Analysis and Visualization of User Navigations on Web

Honey Jindal, Neetu Sardana and Raghav Mehta

Abstract The web is the largest repository of data. The user frequently navigates on the web to access the information. These navigational patterns are stored in weblogs which are growing exponentially with time. This increase in voluminous weblog data raises major challenges concerning handling big data, understanding navigation patterns and the structural complexity of the web, etc. Visualization is a process to view the complex large web data graphically to address these challenges. This chapter describes the various aspects of visualization with which the novel insights can be drawn in the area of web navigation mining. To analyze user navigations, visualization can be applied in two stages: post pre-processing and post pattern discovery. First stage analyses the website structure, website evolution, user navigation behaviour, frequent and rare patterns and detecting noise. Second stage analyses the interesting patterns obtained from prediction modelling of web data. The chapter also highlights popular visualization tools to analyze weblog data.

Keywords Navigation · Visualization · Pattern · Website · User · Weblogs · Analysis

1 Introduction

In the last 20 years, the web has become the largest source of information. The web is expanding exponentially every year. This increase in the web has raised a lot of challenges like handling the large volume of data, handling the structural complexity of web sites, understanding user navigations, etc.

H. Jindal (✉) · N. Sardana (✉) · R. Mehta (✉)
Jaypee Institute of Information and Technology, Noida, India
e-mail: honey.cs0990@gmail.com

N. Sardana
e-mail: neetu.sardana@jiit.com

R. Mehta
e-mail: raghav.mehta.17@gmail.com

J. Hemanth et al. (eds.), *Data Visualization and Knowledge Engineering*,
Lecture Notes on Data Engineering and Communications Technologies 32,
https://doi.org/10.1007/978-3-030-25797-2_9

Users navigate on the web to access the information of their interest. These navigations patterns are stored in web log files. These files can help in extracting useful hidden facts. Initially, during the pre-processing stage, weblogs are cleaned, in which irrelevant information is removed, and noise is filtered. Post this stage the weblogs can be used to discover interesting and useful patterns by applying supervised or unsupervised learning techniques. Data present in the weblogs can be analyzed in various dimensions by visualizing it graphically. Visualization is a process to present information in the varied visual forms, i.e., graphs, trees, maps, charts, etc. This would help people to understand the insights of the large volume of data. Patterns, correlations, and trends which were undetected can be discovered using data visualization. Discovering web navigational patterns, trend and analyzing their results is undeniably gives advantages to web designers and website owners.

The analysis of web navigation patterns would steer ample web application like website design [1], business intelligence modelling [2, 3], promotional advertisement and personalization [4], and e-commerce [5].

Visualization of user navigation patterns is an essential prerequisite for generating effective prediction system. Visualization is an important part of pre-processing, pattern discovery and analyzing. It may be a part of the exploratory process [6]. Patterns can be visualized at two stages. Initially, data can be visualized once data is pre-processed and finally we can visualize patterns after data modelling. Both stages give insights into data with different perceptive. For example, to identify outliers, to discover new trends or patterns, check the quality of patterns, an efficiency of the model and evaluate the strength of evidence. The overall framework of the visualization of web navigation is shown in Fig. 1. The figure shows all the components of Web Navigation Mining and Visualization. The components are web log files, Data Cleaning and Pre-processing, Pattern discovery and stage A & B data visualizations.

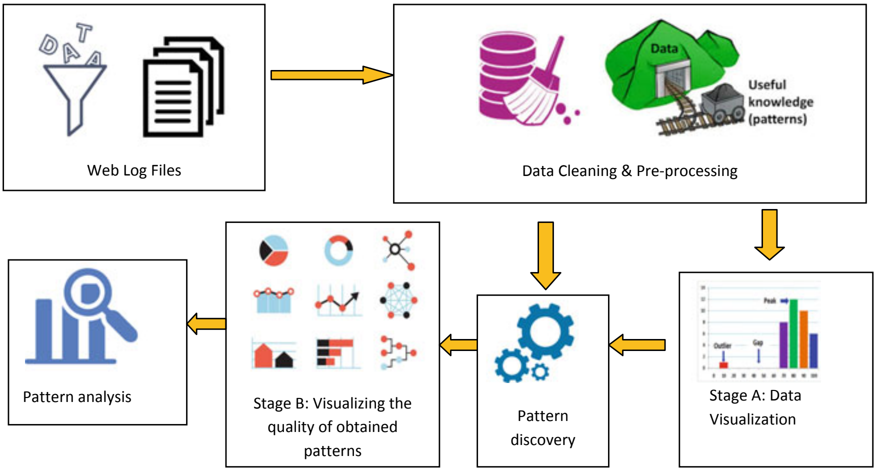

DATA
Web Log Files
Data
Useful knowledge (patterns)
Data Cleaning & Pre-processing
Peak
Outlier
Gap
Stage A: Data Visualization
Pattern discovery
Stage B: Visualizing the quality of obtained patterns
Pattern analysis

Fig. 1 Overall framework of web navigation mining

The following section describes the components of the given framework in a detailed manner. It also explains which questions can be addressed in two identified stages of visualizations.

2 Framework for Web Navigation Mining

2.1 Web Log File

Weblog file [7] is the record of user(s) activities performed during website navigation. The web server automatically creates this log file. Each hit to the website, including web page view, an image is logged in this file. It consists of information about users visiting the site, the source from which they came, and their objective to surf the web site. For every request, raw web log file store one line. A line is consist of various tokens separated by spaces. Hyphen (-) is placed when the token has no value. A line in a log file might look like the following:

```
192.168.1.3 - - [18/Feb/2000:13:33:37 -0600] "GET /HTTP/1.0" 200 5073
```

Some common log file types supported by the web server:

(a) **Common Log Format (Access)** [8]: This format is also known as the NCSA Common log format. The W3C working group defines the format. It is a standard text file format used for generating server log files. The Common Log Format has the following syntax:

```
"%h %l %u %t \"%r\" %>s %b"
```

where h is host id, l is client identity, u is user id, t is time and date, r is request, s is status and b is bytes

For example [9],

```
27.0.0.1 user-identifier zoya [11/Aug/2013:10:50:34 -0700] "GET
/apache_pb.gif HTTP/1.0" 400 2326
```

Where 27.0.0.1 is the IP address of the user, user-identifier is the identity of the user, zoya is the user id, 11/Aug/2013:10:50:34 -0700 denotes date and time of the request, GET /apache_pb.gif HTTP/1.0 is the http request method, 400 is the http status code, 400 is client error, 2326 is the size of the object returned to the client which is measured in bytes.

(b) **Extended (Access, Referrer, and Agent)** [8]: This format has two types: NCSA (National Centre for Supercomputing Applications) and the W3C (World Wide Web Consortium). The NCSA extended log format is the common log format appended with the agent and referrer information.

- **NCSA's extended format** is defined by the following string:

```
"%h %l %u %t \"%r\" %>s %b \"%{Referrer}i\"\%{User-agent}i\"
```

- **W3C extended log format** [10]: This log file is handled by Internet Information Server (IIS). W3C is a extended and customizable ASCII format with several fields like web browser, time, user IP address, method, URI Stem, HTTP Status, and HTTP Version. The time used here is Greenwich Mean Time (GMT).

```
#Software: Microsoft IIS 5.0 #Version: 2.0 #Date: 2004-06-05 20:40:12
20:40:12 183.18.255.255 GET /default.html 200 HTTP/2.0.
```

3. IIS (Internet Information Service) [10]: This log file format is simple. It consist several fields like time of requesting the web file, user IP address, request command, requested file, and response status code.

```
02:40:12 128.0.0.1 GET/404
02:50:35 128.0.0.1 GET/homepage.html 200
02:51:10 128.0.0.1 GET/graphs/prediction.gif 200
03:10:56 128.0.0.1 GET/search.php 304
03:45:05 128.0.0.1 GET/admin/style.css 200
04:10:32 128.0.0.1 GET/teachers.ico 404.
```

2.2 Pre-processing

The weblog contains large information of user navigation history. Sometimes, all information is not meaningful. For example, a user while browsing a web page, downloads images, videos, JS and CSS files. Therefore, we need to filter out this irrelevant information from the weblog file. The objective of the weblog file pre-processing is to remove irrelevant data from the web log file and reducing the amount of data. This process includes three steps: Data Cleaning, Users Identification, Session Identification and Session Reconstruction.

(a) **Data cleaning**: There are three ways to filter the data:

URL or website. The user session consists of information of user HTML navigated web page. The suffix for a gif, jpg, js files has no significance in session formation; thus it can be removed from the log file. For some special sites which generate content dynamically, filtering rules must be adjusted according to the requirement.

Request action and return status. The GET request actions are retained. Also, successful browsing records are retained while browsing record with error code is removed. Sometimes, the error logs of the website is useful for the administrator to analyze security.

Request IP: To get rid of the access from the automated requester, a robot IP filter list can be created.

