INSITE: A Tool for Real-Time Knowledge Discovery from Users Web Navigation*

Cyrus Shahabi Adil Faisal Farnoush Banaei Kashani Jabed Faruque

Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089, USA
{shahabi, faisal, banaeika, faruque}@usc.edu

Abstract

The major challenges in web mining are a) tracking the data accurately (as not everything is reported to the web server), b) real-time acquisition of the huge volume of data (435 Million visits to yahoo per day, 2-4 GB clickstream data per hour), c) real-time interpretation of the data without compromising the privacy of the user (order of seconds for personalization and targeting information), and d) visualization of the data to facilitate policy making. To address these challenges, we demonstrate an integrated software platform, called INSITE – a) to accurately track users interactions with a web space with minimum overhead and no voluntary user participation, b) to generate individual and aggregate user profiles in real time (or off-line) through the use of a unique Connectivity Matrix Model (CM-model), c) to show the efficacy and scalability of the CM-model in capturing the essence of the users' participatory attributes in the context of the web, d) to visualize the result of clustering of users navigation paths in real time by leveraging on the CM-model, and e) to execute a suite of queries (including temporal ones) and prove the utility of the captured data in making meaningful decisions about user interaction with a web site.

Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000

1. Introduction

Understanding and modelling users' behaviour by analysing users' interactions with digital environments, such as web sites, is a hot topic that has resulted in vast recent commercial interests. Commercial products such as Personify.com™, Verbind™, WebSideStory™, and BlueMartini™, and acquired companies such as Matchlogic™, Trivida™, Andromedia™, Rightpoint™, and DataSage™ are all witnesses of such interests. Meaningful interpretation of the users' digital behaviour is necessary in the disparate fields of e-commerce, distance education, online entertainment and management for capturing individual and collective profiles of customers, learners and employees; for targeting customized/personalized commercials or information, and for evaluating the information architecture of the site by detecting the bottlenecks in the information space.

Several researches in various industrial and academic research centres are focusing on this topic. Among them, Lee's et al. [1] have deployed a *parallel coordinate* system for interpretation and analysis of user clickstream data of online stores; they define "micro-conversion rates" as metrics in web merchandising analysis to understand effectiveness of marketing and merchandising efforts. Heckerman et al. [2] take a model-based approach and use a mixture model to predict behaviour of user clusters and visualize the classification of users. Mobasher et al. [3] propose a hybrid approach to real-time web personalization by combining web usage mining and context- based mining, in a try to overcome the inabilities of each approach individually.

** This research has been supported in part by NASA/JPL contract nr. 961518, and unrestricted cash/equipment gifts from Intel, NCR and NSF grants EEC-9529152 (IMSC ERC) and MRI-9724567.*

With INSITE, we have developed a system for knowledge discovery from users web site navigation in a real-time, adaptive and scalable fashion. INSITE consists of the following layers: a) tracking of user interaction, b) acquisition of the data, c) analysis of the data (includes extraction of navigation paths and clustering of the paths), d) interpretation, and e) visualization of the result of analysis. Our system is novel because it demonstrates a) accurate, unobtrusive tracking of users navigation through a web site, b) real-time, scalable and adaptive clustering of navigation paths by leveraging on a new path/cluster model, and c) a role-based recommendation engine that allows the web site to react to the user in real time with customized information (e.g. target advertisement). In Section 2, we introduce the ideas behind the layers of INSITE. Section 3 discusses the features of INSITE that are demonstrated in this presentation. Finally, in Section 4 we present our conclusions.

2. INSITE: Research Issues

In this section we introduce three out of five layers of INSITE and describe the novel ideas behind each layer. The other two layers, acquisition and visualization, mostly involve implementation issues and are discussed in Section 3.

2.1 Tracking

Tracking is defined as following and recording user interactions with a web site. We keep track of both selected hyperlinks and visited pages. The major challenges in tracking are being accurate and unobtrusive.

