An Interactive Visualization Technique for Access Patterns of Web Sites

Makiko Kawamoto, Takayuki Itoh

Ochanomizu University

ABSTRACT

Visualization of Web information is an active research topic. We have researched on a visualization technique for access patterns and link structures of web sites. This poster presents an interactive visualization technique which visualizes user-selected access patterns. This enables users to compare access patterns from each user's viewpoint, and the users can analyze a visualization result united with each user's concern. Moreover, it enables to choose access patterns extracted from the access log stored over a long time. This will enable us to analyze more detail, for example, we can compare access patterns in the same month or same patterns through every year.

KEYWORDS: Visualization, access pattern, access log.

1 INTRODUCTION

Research on visualization of Web sites has already presented widely, and we have researched on a visualization technique for access patterns and link structures of web sites [1].

This poster presents an interactive visualization technique which visualizes access patterns that user selected. Here, this poster defines an access patterns as a set of web pages which are accessed by multiple browsers. Temporarily, there are *m* Web pages $P = \{p_1, p_2, \cdots, p_m\}$ and *n* browsers $B = \{b_1, b_2, \cdots, b_n\}$. When a subset of Web pages $P' = \{p_i, p_j, p_k, \cdots\}$ are accessed by a subset of browsers $B' = \{b_a, b_b, b_c, \cdots\}$, this poster defines the subset of Web pages belonged to P' an access pattern.

Users can compare access patterns from each user's viewpoint by visualizing access patterns interactively, and we think that each user can analyze a visualization result united with each user's concern. Moreover, it enables to choose access patterns extracted from the access log stored over a long time. This will enable us to analyze more detail, for example, we can compare access patterns in the same month or same patterns through every year.

2 RELATED WORK

Web-related visualization has been an active research topic since 1990s, and several survey papers and Web sites have been published [2].

Behaviour of browsers is very interesting information for Web designers and administrators, and therefore analysis and mining of such behaviour are also active research topic. Nasraoui et al. defined a similarity calculation scheme between behaviour of two browsers, and applied fuzzy clustering to them [3]. Pitkow et al. presented a prediction model of user surfing paths based on Markov models [4]. Davison presented a technique to predict actions of browsers by detecting their interest from textual contents [5].

3 PRESENTED TECHNIQUE

This technique constructs the input data by integrating the following information,

- access patterns extracted from Web access log files, and
- link structures constructed using Web crawler software,

and visualizes by our network visualization technique.

3.1 Access Pattern Extraction

This section describes our implementation of Web access pattern extraction.

The implementation first parses a Web access log file, and constructs lists of IP addresses of browsers and accessed URLs. Here, our implementation records URLs which do not points to multimedia contents files (e.g. images, sounds) to the list. It then constructs a matrix where rows correspond to n IP addresses of browsers b_1 to b_n, and columns correspond to m accessed URLs of Web pages p_1 to p_m. It also fills elements of the matrix a_{ij} by the total number of accessed from the i-th IP address to the j-th URL.

The implementation then applies a hierarchical clustering algorithm to divide the browsers based on the sets of accessed Web pages. Here, it treats numbers of accesses to each Web page by a browser as m-dimensional vectors, and calculates cosine between every pair of the browsers. When numbers of accesses to each Web page by browser x and y are expressed $x = (x_1, x_2, \cdots, x_m)$ and $y = (y_1, y_2, \cdots, y_m)$, cosine between vectors is denoted by the following formulas.

$$S_{\cos}(x, y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \quad (1)$$

It then constructs a dendrogram by recursively coupling the browsers. The process first treats the browsers as nodes, and converts the pair of nodes which has the largest cosine value into a single node. Recursively finding the pairs of nodes which have the largest cosine values and converting into the single nodes, the process constructs the dendrogram. The implementation then generates clusters of the browsers c_1 to c_l, where l is the number of clusters, by cutting the dendrogram by a user-defined threshold α.

Finally, the implementation extracts sets of Web pages as Web access patterns. It selects the clusters which contains β or more browsers, where β is user-defined number. It then extracts a set of Web pages which are accessed by γ percent or more of the browsers, where γ is also a user-defined value.

3.2 Network Visualization

We apply a multiple-category-embedded network visualization technique using hybrid force-directed and space-filling graph layout [6], to visualize link structure and access patterns on the same screen. In our implementation, nodes correspond to Web pages, edges correspond to hyperlinks, and hierarchy corresponds to the directory structure of the Web site. The implementation assigns independent colors to the access patterns, and draws nodes which belong to access patterns as colored circles. When a node belongs to multiple patterns, it divides the circle into multiple fans, similar to a pie chart.

