WebKIV: Visualizing Structure and Navigation for Web Mining Applications

Yonghe Niu, Tong Zheng, Jiyang Chen, Randy Goebel
Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada T6G 2E8
{yonghe, tongz, jiyang, goebel}@cs.ualberta.ca

Abstract

A significant part of the web mining problem is simply in understanding the value of any mining method. For example, the value of web mining to improve user navigation is even more challenging if one can't visualize the differences over a large collection of web pages or a significant structure within the existing web.

We present WebKIV, a tool we've developed to help us visualize our own results in web mining. WebKIV combines strategies from several other web visualization tools, to provide a single method of visualizing web structure, and the results of web mining on that structure.

We summarize the value of web visualization tools along the dimensions of scale (can one visualize small and large structures), navigation dynamics (can one visualize navigation dynamically or statically), and cumulative usage (can one distinguish individual and aggregate web usage). We then show how WebKIV provides a way of visualizing the results of web mining in a way that distinguishes properties along all three of these dimensions.

Keywords: visualization, web mining, web structure, web navigation

1 Introduction

Several problems stand between the interpretation and application of web mining results. First, results are often hard to understand [5], so considerable effort must be directed at interpreting them. Second, the volume of potentially useful mining results can be huge, so one has to understand how to identify useful patterns. Third, applying and evaluating results is not trivial [10, 9].

All of these problems can benefit from visualization tools, and several such tools have been developed to address them. For example, Mosaic Plot [5] is designed to visualize association rules; webCANVAS [2] is used to visualize user navigation patterns. However, most existing visualization tools visualize web site structure, web content and web navigation separately [3]. We have found that it is extremely important to simultaneously visualize web structure, web content, and user navigation paths.

Here we present a Web Knowledge and Information Visualization tool (WebKIV) designed to experiment with this simultaneous visualization of structure, content, and navigation. WebKIV combines several visualization strategies from other web visualization tools, to provide a single method of visualizing web structure, and the results of web mining on that structure. Specifically, WebKIV provides the following functionalities:

1. Web structure visualization. WebKIV provides tools for visualizing small and large web structures, with controls that support the display of both detailed and abstract structure.

2. Web navigation visualization. WebKIV provides static and dynamic display of both individual and aggregate user navigation patterns.

3. Web mining results comparison. To support our primary web mining research, WebKIV provides a way of overlaying web navigation patterns, and comparing those constructed from the application of machine learning to navigation improvement.

The rest of our paper is organized as follows. Section 2 reviews the related visualization work, and provides a simple framework for comparing the features of a number of web visualization tools. Section 3 introduces the WebKIV architecture, and summarizes the techniques used to achieve the functionality required to support our web mining research. We demonstrate the use of WebKIV in several different contexts in Section 4, and then summarize the results in Section 5.

Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03)
0-7695-1932-6/03 $17.00 © 2003 IEEE

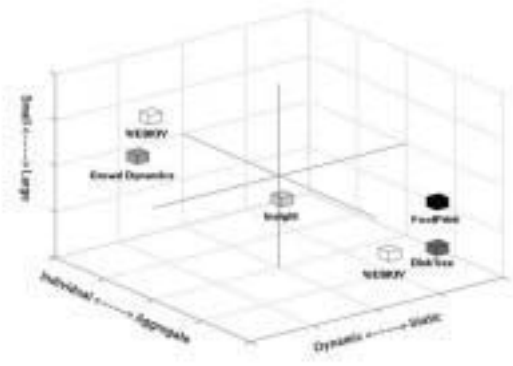


Figure 1. Web Visualization Dimensions

2 Dimensions of Web Visualization

Most of the tools we have assessed have been designed to excel in some dimension of web visualization. For example, Minar's visualization of the dynamics of a collection of web users provides a nice animation of a small number of users [12]. But Minar's visualization doesn't provide much in the way of web structure visualization, and its representation doesn't scale to large web structures or large numbers of users.

