Proceedings of the 2nd Workshop on Adaptive Systems and User Modeling on the WWW (1999)

Adaptive Web Prefetching

Brian D. Davison
Department of Computer Science, Rutgers University
110 Frelinghuysen Road
Piscataway, NJ 08854-8019
davison@cs.rutgers.edu
http://www.cs.rutgers.edu/~davison/

Overview

Many factors contribute to a less-than-speedy web experience, including heterogeneous network connectivity, real-world distances, and congestion due to unexpected network demand. Web caching, along with other forms of data dissemination, has been proposed as a technology that helps reduce network usage and server loads and improve average latencies experienced by the user. When successful, prefetching web objects into local caches can be used to further reduce latencies [KLM97], and even to shift network loads from peak to non-peak periods [MRGM99].

Our interest is in prefetching interactively, so that by dynamically prefetching web objects likely to be of interest, we may invisibly improve the user experience by improving the response time. In order to maximize its potential, such a prefetching system is likely to need to adapt to the user's browsing habits and interests.

Relevant Research

In previous work we have considered the task of anticipating the next user action taken at a UNIX shell [DH97, DH98]. That work focussed on recognizing patterns in the user's history to predict future actions. Additionally, it allowed for the user's profile to change over time (by emphasizing recent actions over those in the past). However, IPAM (Incremental Probabilistic Action Modeling), proposed in that work, did not take into consideration additional sources of information that might have been relevant (e.g., models of the average user, or additional context). As such, it could not predict actions that it had not seen in the past.

Our current research focus is to apply similar machine learning mechanisms to the problem of action prediction on the web. In particular, we wish to be able to predict the next web page that a user will select. If one were able to build such a user model, a system using it could anticipate each page retrieval and fetch that page ahead of time into a local cache so that the user experiences very little retrieval latency, and so reduce widespread complaints about the ''World-Wide Wait''. Thus, performance measurement of such a system is primarily in terms of user-perceived latency. However, since a perfect prediction system is impossible, we must also consider side-effects of prefetching incorrectly, such as increased server loads and bandwidth usage. Elsewhere [Dav99a], we have surveyed existing

techniques for evaluation and proposed a mechanism [Dav99b] for simultaneous evaluation of black-box proxies, including those that implement prefetching, to measure latencies and bandwidth usage.

Naturally, prefetching objects into a cache is not a new concept, and has already been incorporated into a few proxy caches and into a number of browser extensions (see our web site on web caching [Dav99c] for pointers to caching products and browser extensions). The (expected) contributions of this ongoing work are the incorporation of a variety of sources of information for prediction, and the principled evaluation and comparison of such systems. We believe that multiple sources are necessary in order to incorporate desired characteristics into the system. Such sources of information would certainly include client history so that an individual's pattern of usage would serve as a strong guide. But we would also want to include community usage patterns (from proxy and origin servers) so that average usage patterns may be used as intelligent defaults for points in which there is no individual history. Context is also important when history is not relevant --- we plan to use the textual contents of recent pages as a guide to the current interests of the user and the link contents of those pages as significant influences to what may be chosen next. Finally, we hope to capture the contents of related applications (such as net-news and electronic mail) which also present URLs that can be chosen as pages to be retrieved.

We have described a variety of sources of information; some of these can be considered aspects of a user model, but others are better viewed as separate models of the average user. Each model reflects the source under which it is built, and can be generated using technology independent of the others. For example, the model for a particular user's history might be calculated using Prediction by Partial Match [BCW90], while a web server's model of the average user might simply be a list of the most popular pages at that web site. (Elsewhere [LD99] we have considered the potential for adaptively, but unobtrusively pushing likely content from server to client.) In the current work, predictions from each of these models is to be collected at the client's system so that they can be merged into a single ordered list of objects to be prefetched, but alternately this task could be performed at the proxy.

At present we are collecting full-content web traffic logs (for a small number of users) using a custom proxy for off-line analysis. These logs include all HTTP request and response headers and the content of all HTML pages, since traditional logs are insufficient for analysis of content-based prefetching systems [Dav99d]. By combining different sources of information, we expect to be able to make predictions of actions that have never been taken by the user and to to make predictions that reflect current user interests. Our conjecture is that the appropriate combination of information from sources such as these will make more accurate predictions possible via a better user model, and thus reduce the amount of extra bandwidth required to generate adequate improvements in latency.

References

[BCW90] T. C. Bell, J. G. Cleary, and I. H. Witten. (1990) *Text Compression*, Prentice Hall, Englewood Cliffs, NJ.

[DH97] B. D. Davison and H. Hirsh. (1997) **Toward An Adaptive Command Line Interface**. Presented at the Seventh International Conference on Human-Computer

Interaction, August 24-29, 1997, San Francisco, CA. Proceedings published as *Advances in Human Factors/Ergonomics Volume 21B - Design of Computing Systems: Social and Ergonomic Considerations*, pp. 505-508: Elsevier Science Publishers.

[DH98] B. D. Davison and H. Hirsh. (1998) **Predicting Sequences of User Actions**. Presented at the AAAI-98/ICML'98 Workshop on Predicting the Future: AI Approaches to Time Series Analysis, Madison, WI, July 27, 1998, and published in *Predicting the Future: AI Approaches to Time Series Problems*, Technical Report WS-98-07, pp. 5-12, AAAI Press.

[Dav99a] B. D. Davison (1999) **A Survey of Proxy Cache Evaluation Techniques**. Published in the *Proceedings of the Fourth International Web Caching Workshop (WCW99)*, March 31-April 2, San Diego, CA.

[Dav99b] B. D. Davison (1999) **Simultaneous Proxy Evaluation**. Published in the *Proceedings of the Fourth International Web Caching Workshop (WCW99)*, March 31-April 2, San Diego, CA.

[Dav99c] B. D. Davison (1999) **Web Caching Resources**, URL: http://www.web-caching.com/.

[Dav99d] B. D. Davison (1999) **Web traffic logs: An imperfect resource for evaluation**. To be published in the *Proceedings of the Ninth Annual Conference of the Internet Society (INET'99)*, June, San Jose, CA.

[LD99] V. Liberatore and B. D. Davison (1999) **Data Dissemination on the Web: Speculative and Unobtrusive**. UMIACS Technical Report 99-23, University of Maryland, College Park, MD.

[KLM97] T. M. Kroeger, D. E. Long, J. C. Mogul (1997) **Exploring the Bounds of Web Latency Reduction from Caching and Prefetching**. Published in *USENIX Symposium on Internet Technologies and Systems (USITS)*, December 8-11, Monterey, CA.

[MRGM99] C. Maltzahn, K. J. Richardson, D. Grunwald, and J. H. Martin (1999) **On Bandwidth Smoothing**. Published in the *Proceedings of the Fourth International Web Caching Workshop (WCW99)*, March 31-April 2, San Diego, CA.