

Interactive Web Usage Mining with the Navigation Visualizer

Eelco Herder

Department of Computer Science
University of Twente, P.O. Box 217
7500 AE Enschede, The Netherlands
herder@cs.utwente.nl

Harald Weinreich

Department of Informatics
University of Hamburg, Vogt-Kölln-Straße 30
D-22527 Hamburg, Germany
weinreich@informatik.uni-hamburg.de

ABSTRACT

Web usage mining, the analysis of user navigation paths through web sites, is a common technique for evaluating site designs or adaptive hypermedia techniques. However, often it is hard to relate aggregated clusters or measures to actual user navigation behavior. By contrast, basic graph-based visualizations of user navigation paths are easier to interpret, but it is difficult to find effective views that convey all the required information. In this paper we present the *Navigation Visualizer*, a web usage analysis tool that combines the two approaches. The Navigation Visualizer makes use of the rich data set that is collected by the *Scone* proxy-based web enhancement framework and facilitates dynamic selection of the data and interactive exploration with various layout mechanisms, color codings and markers. Several aggregated measures can be calculated and exported to statistical and data mining packages.

Author Keywords

User navigation, web sites, access tracking, interactive visualization, web usage mining

ACM Classification Keywords

H5.4. Hypertext/hypermedia: Navigation

INTRODUCTION

The data that can be collected from web usage logs is highly multidimensional, varying from basic statistics – such as the relative amount of back button usage – to more aggregate data, such as the linearity of user navigation paths and semantic similarity of documents to a query [7]. While it is possible to put these data into tables and to analyze the data with statistical software or data mining applications, there is often still the need for visual analysis to enable researchers to interpret the meaning of the available data [3]. Unfortunately, it is hard – if not impossible – to visualize a large number of dimensions of numerous items in a workable manner, as discernability and understandability decrease with the amount of data that is displayed [8]. Several approaches to

reduce the amount of data, such as clustering, binding, filtering and hierarchization of documents or usage patterns, have been applied in various systems [2, 3]. However, the resulting abstract views are often too complex to interpret or to compare. By contrast, traditional graph visualization of the node-and-link structure [e.g. 13] is easy to understand, but badly scalable and it hardly allows visualizing additional metadata, such as page view times, back button usage, or page relevancy. Especially when evaluating various possible effects of adaptive hypermedia techniques or site redesigns, the interpretation of the recorded data has to be supported.

In this paper we present a graph-based web usage visualization tool – the *Navigation Visualizer* – that allows researchers to dynamically select and match the data to be shown and to interactively explore the visualization. Selection criteria for users, web sites and time slices are used to narrow down or to broaden the number of items; various color codings and markers are available to compare several dimensions side by side; zoom and pan tools [2] are used to explore larger graphs. The tool also offers means to display and export data in a tabular format to be further processed by statistical software.

The remainder of this paper is structured as follows: In the next section we present related work on graphical overviews, web usage mining and web usage visualization. Then we describe the design rationale, followed by an overview of the actual system. We continue with a discussion on the implementation: how data is captured, processed, modeled and represented. In particular, the advanced web logging functionality of the *Scone* framework is introduced.

RELATED WORK

Graph drawing is a common technique for information visualization and it has been subject of research since decades [2, 8]. Graphs are a natural means to model the structure of the web, with the pages represented by nodes and the links represented by edges. Many graph algorithms are used, in original or adapted form, to calculate and express properties of web sites and individual pages [1, 7]. Although to a lesser extent, graph theoretic methods have also been applied to the user navigation paths through web sites. Some interesting results include the observation that more

successful users navigate sites in a non-linear manner [10] and that the time spent on a web page is a good indicator of the user's interest in the page [12]. In earlier research we found that more focused patterns of page revisits indicated that users knew about and were better able to use the site structure – and that these patterns can be captured in measures [6, 7].

Several systems for visualizing user navigation have been developed, some of which are mentioned in [13], in which the WebQuilt visualization system is introduced. WebQuilt is interactive in the sense that it provides semantic zooming and filtering, given a storyboard. However, this is only useful for predefined, goal-directed tasks. Cockburn and McKenzie [4] mention various issues related to graphical representations of web browsers' revisit tools. Although these issues originally are raised from the end user's point of view, they are valid for researchers examining user logs as well.

In short, from the related work it can be observed that the data to inspect is manifold and that graphical representations provide limited views on the data. More interactive graphical representations are expected to perform better in providing researchers with means to find the matters that they are interested in.