With this kind of informal assessment, we have found it useful to assess web visualization tools along three major dimensions, as indicated in Figure 1. Those three include navigation behavior (static to dynamic), cumulative usage (individual to aggregate), and relative size of web structures (small to large).

As we stated above, we know of no visualization tool that fulfills all requirements. However, it is easy to see how various tools fit into the dimensions of Figure 1. In a sense, Figure 1 is itself a crude visual summary of the tools we have assessed.

For example, Site Manager [13] focuses on visualization scalability, and so is further along the scale axis to "large," than Minar's animation. Similarly, Inxight [8] is developed to solve the "lost in hyperspace" problem [7], and so provides better usability for large web structures by providing a tool to move around in large structures. In fact, most web visualization tools focus on either visualizing web site structures (e.g., [8]) or web usage data visualization (e.g., [17, 1, 3, 11, 4, 14]).

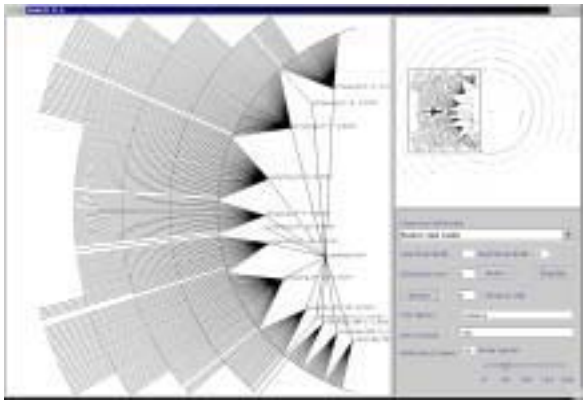
Minar's Crowd Dynamics [12] treats a web site as a inhabited, social, active space. The author implements an animation visualization system to illustrate how the web users walk, loiter in the web, web site popularity, which paths the users like to use, etc. In this case, Minar's system is further along the "static to dynamic" scale than most of the web structure visualization tools, who provide only static displays of structure.

Crowd dynamics animation is simple but information-rich. Most other visualization systems show user activities in a static way, but [12] is like a documentary movie and replays the real-time user behavior. While the system is informative, it doesn't scale well along the "small to large" dimension. First, the web site structure is hand made, and web site specific; to automatically generate a web site map is a major challenge. For example, one needs to know how to group web pages and how to place them. Second, the speed of animation in a big web site map is critical to the success of the animation. People can assess animation as continuous if the frame rate is less than 200ms [13], so very careful design is required for a large web site's animation. Finally, rather than showing "what web users are doing," the system would be better if users' activities were recorded in an aggregate way. For example, to show "what web users have done" is potentially more revealing.

Back to our crude visualization of Figure 1, it is not easy to argue whether crowd dynamics portrays individual or aggregate behavior, as it depends on the perception of the user (try to watch all users move around at once, versus tracking just one). But web usage systems (e.g., [3]) can use other more cognitively manageable techniques to indicate aggregate usage patterns (e.g., like thicker navigation traversal lines or colored lines).

Another visualization tool, FootPrint [16, 15], focuses more on the value of the user experience in a web site. FootPrint applies a hyperbolic tree technique to visualize user traversal patterns. The tree shows well-used hyperlinks and the web pages that were frequently visited by the previous users. The concept of the "FootPrint" is based on the idea that the interaction history with the web site and the work done by previous users can help future users navigate in a complex web site. In this way, the individual surfing activity is correlated to that of other web users. The visiting patterns (foot prints) of the previous users are accumulated and saved for use by new web surfers. Although FootPrint is good in helping solve the lost in "hyperspace" problem [7] without showing the whole web site structure, it is difficult to use for pattern observation. With respect to Figure 1, this system moves along the cumulative usage dimension as it visualizes aggregate behavior, and does it in a relatively large space. However, it does not provide any animation feature to get a sense of usage over time.

