geo-spatial Clickstream data, (a) The map of the USA without distortion; (b) cartogram drawn according to the metric "new visit"; (c) cartogram drawn according to the metric "purchases" "taken from [19]".

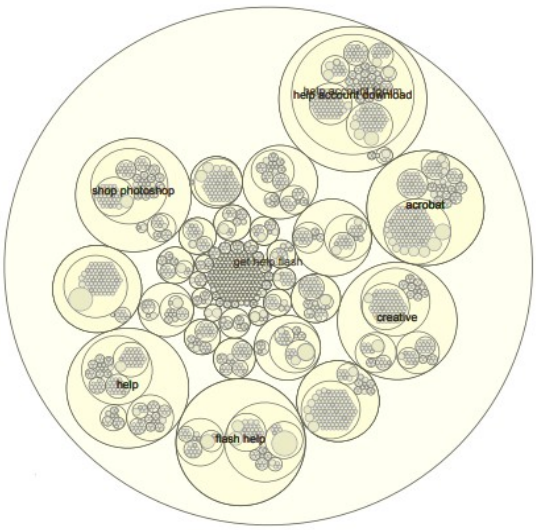

Fig. 19. Bubble graph of nested pages showing the topography of www.adobe.com "taken from [19]".

Table. 9 Analysis summary for Figure 19

Idiom	Bubble Graph
What: data	Tree
How: encode	Parent bubble expresses website root, internal bubbles express group of pages, leaf bubbles express single pages, size of bubble expresses traffic
Why: task	Explore the structure of the website according to traffic
Scale	Hundreds of pages

Video Clickstream data has unique characteristics that add an additional challenge to the analysis process. VisMOOC is a system proposed to analyze Clickstream video data generated from MOOC websites [26]. Analysis of MOOS data is used to learn the behavior of users through two types of analysis; Video-interaction analysis and content-based analysis. To conduct the research, the authors used two Clickstream datasets from two online courses. The data contained video Clickstream data, forum data, and grading data. The data consists of thousands of users and millions of events. The authors categorized events of videos into 6 categories: play, pause, seek, stalled, rate-change, and error. The authors collaborated with five experts and identified five visualization tasks:

1. The overall statistics of the dataset.
2. Types of events that happened at a certain position of a video.
3. The differences between the viewing behavior of different users.
4. The change of learning behavior over time.
5. Factors that affect the viewing behavior of users.

VisMOOC is designed to visualize data in three different views: list view, content-based view, and dashboard view. Each view presents different visualizations based on the tasks identified. In the list view, analysts can select from a list of videos the video that they are interested in analyzing. In the content-based view, analysts have the ability to know the distribution of events of a selected video through a graph event. Additionally, analysts can learn which parts of the video are skipped, and which parts of the video are likely to be watched again through a seek graph that visualizes seek event data. Among the 6 event types, authors found "seek" events to be very important for extracting insights from video data. In the dashboard view, analysts can see different statistics of videos and learners. Figure 20 shows a histogram from the dashboard view. The x-axis represents weeks and the y-axis represents the number of users. The graph shows the popularity of videos based on video type over the length of the course.

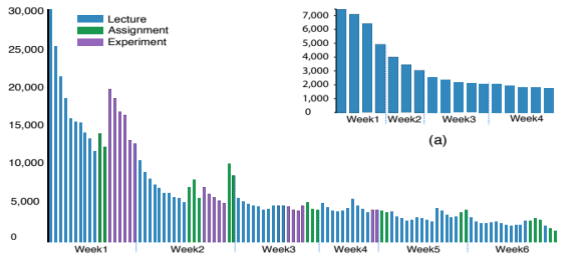

Fig. 20. Histogram showing popularity of educational videos based on type along the course duration, (a) popularity histogram for a single video "taken from [26]".

