

Visual Web Mining

Amir H. Youssefi
Rensselaer Polytechnic
Institute
110 Eight St.
Troy, NY 12180
youssefi@cs.rpi.edu

David J. Duke
University of Bath
Bath, BA2 7AY, U.K.
d.duke@bath.ac.uk

Mohammed J. Zaki
Rensselaer Polytechnic
Institute
110 Eight St.
Troy, NY 12180
zaki@cs.rpi.edu

ABSTRACT

Analysis of web site usage data involves two significant challenges: firstly the volume of data, arising from the growth of the web, and secondly, the structural complexity of web sites. In this paper we apply Data Mining and Information Visualization techniques to the web domain in order to benefit from the power of both human visual perception and computing; we term this Visual Web Mining. In response to the two challenges, we propose a generic framework, where we apply Data Mining techniques to large web data sets and use Information Visualization methods on the results. The goal is to correlate the outcomes of mining Web Usage Logs and the extracted Web Structure, by visually superimposing the results. We propose several new information visualization diagrams and analyze their utility and elaborate on the architecture of a prototype implementation.

1. INTRODUCTION

Analysis of web site regularities and patterns in user navigation is getting more attention from business and research community as web browsing becomes an everyday task for more people around the world. Various efforts have been made to discover and explore these regularities, some of which employ visualization; see [19, 11, 29, 32, 30, 12] for examples.

Information Visualization is the use of computer-supported, interactive, visual representations of abstract data to amply cognition [27]. The field of *Information Visualization* is about creating tools that exploit the human visual system to help people explore or explain data, or simply visualization applied to abstract data. Computer support brings the opportunity to explore radically different representations, including the use of 3D models.

The capability to interact with a visual representation is significant in allowing users to explore large-scale datasets, where it is infeasible to provide both an overview of the space plus information about points of focal interest (the so called “focus plus context problem”). Different approaches have been taken by researchers to visualize information; some notable ones are [26, 14, 5, 6, 16, 13]. A key challenge in Information Visualization is finding a spatial mapping for an abstract data set that is cognitively useful for a specific task. To address this, information visualization draws on ideas from several intellectual traditions, including computer graphics, human-computer interaction, cognitive psychology, semiotics, graphic design, and cartography.

Web Mining as application of data mining techniques to the World Wide Web, can be divided into Web Content Mining, Web Usage Mining and Web Structure Mining. In this paper, our focus is on

Copyright is held by the author/owner(s).
WWW2004, May 17–22, 2004, New York, NY USA.
ACM xxx.xxx.

last two parts. We define the notion of a *user session* as a temporally compact sequence of web accesses by a user. The goal of our web mining, in part, is to work on these sessions for better visualization toward useful information.

A common theme underlying the use of visualization in website analysis is the graph metaphor, that is, the organization of a web site and/or patterns of access are treated as a node-link graph. There is a considerable literature on algorithms for drawing graphs, however, making *aesthetically pleasing* drawings of graphs (e.g., with a minimal number of edge crossings) is computationally expensive.

One issue is that drawing general graphs is harder than drawing trees, for which a number of efficient approaches are known, for example cone trees [26]. Where a graph is not itself a tree, tree layout can be applied to a spanning tree of the graph, with non-tree edges either added to the resulting structure [3, 2], or not included if the spanning structure is sufficient for a given task. For example, the structure of a web site is in general a graph, with pages corresponding to nodes, and links to edges. The high-level organization of a site is often hierarchical, i.e., like a tree, with one topic branching into a number of sub-topics, as in linear printed media. However, a strength of hypermedia is the ability to cross-link topics, and also provide navigational paths back to higher-level areas, and these kind of links then form non-tree edges of the corresponding graph.

1.1 Visual Web Mining Framework

We propose *Visual Web Mining* (VWM) as the application of Information Visualization techniques on results of Web Mining in order to further amplify the perception of extracted patterns, rules and regularities, or to visually explore new ones in web domain.

The *Visual Web Mining Framework*, initially described in [33], provides a prototype implementation for applying information visualization techniques to the results of Data Mining, in particular the output of algorithms such as Sequence Mining[35] and Tree Mining [36], when applied to large web-site graphs and access log files. The former (site graphs) are semi-static snapshots of the web site structure; the latter (access logs) capture the dynamic behavior of surfers visiting a web site. We anticipate that visual superimposition of the static and dynamic views will enable web analyzers or even web masters to gain better insight on what actually happens on a website.

The organization of large websites reflects a diverse range of design criteria and rationale; e.g., personal preferences of web designers, and the breadth and structure of the topics presented on the site. We use visualization to obtain an understanding of the structure of a particular website, as well as web surfers’ behavior when visiting that site.