The third visualization tool, Disktree [3], which was built by Chi et al. at Xerox, has been used to visualize web site evolution, web usage trends over time and evaluation of information foraging. This tool fares pretty well in the overall placement with the visualization dimension space, for it does provide insight into large structures, and does provide some static images of aggregate navigation behav-



Left: Focus View Window Right: Context Window

Figure 2. WebKIV — System Overview

ior. In [3], several radial trees constructed at different time were aligned together. This provides a kind of static time series evaluation, which moves away from the "static" side of the "static to dynamic" dimension. Similarly, as a web site structure changes over time (e.g., addition and deletion of web pages), a web surfer's favorites are displayed together. This helps reveal the interaction between the web site structure and web surfing patterns.

This brief discussion of where existing tools lie in our three dimensions is still pretty crude, but it does provide an abstract visualization of where the tools sit, and motivates the design of our WebKIV tool.

3 WebKIV System

The WebKIV system was designed as a visualization tool which provides the essential functionality to be able to visualize in the three dimensions (individual vs. aggregate, static vs. dynamic, small scale vs. large scale) discussed in Section 2.

In the WebKIV system, a radial tree algorithm [3] is used to construct the web site structure in a 2D plane. The reason for using two dimensions instead of three dimensions is to avoid the occlusion problem [6]. Zooming and panning techniques are used to view the detailed information when necessary. In order to alleviate the context switching problem inherited from the zooming and panning techniques, WebKIV provides two display windows: a *Context window* is used for the context overview, and a *Focus window* is used for drilling down. Moreover, a rectangle in the Context window shows which part of the web site is displayed in detail in the Focus window (see Figure 2).

Figure 3. WebKIV — Surfing Animation and Usage Aggregation

3.1 WebKIV Functions

The current version of WebKIV can be scaled up to visualize 70,000 nodes. In addition to the web structure, WebKIV can be used to visualize web usage data. It not only displays the static view of web usage data, but also employs various animation techniques to replay real-time user behaviors. Distinct from Crowd dynamics, WebKIV not only shows the individual web user visiting patterns, but also records the history of multiple users' activities. In addition to showing "What people are doing," it can also display "What web users have done."

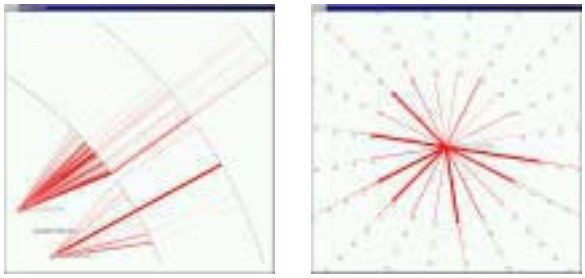
WebKIV has three major functions described as follows:

Web Surfing Animation and Usage Aggregation

An example of the Web surfing animation is shown in a static snapshot in Figure 3. Each black dot represents an individual surfer in the web site. A dot moving from one place to another indicates the traversal path the surfer followed. A dot stopping at a tree node means the surfer is viewing the specific web page, and the time a dot stays at a node indicates the time the surfer spends on that web page. And a dot will disappear if the user has stopped for a certain amount of time (base on the session cut-off time). In this case, WebKIV maintains a count for each hyperlink traversal by all users. A grey scaled line is drawn for each hyperlink based on this count. The more a hyperlink is traversed, the darker its corresponding line will be.

Web Mining Evaluation

Though web mining results can be visualized directly (e.g., visualizing association rules with Mosaic Plot [5], or visualizing user navigation patterns with webCANVAS [2]), they are handled differently by WebKIV. Instead of trying to visualize the meaning of web mining results, WebKIV aims at visualizing their effects in improving certain applications.



(a) Rectangled Zoom-in (b) Subtree Zoom-in

Figure 4. Rectangled Zoom-in and Subtree Zoom-in

For example, we can apply web mining techniques to the web log data, and then use the mining results (e.g., association rules) to improve user navigation by creating and using Navigation Compression Models [18]. The value of this process can then be visualized in WebKIV by first generating compressed web log data assuming the mining results were applied, and comparing it with the original web log data (see Figure 7 in Section 4).