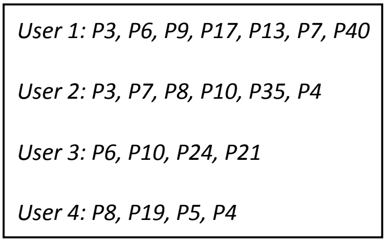
(b) **User Identification**

The user refers to an individual accessing server through a web browser [11]. A weblog file can distinguish the user's through user IP, user agents, and session cookies. This heuristic assumes that if two weblog entries are having the same IP address but different user agents, then these entries may belong to two different users. Figure 2 illustrated four user and their navigation paths. These users are identified through their IP address. Here, P_i represents the web page.

(c) **Session Identification**

A session identification is a technique which records user navigations during one visit into a session. In web navigation log files, sessions are defined as the sequence of web pages traversed by the users. The session identifier stops working when the users' process terminates, when the internet connection is lost or when a timeout. The session identification or construction techniques use three fields when processing web server logs: the IP address of the user, request time of the web page and requested URL address. Session construction techniques may differ because some use time information while other use navigation information of the user. The session identification heuristics are categorised into four; time-oriented, navigation oriented, integer programming. Based on session construction heuristics the outcome of this

Fig. 2 Pages navigated by the users

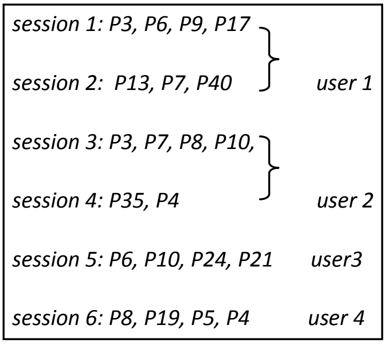


User 1: P3, P6, P9, P17, P13, P7, P40

User 2: P3, P7, P8, P10, P35, P4

User 3: P6, P10, P24, P21

User 4: P8, P19, P5, P4



session 1: P3, P6, P9, P17
session 2: P13, P7, P40 user 1
session 3: P3, P7, P8, P10,
session 4: P35, P4 user 2
session 5: P6, P10, P24, P21 user3
session 6: P8, P19, P5, P4 user 4

Fig. 3 Session identification from weblogs

phase is presented in Fig. 3 which generates sessions from Fig. 2. Herein six sessions are formed corresponding to four users.

Time-Oriented heuristics: Time-oriented heuristics have used the time-based threshold to create sessions: session time and page-stay. Catedge and Pitkow [12] measured mean time spend in a website is 9.3 min. Later, he derived a new duration time 25.5 min by adding 1,5 standard deviations. This has been rounded to 30 min. This becomes the thumb rule for many applications [13, 14] and used widely as session duration cut-off. During navigation with session duration 30 min, if a time between one accessed page is longer than session duration cut-off, then the next request webpage will be considered as a new session. This observation results in another type of heuristic known as page stay time. Generally, this threshold is set to 10 min according to [12]. If the difference between the timestamp of current navigated web pages and next navigation web page is greater than the specified threshold, then the current session is terminated, and the next session will start. The page stay-time [15] is affected by the page content, the time needed to load the page components and the transfer speed of communication line.

Time-oriented heuristics do not follow link information of the website, i.e., some useful information of the navigation paths are missed. Moreover, the order of web pages navigated by the users' may not be recorded correctly in the time based heuristics due to proxies or browser cache.

Navigation-Oriented: This heuristic does not consider time-based thresholds; rather it uses graph-based modelling. In navigation-oriented, web pages are considered as nodes and hyperlinks are considered as the directed edges between the nodes. According to Cooley [14], a requested web page P_i which is not reachable from the navigated web pages should be assigned to the new session. This heuristic is taken into account that P_i need not to accessible from the visited web pages; rather a user may backtrack to visited pages to reach P_i. This backward movement may not be recorded in the log file due to user cache. In this case, the heuristic discovers the shortest subsequence leading to P_i and add it to the user session.

The sub-sessions generated from clicking the back button do not imply strong correlations among web pages [16]. Therefore, it becomes a major issue. Cooley et al. [13, 14] show that backward navigation occurs due to the location of the web page rather than their content. The occurrence of this backward navigation produces noise in the session. Another drawback of this method is the increase in session length. Addition of backward navigations in the session results into longer patterns which will further arise the network complexity and become computationally more expensive.

Integer Programming Approach: Dell et al. [17, 18] proposed the integer pro gramming approach which partitioned sessions into chunks using the IP address and agent information. To obtain longer session, these chunks are divided into sessions using logarithm function. The objective of this logarithm function is to assign web pages in a chunk to different web session. This implies that the reconstructed session have unique web pages. In an improved version [19], a backward pattern is added which allows repetition of web pages at *K*th and (K + 2)th position. However, this technique does not cover all combination of user return to the page like (k − 3)th, etc. This addition also increases noise in the session and page repetition.

(d) **Session Reconstruction**

Session reconstruction is essential to provide correct assignments of navigations to the prediction model. The performance of web navigation prediction models relies on the quality of session inputs [16, 20]. There are some methods to regenerate sessions: Dynamic time-oriented heuristic [21], Navigation Tree [5, 22], Smart-SRA [16], BBCom [20], BBDcom [20]. Dynamic time-oriented heuristic [21] uses session duration and page-stay time thresholds. This function generates more effective site-specific thresholds and replaces either session duration or page stay threshold during session construction. Navigation Tree is formed using maximal forward length sequences. The sessions are generated from the root node to leaf node traversals. The sub-sessions obtained from this technique may consist of repeated web pages which require high storage and computational cost. Another technique called Smart-SRA (Smart Session Reconstruction Algorithm) [16] was proposed which takes session duration, page-stay-time, maximum length sessions, web topology collectively to eliminate backward browsing sequences. This technique produces correlated web sessions, but the process to reconstruct the session is quite long. To find the shortest path between the source (entry) web page and destination (exit) web page, Jindal et al. [20] proposed two backward browsing elimination techniques, Compressed backward browsing, BBcom and Decompressed backward browsing, BBDcom. BBcom compresses the navigation sequences and reduces the redundancy of the web pages. Whereas BBDcom decomposes the session into multiple sub sessions such that each sub-session consists web pages which were traversed in forward directions.

Once the session is generated, visualization tools are employed to check their effectiveness. Session 2.3 presents challenges of web navigation mining. It also describes how the visualization of data can help to provide solutions for the identified challenges.

2.3 Post Pre-processing Visualization: Stage 1

Once the data residing in weblogs is pre-processed, the visualization can be performed to know useful facts about the website. There are varied charts available to visualize the website with varied perspectives. The visualization can help to know the frequent and infrequent navigation patterns, depicting website structure from website navigations, explore the evolution of website with time, Identify entry and exit point from the navigational structure, Visualize the traffic and popularity of the Website(s) from Web Server logs, Classifying users based on navigations and noise detection to identify usual & unusual users.

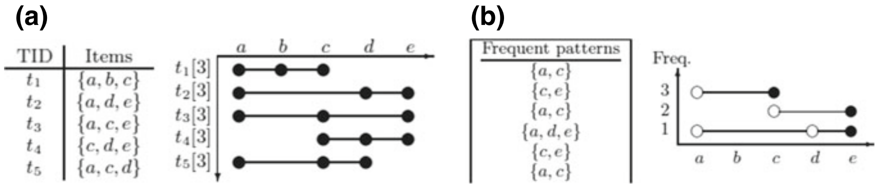
(a) *Identify Expected and Unexpected Frequent Patterns*

While analyzing the website navigation patterns, it's imperative to know the set of pages that are frequently accessed by the users as it signifies the popularity of the pages and also helps the website owners to make it richer in terms of content. Similarly, it's vital to know the rarely accessed set of pages as it helps in restructuring the website.

An interactive visual tool to analyze frequent patterns, called FpVAT [23] provides effective visual for data analysis. FpVAT consists of two modules: RdViz and FpViz. RdViz visualizes raw data, which helps the users to derive insight from the large raw data quickly. The second module, FpViz visualizes frequent patterns which are used to detect the number of frequent patterns expected and to discover the number of frequent patterns unexpected. Figure 4 presents the visualization of raw data and processed data. The difference between processed and unprocessed data can be observed from the Fig. 4a, b. The FpViz is an interactive tool used for large data (shown in Fig. 5) where nodes are represented with red color. Line between these nodes denotes connectivity. The connected path is the resultant frequent pattern obtained from the tool.