Due to the scale of the dataset on which we work, and the struc-

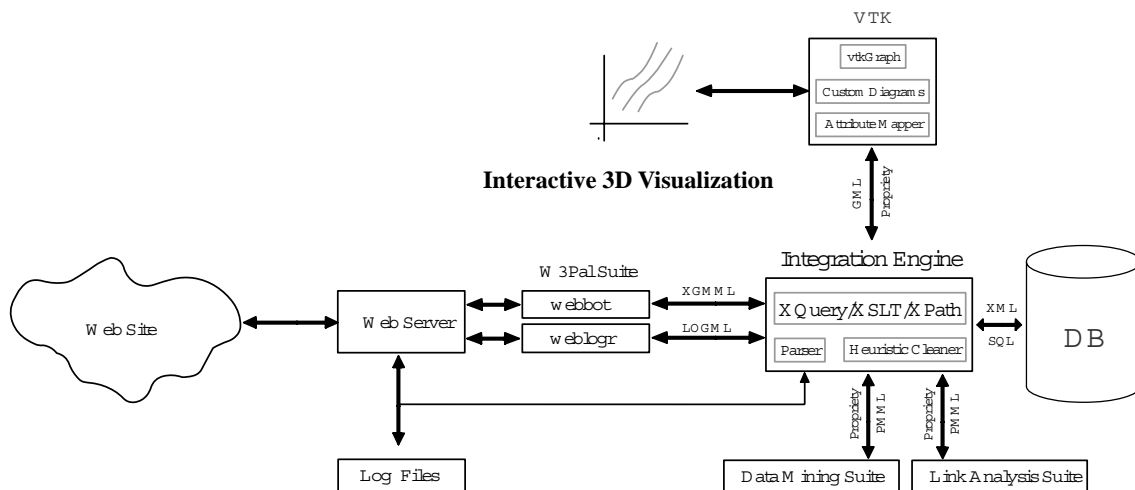


Figure 1: Sample implementation architecture of VWM

tural complexity of the sites from which the data is obtained, we have decided to use 3D visual representations, with user interaction to support navigation and exploration. We have implemented this using an open source toolkit called the Visualization ToolKit (VTK) [28]. The decision to implement website visualization using a toolkit developed, at least initially, for engineering and medical applications, deserves further comment. A number of public-domain tools for the visualization of graphs do exist, and some can cope with large graphs, with up to one million nodes. However, these tools typically provide a fixed representation of a graph (a node-edge diagram), and have limited modularity. In the case of visual web mining, we wish to experiment with novel combinations of metaphors, including spatial positioning, tubing of edges, node glyphs, and color mapping. These techniques are provided in a modular way within general visualization tools, such as VTK, and we have therefore chosen to extend VTK with specific support for working with node-link graphs.

1.2 Contributions

To enable visual superimposition of web usage mining and extracted web structure, we have designed new visual representations of site structure and user access patterns. Aspects of our representations have been motivated by metaphors developed initially within scientific visualization, in particular the representation of vector-fields in fluid dynamics. The results are interactive 3-dimensional (3D) visualizations implemented in the VTK toolkit, providing exploratory tools for supporting Visual Web Mining (VWM).

Our results already show different notable phenomena like cones, funnels and debris, which correspond, to user behaviors as (s)he browses the website (see figure 5). Different mappings are made from data attributes to visual attributes on each diagram. Visualization techniques are used to portray mined structures via visual features such as splines, trees, and stream tubes, with visual attributes such as color, and size used to convey interesting data features, for example users leaving the web site.

As such our work represents the state-of-the-art in visual web mining. Implementation of this visual web mining framework has been applied to the website and access logs of the Computer Science Department at RPI¹.

Given the interactive and graphical nature of our work, we en-

courage the reader who has time to browse through the figures and video clips of the system in action, which are available from the webpage of the VWM project [34].

2. VISUAL WEB MINING ARCHITECTURE

Figure 1 shows the architecture of our implementation. We target one or a group of websites for analysis. Input of the system consists of web pages and web server log files. Access to web log is either through the local file system, or by downloading it from a remote web server (possibly as a password protected resource).

A web robot (weebot) is used to retrieve the pages of the website. For more information see Libwww Robot² and WWWPal Suite [22]. In parallel, Web Server Log files are downloaded and processed through a sessionizer and a LOGML [23] file is generated. Both input and output of different parts are in standard (XML) languages, so output of other programs can be utilized by our system, and so can our results be used by other systems.

The *Integration Engine* is a suite of programs for data preparation, i.e., cleaning, transforming and integrating data. The engine uses XQuery, XSLT and XPath on both standard and propriety data formats. Support for PMML³ is readily available as well. A connection module is in charge of Bulk Loading of XML data into database and executing SQL commands against the database. Schema matching is done using external tools as well as code snippets which map different schema and import/export different XML or propriety file formats into/from relational tables in our database server. Web access log files are also imported into the database for exploring and comparing data mining algorithms as well as for verification of data integration. Since for some analyzed web sites data generated as web logs are in order of gigabytes per month, much effort has been put into efficiency of the transformation system in the *Integration Engine*, and the database.

Data Mining Suite and also *Link Analysis Suite* need special data format as input and give output in propriety formats hence the *Integration Engine* is required to covert data in these formats as well.

The visualization stage of this pipeline, which maps the extracted data and attributes into visual images, is realized through VTK extended with support for graphs. VTK[28] (and this extension) is a set of C++ class libraries, available on a range of architectures

¹<http://www.cs.rpi.edu>

²<http://www.w3.org/Robot/>

³Predictive Model Markup Language (<http://www.dmg.org>)

including Windows and Linux. The class library is accessible either through linkage with a C++ program, or via wrappings supported for scripting languages (Tcl, Python or Java). For our work, the visualization engine uses a Tcl script. Results are interactive 3D/2D visualizations which could be used by analysts to compare actual web surfing patterns to expected patterns, or more generally, to compare the intended purpose and role of the website.

Web Mining has no single recipe for satisfying requirement of analysts or business managers. Businesses ask high level, diverse questions[31], for example:

- Is our site sticky? Which regions in it are not?
- How adept is our conversion of browsers to buyers?
- What site navigation do we wish to encourage?
- What attribute describes our best customers?
- What makes customers loyal?
- How can profiling help use cross-sell and up-sell?

These questions are ‘semantically distant’ from the data available for analysis, and suitable responses are beyond the scope of either simple data mining or visualization. What VWM can provide, however, is insight into more specific, focused, questions that form a bridge between high-level domain concerns (such as the above), and the raw data:

- What is the typical behavior of a user entering our website?
- What is the typical behavior of a user entering our website in page A from ‘Discounted Book Sales’ link on a referrer web page B of another web site?
- What is the typical behavior of a logged in registered user from Europe entering page C from link named “Add Gift Certificate” on page A?
- What is the typical behavior of a user who comes in our website 1 day to 3 weeks before Christmas and buys something, versus one who didn’t buy anything?

In order to partially support this kind of analysis we take the following approach in our example implementation of the VWM framework:

- Make personalized results for targeted web surfers (users, customers) as opposed to blindly aggregating on all users.
- Use data mining algorithms for extracting new insight and measures.
- Employ a database server and relational query language as a means to submit specific queries against data, e.g., projection and aggregations, joins, etc.
- Utilize visualization to obtain an overall picture correlating static site structure with (dynamic) access patterns; provide ‘at a glance’ overviews using visual aggregation and filters; amplify special features via interactive control.

A key issue for the data mining part of this process is how one translates the notion of the *typical behavior* of a user into actual queries on data sets to yield sufficient insight for an analyzing team to identify whether (or how) the (business) goals of the website have been satisfied relative to those targeted customers.