Subtree Zoom-in

In addition to the simple rectangled zoom-in function, subtree zoom-in is another operator used to view and compare the data in detail. With subtree zoom-in, WebKIV can generate a new radial graph centered with the root of the subtree (Figure 4 (b)). Subtree zoom-in is really useful when the data (patterns) need to be observed in detail and are isolated in part of the hierarchy. In comparison with rectangled zoom-in (Figure 4 (a)), subtree zoom-in is more efficient in using the precious display space. However, the problem with subtree zoom-in is that it completely changes the geographical patterns, so users might feel confused.

3.2 WebKIV architecture

Our WebKIV system consists of four basic components, as shown in Figure 5. Based on this architecture, the process of the visualization can be described as follows:

1. *Data Collection*. Three different kinds of data are used in our WebKIV system. The web site topology data is collected by a web crawler using a breadth-first traversal algorithm started at a given root node. Web log data is retrieved directly from server log files. Web mining results can be obtained from various web mining techniques applied to various web data (e.g., web log, web content, etc.).

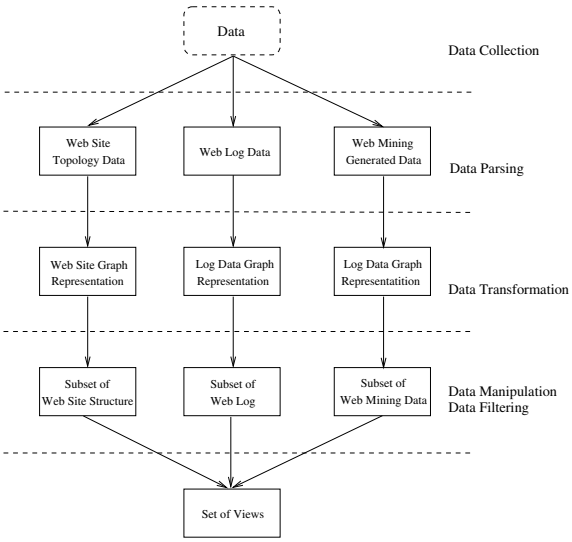


Figure 5. WebKIV — System Architecture

2. *Data Parsing*. In this phase, a radial tree algorithm [3] is applied to the web site structure data to generate the graph representation of the web site. The radial tree is a hierarchical acyclic tree, where each level of the hierarchy is represented by a concentric ring of page nodes distributed around a central focus node. Typically the central node is the homepage of a web site.

3. *Data Transformation*. There are three different shapes used to represent web data attributes in WebKIV:

   node representation Each web page is represented by a square. The size and position of the square is a function of the distribution of page nodes in the radial tree. The square has no color representation because it is often too small to make a difference to the user.

   line representation Each stroke has two attributes: width and color. Depending on user parameters, strokes can indicate aggregate or individual user traversals, and can be colored to distinguish individuals or volume of traffic.

   dot representation A dot is used to represent a web surfer in the web surfing animation function. The dots will move around during the animation process, which indicates that the specific user is moving from page to page through hyperlinks.

4. *Data Manipulation and Data Filtering*. There are a number of data operators in our WebKIV system which, as a whole, provide the flexibility for manipulating the web data and observing different data attributes from different perspectives, e.g., tree rotation,

zooming and panning, subtree generation, threshold setting, web usage data comparison, web usage animation, etc.

4 Examples of WebKIV Visualizations

Here we illustrate some of the WebKIV functions based on the experiments we conducted on a music web site: *http://machines.hyperreal.org*. This web site introduces musical machines such as synthesizers, drum machines and recording equipment, etc. The access logs are obtained from *http://www.cs.washington.edu/ai/adaptive-data/*.