Textual form Graphical form Textual form Graphical form

A frequent pattern can also be visualized using Heatmaps [24]. Heatmaps are used to identify hot and cold regions. Figure 6 illustrates three different heatmaps: link position, click position, clicks/links. Herein, the links and clicks positions are visualized. Dimitrov et al. [24] divides the website screen into regions and gives



TID	Items
t_1	$\{a, b, c\}$
t_2	$\{a, d, e\}$
t_3	$\{a, c, e\}$
t_4	$\{c, d, e\}$
t_5	$\{a, c, d\}$

Textual form Graphical form

Frequent patterns
$\{a, c\}$
$\{c, e\}$
$\{a, c\}$
$\{a, d, e\}$
$\{c, e\}$
$\{a, c\}$

Textual form Graphical form

Fig. 4 **a** Representation of raw data navigations. **b** Representation of mining results of frequent patterns in both textual and graphical forms [23]

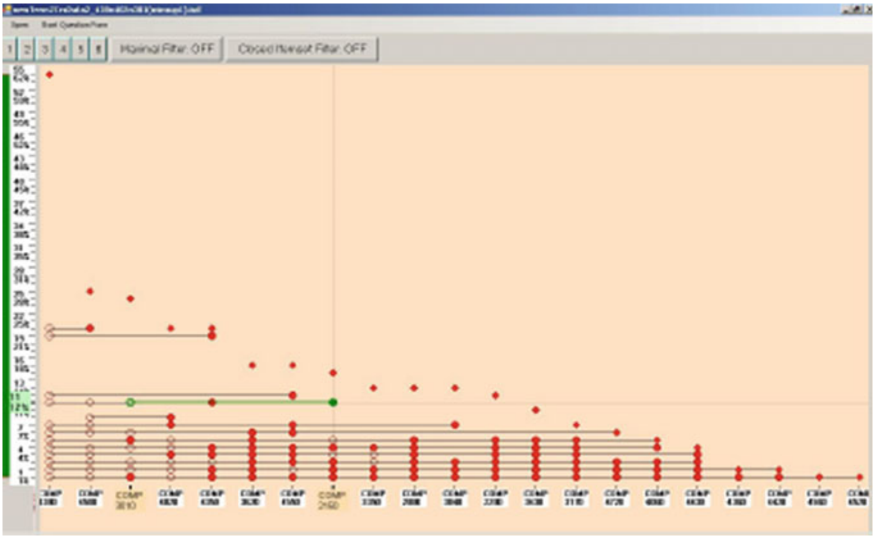


Fig. 5 FpViz module showing the visualization of frequent patterns mined [23]

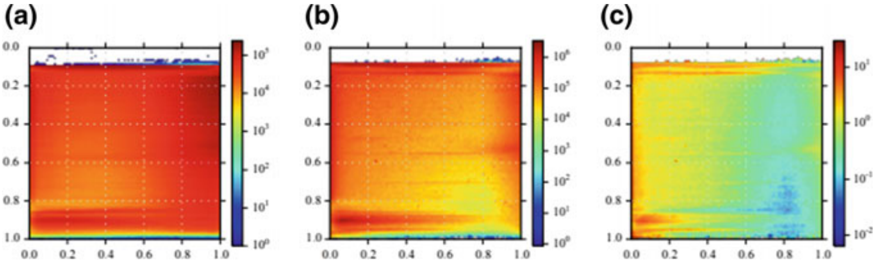


Fig. 6 Heatmaps [24]. **a** Links positions, **b** clicks positions, and (**c**) clicks/links

insights of frequently preferable regions by the users. They observed the region where links are placed. He analysed the region where users click on the links and found which regions is frequently (or rarely) clicked by the user. Figure 6a shows the position of links on the screen. This heatmap indicates high link density. Figure 6b presents heatmap of click position which indicates high click frequency regions. Figure 6c displays the number of clicks on a region. Here, dark (hot) colors indicate high frequency regions whereas light (cool) colors indicate a low frequency regions. Through heatmap, the author highlights that most of the users prefer to click on the left side of the screen.

(b) *Understanding Website Structure from Web Navigation Files*

Generally, a website structure is represented by the navigation tree. There are other layouts to visualize the navigation structure of the web or website. H-tree layouts (Fig. 7) representation are similar to binary trees which is suitable for balanced trees.

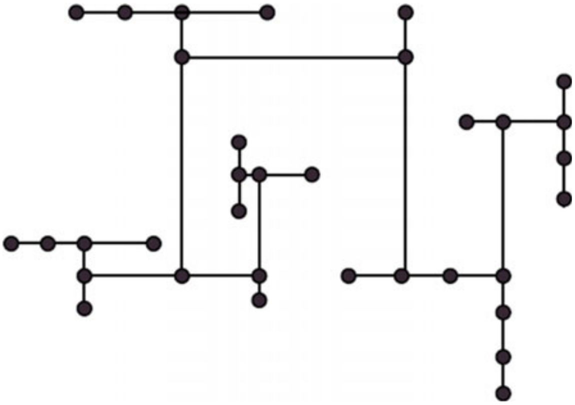
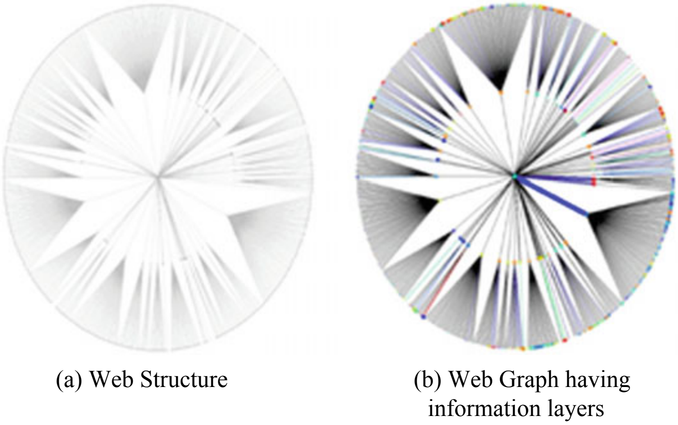


Fig. 7 H-tree layout [25]

This representation is suitable for those websites which have a balanced tree-like structure. In the web, usually, websites have complex structure i.e., high connectivity. Therefore, this complex structure could not be obtained from H-tree.

To obtain a graphical representation of complex website structures Eades [26] and Chi et al. [27] suggested a variation of tree known as the radial tree or disk tree (Fig. 8a). The nodes are placed in the concentric circles based on tree depth. This representation is known as a radial tree. The node of the radial tree indicates the web page, and the edge indicates the hyperlink between two web pages. The root node is placed at the center, and concentric hierarchical circular layers with root are formed. Parent to children links is connected from inner to the outer circle. Using Disk tree, data layers can be obtained from the web log file which is helpful to generate user sessions.



(a) Web Structure

(b) Web Graph having information layers

Fig. 8 Visualizing Web Navigation Patterns [28]: **a** visualize the web structure. **b** Visualize superposition of web navigations on web structure

WebViz [28] is an interactive tool to analyse the web navigation data through the structure, pattern discovery on web navigation behavior. By default, in WebViz node size represents the frequency of page visit, node color shows average page view time, edge thickness indicates a number of the hyperlink, edge colour shows the percentage of hyperlink usage count (i.e., hyperlinks sharing the same start page divided by the total number of hyperlinks). Figure 8b displayed a web graph having several information layers. In addition to this, outputs of data mining models like classes or cluster of web pages or links can be visualized.

Figure 9 shows a closer look at the web navigation sequences using Radial Tree. While visualization two different sequences are observed. The first sequence shows a path 0->1->2, and the second sequence shows a path 0->3->4. These two sequences indicate the highest popular sequences navigated by the users. The frequent pattern traversed by the users can be determined through Radial Tree. The popularity or frequency of navigations is determined through support and confidence measures. These measures will be used to filter less frequent navigation pairs. The visualization of navigation pairs and frequently accessed navigation path is highlighted in the Fig. 9.

A tree representation of navigations can be viewed as the collection of sub-trees, to understand the hierarchal information structure cone tree [30] was defined. Cone tree projects the tree on a plane where sibling subtrees are attached to their parent node. Another simpler view of cone tree is known as the balloon tree. Figure 10 shows a balloon view of the navigations. Here, the root node is placed at the center. The group of sub-tree nodes is placed in the second layer. There are six sub-trees which are represented by the circles of nodes. However, a single child who is not a part of any subtree is represented as a node and directly linked to its parent node. Figure 10 indicates six child nodes. Using the balloon view of the cone tree, the position of nodes can be determined directly, without reference to the trees [31, 32].

(c) *Evolution of website structure with time*

With time new information is added to the website while outdated web pages are deleted. This updation of web site structure causes many changes in the website struc-

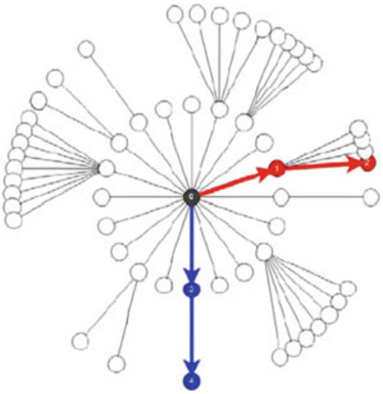


Fig. 9 Visualizing sequences using radial tree [29]

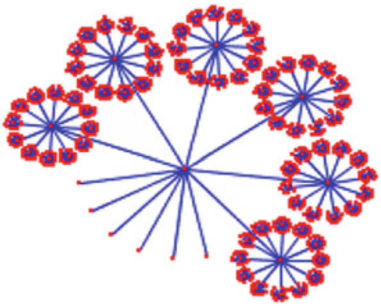


Fig. 10 Balloon view [30]

ture; i.e. adding and deleting the nodes or connections. Understanding the structural changes in the website becomes a major challenge for experts. To address this, Chi et al. [27] proposed a new visualization technique called the Time Tube. These tubes record several Disk tree over a period of time. This representation gives information about the change in website structure with time. Figure 11 illustrates four disk trees which show the evolution of web structure with time. This new visualization will guide the user to understand complex relationship and connection between production and consumption of information on the websites.