3.3 Interactive Visualization

Access patterns extracted by the technique described in Section 3.1 are placed as colored buttons on a GUI screen, as shown in Figure 1. When a user selects interested patterns, this technique paints nodes contained in the selected patterns. Moreover, buttons are colored according to five levels, determined based on the number of average accesses in the access patterns. This coloring system works as an access pattern recommendation function.

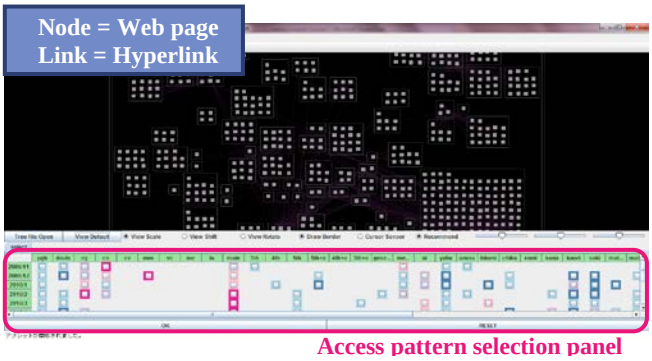

Figure 1. GUI window

4 Case Study

This section introduces a case study that visualized Web access patterns of Web sites of authors' laboratory (http://itolab.is.ocha.ac.jp/), where the dataset is constructed from the Web access log files from November 2009 to October 2010. We implemented the technique with Java JDK 1.6.0, and tested on Windows7 (CPU 1.2GHz, RAM 4GB).

Looking at the access pattern selection panel, it turned out that an access pattern of Lecture A has many accesses only in February through a year. Interested in the fact, we selected all pattern of Lecture A, and we got the visualization result shown in Figure 2.

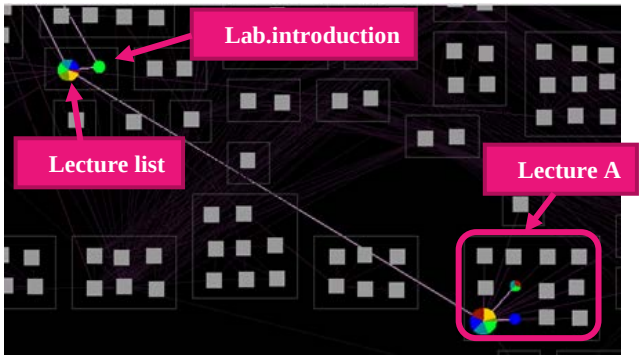

Figure 2. Visualization of access patterns

Looking at the colors of each page in detail, we found that the laboratory introduction page is also colored in green. The green represents an access pattern in February. Therefore, we found that the laboratory introduction page is contained in the pattern well-accessed only in February. Since laboratory assignment is performed in February in our department, we can suppose that visitors have also accessed the laboratory introduction page together with pages of Lecture A. Thus, discovery of the pages which is not contained in the usual access pattern leads us to grasp which pages are accessed in special seasons. It is especially useful for the planning of Web site updates, and examination of the contents of pages.

5 Conclusion

This poster presented an interactive visualization technique for access patterns of Web sites. Interactive visualization of access patterns using this technique brings useful for Web site improvements, such as grasp of the access tendency in special seasons.

Our potential future work includes the following issues. Firstly, we need to improve the scalability of the technique; we would like to speed up the technique for large scale Web sites containing ten thousands of Web pages and hundreds of access patterns. Another issue of our current implementation is access pattern extraction process because it is so naïve that we do not expect it can perfectly discover meaningful access patterns. We would like to improve the implementation. Finally, we are interested in visualizing the correlation between access patterns and contents of Web pages. We would like to extract keywords that are distribution of keywords as well as access patterns.

References

[1] M. Kawamoto, T. Itoh, A Visualization Technique for Access Patterns and Link Structures of Web Sites, International Conference on Information Visualization, pp. 11-16, 2010.

[2] An Atlas of Cyberspace, http://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography/atlas/

[3] O. Nasraoui, H. Prigui, A. Joshi, R. Krishnapuram, Mining Web Access Logs Using Relational Competitive Fuzzy Clustering, Eight International Fuzzy Systems Association World Congress, 1999.

[4] J. Pitkow, P. Pirolli, Mining Longest Repeating Subsequences to Predict World Wide Web Surfing, 2nd conference on USENIX Symposium on Internet Technologies and Systems, pp. 139-150, 1999.

[5] B. D. Davison, Predicting Web Actions from HTML Content, 13th ACM Conference on Hypertext and Hypermedia, pp. 159-168, 2002.

[6] T. Itoh, C. Muelder, K.-L. Ma, J. Sese, A Hybrid Space-Filling and Force-Directed Layout Method for Visualizing Multiple-Category Graphs, IEEE Pacific Visualization Symposium, pp. 121-128, 2009.