These points serve to emphasize how visualization complements data mining. The latter is about utilizing capabilities of the machine to find and/or compute patterns within the dataset, based on notions of pattern derived from domain knowledge or statistics. Visualization, in contrast, is about using human capabilities to detect patterns or discern trends within visual representations.

In our approach we first extract user sessions from web logs, this yields results related to a specific person as much as possible. User sessions are then converted into a special format for sequence mining using cSPADE [35]:

```
<customer_id> <time_id> <item1> <item2>
```

Here *item1* is the smaller numerical value of SourcePageID and TargetPageID, while *item2* is the larger numerical value of TargetPageID and SourcePageID; this ordering is needed as input to cSPADE needs to be sorted. The interpretation is that a web surfer visits *item1* and *item2* (the source and target pages) together at her click. A list of such source-target pairs form the user session for each user.

For each click of user we need to have both source page and target page, not just source or target page alone, in order to deal with different cases in user navigation, such as hitting browser’s “back” button or finding the entry or exit page. Furthermore, source pages (target pages) out of our website could either be considered as a single source page (sink page) or for more detailed analysis of referrer web sites one can merge classes or clusters of such web sites and use a representative PageIDs for each class or cluster.

We convert text input data into binary, vertical format, and make indices for running cSPADE algorithm, which outputs the *contiguous* (without-gaps) frequent sequences with a given minimum support threshold (all output sequences are at least frequent as the minimum threshold). Results are imported into a database, and non-maximal frequent sequences are removed, i.e., we consider only the maximal (based on subsequence relation) frequent contiguous sequences. Later different queries are executed against this data according to some criterion, e.g., support of the patterns, length of patterns, etc. This gives the flexibility to translate *typical behavior* into different appropriate queries according to subjective requirement set forth by analysis team without going all the way down to web log analysis again. The combined outcome of mining is actually a graph rather than a simple sequence since a user can go back from one page to previous ones or follow anchors in the same web page or simply follow a closed path as sequence of pages.

We attempt to present abstractions of large amounts of data in tuned diagrams in order to maximize opportunities for pattern detection via human perceptual capabilities. We have designed several visualization diagrams, some of which have been implemented. The contribution of some of those diagrams is discussed and illustrated shortly; video clips of the system in action are available from the webpage of the VWM project [34].

3. FROM FLUID DYNAMICS TO WEB DYNAMICS

Advances in scientific visualization techniques, particularly in the domain of fluid dynamics [24, 25, 21], have inspired us to apply similar techniques in information visualization on the web domain. We use an analogy between the ‘flow’ of user click streams through a website, and the flow of fluids in a physical environment, in arriving at new representations. As with most problems of information visualization, our representation of web access involves locating ‘abstract’ concepts (e.g. web pages) within a geometric space. With this as a base, the following ideas motivate our design of the visualization diagrams discussed in the next section:

- Iconic visualization on extracted features from Data Mining.
- Particle tracking for tracing the path of user navigation.
- Feature tracking/detection, for instance turbulence vortex detection techniques, for identifying a user cycling around a set of web pages. A similar phenomena happens for a user navigating pages and this could be interpreted both as confusion and misled tracking, or quite reverse, a case where a user is focusing on a cluster which (s)he has found interesting.
- Splatting for simplifying visualization implementation in dense parts. Both ‘hot’ and ‘cold’ parts of a graph could be splatted in order to focus users’ attention on edges connecting these two regions.
- Fluid event detection on navigation click-streams to capture special (interesting) events.

In [20] visualization techniques were grouped in three categories:

1. *Global Techniques* give a qualitative, global visualization of the data at a low level of abstraction.
2. *Geometric techniques* extract geometric objects (curves, surfaces, solids) from the data. They can be considered as intermediate-level representations, both with regard to locality and level of abstraction.
3. *Feature-based techniques* extract high-level, abstract entities from the data. The emphasis is on quantification for more precise evaluation and comparison.

The second and third of these are used for our visualization methods.

3.1 Structures

3.1.1 Graphs

Much work has been done on two and tree dimensional embedding of graphs, in the form of both algorithms and tools, see for example [10, 3, 17]. We have developed a library under VTK for graph visualization⁴; the rationale and design issues underlying this approach can be found in [8] Two approaches to visualizing the output of data mining have so far been implemented using this library. In the first approach, mentioned in [33], we extract a spanning tree from the site structure, and use this as the framework for presenting access-related results through glyphs and color mapping. In the second approach, the link analysis is treated as a directed graph. This graph is then laid out, currently via one of the following two approaches:

1. A form of Sugiyama layout, described by Auber [2]. This works by extracting a spanning tree from the graph, laying out the tree (for this step we use a variation of the cone tree layout first proposed by Robertson et. al. [26]), and then re-introducing non-tree edges.
2. By using higher dimensional embedding, introduced by [9]. Here the graph is first drawn within a higher dimensional space, and then projected into 2- or 3-dimensions via the principal components of this space.

⁴For more details, or to download the library, see <http://www.cs.bath.ac.uk/~djd/graphs.html>

Attributes obtained by data mining can then be superimposed on the underlying graph structure, for example as variable-width tubes showing particularly strong access paths. In either approach, the visualization designer has freedom to use the visual attributes of the structural element of the graph to convey properties of the underlying data. We will shortly provide a number of examples, and explain the attribute mappings and meanings for each.

3.1.2 Stream Tubes

Variable-width tubes showing access paths with different traffic are introduced on top of the web graph structure. Here, depending on the type of visualization diagram, particular weights, e.g., support of a single click-stream, total sum of support on all click-streams, or the support extracted from Tree Mining algorithm [36] can be mapped onto the width (radius) of the stream tube. Color mapping could be used on the number of users leaving the website (or a cluster of these in a zoomed view), a property of the graph structure (such as the Strahler value), or simply the number of hits of some branch.

3.2 Design and Implementation of Diagrams

Building on the intuition and heuristics set out above, we have designed new visualization diagrams; these are summarized below, after which we will consider issues of utility.

