Mining the Browsing Context: Discovering Interaction Profiles via Behavioral Clustering

Luis A. Leiva

ITI – Institut Tecnològic d'Informàtica
Universitat Politècnica de València
Camí de Vera, s/n - 46022 (Spain)
llt@acm.org

Abstract. Web clustering usually groups semantically related pages, often including browsing usage information from server logs. However, this is rather limited when it comes to getting deep information about user behavior. Here we explore a different perspective. Behavioral clustering is about modeling the website, that is, finding interaction profiles according to how users behave while browsing. By using a client-side logging tool, we gathered interaction data on three websites. Then we applied a partitional clustering algorithm with interaction-based features as input vectors. We describe our approach, reporting preliminary results, and envision some applications for further research. Behavioral clustering helps to find common interaction profiles as well as to easily identify outliers.

Keywords: unsupervised learning, user profiling, implicit modeling

1 Limitations of Server-side Logs for User Modeling

Modeling the users has long been identified as the key factor of every adaptive hypermedia system. The usual approach is resorting to automatic analysis tools and Machine Learning (ML) techniques, since "manual" work (e.g., preparing usability tests or filling in online questionnaires) is certainly not scalable on the long term due to the highly dynamic nature of user interests and preferences.

Web clustering is a ML technique that aims to group web pages by similarity, by mining features that are document- or transaction-centered (i.e., based on text, link, and usage analysis). Grouped data are then used to make inferences about what users have read, are interested in, etc. Unfortunately, web clustering has been traditionally limited from the user interaction's point of view. Apart from the above mentioned dynamism of users' interests and preferences, websites are constantly updated, and newer paradigms (e.g., caching, Ajax) have substantially altered the traditional *client-request* ↔ *server-response* model. Thus, the web server provides limited knowledge when it comes to getting deep information about the user behavior.

We claim that the browsing context should be added to better contribute to such behavior understanding; that is, we need to move to the client side. Also, mining the browsing context may enhance, complement, and strengthen current approaches to user modeling, adaptation, and personalization (cf. [2]).

2 Contributions

Our proposal has been entitled *behavioral clustering* because its main aim is to cluster websites by *how* users behave, *what* do they do, and *when* they access and leave a site. In the same way as browsing gives information about what is interesting (or not), behavioral clustering can be used to corroborate the coherence of a website from the user interactions' point of view. In addition, mining the browsing context could be used to augment known web clustering techniques, such as page ranking or relevance classification [1]. It also may add a new vision to describe web pages, e.g., *"document A is handled in the same way as document B, where users usually hesitate over the site logo and then click on the first link of the aside menu."* Finally, another contribution of this paper is the empirical validation of the proposed approach through a field study, which also replicates as well as extends previous work in the field [4].

3 Methodology

We collected usage data remotely on three *informational* websites (tendenciashabitat.es, lakq.es, sivaris.eu) for approximately a month, by using an Open Source tracking tool [3]. Such a tracking tool logged user interactions via DOM events (e.g., `mousemove`, `click`, `blur`, or `resize`). It also reported other useful information about interaction, such as scrolling, motion frequency, etc.

Users were chosen by random sampling, which means that only a fraction of all visitors (with equal probability of selection) was collected. Each interaction log was stored in a MySQL database and then exported in XML format. We used Octave[1] to process all logs (11636 files, see Table 1), which were modeled as normalized interaction-based feature vectors. (For an overview of the chosen features as well as a previous pilot study one may consult [4].) We then applied the well-known K-means algorithm to group them in an unsupervised way, using random convex combination as initialization method. The optimal number of clusters was determined as the less distorted grouping in terms of the intra-cluster variance. Finally, we looked at the features that logs assigned to each cluster had in common.

4 Experimental Results

As observed in Table 1, we found some clusters that were clear outliers. This fact reinforced the idea of using behavioral clustering for isolating sub-populations. Looking at these outliers we found that logs belonging to those clusters had extremely unusual behaviors (e.g., browsing time greater than ten hours in the same page, almost no mouse motion, etc.). On the other hand, though, the remaining clusters showed more consistent behaviors. Looking at these groupings we could identify which pages were clubbing active users (e.g., rapid mouse

[1] `http://www.gnu.org/software/octave/`

movements, slight scroll reach, few clicks, etc.) or which ones caused people to hesitate most (e.g., repeated patterns of 'move-stop-move'). These results led us to conclude that each user sample we tracked was in fact a mixture of populations. This evidence encourages to be cautious in using logging tools or intuitions that assume a normal distribution for all users.

Table 1: Clustering results for the evaluated datasets.

Corpus (+ size)	OTH (4803 logs)						NM (5601 logs)					LAKQ (1232 logs)				
Cluster No.	1	2	3	4	5	6	1	2	3	4	5	1	2	3	4	5
Population (%)	0	46	15	0	0	37	25	0	16	27	29	43	7	14	20	13
Distortion	0	0.1	0.72	0	0	0.17	0.15	0.31	0.28	0.14	0.9	0.21	0.16	0.16	0.18	0.27

5 Summary, Conclusions, and Future Work

We have introduced the behavioral clustering methodology, which was evaluated on three real-world datasets, to discover "hidden" profiles on websites. This technique can be used as a measure of similarity between web pages or to evaluate their design. It is also suitable for discovering outliers or "wild-shots". Although we have used only behavior data generated by browser events, we have demonstrated that ours is a useful approach to organize and describe websites from the user interactions' point of view.

As observed, mining the browsing context from user behavior may serve as a complement to current web mining techniques. Future work includes verifying if behavioral clustering results are indeed better than traditional web clustering (i.e., based on clickthrough data only). Further suitability of this work relates to any system that taps knowledge about the user, e.g.: Information Retrieval, Relevance Feedback, Document Organization, or Usage Inference, just to name a few. We believe that we have barely scratched the surface of potentially novel research on user modeling and related applications.

Acknowledgements. Work supported by MIPRCV project (CSD2007-00018).

References

1. Google Inc.: System and method for modulating search relevancy using pointer activity monitoring. US Patent 7,756,887. Available: http://patft.uspto.gov (2010)
2. Hauger, D., Paramythis, A., Weibelzahl, S.: Your browser is watching you: Dynamically deducing user gaze from interaction data. In: Adj. Proc. UMAP (2010)
3. Leiva, L.A.: smt2 – A tool for understanding the user behaviour. Available at: http://smt2.googlecode.com (2009)
4. Leiva, L.A., Vidal, E.: Assessing users' interactions for clustering web documents: a pragmatic approach. In: Proc. HT. pp. 277–278 (2010)