Web Surfing Animation and Usage Aggregation

Web surfing animation snapshots are shown in Figure 6. The data is obtained from Jan 1 1999, starting from midnight. The frame refresh rate is set to 10 frames per second, and the animation rate is 50 times faster than the real time, i.e., one animation second equals 50 seconds in real time.

Figure 6 (a) shows a static view of web users' activities at midnight, when there were several surfers browsing the web site, and there was no obvious traversal patterns available. The web site started becoming busier in the morning, with around 20 surfers wondering at the web site. Several heavily traversed hyperlinks were starting to "stand out", as shown in Figure 6 (b). The most crowded time is around evening, while around 30 people visiting the web site, and the "hot" web pages and "heavily" used hyperlinks are clearly revealed as shown in Figure 6 (c). Figure 6 (d) shows the result of one-day trace history the web users left.

Compared with static web surfing visualization tools, WebKIV not only shows the web access results, but also shows when and how the web site are being used over time.

Web Mining Evaluation

As previously discussed, Navigational Compression Models (NCMs) [18] can be applied to improve user navigation. By applying NCM knowledge to the web data, surfers are led to information using potentially shorter traversal paths. Different parameter settings of NCMs can change user traversal patterns drastically, so one way to evaluate the web mining results is to use the WebKIV's "subtraction" operator to emphasize the difference between the NCM generated web logs and the original web access log, as shown in Figure 7, in which the red lines (pointed to by arrows) represent the NCM suggested links and blue lines (not pointed to by arrows) are "saved" links.

In this example, two experiments were conducted to compare the effectiveness of NCMs, by showing the difference between the web logs before and after applying NCMs, as shown in Figure 7 (a) and Figure 7 (b) respectively. In Figure 7 (a), the 100 most visited web pages are treated

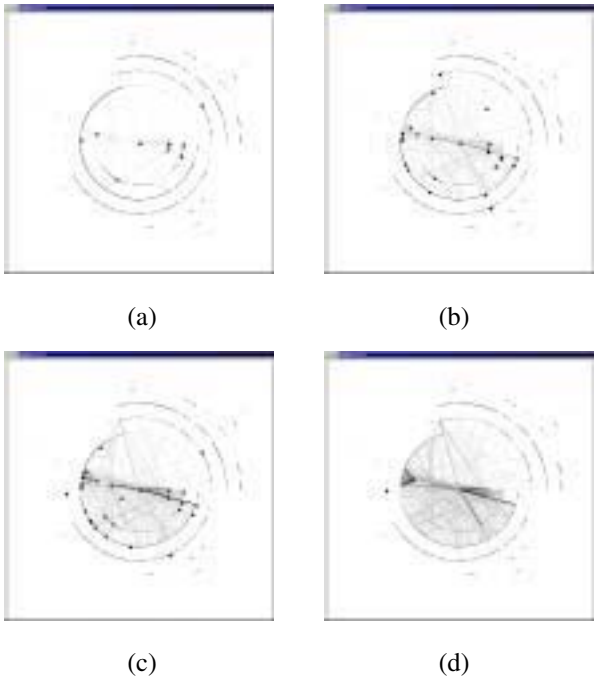
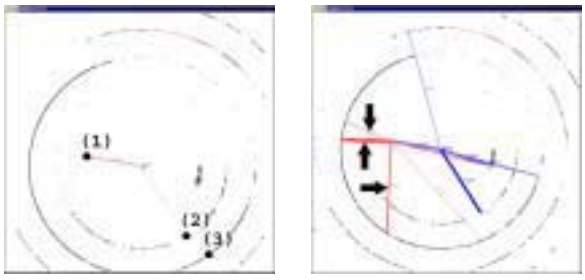


Figure 6. Music Machine Web — Surfing Animation

as content pages in NCMs. Most of the red lines representing the more used hyperlinks are related to the page *http://machines.hyperreal.org/manufacture.html*. Because this page is visited most, the NCMs treat this web page as the content page which surfers want to look at. But this page is actually a pass-through ("hub") web page. The outer circles contain the real content of the music machine website, and where the user might be suggested to go, but the graphs indicate that the outer hyperlinks (the blue lines representing the saved paths) are recommended by the NCMs as non-content pages and thus are "saved." For example, at the right bottom of Figure 7 (a), the web page */Roland/Overview* and */Roland/TR-33* are the content pages, but the NCM thinks these two pages are not "important" to users, thus these two pages are not recommended to the users.