Figure 21 shows the event graph for four types of videos. The x-axis represents time and the y-axis represents the number of users. Color hue represents the type of event. Through the event graph, analysts can know how many users did a certain event along the video duration. Analysts can know where users paused or "seeked" the most and can locate errors. Figure 22 illustrates the second graph on the content-based view which is the seek graph. The seek graph consists of two parallel coordinates plots, one for forward seek at the top, and one for backward seek at the bottom. Each plot consists of two horizontal parallel axes that encode the starting and ending positions of the seek events. The upper axis represents the starting points and the lower axis represents the ending points. For each seek event, a line is drawn between the starting and the ending points. Color hue denotes viewing time: blue lines denote the seek events that happened on the first watch, while orange lines denote seek events that happened on a review after the first watch. Figure 22 compares seek events of the same video captured at two different

times. Figure 22a) shows data captured on the first week of the video release, while Figure 22(b) shows data captured on the week of the release of a related assignment. From the graph, we can notice that most of the users "seeked" the video from the starting time to many random locations. Besides, we can notice that some users watched the video for the first time after the release of the assignment. Lastly, we can notice multiple backward seeks happened at a certain location in Figure 22 (b) which indicates users watching the video again at a certain location to extract information to answer the assignment.

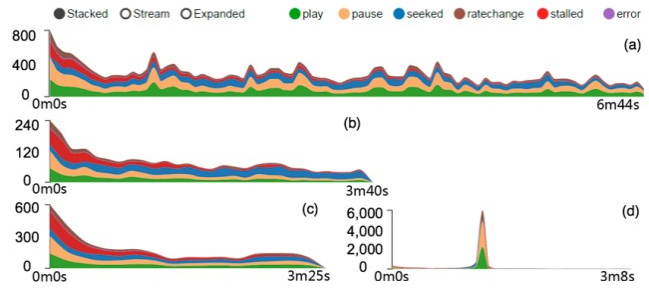

Fig. 21. Event graph for four educational videos along their duration, (a) lecture video; (b) assignment video; (c) experiment video; (d) experiment video with in-video question "taken from [26]".

Table. 10 Analysis summary for Figure 21

Idiom	Event graph
What: data	Multidimensional Table
How: encode	Layers express user count over time, color express event type
Why: task	Detect the time of peaks and bottoms of events to analyze user behavior based on content
Scale	Thousands of users, dozen of events

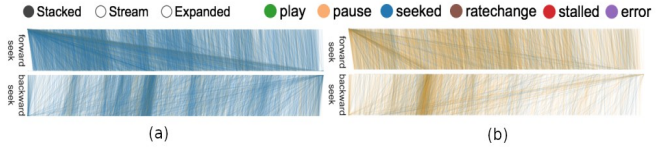

Fig. 22. Seek graphs for the same video captured on different times, (a) captured on the week of first release; (b) captured on the week of a related assignment released "taken from [26]".

Table. 11 Analysis summary for Figure 22

Idiom	Seek Graph (parallel coordinates)
What: data	Multidimensional Table
How: encode	Parallel Coordinates at the top express forward seek, Parallel Coordinates at the bottom express backward seek, axes at the top express starting points of seek, axes at the bottom express ending points of seek, color express watching for the first time (blue) and watching after the first time (orange)
Why: task	Track dense locations to understand the seek behavior of users based on content
Scale	Thousands of users

3.4 Analysis Architectures

As the process of analyzing and visualizing Clickstream data is integral, we included studies on analysis architectures of Clickstream data in our review. We shed the light on analysis architectures that proposed full architectures to convert raw Clickstream data into meaningful visualizations. The focus of these architectures is on the technologies that are used to analyze Clickstream data.

A system proposed analysis and visualization of gathered and stored Clickstream data [23]. The authors used a variety of tools to process and extract insights from the data of an e-commerce website. The gathered data is stored in the form of logs in a database. Figure 23 displays the proposed architecture which consists of 3 stages. In the first stage, Sqoop is used to import data from the database to HDFS (Hadoop Distrusted File System) and vice versa. Hadoop is a collections of open source software that is used to process big data [33]. HDFS is a distributed files system used in Hadoop and Sqoop is an application used to transfer data between databases and the Hadoop package. The data is replicated and stored in multiple locations in Hadoop, then it is preprocessed using Apache PIG. PIG is a high-level tool to analyze large data in simple programming [34]. In the second stage, the extraction of insights is done using Apache HIVE, a tool that has an interface similar to an SQL interface which is used to query and analyze data. Examples of extracted insights are a list of the top 10 sold items on the website and timely-based sales reports. The third stage is the visualization stage where the system uses Hadoop streaming of R scripts to visualize the insights. Examples of visualizations are simple bar plots, pie charts, and histograms. Even though there isn't much attention given to the visualization interface, the system has the capability to process big data of Clickstreams in a cost-effective and easy way.