DESIGN RATIONALE

The idea of an interactive web usage visualization tool arose when we were looking for specific user navigation patterns that indicate users' disorientation while carrying out tasks on various web sites [6]. After having calculated a large number of measures, we wanted to see what these measures really meant. Despite the availability of several web usage visualization tools, no tool was found that allowed for visualization and exploration of the various patterns in web navigation.

Visual analysis usually starts with getting the 'big picture', followed by gradually zooming in to the interesting details [2]. Concretely, this means that it should be possible to define the data set that one is interested in by selecting the users, the web sites and time slices to be displayed. As mentioned before, comprehensibility is increased by limiting the number of items to be displayed.

As mentioned in the preceding section, navigation paths are commonly modeled as directed graphs [1]. These navigation graphs can be regarded as an overlay of the web site's structure, consisting of the pages visited and the links followed. A tree layout [8], with pages ordered from top to bottom by its shortest distance from the start of any session and from left to right by its first occurrence turned out to provide the most stable and comprehensive layout.

The navigation graph should be able to represent various dimensions of the data. Color coding is an obvious candidate for this purpose. As can be observed from the card game Set, it is quite hard to examine three continuous scales at once. Therefore, we opted for a combination of color codings and interactive markers for both nodes and edges. Finally, as it is sometimes easier to notice patterns and outliers by means of statistical analysis, the system should provide means to calculate relevant measures that can be exported to a statistical package or a data mining tool.

SYSTEM OVERVIEW

This section provides an overview of the Navigation Visualizer, followed by an example of how the system is used for data analysis.

Functionality

Figure 1 displays the main screen of the Navigation Visualizer. The largest part of the window is occupied by the navigation graph itself. The upper right part offers various manipulation options:

