

Fast combinatorial algorithm for Web User Session Reconstruction

Robert F. Dell¹, Pablo E. Román^{2*} and Juan Velásquez²

¹ Naval Postgraduate School, Operations Research Department,
Monterey, California, USA

² University of Chile, Department of Industrial Engineering,
República 701, Santiago, Chile

1 Introduction

Diverse approaches have been used for categorizing the web site content and structure of that are more attractive to users [1]. E-business are currently profiting of those advances using the battery of tools for web intelligence, but all the machinery are built with the supposition that each user's visiting page sequence (sessions) on the site are known. Sessions could be tracked explicitly registered but some confidentiality and legal issues are raised, in some countries web user session storage are not allowed by some local legislation. When such sessions are not available, the process of extracting sessions from web data depends on web logs, resulting in approximated sequences retrieval. By the other hand the quality of the mining process is largely dependent on the quality of its data sources. For this reason the sessionization plays a key role in the web intelligence process. Web log are a noisy data repository, and the origin of this randomness comes from the following source: the cache web browser, the proxy cache and concurrency of anonymous request to the web site. Prior works on sessionization relied on simple heuristics [2]. These heuristics consist basically on grouping sessions by IP/agent and identifying sessions by considering that session's duration doesn't take more than 30 minute [2]. This time driven heuristics has been popular on web mining applications, obtaining good results for web mining. Better approaches have been used for sessionization based on integer programming [3] but with high computational cost. Several authors have looked at the overall characteristics of sessions. They found that the size n of a web user session follows a power law ($n^{-\alpha} / \sum_k k^{-\alpha}$) distribution [4, 5]. The *size* of a session is the total number of registers in the session. In early works [3] we propose to use this verified empirical property as a novel measure of quality of the result of sessionization, since real session are not available.

2 Integer programming for sessionization

We build a network flow optimization problem for extracting sessions from web log registers. The problem is constrained by network flow conservation. For these

* Corresponding Author

purposes each web log register are mapped to two nodes (r, r') where a directed edge with flow $X_{r,r'}$ that transport a fixed one unity of flux. In this representation flux are restricted to be binary and correspond to a user session visit and we ensure that every node are visited. We allows jumps between registers (r_1, r'_1) and (r_2, r'_2) including a directed edge with a binary flow X_{r_1, r'_2} if there exist a link from the web page visited on register 1 to the page visited on register 2 and the register 2 occur in the near future of 1 within a time tolerance. Two artificial nodes are included connected to the set of registers $\{(r, r')\}$, a source node I with a directed binary flow connection $X_{I,r}$ with each register and a sink node O that receive from each register the collected flow $X_{r',O}$. The total flow received at the sink I is minimized on variables $X_{i,j}$, and a session is reconstructed following each connected flows with $X_{i,j} = 1$ from source to sink. As a proof of concept, we consider a university web site (<http://www.dii.uchile.cl>) that hosts the main page of the Industrial Engineering Department of the University of Chile, sub-sites of research groups, personal homepages, a web mail site, academic programs and related project sub-sites. 3,756,006 raw registers were collected over a time window of one month, April 2008. We compare our results with a traditional sessionization timeout heuristic on all clean registers. The timeout heuristic is substantially faster (only 13 seconds) but results shows a distribution of sessions with a $R^2 = 0.92$ correlation coefficient (not as good as the $R^2 = 0.98$ found by the integer program) and a standard error of $err = 0.64$ (nearly twice the standard error of 0.39 found by the integer program). Earlier works [3] relate to integer models for sessionization, the problem of these models relate on the computer time take for processing (between 3 and 6hrs), with the network model it take less than 15 minute to process all data. Comparing the total number of session found by the model and the earlier optimization algorithm, it shown a 0,3% gap difference from both approaches confirming similarity of sessions reconstructed. Future works relate with using the network model for exploring the variance of session found and exploring the analysis of the back and forward button.

References

1. J.D.Velásquez, Palade, V.: Adaptive Web Sites: A Knowledge Extraction from Web Data Approach. IOS Press, Amsterdam, NL (2008)
2. Cooley, R., Mobasher, B., Srivastava, J.: Towards semantic web mining. In: Proc. in First Int. Semantic Web Conference. (2002) 264–278
3. Roman, P., Dell, R., Velásquez, J.: Web user session reconstruction using integer programming. In: Procs. of The 2008 IEEE/WIC/ACM International Conference on Web Intelligence, Sydney, Australia (December 2008) 385–388
4. Huberman, B., Pirolli, P., Pitkow, J., Lukose, R.M.: Strong regularities in world wide web surfing. *Science* **280**(5360) (1998) 95–97
5. Vazquez, A., Oliveira, J.G., Dezso, Z., Goh, K.I., Kondor, I., Barabasi, A.L.: Modeling bursts and heavy tails in human dynamics. *PHYSICAL REVIEW E* **73**(3) (2006) 036127