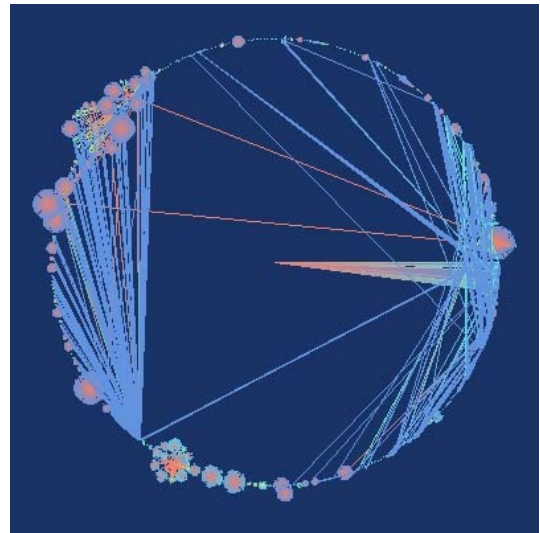


Figure 2: 2D visualization with Strahler Coloring.

- Figure 2 is a visualization of the web graph for the Computer Science department of Rensselaer Polytechnic Institute (<http://www.cs.rpi.edu>). Strahler numbers are used for assigning colors to edges. The Strahler numbers (Horton-Strahler numbers) were originally an attempt to give quantitative information about the complexity or *shape* of a tree. They have been used in various fields, such as hydrology, molecular biology and botany. Herman et. al [10] were the first to propose using them for graph visualization, when they described how these numbers could be used to provide visual cues about tree structure, and generalized the definition to work on arbitrary trees (the original definition assumed binary trees). The Strahler numbers were then generalized further by Auber [2] to provide a means of characterizing the structural complexity of a graph, the insight here being that a graph can be viewed as an expression evaluator. One can

see user access paths scattering from first page of website (the node in center) to cluster of web pages corresponding to faculty pages, course home pages, etc.

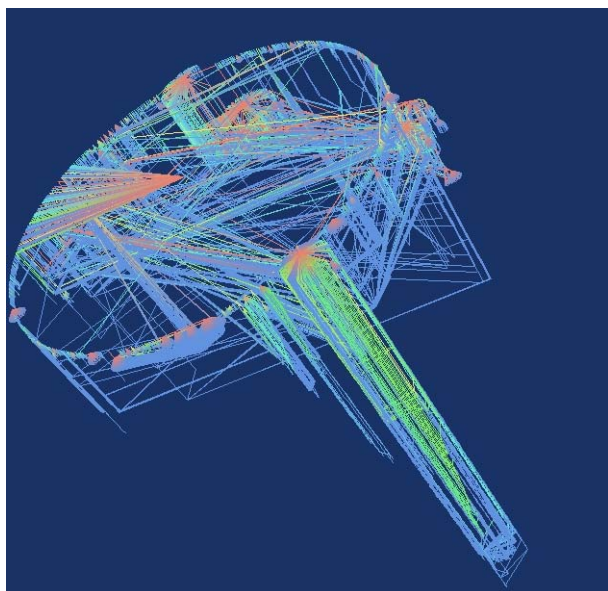


Figure 3: Circular basement is similar to figure 2, adding third dimension enables visualization of more information and clarifies user behavior in and between clusters. Center node of circular basement is first page of web site from which users scatter to different clusters of web pages. Color spectrum from Red (entry point into clusters) to Blue (exit points) clarifies behavior of users.

- Figure 3 is a 3D visualization of web usage for above site. The *cylinder* like part of this figure is visualization of web usage of surfers as they browse a long HTML document converted from \LaTeX by LaTeX2HTML software. Figure 4 gives another example of long “strings” or sequences.
- Figure 5 is presentation of the same site in which color mapping is used to highlight access paths of mass of surfers scattering into clusters making a *cone* shape and later coming back to main pages of clusters, and further to the first page of the website, constructing a *funnel* like shape. User sessions with many visited nodes drilling down into the cylinder shaped cluster is observable.
- One response to the problem of structurally complex graphs is finding new representations. Figure 6 is an example of using hedgehogs, an idea originally developed for vector fields, to capture the distribution and direction of edges. The enclosing blue/gray field represents the extent of space occupied by graph nodes.

A web master can find out where the load of the website and bandwidth goes by a glance at above diagrams, e.g., figure 3. Also *Cube Glyphs* with proportional sizes are put on top of nodes to make them easily clickable (see figure 7). User can pick (click on) those cubes and find out which web page they represent. Glyph size adds another visual dimension that can be used for encoding attribute data. The efficiency of the underlying VTK graphics interface (in turn built on an interface to OpenGL or Mesa) is such that users can easily zoom, pan and rotate these specific diagrams,

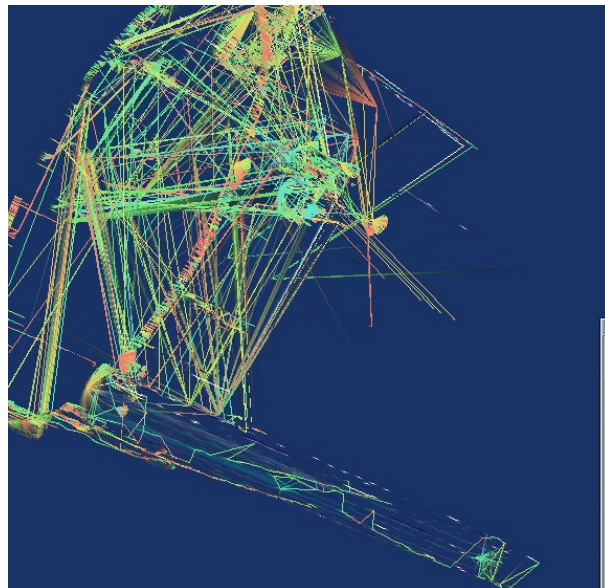


Figure 5: User’s browsing access pattern is amplified by a different coloring. Depending on link structure of underlying pages, we can see vertical access patterns of a user drilling down the cluster, making a *cylinder* shape (bottom-left corner of the figure). Also users following links going down a hierarchy of webpages makes a *cone* shape and users going up hierarchies, e.g., back to main page of website makes a *funnel* shape (top-right corner of the figure).