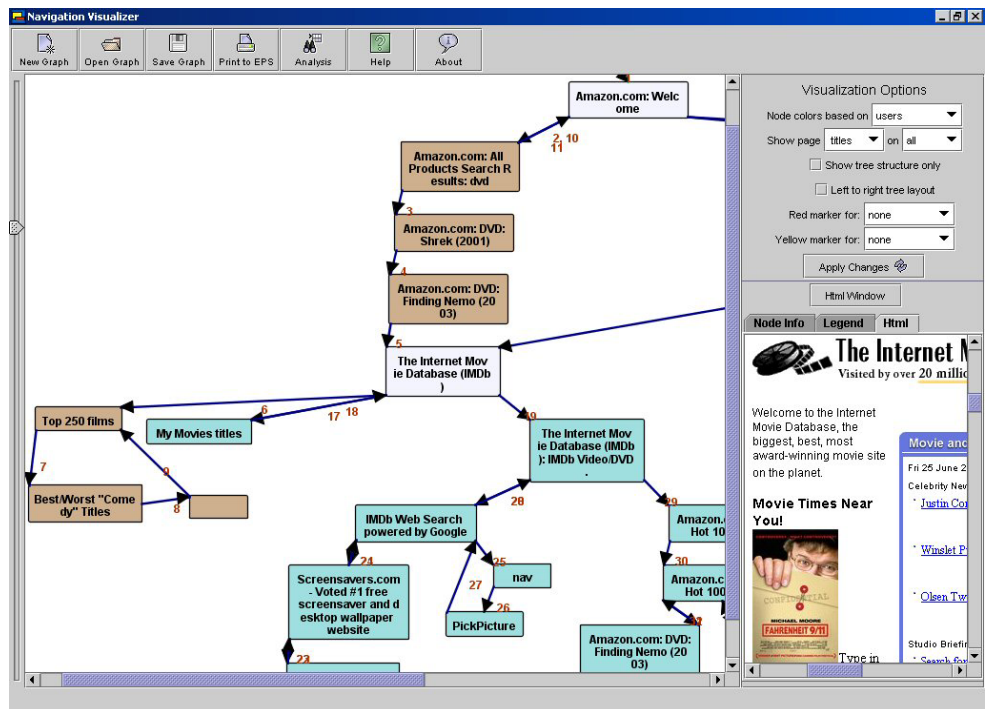


Figure 0: The main window of the Navigation Visualizer

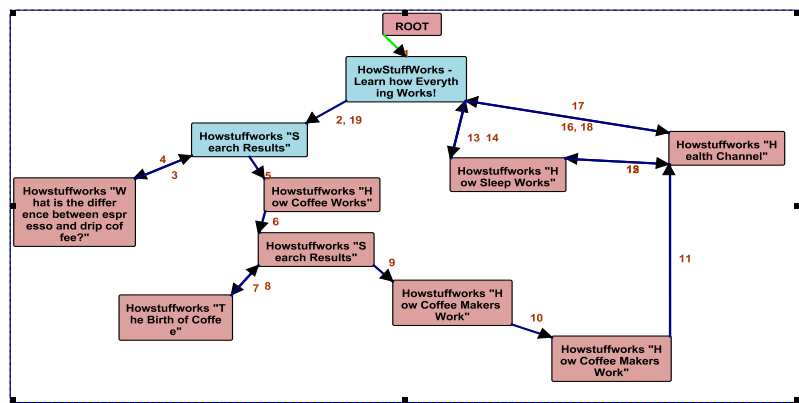
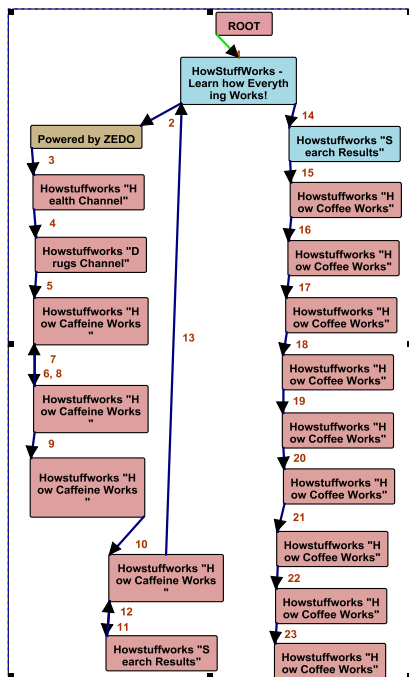


Figure 2. In the picture above the navigation path of a subject working on a task without link suggestions is displayed. The cycle in the right side part of the graph shows how the subject repeatedly arrives at a page visited before and finally gives up.

In the picture on the left the navigation path of a subject working on the same task, but with link suggestions is displayed. After having solved one subtask (left side), she returns to the site's home page for the next subtask.

- *Color coding options:* Switching between colors is a powerful way to distinguish between dimensions such as user characteristics, page access statistics and textual similarity;
- *Marker options:* Selection of markers that are used in addition to the color coding, to facilitate comparison between multiple dimensions;
- *Page identifier options:* Neither the page title, url or thumbnail is sufficient for easily distinguishing between pages [4]. Therefore, we allow researchers to change between these options;
- *Graph layout options:* By default, the spanning tree graph layout is used. Several alternatives to build the tree (e.g. depth-first or breadth-first) can be selected. Alternative graph layout mechanisms, such as force-directed layout [8], are available as well.

The bottom right part provides details on a selected node. This is either a list of descriptive statistics or a preview of the page, which can be displayed in a separate window on request. A separate dialogue allows selection of relevant statistical information on user navigation paths and site characteristics. At the moment, primarily measures on navigation complexity, page view times and patterns of revisits [6, 7] are calculated. When required, other export options can be easily defined.

Example Analysis

In this subsection we illustrate the use of the Navigation Visualizer with an excerpt from a study on the effect of link annotation on user navigation patterns [6]. Thirty-two subjects were asked to carry out five everyday web tasks, such as planning a weekend trip to London. Half of the subjects were provided with previously defined navigation suggestions. We found significant differences in navigation

complexity, use of the back button and home page visits – all measures were lower in the link annotation condition. This could mean that:

- either the subjects were simply following the link suggestions, which would be a negative effect;
- or the subjects were less prone to get stuck and consequently visited the home page – to start a new trial – to a lesser extent.

Visual exploration of the navigation paths of subjects that scored very high or low on the measures suggested a pattern of random, unexpected visits to pages visited before – see figure 2 – which suggested that the latter was the case. This was confirmed by the results of the post-questionnaires.

IMPLEMENTATION

In this section we explain some implementation aspects of the Navigation Visualizer. In the first subsection we briefly describe the framework *Scone* that is used to capture the data. In the second subsection we describe the object-oriented graph model and its visualization.

Capturing the Data with Scone

For successful web usage mining four datasets are required [3]: the contents of each page visited, the link structure between the pages, the web usage recordings and user characteristics, such as demographics and tasks.

In order to avoid common problems associated with processing server logs [5] and to capture additional and more detailed data, we employed and extended *Scone*, our framework for rapid development of web enhancement and evaluation of web tools [14]. *Scone* is a Java-based framework for rapid development of web enhancement and evaluation of web tools. A core element of the framework is IBM's intermediary WBI [9] that provides means to moni-

tor and manipulate any data transferred between web browser and server.

Scone also offers an *access tracking* component that registers the most relevant user actions and records them in a database. This is achieved by adding JavaScript code that captures and submits many browser events, such as a click on a link, a form submission, and the opening and closing of browser windows.