(d) *Identify the entry and exit point from the navigational structure*

The entry point is the web pages where users' typically enter the website. Exit points are the web pages where users' leave the website. These points provide insight into those web pages from which browsing starts and ends. To understand user entry and exit behavior, Chen et al. [33] had defined two operators: MINUS IN and MINUS OUT.

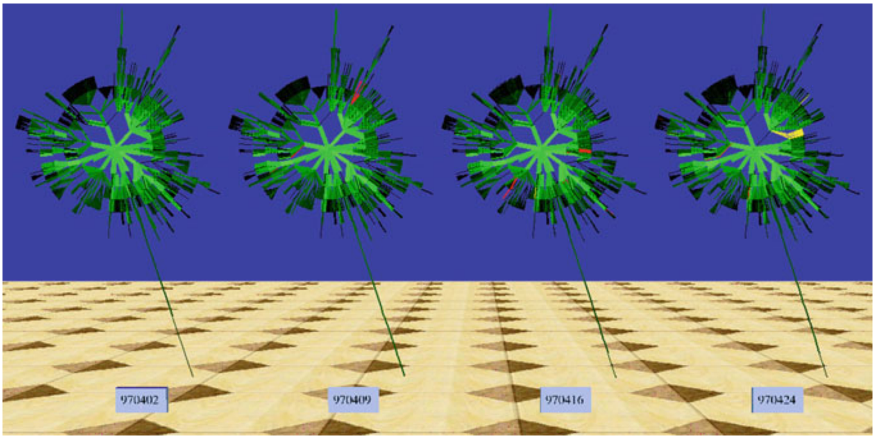


Fig. 11 Four disk trees shows the evolution of web struction with time [27]

To analysis, the entry point a MINUS IN operator is used which subtract the sum of the access to the page from any other pages that have a link to it from the number of visits of that page. This difference represents how often users enter the page directly. To analysis the exit point a MINUS OUT operator is used which subtract the sum of the access to any other pages from the specific page from the number of visits of that page. This difference indicates how often a user stops the navigation and leaves the site. Figure 12 illustrates the process to get Entry Points and Exit Points of the website.

(e) *Visualize the traffic and popularity of the Website(s) from Web Server logs*

The traffic on a web grows exponentially with time. A website having more traffic indicates its popularity among other websites. Understanding the traffic will give the insight of the incoming and outgoing users on the website. In addition, it provides information about the popularity of the website. Mark [34] presented some plots to understand the distribution of web traffic i.e., in-degree and out-degree of the node. According to him, the in-degree and out-degree must satisfy the power law (shown in Fig. 13a, b). This distribution depicts the strength of incoming and outgoing degree. Furthermore, the distribution is used to understand user traffic visiting the website or leaving the website. Figure 14a presents inbound traffic of the web server users. Figure 14b presents outbound traffic of the users contributing to the traffic of the web server using power law. These distributions measure popularity of a Website and stated that most popular sites have unbounded audience.

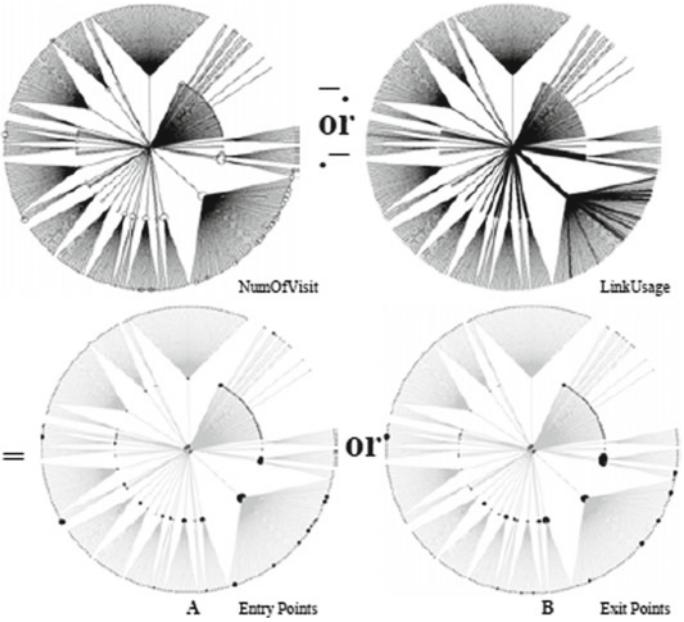


Fig. 12 Process to obtain entry and exit points using disk tree [33]

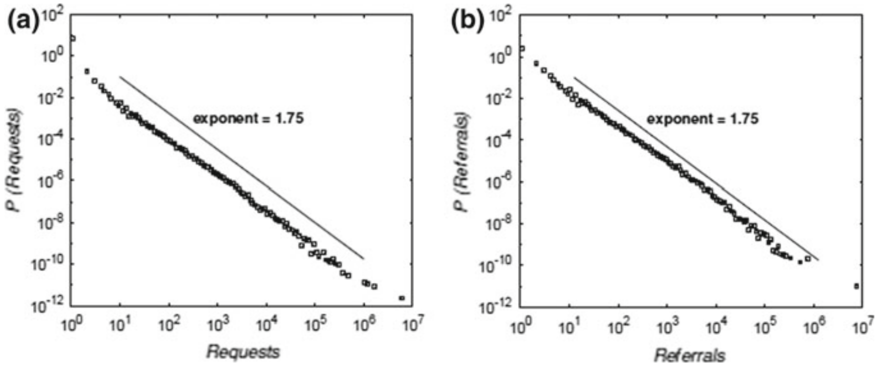


Fig. 13 Distributions of **a** in-degree strength and **b** out degree strength the web server data [34]

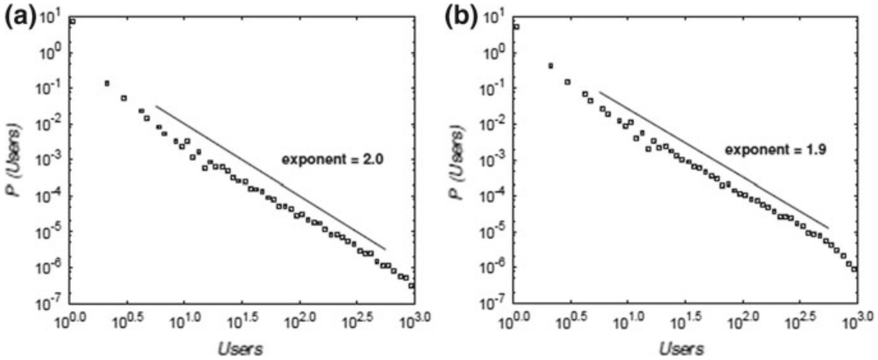


Fig. 14 Distributions of **a** unique incoming users and **b**) unique outgoing users from web server data [34]

(f) *Classifying searching-oriented user or browsing-oriented user*

There is two types of users navigation behavior: search-oriented and browsing-oriented. Search-oriented users search the desired information on the web while browsing-oriented users surf the web randomly. The navigation data of users gives useful insights to classify these users. Figure 15a shows the distribution of the active user's requests per second. A user behaviour can be understood by a ratio of the number of unique referring sites to the unique target sites. User browsing behavior can be obtained by comparing referring host and servers. If the referring hosts is less as compared to the servers, then the user browses through search engines, social networking sites, or a personal bookmark file. If the referring hosts is high compared to the servers, then a user is having surfing behavior: Fig. 15b presents bimodal distribution which states the existence of two user groups: search oriented and browsing oriented. Search-oriented users visit more sites as compared to surfers for each referrer.