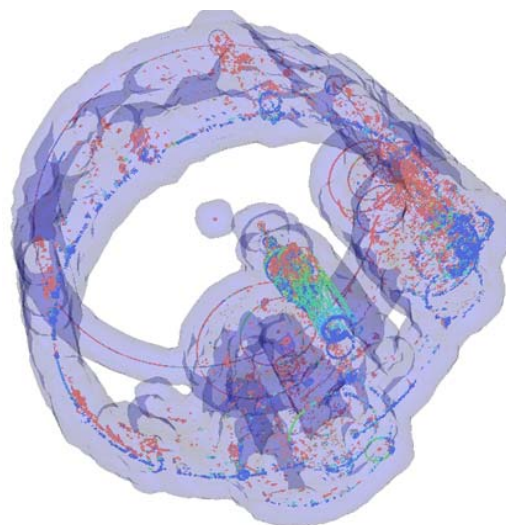


Figure 6: Combining hedgehog representation of edges with a surface providing a cue on the distribution of nodes: the graph represented here contains approximately half a million nodes.

even on modest hardware (e.g., a 433Mhz Pentium II). Thus users can explore different parts of the sample website as extracted from the weblogs. Strategies for real-time interaction with much larger datasets are already being explored.

Web mining results in patterns of most frequent access which are visualized in 7 as white edges superimposed on the remainder of the figure. Figure 8 extends this approach by using thickness of

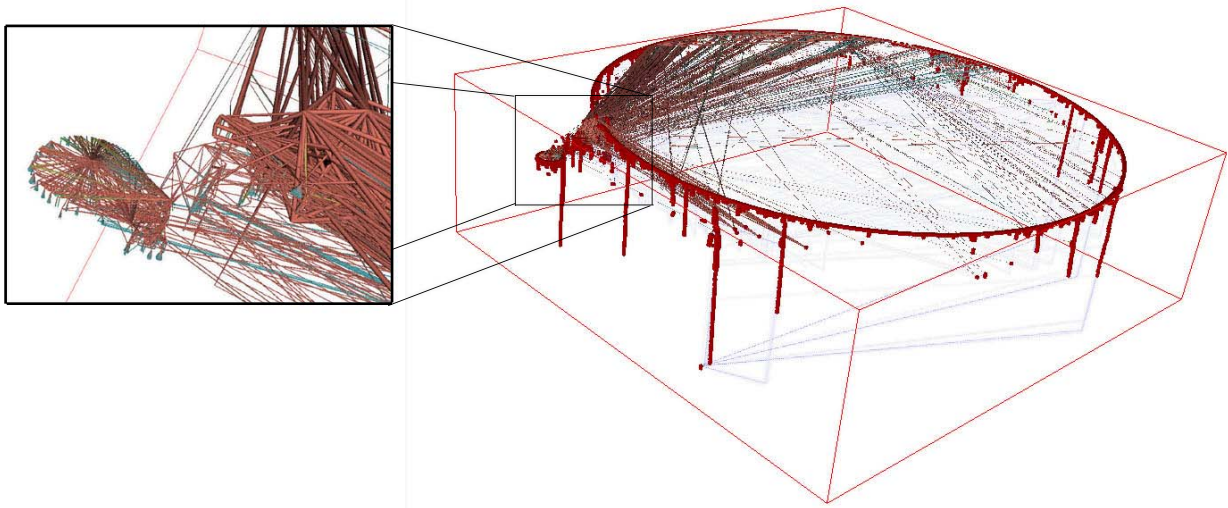


Figure 4: Right: One can observe long user sessions as strings falling off clusters. Those are special type of long sessions when user navigates sequence of web pages which come one after the other under a cluster, e.g., sections of a long document. In many cases we found web pages generated by LaTeX2HTML with many nodes connected with Next/Up/Previous hyperlinks. Left: A zoom view of the same visualization

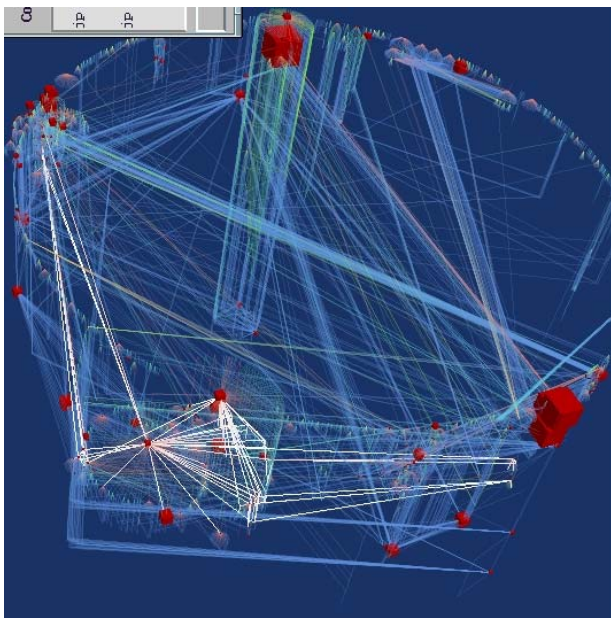


Figure 7: Frequent access patterns extracted by web mining process are visualized as a *white graph* on top of embedded and colorful graph of web usage.

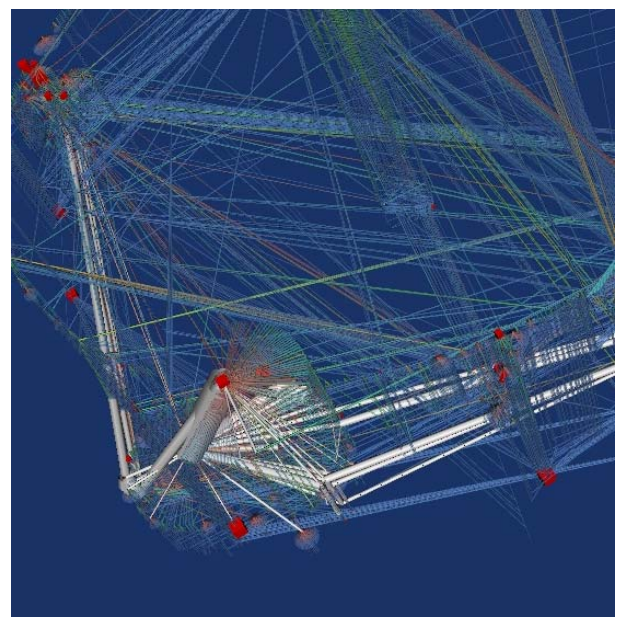


Figure 8: Similar to figure 7 with addition of another attribute, i.e., frequency of pattern which is rendered as thickness of *white tubes*; this would significantly help analysis of results.

stream tubes to add another visual dimension, in this case encoding how frequent those access patterns are.