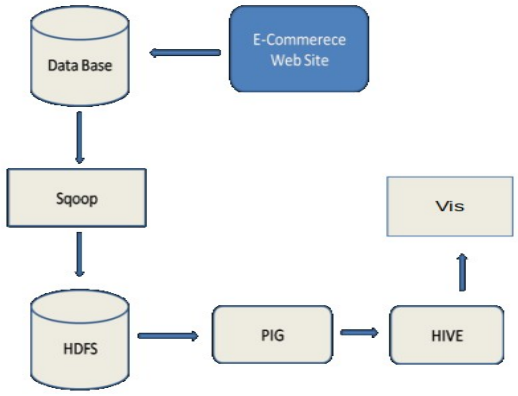

Fig. 23. A stored Clickstream data analysis architecture "adapted from [23]".

Another architecture we encountered proposed a system to analyze and visualize real-time Clickstream data [21]. The advantage of analyzing real-time Clickstream data over stored Clickstream data is the ability to predict user behavior in real time. The authors used Clickstream datasets from online education portals to test their system. They mentioned four benefits that are obtained from analyzing Clickstream data for online education websites in general; First, the ability to optimize the navigation path. Second, it helps analysts to know what courses users have the tendency to enroll in together. Third, it helps analysts to know what aspects of the website are well utilized in terms of user usage and what aspects are not. Fourth, it helps analysts to know how certain segments of users are utilizing the website. The architecture proposed consists of five stages: data collection, data processing, offline batching, data search and analysis, and data visualization. Figure 24 displays the full proposed architecture. The authors used a collection of Apache tools to build their system; Apache Kafka for collecting data, Apache Spark for processing data, Elasticsearch for real-time analysis of data, Apache Hadoop to store data, and

Kibana to visualize the data. Kibana is a source-available software that enables users to visualize Elasticsearch data and track the load of queries [35]. Even though the authors used commercial technologies like Elasticsearch and Kibana to build their architecture, they introduced analysis of real-time Clickstream data which is less discussed in the reviewed papers than stored-data analysis.

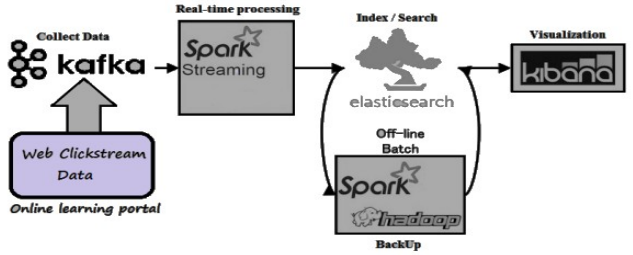

Fig. 24. A real-time Clickstream data analysis architecture “taken from [21]”.

4 Milestones

The work on the survey was divided into tasks in order to make the process systematic and organized. At first, we learned about systematic reviews, scoping reviews, and PRISMA. While writing the initial writeup, we developed the method for searching and selecting of papers, then we did a preliminary survey for papers. We optimized the method and finalized the selection of papers to conduct the review. We surveyed papers by categorizing their main topics and summarizing their main findings. Table 12 displays all the tasks required to conduct the review and the estimated time spent on each task.

Table 12. The distribution of milestones for conducting the review

Task	Est. Time	Deadline	Description
PRISMA	6 h	Nov. 11	Systematic reviews, PRISMA checklist and flow diagram
Initial Writeup	8 h	Nov. 15	Introductory to the topic, scope and focus points
Preliminary survey	12 h	Nov. 15	Method, preliminary search, record of relevant papers
Peer project reviews	2 h	Nov. 16	Receive feedback and suggestions for improving the paper
Optimize the methods	6 h	Nov. 22	Selection method of papers, revise initial writeup
Post-update meetings	2 h	Nov, 23	Discuss project direction and focus points, receive feedback on the paper structure, writing, and methods
Survey	10 h	Nov. 27	Identify papers, group papers, find patterns
Synthesize	24 h	Dec. 7	Categorize papers, distill topics, discussion and conclusion
Final presentation	2 h	Dec. 14	Prepare slides
Final writeup	8 h	Dec. 15	Finalize writing
	80 h		