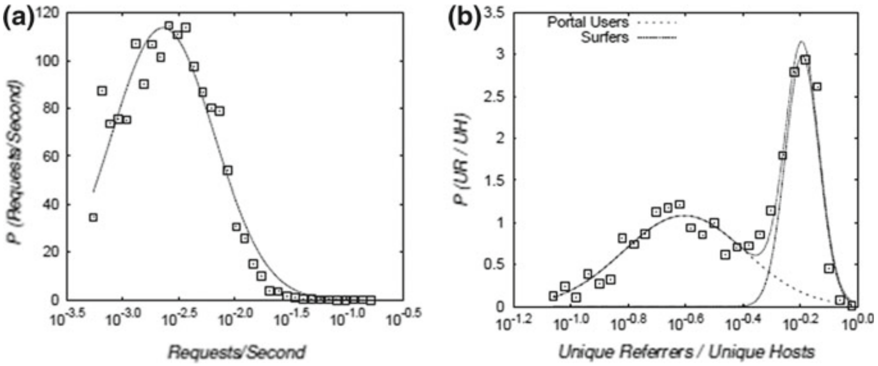


Fig. 15 **a** Distribution of the requests made per second by each user using log-normal fit. **b** Distribution of the ratio of the unique referring sites to unique target sites for each user using bimodal distribution [34]

Catledge and Pitkow [35] categorized search oriented browsing and surfer browsing into three using user browsing behavior. These three dominant types of behavior are still applied in many research areas. The first browsing behavior is known as search browsing, which is goal specific browsing. This is a directed search with the specified goal. The second browsing behavior is referred to as general purpose browsing. The idea of the browsing goal exists, but the path to the goal is undefined. In general purpose browsing, the user consults different sources, which are supposed to contain the desired information. The third browsing type is serendipitous browsing which is truly random browsing behavior. Figure 16 illustrates the type of users based on their browsing behavior [36]. The search browser is known as goal-oriented users, and the serendipitous browser is known as exploratory users. Figure 17 depicts an example of the goal-orient browser. Herein, the user follows a path towards blogs. There is less variation in the navigation path. Hence, the user is categorized into the goal-oriented browser. Figure 18 illustrates the exploratory browsing behavior. In this typical user browse a mixture of interesting content i.e., stories or blogs and ends in the comments section frequently.

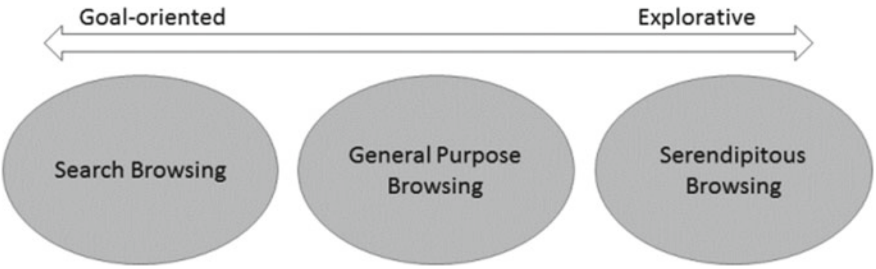


Fig. 16 Type of users in online browsing [36]

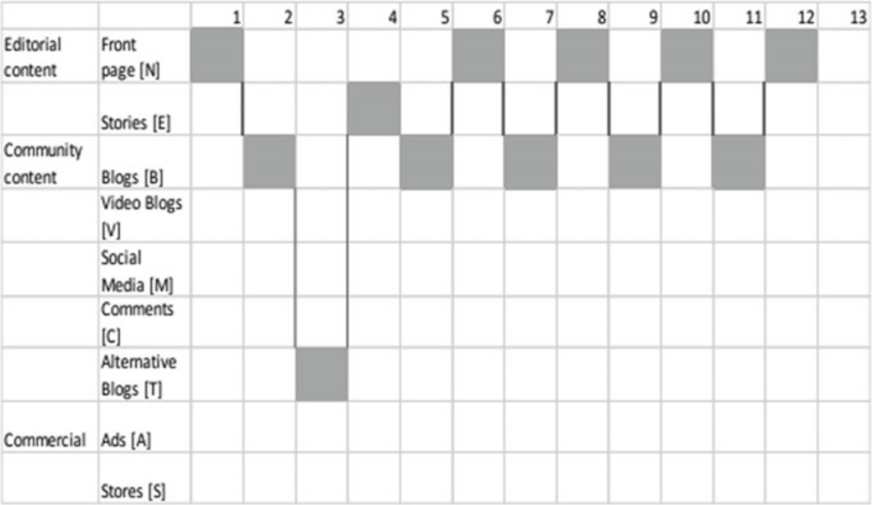


Fig. 17 Navigation path for goal-oriented browser [36]

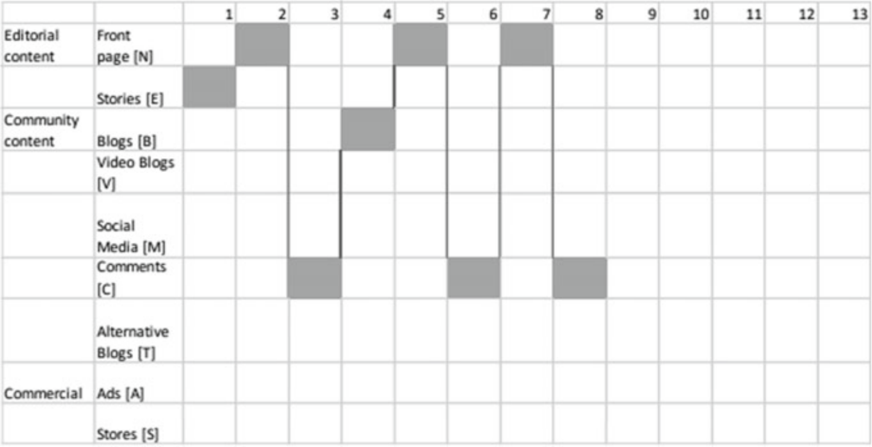


Fig. 18 Navigation path for exploratory browser [36]

(g) *Correlation of Sessions or Users*

Finding a correlation between users or sessions or features is important to find how similar or dissimilar properties they have. Figure 19 presents the correlation between the features using a heat map [37]. The color code indicates the strength of the correlation. The red colour indicates high correlation whereas blue colour indicates low correlation.

(h) *Noise Detection*

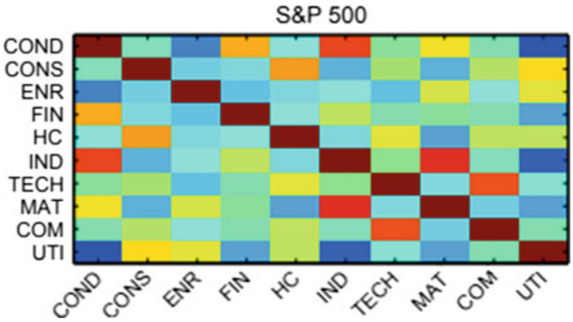


Fig. 19 Correlations of the features [37]

Data visualization gives a view to identify the normal and unusual users easily. Mark et al. [34] examines click stream sizes of the users'. Distribution of empty-referrer request indicates the user frequency who has jumped directly to a specific page (e.g., a homepage, a bookmark, etc.) instead of following web pages from the visited web pages. The resulting distributions are shown in Fig. 20a, b.

In web navigation, the user can either follow a forward path or backward path. The backward path shows the presence of noise as navigation doesn't follow the website topology. Forward path occurs when the user is interested in desired information from the web pages which has not been traversed. However, the backward path occurs when the user attempts to follow visited web pages. Jindal et al. [20] presented a study on backward navigations over varied session length. They have examined how frequently the user follows backward navigation path. According to them, the backward path are the repeated sequence of web pages in the session. The presence of a backward path produce longer sessions and has noise. For instance, a navigation path E->B->J-> H->B has noise as it consists backward navigation from H to B and it does not follow web topology (there is no link from H to B in the website structure).

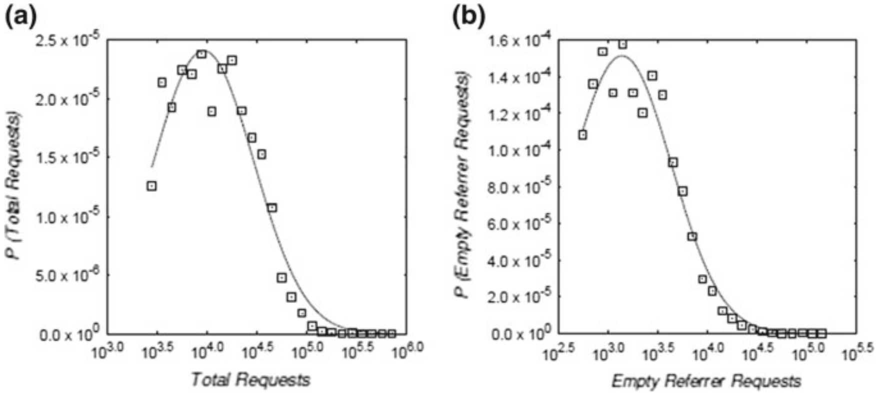


Fig. 20 Distributions of the **a** number of requests per user **b** number of empty-referrer requests per user [34]

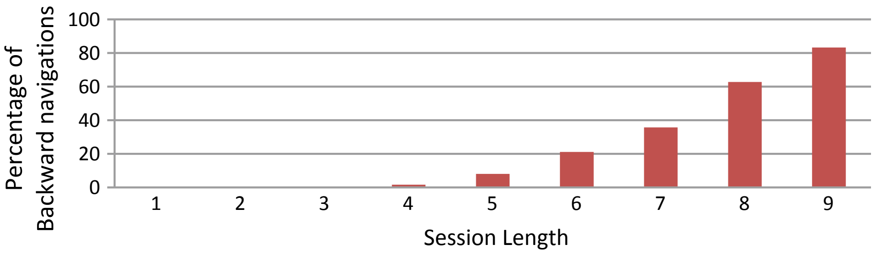


Fig. 21 Distribution of backward navigations with session length [20]

Figure 21 shows the number of backward navigations increases with session length [20]. The sessions with length one and two have no backward navigations. The backward navigations of higher order consist more repetition ranging from 0 to 3 which means session in the dataset may consist 0, 1, 2 or 3 cycles. This analysis gives insight to the users about variations of backward navigations with session length. This implies that user with more session length usually performs backward navigations. Presence of the backward paths makes prediction models difficult to learn user intentions.