In Figure 7 (b), the content page setting criteria were changed. Not only the visit count of a web page must be considered, but also the time spent on the web page. A web page needs to fulfill the following requirements to be considered as a content page: it was visited more than 200 times, and each time the user spent at least one minute on it. As can be seen in the left part of Figure 7 (b), several suggestions (red lines; pointed by arrows) are correctly made to direct the web users to the content pages. The web page *http://machines.hyperreal.org/manufacture.html*, however, is still treated as a content page and recommended



(a) (b)

(1) *manufacture.html*
(2) */Roland/Overview*
(3) */Roland/TR-33*

Figure 7. Music Machine Web — NCM Comparison

by NCM although we know it is a pass-through page. The reason might be that this page has more than 50 links, so users may need more than one minute to finish browsing this page.

Of course the point here is that by displaying data generated by different NCMs as well as the data generated by the same NCMs with the different parameters, the visualization system can support comparison of the effectiveness of the web mining results and provide some insights into domain-specific adjustments. For example, to decide the content pages, one minute to finish browsing a web page and counts of visits are better parameters to determine content pages rather than using counts of visits only for the Music Machine web site.

Subtree Zoom-in

Some of the patterns are not easily identified without viewing the data from different perspectives. One advantage of subtree zoom-in is to display the different data attributes side by side. In the example of Figure 8, we are looking to answer the questions: (1) "What are the web pages the users visited most?", (2) "Which hyperlinks were used most?" and (3) "From where did the users reach those web pages?"

Note that this functionality also provides one aspect of individual and aggregate user navigation behavior, in a static fashion.

For example, in the left part of Figure 8 (a), there are four web pages that were well visited (see Figure 8 (b) for an enlarged view). However, it is hard to tell which route the surfers followed to reach those four web pages merely from Figure 8 (a) itself. For this question, a hyperlink usage view can be created by right-clicking the root node in the WebKIV, as shown in Figure 8 (c). We then found that users were not browsing via the hierarchical web site structure. Note in Figure 8 (c) that the hyperlink from *manufacture.html* to *samples.html* is scarcely used (31 times; since it is below the threshold, it is not shown in the view). Given that the *samples.html* page is heavily visited (660 times) and that no link to this page within the web site contributes significantly to the number of visits, the implication is that either the web page *samples.html* is bookmarked by the users or it is linked by other related web sites.

But without viewing the web page usage view as well as the hyperlink usage view, it is almost impossible to find the answer for question (3).

5 Summary and Conclusions

Web visualization is a complex task, but one can begin to understand the requirements by decomposing the task into the three dimensions of scale of web structure visualization, representation of aggregate and individual navigation usage behavior, and the comparative display of navigation improvement methods.

While many visualization tools address one or two of these aspects, they don't always provide the complete suite of tools to both visualize a large web structure, and support question answering for navigation improvement research.

Our WebKIV tool is a first approximation to combining what we believe are necessary components of such a tool, and we have provide a brief sketch of its functionality, with glimpses of how to deploy the tool for both structure understanding and the interpretation of web navigation mining.

Acknowledgments

Our work is supported by the Canadian Natural Sciences and Engineering Research Council (NSERC), and by the Alberta Ingenuity Centre for Machine Learning.

Thanks to *http://machines.hyperreal.org* for providing the log data used in this paper showing examples of WebKIV visualization.