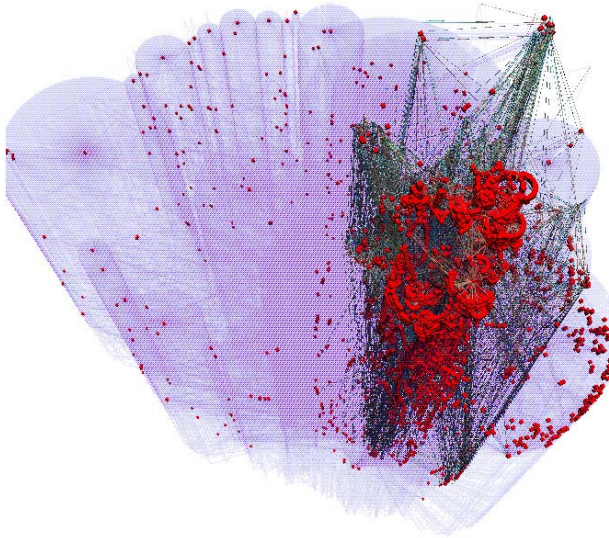


Figure 9: Superimposition of Web Usage on top of Web Structure with span tree layout [2]. One can easily see what parts of the web site was visited by users and what parts are not frequently used. Semi-Static Web Graph extracted by webbot makes a gray basement for colored frequent paths extracted from Web Logs using Web Mining Algorithms. Coloring gives visual cue of entry and exit points of access paths. One can click on glyphs and see details of corresponding node.

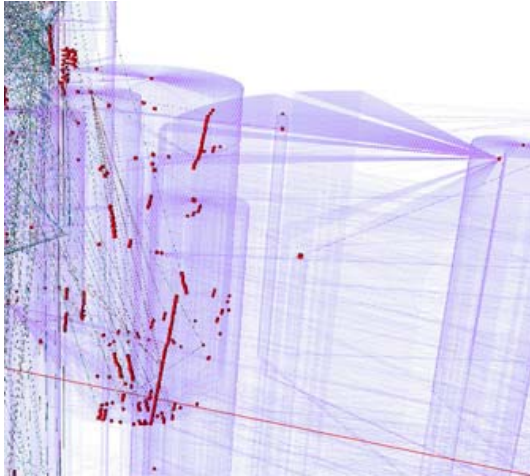


Figure 10: Further exploration is possible through interactions in 3D environment , e.g. zoom, pan and rotation which is reasonably fast. Clusters with/without frequent access are easily identified and details of user access can be analyzed in this zoom view.

In figure 9 we have visualized the static structure of website taken from the webbot (gray edges) to make a basement on which actual dynamic behavior of users is superimposed (non-gray edges). There are about half a million visualized nodes. A webmaster or a web analyzer can easily see which parts of the website are ‘cold’ parts with few hits and which parts are ‘hot’ parts with many hits. This also paves the way for making exploratory changes in web

site and to analyze the changes in user access. For instance a webmaster can change link structure (e.g., by adding a hyper-link to a cold cluster in the first page of website or any page in hot cluster), and observe what path users would choose in real world. Another example is change of content, e.g., by highlighting existing anchor text or putting it on a more visible location of a web page, or adding advertisements/banners and then analyzing changes to the user behavior.

Two further examples of superimposition are given in figure 10 and figure 12. The latter uses the high order layout [9] mentioned in Section 3.1.1.

Although the visual images represent preliminary results from our fusion of mining and visualization, we have endeavored to use suitable guidelines and heuristics based on perceptual, cognitive and aesthetic criteria, for example using psychological notions of hot or cold color to guide assignment of color to nodes and edges having high/low numbers of hits, respectively. Using different diagrams we can focus on each of three link classes we find interesting: hot nodes/edges, cold nodes/edges, and edges connecting two different types of clusters, i.e., hot clusters with high number of hits as opposed to cold clusters with low number of hits. This is valuable for web masters to make decisions and analyze changes of dynamics by informed or exploratory addition of edges between these two clusters. In future work we intend to use techniques such as splatting, spot noise, and/or filtering edges/nodes which are less important, in order to refine and clarify our understanding. As Hamming (1973) said, “The purpose of computation is insight not numbers”. Likewise “The purpose of visualization is insight, not pictures” [27]. A detailed assessment of the utility of our diagrams is beyond the scope of this paper and we leave it for our future work. Recent studies lend support to our approach of discriminating different types of cluster structures [18].

4. FUTURE WORK

There is considerable room for improving the visualization of these results. First, at the algorithmic level, the scale and complexity of the graphs produced from the data mining stage still have the potential to embarrass available graph layout algorithms. There is often a tension in the design of algorithms between accommodating a wide range of data, or customizing the algorithm to capitalize on known constraints or regularities, and in the case of web log data, knowing more about the kind of graph that is to be drawn may help in simplifying the layout process. On the human side, further thought is needed on the mapping from data attributes to visual attributes, in particular where the visualization is superimposing access properties above the basic site structure. Part of this work can and should be based on known characteristics of perception and principles of visualization design, however, the ultimate utility of the representation will only become apparent once it is assessed through controlled experiments, and this will require time and a more polished version of the user interface.

A number of further tasks have already been mentioned in the text, the following could be added:

- Demonstrating the utility of web mining can be done by making exploratory changes to web sites, e.g., adding links from hot parts of web site to cold parts and then extracting, visualizing and interpreting changes in access patterns. This may also require running our implementation on logs obtained over longer period of time.
- Our visual web mining architecture (figure 1) sets the stage for using the latest algorithms in our system. In particular we plan to use new graph mining algorithms.