An anomaly detection tool, OPAvion [38] is composed of three sub-parts: Data Summarization, Anomaly Detection, and Visualization. Summarization gives an overview of large network structures and their features. Anomaly detection phase extract features of the graph, node and neighborhood information. The graph induced with this information forms egonet. To search for the anomaly, the majority of normal neighboring nodes should be understood. The deviation (if any) has been recorded and analyzed to capture anomaly. Figure 22a the 'Stack Overflow' Question & Answer network is illustrated as Egonet. Redline is the least squares on the median values (dark blue circles) of each bin. The top anomalies deviating are marked with trian-

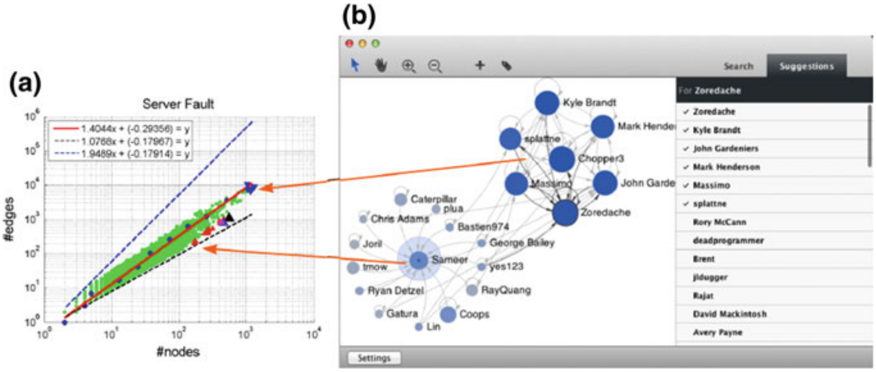


Fig. 22 **a** Illustrating Egonet of the 'Stack Overflow' Q&A graph. The top anomalies deviating are marked with triangles. **b** Visualization module working with the anomaly detection module, showing as a star [38]

gles. Figure 22b presents the anomaly detection visualization module. The anomaly (Sameer, at its center) is presented with the star in Fig. 22b. The Nodes represent Stack Overflow users and a directed edge from a user who asks a question, to another user who answers it. The user Sameer is a maven who has answered a lot of questions (high in-degree) from other users, but he has never interacted with the other mavens in the near-clique who have both asked and answered numerous questions. Therefore, Sameer is the anomaly of this network.

2.4 Pattern Discovery

Pattern discovery has attracted many experts to find interesting patterns, to understand user behaviour, etc. Wide research has been carried out to find interesting patterns varied according to the application. Hu et al. [39], Pani et al. [40], Singh and Singh [41], Facca and Lanzi [42], Jindal and Sardana [43] present an overview of web usage mining system and survey on pattern discovery techniques. These papers present the detail of the components of web navigation mining. Each component requires a considerable attempt to accomplished desired objectives. The pattern discovery component is the most important part of the web navigation system. Several data mining techniques were proposed like clustering, classification, association rule mining, sequential mining used for Web Navigation Mining. Tseng et al. [44] proposed association rule-based techniques to discover temporal patterns using temporal properties like weekdays or hours. Chordia and Adhiya [45] proposed the sequence alignment method for clustering sessions based on certain criteria. Abrishami et al. [46] and The [47] integrate three web navigation prediction techniques and formed new hybrid approaches. The proposed techniques combine association rule mining, clustering and sequence mining techniques to improve the accuracy of the model. It has been seen that the Markov Model is extensively used to discover web navigation patterns. Markov models [48–54] are stochastic processes and well suited for modelling and analyzing sequential patterns. The detail description of Markov based models is highlighted by Sunil et al. [55] and Jindal [56]. Based on application and objectives Markov models are combined with data mining techniques. Thwe [47] and Vishwakarma et al. [57] analyzed the application of Markov based Models on clusters. Sampath et al. [58–60] present the integrated model to improve the prediction accuracy. They observed the effect of association rule mining on Markov based Models.

2.5 Post Pattern Discovery Visualization: Stage 2

The second stage of visualization is used to visualize the interesting patterns obtained from pattern discovery phase. It help to understand user navigation behavior after data mining algorithms.

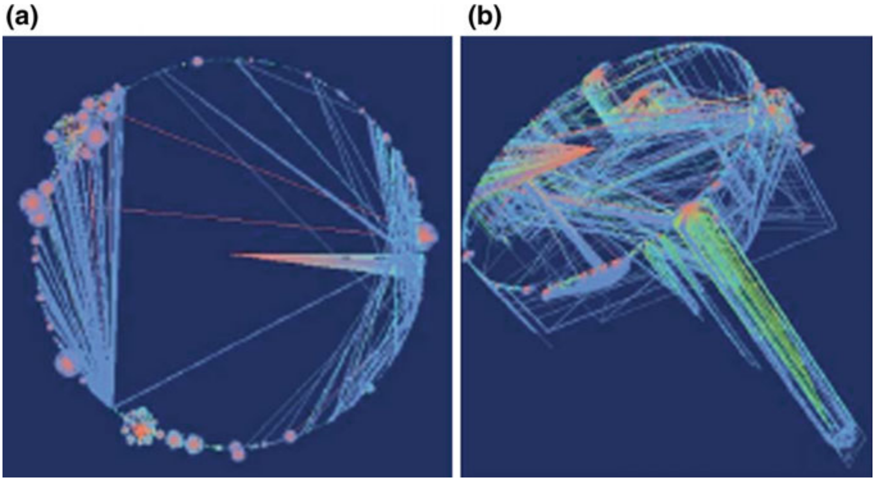


Fig. 23 **a** 2D visualization of web logs using Strahler coloring technique. **b** 3D visualization of more information which clarifies user behavior in and between clusters [61]

(a) *Visualizing the user behaviour after clustering*

To visualize the users and their connections in and out to the clusters, Amir et al. [61] presents the visualization of weblog using Strahler numbering (Fig. 23a).

Strahler number provides quantitative information of the tree shape and complexity by assigning colours to the tree edges. The leaf node has Strahler number one. The node having one or more childs have Strahler number K or K + 1 depending upon the strahler number of their childrens. Herman et al. [25] described how Stracher numbers can be used to provide information about tree structure. The cylinder part of Fig. 23b gives insights of web usage visualization of users in and between the cluster. The Center node of the circular basement is the first page of the web site from which users scatter to different clusters of web pages. The Red spectrum denotes the entry point into clusters and Blue spectrum illustrates the exit point of the users.

Figure 24 indicates that the combination of some visualizations techniques can be used to find new patterns and analyze the patterns obtained through mining algorithms. Through this framework, frequent patterns can be extracted and visually superimposed as a graph of node tubes. The thickness of these tubes indicates aggregated support of navigated sequences. The size of clickable Cube Glyphs [61] indicates the hit of web pages. In Fig. 24b shows the superimposition of Weblogs on top of Web Structure with higher order layout. The top node represents the first page of the website. This hierarchical output of layouts makes analysis easier.

