

Relating web analytics to system performance and availability metrics

IBM University Partnership Program Proposal

Joydeep Ghosh

Professor, Dept. of Electrical & Computer Engineering

ACES 3.118, Univ. of Texas, Austin, TX 78712-1084

IBM Technical Contact: Binh Q Nguyen,

Senior Architect WebSphere Site Analyzer,

Tivoli Software, IBM Software Group

IBM Sponsor: Christopher O'Connor,

Director of Market Management, Performance & Availability Solutions,

Tivoli Software, IBM Software Group

1 Objectives and Approach

Current web analytics tools such as weblog analyzers and recommender systems are largely focussed on usage measurements, Customer Relationship Management and other web personalization technologies. On the other hand, performance and availability of computer systems are monitored and evaluated using a wide variety of modeling techniques, ranging from queueing theory to benchmarking, but are typically not directly related to weblogs. In this project, we propose to investigate several ties between these two areas, most specifically how web-logs can be used to better predict system behavior. The key procedure will be to determine how a web-site's "visitors" (e.g. software agents, remote processes) can be clustered or segmented into relatively uniform groups in terms of their system requirements and anticipated system demands. Individual prediction models can then be built for each segment. Finally, by labelling incoming traffic into the segments (behavior modes) they most likely belong to, one can do more precise and efficient forecasting of system demands.

This approach to linking predicted visitor requirements with QoS issues will use an array of techniques with a common underlying theme: if one can determine an appropriate (domain specific) measure of similarity between two visitor behaviors, then the problem can be translated into similarity space wherein problems of irrelevant attributes, high amounts of noise etc, can be avoided, and the visitors can be segmented more easily.

This understanding has emerged from our recent work on clustering visitors to a website based on clickstream analysis [BG01], market basket studies on large retail data [SG00], and clustering web documents [SGM00]. For example, current leading clickstream analysis tools (including those from IBM, E.piphany, Accrue, Webtrends) operate either via OLAP or through direct counting and breakouts. Visitors are labelled based on simple metrics such as which subdomain was most visited or at what stage was the shopping cart abandoned. A richer description of a visit can incorporate the actual *sequence* of pages visited, times spent on each page, content of pages, typed in entries etc. But this leads to too many parameters, and cannot be handled by current means. We approached the problem by deriving a similarity measure between two sessions, that first identifies the longest common subsequence of pages visited between two sessions, and then weights it by the importance of this common behavior relative to the entire trace, as well as by the similarities in the times spent at the common pages. Applying this to a portal's traffic, we found that it could make finer but important distinctions, and was much more

accurate in characterizing behavior of visitors as compared to commercially available software.

Similar approaches also enabled us to obtain finer grain clustering, with consequent better predictive modeling, in other difficult domains as well [SG00, SG00].

2 Proposed Research

The project centers around exploring linear, non-linear, multi-dimensional, sequence relations between web usage and Performance availability metrics. Web usage metrics include hit counts, data download volumes, page views and session statistics. These measures can be further broken down in terms of the visitors/computers that generated them. Performance / availability metrics include end-user reponse time, server reponse time, server characteristics (e.g. system metrics such as CPU and memory), client-side network load, etc. Currently, Tivoli Web Log Analyzer linearly correlates certain pairs of high level measures, such as relating user response time to web site traffic. Are there non-linear or multivariable interactions as well that are significant? Two main approaches will be taken to address this question:

1. Similarity based Prediction. One way to determine this issue is to augment each historic clickstream session with its demands on system resources (bandwidth, memory, QoS requirements) and associated performance and availability metrics. We shall then determine appropriate similarity measures for this extended representation. Characterizing visitors in this extended space will provide an alternative and perhaps superior method for predicting system demands, capacity planning and improving performance. It can also help in improving caching policies, and better anticipating need for other web services. Predictions can be validated with actual response time metrics from Tivoli Monitoring for Transaction Processing (TMTP), and the impact of of similarity-based / clustering techniques assessed.

2. Associating Sequences: is useful for a wide variety of applications. For example, if a sequence of system requests is a known indicator of a network intrusion, one would be interested in identifying other sequences that are similar in the relevant behavioral aspects of the intrusion sequence. One approach will be to identify core sequences, and then using an efficient heuristic to associate all other sequences with one of the cores, an approach that scales linearly with number of samples and has been successfully used in studying gene expression sequences. Note that sequence analysis required a different set of tools, and is a formidable project in itself, but needs to be investigated since it is better able to capture recent history and context and its effects on performance/availability.

In the long run, we would like to use the information above to derive a statistical model that can relate server configuration to the web traffic it can support. This in turn can be tied into a comprehensive and proactive personalization system built as an application on top of the web management system.

Biosketch: Joydeep Ghosh is a Full Professor and Archie Straiton Fellow in the Department of Electrical and Computer Engineering at the University of Texas, Austin. He has over 200 refereed publications related to the theory and applications of intelligent data analysis. His 13 years of teaching at UT include data mining courses taught to students from engineering, business and industry, including a new course on web mining last semester. Dr. Ghosh received the 1992 Darlington Award for best journal paper from IEEE Circuits and Systems Society, and also six other awards for papers on neural network theory and applications. He co-organized the Web Mining Workshop in SDM, April 2001, and serves on the program committees of several data mining related conferences.

References

- [BG99] Kurt D. Bollacker and Joydeep Ghosh. Effective supra-classifiers for knowledge base construction. *Pattern Recognition Letters*, 20(11-13):1347–52, November 1999.
- [BG01] A. Banerjee and J. Ghosh. Clickstream clustering using weighted longest common subsequences. In *Workshop on Web Mining : 1st SIAM Conference on Data Mining*, pages 33–40, April 2001.
- [SG00] Alexander Strehl and Joydeep Ghosh. Value-based customer grouping from large retail datasets. In *Proceedings of the SPIE Conference on Data Mining and Knowledge Discovery : Theory, Tools, and Technology II, 24-25 April 2000, Orlando, Florida, USA*, volume 4057, pages 33–42. SPIE, April 2000.
- [SGM00] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. Impact of similarity measures on web-page clustering. In *Proceedings of Seventeenth National Conference on Artificial Intelligence : Workshop of Artificial Intelligence for Web Search (AAAI 2000), 30-31 July 2000, Austin, Texas, USA*, pages 58–64. AAAI, July 2000.