A *robot* component is available for exploring and analyzing web sites. The Navigation Visualizer utilizes the robot to gather pages that have not been visited yet, in order to build a web server's link topology. In user-adaptive systems the robot can also function as a scout gathering information on the user's behalf.

The data that are collected by the user tracking component and the robot are mapped onto an object model called *netObjects*. It represents the pages and links relevant for a user session as well as the users' attributes and their actions. Scone can also automatically store the objects of this model in a relational database in real time.

Modeling and Visualizing Data with the Navigation Visualizer

The advantage of the object-oriented data model is that various restrictions on the users, page characteristics and time slices can be specified for building the *navigation graph*. The graph consists of the pages, which are represented by the nodes, and the links, which are represented by the edges. The page characteristics and page access information are contained in the nodes, link traversal data in the edges.

Each object in the navigation graph has its own visualization object. The graph visualization object keeps track of the aspects that the researcher aims to visualize and instructs the node and edge visualizations to base their appearance on specific attributes in their model counterparts. The strict separation of model and view facilitates extension of the visualization possibilities without having to change the model.

The graph model provides various methods for calculating distances between nodes [1]. Overall graph characteristics that are relevant for navigation path analysis can be calculated as well, among other things its diameter, fan degree, density, stratum and compactness [6].

SUMMARY

In this paper we presented the Navigation Visualizer, a web usage analysis tool that combines the mathematical, high level approach of web usage mining with interactive graph-based visualizations. It facilitates tracing *and* understanding user actions, which would be harder to do with either one of the approaches. Furthermore, the Navigation Visualizer provides means for preprocessing complex user data for further analysis in statistical packages.

ACKNOWLEDGMENTS

The first author's work is carried out in the context of the PALS project, which is sponsored by the Dutch Innovative Research Program IOP-MMI.

REFERENCES

1. Baldi, P., Frasconi, P. and Smith, P. Modeling the Internet and the Web: Probabilistic Methods and Algorithms. Wiley, ISBN 0-470-84906-1, (2003).
2. Chen, C. Information Visualisation and Virtual Environments. Springer-Verlag, ISBN 1-85233-136-4, (1999).
3. Chi, E.H. Improving Web Usability through Visualization. IEEE Internet Computing 6 (2), (2002), 64-71.
4. Cockburn, A. and McKenzie, B. What Do Web Users Do? An Empirical Analysis of Web Use. Intl. J. Human-Computer Studies 54 (6), (2000), 903-922.
5. Cooley, R. The use of Web structure and content to identify subjectively interesting Web usage patterns. ACM Trans. Internet Technology 3 (2), (2003), 93-116.
6. Herder, E. and Juvina, I. Discovery of Individual Navigation Styles. Proc. of Workshop on Individual Differences in Adaptive Hypermedia at Adaptive Hypermedia 2004 (2004).
7. Herder, E. and Van Dijk, B. Site Structure and User Navigation: Models, Measures and Methods. In Adaptable and Adaptive Hypermedia Systems, edited by S.Y. Chen and G.D. Magoulas, (2004), 19-34.
8. Herman, I., Melançon, G. and Marshall, M.S. Graph Visualization and Navigation in Information Visualization: a Survey. IEEE Trans. Visualization and Computer Graphics 6 (1), (2000), 24-43.
9. Maglio, P. and Barrett, R. Intermediaries personalize information streams. Communications of the ACM 43 (8), (2000), 96-101.
10. McEneaney, J.E. Visualizing and Assessing Navigation in Hypertext. Proc. Hypertext '99, (1999), 61-70.
11. Mukherjea, S. and Foley, J.D. Visualizing the World-Wide Web with the Navigational View Builder. Proc. 3rd Intl. WWW Conference on Technology, Tools and Applications (1995), 1075-1087.
12. Shahabi, C., Zarkesh, A.M., Adibi, J. and Shah, V. Knowledge Discovery from Users Web-Page Navigation. Proc. 7th Intl. Workshop on Research Issues in Data Engineering RIDE '97 (1997), 20-29.
13. Waterson, S.J., Hong, J.I., Sohn, T. and Landay, J.A. What Did They Do? Understanding Clickstreams with the WebQuilt Visualization System. Proc. Advanced Visual Interfaces (2002).
14. Weinreich, H., Buchmann, V., Lamersdorf, W. Scone: Ein Framework zur evaluativen Realisierung von Erweiterungen des Webs. Proc. KiVS '03, (2003), 31-42. Project website: <http://www.scone.de/>.