(d) *Understanding Collaborative Customers Buying Patterns*

The rules discovered by pattern discovery phase can be visualized to understand customers buying patterns [62, 63]. Figure 25 presents MineSet's Association Rule Visualizer [62] which maps features of the market basket in x-axis and y-axis. The

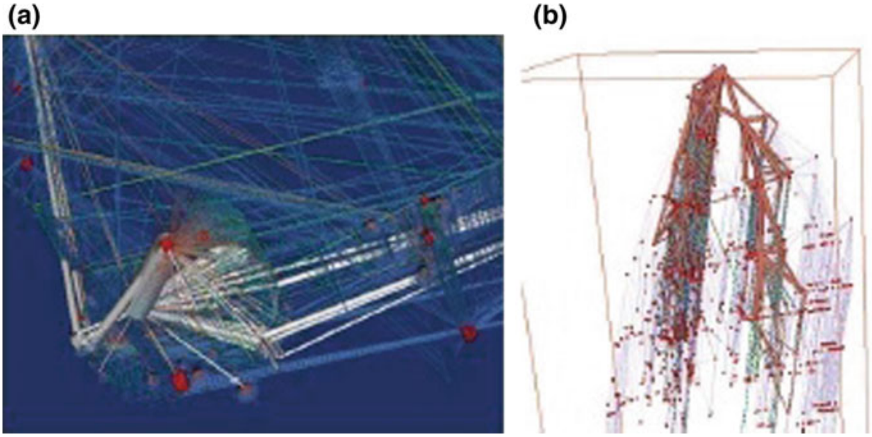


Fig. 24 **a** Frequent access patterns extracted by web mining process using *confidence threshold* are visually superimposed as a graph of node-tubes (as opposed to node-links). **b** Superimposition of web usage on top of web structure with higher order layout [61]

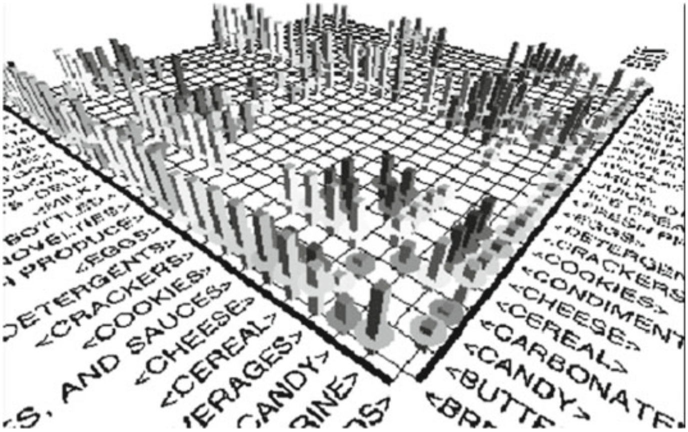
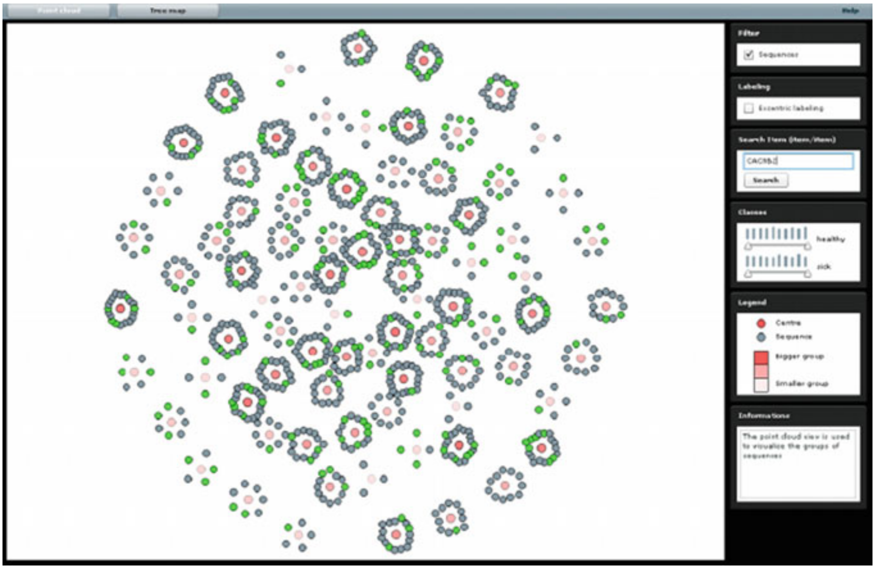


Fig. 25 MineSet's association rule visualizer [62] maps rules to the x- and y-axes. The confidence is shown as the bars and the support as the bars with discs

confidence of association rule is highlighted with bar whereas support as the bar with the disc. The colour on the bar signifies the importance of the rule.

Sequences Viewer [38] helps experts to navigate and analyze the sequences identified by pattern discovery techniques (see Fig. 26). The Point Cloud representation is a sequence viewer which allows users to visualize groups of sequences. Through this view, the centers of the groups, the distance from the centers, and associated sequences can be easily determined.



The point cloud view is used to visualize the groups of sequences

Fig. 26 Point cloud with sequences and highlighted searched items [38]

3 Research Questions

This section highlights research challenges which mainly occurred in web navigation mining. The research challenges presented visualization aspect in web usage mining. Table 1 categorizes visualization aspects into two: (a) Web Usage Mining and (b) Web Structure Mining through navigation files. The solution to these challenges is described in Sect. 2.

Table 1 Visualization aspects in web navigation mining

Visualization aspects from web navigation mining	Visualization aspects from web structural mining through navigation files
(a) How has information been accessed? (b) How frequently? (c) How recently? (d) What is the most and least popular information? (e) Which web page user follows to enter the site? exit? (f) In which region user spend time? (g) How long do they spend there? (h) How do people travel within site? (i) What is the most and least popular path? (j) Who are the users visiting the website?	(a) How many web pages has been added? Modified? Deleted? (b) How is new web page accessed? When does it become popular? When does it die? (c) How does the addition of web page change the navigation patterns within the site? Can people still navigate to desired web page? (d) Do people look for deleted web page? How relevant is the deleted web page?

4 Visualization Tools

The tools used for Visualization has been described in this section. There is a wide range of data mining tools available which presents different visualization aspects. Following Tables 2, 3 and 4 presented visualization tools in data pre-processing, pattern discovery and pattern analysis.

Table 2 Data Pre-processing tools

Tools	Features
Data preparator [64]	It is free software used for the common task in pattern mining. It performs data filtering, pattern discovery and data transformation
Sumatra TT [65]	It is a platform independent data transformation tool
Lisp miner [66]	Analyzing the navigations and preforms pre-processing
Speed tracer [67]	It mines web log files, visualizes and identify the sessions to reconstruct user navigational sessions

Table 3 Pattern discovery tools

Tools	Features
Sewebar- Cms [68]	Selected rules among large set rules discovered from association rule mining. It provides interaction between data analyst and domain expert to discover patterns
i-Miner [69]	Discover cluster through the fuzzy clustering algorithm which is further used for pattern discovery and analysis
Argunaut [70]	Develop interesting patterns by using the sequence of various rules
MiDas (Mining Internet Data for Associative Sequences) [71]	It is an extension of traditional sequential techniques which adds more features to the existing techniques. It discovers market-based navigation patterns from log files

Table 4 Pattern analysis tools

Tools	Features
Webalizer [72]	Discover web pages after analyzing the patterns
Naviz [73]	Visualization tool which combines the 2-D graph of user navigation patterns. It clusters the related pages and describes the pattern of user navigation on the web
WebViz [74]	Analyze the patterns and provides a graphical form of the patterns
Web miner [75]	Analyse useful patterns and provides the user specific information
Stratdyn [76]	Enhances WUM and provides patterns visualization

5 Conclusion

Visualization can be performed to analyze Web Navigations with a different perspective. The chapter presents the framework for having various components of web navigation mining. The components are weblogs consisting of users navigation details, Pre-processing and Cleaning and Pattern Recognition. İn addition, two components are depicting the visualization of web data after preprocessing and after identification of patterns. The visualization can be performed to investigate the website structure, evolution of the website, user navigation behavior, frequent and rare patterns, outlier or anomaly detection, etc. It also determines and analyses the interesting patterns from the clusters or rules. It validates the quality of patterns discovered. Since the web is the largest repository of information which makes users difficult to find the desired information. Visualization is used to provide a summarised view of large data pictorially. Visualization can be used to analyze and discover hidden information from large data. The application of visualization and its metrics varies according to the requirements. The chapter discusses various visualization aspects in the area of web navigation mining. In the end, some tools are listed which makes analysis simpler and easier.

References

1. Flavián C et al (2016) Web design: a key factor for the website success. J Syst Inf Technol
2. Rahi P et al (2012) Business intelligence: a rapidly growing option through web mining. IOSR J Comput Eng (IOSRJCE) 6(1):22–29
3. Jindal H et al (2018) PowKMeans: a hybrid approach for gray sheep users detection and their recommendations. Int J Inf Technol Web Eng 13(2):56–69
4. Kazienko P et al (2004) Personalized web advertising method. In: International conference on adaptive hypermedia and adaptive web-based systems, pp 146–155
5. Ketukumar B, Patel AR, Web data mining in e-commerce study, analysis, issues and improving business decision making. PhD Thesis, Hemchandracharya North Gujarat University, Patan, India
6. McKinney W (2018) Python for data analysis data wrangling with pandas, NumPy, and IPython. O'Reilly
7. https://www.loganalyzer.net/log-analysis-tutorial/what-is-log-file.html, Accessed 13 Feb 2019
8. https://www.ibm.com/support/knowledgecenter/en/ssw_ibm_i_71/rzaie/rzaielogformat.htm, Accessed 14 Feb 2019
9. https://en.wikipedia.org/wiki/Common_Log_Format, Accessed on 14 Feb 2019
10. http://www.herongyang.com/Windows/Web-Log-File-IIS-Apache-Sample.html, Accessed 13 Feb 2019
11. Masand B, Spiliopoulou M (eds) Advances in web usage analysis and user profiling. In: LNAI 1836. Springer, Berlin, Germany, pp 163–182
12. Catledge L, Pitkow J (1995) Characterizing browsing behaviors on the world wide web. Comput Netw ISDN Syst 26:1065–1073
13. Cooley R, Mobasher B, Srivastava J (1999) Data preparation for mining world wide web browsing patterns. J. of Knowl Inf Syst 1, 5–32
14. Cooley R, Tan P, Srivastava J (2000) Discovery of interesting usage patterns from web data