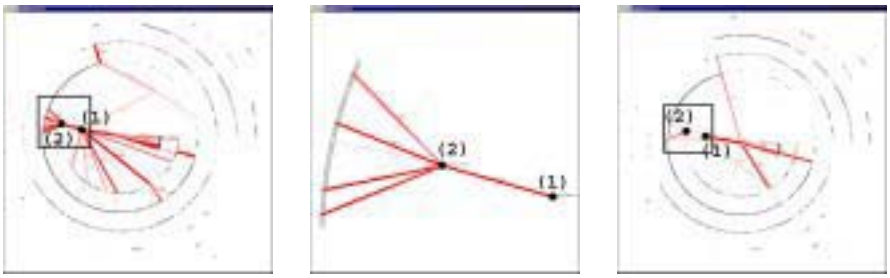
References

[1] http://www.mercuryinteractive.com.

[2] I. Cadez, D. Heckerman, and C. Meek. Visualization of navigation patterns on a web site using model_based clustering. In *Proceedings of KDD 2000*, pages 280–284, August 2000.

[3] E. Chi, J. Pitkow, J. Mackinlay, P. Pirolli, and J. Konstan. Visualizing the evolution of web ecologies. In *Proceedings of CHI'98*, pages 400–407, 644–645, 1998.

[4] R. Hendley, N. S. Drew, A. M. Wood, and R. Beale. Narcissus: Visualizing information. In *Proceedings of IEEE Information Visualization Symposium*, pages 90–97, Los Alamitos, October 1995.



(a) (b) (c)

(1) *manufacture.html* (2) *samples.html*

Figure 8. Music Machine Web — Visits and Traversal Distribution

[5] H. Hofmann, A. Siebes, and A. Wilhelm. Visualizing association rules with interactive mosaic plots. In *the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 227–235, August 2000.

[6] E. H. hsin Chi and S. K. Card. Sensemaking of evolving web sites using visualization spreadsheets. In *Proceedings of the Symposium on Information Visualization*, pages 18–25, San Francisco, California, 1999.

[7] N. Jokob. The art of navigating through hypertext. *Communications of the ACM*, 33(3):296–310, 1990.

[8] J. Lamping, R. Rao, and P. Pirolli. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 401–408, New York, May 1995.

[9] B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *International Conference on Knowledge Discovery and Data Ming*, pages 125–134, San Diego, CA, August 1999.

[10] Y. Ma, B. Liu, and C. K. Wong. Web for data mining: Organizing and interpreting the discovered rules using the web. *SIGKDD Explorations*, 2(1):16–23, July 2000.

[11] Y. S. Maarek, M. Jacove, M. Shtalhaim, S. Ur, and D. Zernik. Webcutter: A system for dynamic and tailorable site mapping. In *Proceedings of WWW6 Conference*, pages 713–722, Santa Clara, California, April 1997.

[12] N. Minar and J. Donath. Visualizing the crowds at a web site. In *Proceedings of CHI'99*, 1999.

[13] T. Munzner. Drawing large graphs with h3viewer and site manager. In *Symposium on Graph Drwaing*, pages 384–393, August 1998.

[14] J. E. Pitkow and K. A. Bharat. Webviz: a tool for world-wide web access log analysis. In *Proceedings of First International WWW Conference*, pages 271–277, May 1994.

[15] A. Wexelblat. History-based tools for navigation. In *Proceedings of the Hawaii International Conference On System Science*. IEEE Computer Society Press, January 1999.

[16] A. Wexelblat and P. Maes. Footprints: History-rich tools for information foraging. In *Proceedings of CHI'99*, pages 270–277, May 1999.

[17] G. J. Wills. Nicheworks - interactive visualization of very large graphs. *Journal of Computational and Graphical Statistics*, 8(2):190–212, 1999.

[18] T. Zheng, Y. Niu, and R. Goebel. Webframe: In pursuit of computationally and cognitively efficient web minging. In *Proceedings of the 6th Pacific-Asia Conference, PAKDD*, pages 264–275, May 2002.