5 Discussion

We utilized the PRISMA flow diagram and followed the direction of several elements from the PRISMA checklist. We used PRISMA to add a sense of structure to our search and selection of papers. We did not take all the elements of the PRISMA checklist into consideration, because the aim of the paper is to be a general review, not a systematic review. After we identified seed papers, we did a backward citation search only but no forward citation search was done. This is one weakness of this paper, especially when the total number of papers selected for review is 12. The research on Clickstream data is still in the early stage [22], which makes the number of reviewed papers in this review understandable. During the review, we noticed authors referring to general visualization techniques in their “related work” section. However, this paper is concerned with research that uses techniques to visualize Clickstream data specifically, in particular e-commerce Clickstream data. Furthermore, we noticed that the order or the sequence of Clickstream data is very important. Clickstream data is a collection of millions of ordered Clicks for thousands of users for different tasks. Therefore many authors considered logs data and event-sequence data visualization in their work. This paper is concerned with Clickstream data only, however, the sentences we used for the survey are very close in vocabulary, and using different terms, such as "event-sequence data", might give us a wider pool of related results.

We proposed a categorization for the reviewed work. We did the categorization based on the trends we found among papers. The papers on "path navigation", "data clustering", and "analysis architectures" were obvious to categorize. The rest of the papers are attempts to analyze and visualize Clickstream data in a novel and customized way. Therefore, we categorized them as “visualization approaches” in order to show the variety of methods that the researchers proposed to visualize Clickstream data according to customized tasks. We summarized the reviewed papers in a way that displays all relevant information of each paper. We chose what to include from reviewed papers according to the scope we identified after the preliminary selection. The hierarchical clustering system used Clisktream datasets for social media websites, and VisMOOC used Clickstream datasets for educational websites. But we included both in this review because we noticed the tasks and the visualizations identified can be easily utilized for e-commerce Clickstream datasets. We noticed the similarity between actions and tasks, and certain characteristics of the used datasets with e-commerce Clickstream datasets.

From the reviewed papers we noticed the importance of having low-level and high-level views on the dataset. A low-level view is concerned more with understanding particular user activities. A high-level view is concerned more with the number of users and clicks count. Examples of low-level views are, tracking the navigation path of users for a single task and visualizing Clicks of data based on the user interaction with the content. Examples of high-level views are, showing traffic volume on pages, displaying the association between Clickstream data attributes, and displaying the statistics of users carrying out certain tasks. Path navigation methods are very efficient in providing low-level views while data clustering methods provide high-level views of data clearly and easily.

Among the tools reviewed, we identify MatrixWave as the tool that provided a balanced view between low-level and high-level views. The interface, which provides a zig-zag layout of multiple transition matrices to display navigation paths and traffic volumes, might be complex but it is efficient. Analysts could compare a dataset captured at different times, and track navigation paths and traffic volumes on pages and links along the paths. The footstep graph with its 2D patterns is a unique technique in path navigation

because it add more attention towards traffic volume. The patterns are designed in order for analysts to extract insights from Clickstream data easily. As shown in 3.3.2, several patterns are identified, and with this identification, it is easy for analysts to identify the behavior of users from any dense and complex paths. Moreover, the association between Clickstream data attributes and the topography of a website has less importance compared to knowing the navigation path of users and traffic volumes. However, the heatmap matrix of association provides analysts with an examining view of Clickstream data and it might help them in deciding where to focus when identifying tasks. Analysts would identify new tasks based on association. On the other hand, the bubble graph that displays the topography of a website according to traffic is a very efficient way to explore ramified and complex websites. Analysts could understand the overall structure of the website and know which parts of the website are utilized the most. Lastly, the content-based view proposed by VisMOOC is unique in its functionality and purpose. The purpose is to analyze the content of video data and the functionality is done through seek and event graphs. The combination of these two graphs to extract insights from content showed how helpful the system is according to the feedback that the authors received from experts.

From the review conducted, we learned and summarized recommendations that can help designers optimally extract insights from Clickstream data. Designers should take into consideration using multiple methods in their visualization systems. Clickstream data is complex and no single method is enough to extract all the insights that Clickstream data hold. Furthermore, designers should aim to provide both low-level and high-level views of Clickstream data in their visualizations. The challenge of Clickstream data comes from its size and type. There are millions of clicks from thousands of users carrying out hundreds of tasks. Designers should aim to find a balanced view between understanding tasks and considering the number of clicks and users carrying out these tasks. Designers might derive visualizations from tasks then display their visualizations in multiple views according to the tasks identified. Lastly, designers should pay attention to both the accuracy of their analysis and the visualization interface. Having sophisticated visualizations without paying attention to the accuracy of the analysis might lead to taking bad decisions by decision-makers of e-commerce websites. Similarly, having high analysis accuracy without proper visualizations makes the process of extracting insights difficult.