15. Myra S, Bamshad M, Bettina B, Miki N (2003) A framework for the evaluation of session reconstruction heuristics in web-usage analysis. Inf J Comput 15(2):171–190
16. Murat AB, Ismail HT, Murat D, Ahmet C (2012) Discovering better navigation sequences for the session construction problem. Data Knowl Eng 73:58–72
17. Dell RF, Roman PE, Valaquez JD (2008) Web user session reconstruction using integer programming. In: IEEE/WIC/ACM international conference on web intelligence and intelligence agent technology, vol 1, pp 385–388
18. Dell RF, Roman PE, Valaquez JD (2009) Fast combinatorial algorithm for web user session reconstruction. IFIP
19. Dell RF, Roman PE, Valaquez JD (2009) Web user session reconstruction with back button browsing. In: Knowledge-based and intelligence information and engineering systems, of Lecture notes in Computer Science, vol 5711. Springer, Berlin, pp 326–332
20. Jindal H, Sardana N (2018) Elimination of backward browsing using decomposition and compression for efficient navigation prediction. Int J Web Based Commun (IJWBC) 14(2)
21. Zhang J, Ghorbani AA (2004) The reconstruction of user sessions from a server log using improved time-oriented heuristics. In: Annual conference on communication networks and services research, pp 315–322
22. Chen Z et al, Linear time algorithms for finding maximal forward references. In: Information technology: coding and computing [Computers and Communications] proceedings. ITCC, pp 160–164
23. Leung CK-S, FpVAT: a visual analytic tool for supporting frequent pattern mining. SIGKDD Explor 11(2)
24. Dimitrov D et al (2016) Visual positions of links and clicks on wikipedia. WWW, ACM
25. Herman I, Melançon G (2000) Graph visualization and navigation in information visualization: a survey. IEEE Trans Vis Comput Graph 6
26. Eades P (1992) Drawing free trees. Bull Inst Comb Appl 10–36
27. Chi EH, Pitkow J, Mackinlay J, Pirolli P, Gossweiler R, Card SK (1998) Visualizing the evolution of web ecologies. In: Proceeding of CHI
28. Chen J, Zheng T, Thorne W, Zaiane OR, Goebel R, Visual data mining of web navigational data
29. Oosthuizen C et al (2006) Visual web mining of organizational web sites. In: Proceedings of the information visualization
30. Carrière J, Kazman R (1995) Research report: interacting with huge hierarchies: beyond cone trees. In: Proceedings of the IEEE conference on information visualization '95. IEEE CS Press, pp 74–81
31. Melançon G, Herman I (1998) Circular drawings of rooted trees. Reports of the Centre for Mathematics and Computer Sciences, Report number INS–9817. http://www.cwi.nl/InfoVisu/papers/circular.pdf
32. Yuntao J (2007) Drawing trees: how many circles to use?
33. Chen J et. al (2004) Visualizing and discovering web navigational patterns. In: Seventh international workshop on the web and databases
34. Mark M, What's in a session: tracking individual behavior on the web. ACM
35. Catledge L, Pitkow J (1995) Characterizing browsing strategies in the world-wide web. Comput Netw ISDN Syst 27:1065–1073
36. Lindén M (2016) Path analysis of online users using clickstream data: case online magazine website
37. Kenett DY et al (2014) Partial correlation analysis. applications for financial markets. Quant Financ
38. Sallaberry A et. al (2011) Sequential patterns mining and gene sequence visualization to discover novelty from microarray data. J Biomed Inform 44, 760–774
39. Hu C et al (2003) World wide web usage mining systems and technologies. J Syst Cybern Inform 1(4):53–59
40. Pani SK et al (2011) Web usage mining: a survey on pattern extraction from web logs. Int J Instrum Control Autom 1(1):15–23

41. Singh B, Singh HK (2010) Web data mining research: a survey. In: IEEE international conference on computational intelligence and computing research (ICCIC), pp 1–10
42. Facca FM, Lanzi PL (2005) Mining interesting knowledge from weblogs: a survey. Data Knowl Eng 53(3):225–241
43. Jindal H, Sardana N (2007) Empirical analysis of web navigation prediction techniques. J Cases Inf Technol (IGI Global) 19(1)
44. Tseng VS, Lin KW, Chang J-C (2008) Prediction of user navigation patterns by mining the temporal web usage evolution. Soft Comput 12(2), 157–163
45. Chordia BS, Adhiya KP (2011) Grouping web access sequences using sequence alignment method. Indian J Comput Sci Eng (IJCSE) 2(3):308–314
46. Abrishami S, Naghibzadeh M, Jalali M (2012) Web page recommendation based on semantic web usage mining. Social informatics. Springer, Berlin, pp 393–405
47. Thwe P (2014) Web page access prediction based on integrated approach. Int J Comput Sci Bus Inform 12(1):55–64
48. Papoulis A (1991) Probability, random variables, and stochastic processes. McGraw-Hill, USA
49. Bhawna N, Suresh J (2010) Generating a new model for predicting the next accessed web pages in web usage mining. In: IEEE international conference on emerging trends in engineering and technology, pp 485–490
50. Xu L, Zhang W, Chen L (2010) Modelling users visiting behaviours for web load testing by continuous time Markov chain. In: IEEE 7th web information systems and application conference. IEEE, pp 59–64
51. Wang C-T et al (2015) A stack-based Markov model in web page navigability measure. In: International conference on machine learning and cybernetics (ICMLC), vol 5. IEEE, pp 1748–1753
52. Speiser M, Antonini G, Labbi A (2011) Ranking web-based partial orders by significance using a Markov reference mode. In: IEEE international conference on data mining, pp 665–674
53. Dhyani D, Bhowmick SS, Ng W-K (2003) Modelling and predicting a web page accesses using Markov processes. In: Proceedings 14th international workshop on database and expert systems applications. IEEE
54. Dongshan A, Junyi S (2002) A new Markov model for web access prediction. IEEE Comput Sci Eng 4(6):34–39
55. Sunil K, Gupta S, Gupta A (2014) A survey on Markov model. MIT Int J Comput Sci Inf Technol 4(1):29–33
56. Jindal H, Sardana N (2018) Decomposition and compression of backward browsing for efficient session regeneration and navigation prediction. Int J Web Based Commun (Inderscience) 14(2)
57. Vishwakarma S, Lade S, Suman M, Patel D (2013) Web user prediction by: integrating Markov model with different features. Int J Eng Res Sci Technol 2(4):74–83
58. Sampath P, Ramya D (2013) Analysis of web page prediction by Markov model and modified Markov model with association rule mining. Int J Comput Sci Technol
59. Sampath P, Ramya D (2013) Performance analysis of web page prediction with Markov model, association rule mining (ARM) and association rule mining with statistical features (Arm-Sf). IOSR J Comput Eng 8(5):70–74
60. Sampath P, Wahi A, Ramya D (2014) A comparative analysis of Markov model with clustering and association rule mining for better web page prediction. J Theor Appl Inf Technol 63(3):579–582
61. Amir H et. al (2004) Visual web mining. WWW. ACM
62. S. G. Inc. Mineset (2001) Mineset. http://www.sgi.com/software/mineset
63. Hofmann H, Siebes A, Wilhelm A (2000) Visualizing association rules with interactive mosaic plots. In: SIGKDD International conference on knowledge discovery & data mining (KDD 2000), Boston, MA
64. http://www.datapreparator.com/
65. Stěpankova O, Klema J, Miksovsky P (2003) Collaborative data mining with Ramsys and Sumatra TT, prediction of resources for a health farm. In: Mladenić D et al (ed) Data mining and decision support
66. https://lispminer.vse.cz/, Accessed 9 Mar 2019

67. https://code.google.com/archive/p/speedtracer/, Accessed 8 Mar 2019
68. Kliegr T, SEWEBAR-CMS: semantic analytical report authoring for data mining results
69. Abraham A (2003) i-Miner: a web usage mining framework using hierarchical intelligent systems. In: The 12th IEEE international conference on fuzzy systems
70. https://www.researchgate.net/figure/ARGUNAUTs-Moderators-Interface-and-some-of-its-shallow-alerts_fig12_220049800
71. Büchner AG, Navigation pattern discovery from internet data
72. http://www.webalizer.org/
73. https://navizanalytics.com/, Accessed 9 Mar 2019
74. Pitkow JE, Bharat KA, WEBVIZ: a tool for world-wide web access log analysis. In: Proceedings of first international WWW Conference
75. https://www.crypto-webminer.com/
76. Berendt B et al (2001) Visualizing individual differences in Web navigation: STRATDYN, a tool for analyzing navigation patterns. Behav Res Methods Instrum Comput 33(2):243–57