- More perceptual and logical appraisal of the visualizations relative to better understanding of specific user tasks, e.g., currently in one case we concentrate on a specific user as (s)he surfs a special cluster. Visualization makes a *debris* like shape which can be analyzed by designer of that cluster of web pages.
- Empirical usability tests of the visualizations are subject of work already in progress.
- Output of different systems can be visually correlated with real user behavior in a non-intrusive manner. For instance visualizing output of navigation predictors, recommender systems, browsing simulators, and user modeling/profiling systems as well as a broad range of AI algorithms, learners, probabilistic algorithms and more can be visually superimposed on user behavior (or vice versa) which would help us explore those systems further. Visualization of new relational probabilistic models e.g. [1] is of particular interest because of inclusion of relational structure of web sites in both techniques. Along the same line, visualization of link analysis and comparison of PageRank [4] and Hub/Authority [15] against actual web usage is possible. Also web content mining [7] can be introduced to implementations of our architecture.
- We remove requests of web crawlers and robots, e.g., Google Web Crawler in early stages of pipeline, one interesting application is to visualize access path of those.
- In most of the superimpositions studied in this text we remove web pages not common to Web Usage (U) results gathered in a period of time and structure of web site i.e. Web Graph nodes (S) taken in a special date i.e. we use intersection of two sets S and U . In our new studies we further explore visualization of changes of both S and U over time especially by visualizing $(S \cap U, S - U \text{ and } U - S)$ with appropriate colorings.

5. REFERENCES

- [1] C. R. Anderson, P. Domingos, and D. S. Weld. Relational markov models and their application to adaptive web navigation. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.
- [2] D. Auber. *Outils de visualisation de larges structures de donnees*. PhD thesis, Universite Bordeaux I, 2002.
- [3] D. Auber. Using strahler numbers for real time visual exploration of huge graphs. *Int'l Journal of Applied Mathematics and Computer Science*, 2002.
- [4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 1998.
- [5] E. H. Chi, J. Pitkow, J. Mackinlay, P. Pirolli, R. Gossweiler, and S. K. Card. Visualizing the evolution of web ecologies. In *ACM Conference on Human Factors in Software, SIGCHI*, 1998.
- [6] J. Cugini and J. Scholtz. Visvip: 3d visualization of paths through web sites. In *Int'l Workshop on Web-Based Information Visualization in conjunction with DEXA*, pages 259–263, 1999.
- [7] B. D. Davison. Predicting web actions from html content. In *Proc. of the thirteenth ACM conference on Hypertext and hypermedia*, 2002.

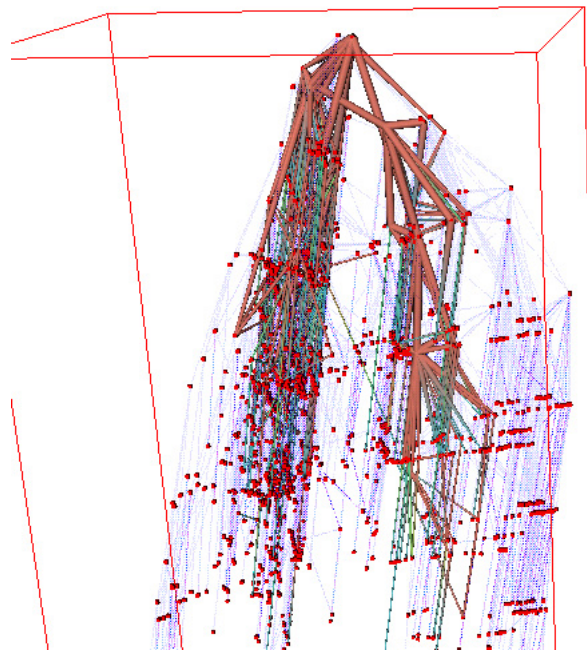


Figure 12: Superimposition of Web Usage on top of Web Structure with higher order layout. Top node is first page of website. Hierarchical output of layouts make analysis easier.

- [8] D. Duke. Modular techniques in information visualization. In *Proceedings of the 1st Australian Symposium on Information Visualization*, volume 9, pages 11–18, 2001.
- [9] D. Harel and Y. Koren. Graph drawing by high dimensional embedding. In *Proceedings Graph Drawing 2002*. Springer Verlag, 2002.
- [10] I. Herman, M. Delest, and G. Melancon. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1), 2000.
- [11] B. A. Huberman, P. L. Pirolli, J. E. Pitkow, and R. M. Lukose. Strong regularities in world wide web surfing. *Science*, 280(3):96–97, April 1997.
- [12] A. Keahey and S. G. Eick. Visual path analysis. In *IEEE Symposium on Information Visualization*, 2002.
- [13] D. A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1), 2001.
- [14] T. M. Kirsten Ridsen, Mary P. Czerwinski and D. B. Cook. An initial examination of ease of use for 2d and 3d information visualizations of web content. *International Journal of Human Computer Studies*, 53(5):695–714, 2000.
- [15] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proc. 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.
- [16] T. Munzner. Drawing large graphs with h3viewer and sitemanager. In *Proceedings of Graph Drawing*, 1998.
- [17] P. Mutton and P. Rodgers. Spring embedder preprocessing for www visualization. In *Proc. Int'l Symposium on Web Graphics and Visualization IV02-WGV*, pages 744–749, 1998.
- [18] M. Nakagawa and B. Mobasher. A hybrid web