6 Conclusion

In this paper, we conducted a review on visualizing Clickstream data of e-commerce websites. We used the PRISMA checklist and flow diagram to add organization to finding and selecting papers to review. After selecting 9 seed papers we used backward citation and selected another 3, which made the total number of papers to review 12. We categorized papers into topics and summarized each paper according to our identified scope. We identified four main topics in visualizing Clickstream data, path navigation, data clustering, visualization approaches, and analysis architectures. Path navigation tools are concerned with tracking the browsing patterns of users navigating a website. Data clustering tools are concerned with clustering Clickstream data based on the actions of users and traffic volume. The topic of visualization approaches shows how different researchers approached handling Clickstream data based on customized tasks. Lastly, analysis architectures are built either based on stored data or analyzed in real time. We provided a "milestones" table to show how we conducted the work on this paper. Then we discussed the reviewed topics and provided recommendations for analyzing and visualizing Clickstream data.

In the future, we will conduct a rigorous systemic review by applying all elements from the PRISMA checklist. We will use forward citation search to track the related built-up work based on the current reviewed work. Furthermore, The review should have a broader scope that will include research on the visualization of event-sequence data. This will allow us to use a wider range of vocabulary for searching for papers which will lead to a wider range of topics and techniques that would be helpful in visualizing Clickstream data. Lastly, we will use other databases beside Google Scholar to broaden search results.

References

[1] Hao, M. C.; Dayal, U.; Hsu, M.; Sprenger, T. & Gross, M. H. Visualization of directed associations in e-commerce transaction data. *Data Visualization 2001, Springer*, 2001, 185-192.

[2] Calibo, D. I. & Rodriguez, C. A. eCommerce Sales Attrition: A Business Intelligence Visualization. *Big Data Technologies and Applications, Springer*, 2018, 107.

[3] Mackinlay, J. D. Opportunities for information visualization. *IEEE Computer Graphics and Applications, IEEE*, 2000, *20*, 22-23.

[4] Negash, S. & Gray, P. Business intelligence. *Handbook on decision support systems 2, Springer*, 2008, 175-193.

[5] Zheng, J. G. Data visualization in business intelligence. *Global business intelligence, Routledge*, 2017, 67-81.

[6] Kauffman, R. J.; Srivastava, J. & Vayghan, J. Business and data analytics: New innovations in the management of e-commerce. *Electronic Commerce Research and Applications, Elsevier*, 2012, *11*, 85.

[7] Yu, C. & Ying, X. Application of data mining technology in e-commerce. *2009 International Forum on Computer Science-Technology and Applications*, 2009, *1*, 291-293.

[8] Kohavi, R.; Mason, L.; Parekh, R. & Zheng, Z. Lessons and challenges from mining retail e-commerce data. *Machine Learning, Springer*, 2004, *57*, 83-113.

[9] Urbaczewski, A.; Jessup, L. M. & Wheeler, B. Electronic commerce research: A taxonomy and synthesis. *Journal of organizational computing and electronic commerce, Taylor & Francis*, 2002, *12*, 263-30.

[10] Albert, B.; Tullis, T. & Tedesco, D. Data preparation. *Beyond the usability lab: Conducting large-scale online user experience studies, Morgan Kaufmann*, 2009.

[11] Dextras-Romagnino, K. & Munzner, T. Segmentifier: Interactive refinement of clickstream data. *Computer Graphics Forum*, 2019, *38*, 623-634.

[12] Tang, J.; Zhou, Y.; Tang, T.; Weng, D.; Xie, B.; Yu, L.; Zhang, H. & Wu, Y. A visualization approach for monitoring order processing in e-commerce warehouse. *IEEE Trans. on Visualization and Computer Graphics, IEEE*, 2021, *28*, 857-867.

[13] Ferreira, T.; Pedrosa, I. & Bernardino, J. Business intelligence for e-commerce: Survey and research directions. *World Conf. on Information Systems and Technologies*, 2017, 215-225.

[14] Prisma [Internet]. PRISMA. Available from: https://www.prisma-statement.org//

[15] Brainerd, J. & Becker, B. Case study: e-commerce clickstream visualization. *IEEE Symp. on Information Visualization (INFOVIS)*, 2001, 153-156.

[16] Waterson, S. J.; Hong, J. I.; Sohn, T.; Landay, J. A.; Heer, J. & Matthews, T. What did they do? understanding clickstreams with the WebQuilt visualization system. *Proceedings of the Working Conf. on Advanced Visual Interfaces*, 2002, 94-102.

[17] Ting, I.-H.; Kimble, C. & Kudenko, D. Visualizing and classifying the pattern of user's browsing behavior for website design recommendation. *International Workshop on Knowledge Discovery in Data Stream, Pisa, Italy, 2004, 24, 101-102.*

[18] Wei, J.; Shen, Z.; Sundaresan, N. & Ma, K.-L. Visual cluster exploration of web clickstream data. *2012 IEEE conf. on visual analytics science and technology (VAST)*, 2012, 3-12.

[19] Kateja, R.; Rohith, A.; Kumar, P. & Sinha, R. VizClick visualizing clickstream data. *2014 International Conf. on Information Visualization Theory and Applications (IVAPP)*, 2014, 247-255.

[20] Wang, G.; Zhang, X.; Tang, S.; Zheng, H. & Zhao, B. Y. Unsupervised clickstream clustering for user behavior analysis. *Proceedings of the 2016 CHI conf. on human factors in computing systems*, 2016, 225-236.

[21] Hanamanthrao, R. & Thejaswini, S. Real-time clickstream data analytics and visualization. *2017 2nd IEEE International Conf. on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 2017, 2139-2144.

[22] Frhan, A. J. Website clickstream data visualization using improved Markov chain modelling in apache flume. *MATEC Web of Conferences*, 2017, *125*, 04025.

[23] Anand, P. M.; Vamsi, G. S. & Kumar, P. R. A novel approach for insight finding mechanism on clickstream data using hadoop. *2018 Second International Conf. on Inventive Communication and Computational Technologies (ICICCT)*, 2018, 446-449.

[24] Hong, J. I. & Landay, J. A. WebQuilt: a framework for capturing and visualizing the web experience. *Proceedings of the 10th international conf. on World Wide Web*, 2001, 717-724.

[25] Zhao, J.; Liu, Z.; Dontcheva, M.; Hertzmann, A. & Wilson, A. Matrixwave: Visual comparison of event sequence data. *Proceedings of the 33rd Annual ACM Conf. on Human Factors in Computing Systems*, 2015, 259-268.

[26] Shi, C.; Fu, S.; Chen, Q. & Qu, H. VisMOOC: Visualizing video clickstream data from massive open online courses. *2015 IEEE Pacific visualization symposium (PacificVis)*, 2015, 159-166.

[27] Lee, J.; Podlaseck, M.; Schonberg, E. & Hoch, R. Visualization and analysis of clickstream data of online stores for understanding web merchandising. *Data mining and knowledge discovery, Springer*, 2001, *5*, 59-84.

[28] Lee, J. & Podlaseck, M. Using a starfield visualization for analyzing product performance of online stores. *Proceedings of the 2nd ACM Conf. on Electronic Commerce*, 2000, 168-175.

[29] Lee, J.; Podlaseck, M.; Schonberg, E. & Hoch, R. Visualization and analysis of clickstream data of online stores for understanding web merchandising. *Data mining and knowledge discovery, Springer*, 2001, *5*, 59-84.

[30] Lee, J.; Lee, H. S. & Wang, P. Analytical Product Selection Using a Highly-Dense Interface for Online Product Catalogs. *IBM Institute of Advanced Commerce Technical Report*, 2001.

[31] Munzner, T. Visualization analysis and design. *CRC press*, 2014.

[32] Wang, W.; Wang, H.; Dai, G. & Wang, H. Visualization of large hierarchical data by circle packing. *Proceedings of the SIGCHI conf. on Human Factors in computing systems*, 2006, 517-520.

[33] Apache Hadoop [Internet]. Hadoop. Available from: https://hadoop.apache.org/

[34] Apache PIG [Internet]. PIG. Available from: https://pig.apache.org/

[35] Kibana [Internet]. Kibana website. Available from: https://www.elastic.co/kibana/