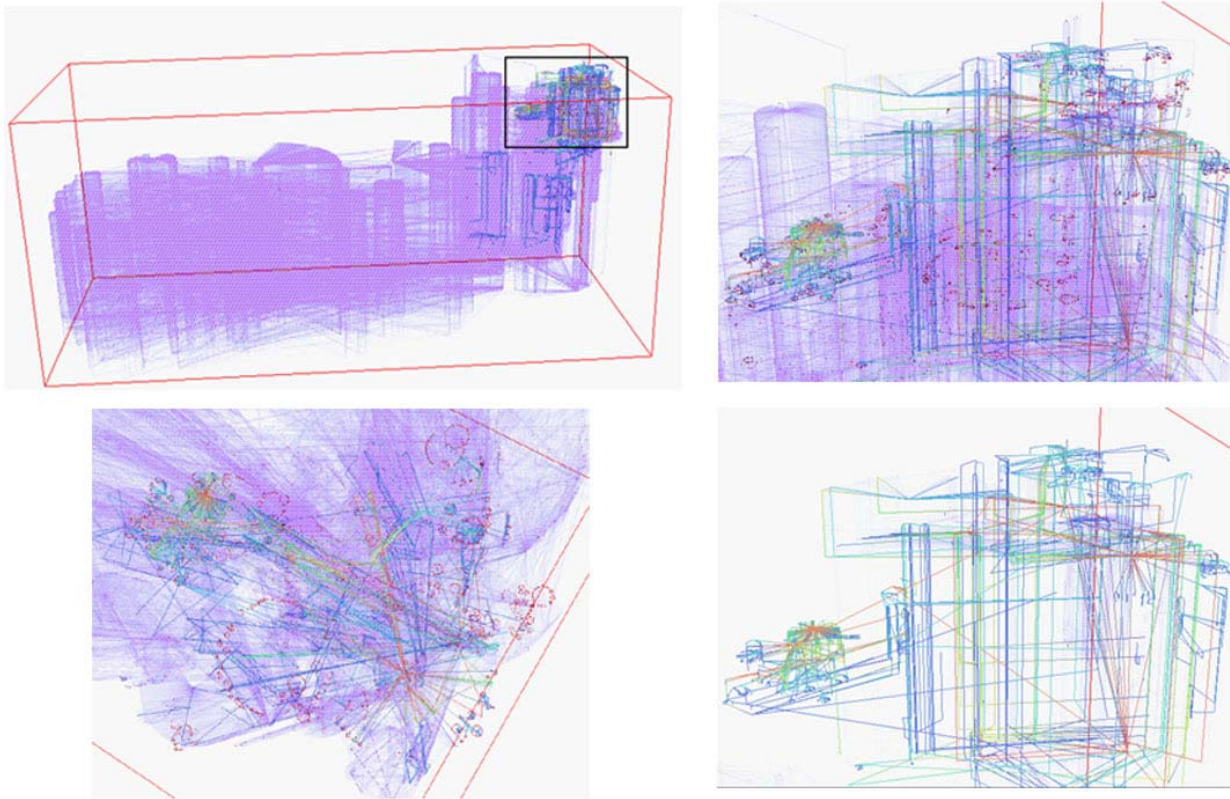


Figure 11: Above figures are similar to figure 9 but edges are rendered as colored lines instead of tubes for clear distant view. Top-left figure indicates the region of zoom from which other three figures are taken. Top-right is simple zoom for comparison. Bottom-right is the same region with layout of Web Usage taken from Web Graph basement but the basement itself is removed for clarity. Bottom-left is top view in which cluster of web pages are clearly observable. We have a spectrum from red to blue and red edges are the ones close to entry points of surfers and similarly blue edges are the ones close to exit points of user sessions.

personalization model based on site connectivity. In *WebKDD workshop SIGKDD*, 2003.

- [19] N. I. of Standards and Technology. <http://www.nist.gov/webmet/>.
- [20] F. Post, W. de Leeuw, I. Sadarjoen, F. Reinders, and T. van Walsum. Global, geometric and feature-based techniques for vector field visualization. *Future Generation Computer Systems*, 15, February 1999.
- [21] F. H. Post, B. Vrolijk, H. Hauser, R. S. Laramée, and H. Doleisch. The state of the art in flow visualisation: Feature extraction and tracking. In *Proceedings of the 1st Australian Symposium on Information Visualization*, volume 22, pages 11–18, 2003.
- [22] J. Punin and M. Krishnamoorthy. wwwpa1 system- a system for analysis and synthesis of web pages. In *Proc. WebNet*, 1998.
- [23] J. Punin, M. Krishnamoorthy, and M. J. Zaki. Logml: Log markup language for web usage mining. In *WebKDD Workshop, ACM SIGKDD*, pages 88–112, 2001.
- [24] F. Reinders, F. Post, and H. J. W. Spoelder. Visualization of time-dependent data using feature tracking and event detection. *The Visual Computer*, 17(1):55–71, February 2001.
- [25] F. Reinders, I. A. Sadarjoen, B. Vrolijk, and F. H. Post. Vortex tracking and visualisation in a flow past a tapered cylinder. *Computer Graphics Forum*, 21:675–682, Nov 2002.
- [26] G. G. Robertson, J. Mackinlay, and S. Card. Cone trees: Animated 3d visualizations of hierarchical information. In *Proc. ACM SIGCHI*, 1991.
- [27] S. Card, J. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, San Francisco, 1999.
- [28] W. Schroeder, K. Martin, and B. Lorensen. *The Visualization Toolkit: An Object-Oriented Approach to 3D Graphics*. Prentice Hall, 1998.
- [29] C. Shahabi, A. M. Zarkesh, J. Abidi, and V. Shah. Knowledge discovery from user’s web-page navigation. In *Proc. Seventh IEEE Intl. Workshop on Research Issues in Data Engineering (RIDE)*, pages 20–29, 1997.
- [30] M. Spiliopoulou and C. Pohle. Data mining for measuring and improving the success of web sites. *Data Mining and Knowledge Discovery*, 5, 2001.
- [31] J. Sterne. Invited talk: Webkdd in the business world. In *WebKDD Workshop in conjunction with SIGKDD*, 2003.
- [32] L. Tauscher and S. Greenberg. Revisitation patterns in world wide web navigation. In *Proceedings of the Conference on Human Factors in Computing Systems, SIGCHI*, 1997.
- [33] A. H. Youssefi, D. J. Duke, M. J. Zaki, and E. P. Glinert.

Toward visual web mining. In *Visual Data Mining Workshop IEEE Int'l Conf. on Data Mining*, 2003.

- [34] A. H. Youssefi et al. Visual web mining web page, <http://www.cs.rpi.edu/~youssefi/research/vwm>.
- [35] M. J. Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine Learning Journal*, 42:31–60, 2001.
- [36] M. J. Zaki. Efficiently mining trees in a forest. In *ACM SIGKDD*, 2003.