As for accuracy, in [4] we proposed the utilization of an agent in the client side to detect and log user interactions in place and send the collected data to an agent server for further analysis. This approach enables us to track the exact time of user interactions. With server-side tracking, in contrast, we have to rely heavily on time approximations because profiler is located at the server, far away from client in the network. In addition, our agent reports visiting of cached pages (whether at the proxy or at the browser) to the server, which results in more accurate tracking as compared with what recorded at server logs. Finally, tracking is performed unobtrusively: no modification is needed to the browser and/or current HTTP, and no collaboration required from users.

2.2 Analysis

To make the analysis layer scalable, we need to cluster similar paths in order to reason about clusters as opposed to paths. There are a handful of clustering algorithms such as CLARANS[5], BIRCH[6], CLUDIS[7], and K-Means that we can use for this purpose. However, regardless of what clustering algorithm we choose, we have to model both path and cluster, and we have to define a distance function to quantify similarity between paths and clusters. In [8], we introduced a unique model to represent both paths and clusters denoted as *Connectivity Matrix Model (CM-Model).* Via this model, we represent a path/cluster by a set of matrices, each representing a spatial or temporal navigational attribute of the path/cluster. The attributes we are considering in our system are Hit, Sequence, Time and Frequency. The choice of attribute is application-dependent and our model is open to deployment of any other suitable attribute. We aggregate each attribute of the path/cluster into its corresponding matrix in the CM-Model of the path/cluster. Moreover, we introduced a new similarity measure in [4], which suffers from overestimating the similarity between two paths. This is due to the fact that the base sub-paths of the two compared paths that are used for computing their inner product are not orthogonal. To overcome this problem, we have proposed a new similarity measure termed *Vector Angel (VA).* To find out how similar a path is to a cluster in context of each specific attribute, we consider the corresponding matrices of the attribute in their CM-Models as two vectors, and find the angle between these two vectors:

$$VA(\vec{u},\vec{v}) = Cos < \vec{u},\vec{v} >= \frac{\vec{u}.\vec{v}}{|\vec{u}||\vec{v}|}$$

The smaller the value of VA, the more similar the path to the cluster. The final distance between a path and a cluster is computed as the weighted average of VA values over all the attributes. Note that although we are still using the inner product of vectors to define this new similarity measure, the base vectors (unit vectors) are now orthogonal. To be able to make real-time decisions, it is possible to treat the similarity measure as Member Probability (MP), which denotes the probability of a path belonging to a cluster.

Another challenging question with clustering is that should we update the clusters as new paths become available (i.e. more users visit the site)? While updating the clusters makes them adaptive to user behaviour changes, it increases the time complexity of the clustering process drastically. One approach is to apply the clustering algorithm periodically to regenerate or update the existing clusters. In popular web sites such as Yahoo, with Terra-bytes of logged data, this approach will be non-affordably costly. As an alternative, we might be able to cluster the paths in a dynamic manner as follows. When a new path is extracted, its model is compared with current model of all clusters and will be assigned to the closest one. If the new path is not close enough to any cluster, a new cluster is generated with the same model as the path. Once the new path is assigned to its cluster, we

use the path attributes represented by the path CM-Model to update the cluster attributes in the cluster CM-Model dynamically. Thus, the cluster model will reflect both the attributes of the newly added path as well as those of the old paths that had already joined that cluster. This second approach has been implemented within INSITE and will be demonstrated.

2. 3 Interpretation

In [9], we defined a methodology to measure the correlation between the structure of the web site and the user profile by borrowing the concept of channel mutual information from the field of *Information theory*. If there is a strong correlation of the users personal profile with his/ her navigation path, we can statistically predict the answer of the user to a given question (i.e. predict the users behaviour in a certain context) from his/ her interaction with the web site. With INSITE, the soft memberships of the users' paths to the clusters (computed as member probability, MP, in Section 2.2) help generate a footprint for the user in real time. The footprint consists of the clusters that the user may belong to with certain probabilities. By associating each cluster with a set of users' behavioural characteristics with certain probabilities we can predict the behaviour of the user. If the user path belongs to cluster C1 with probability p1 (dynamic, determined in real time), and if anyone associated with cluster C1 is supposed to belong to a certain behavioural group (static, determined in training session) with probability p2, then the user is playing the particular role with a probability of p1 x p2. Thus, the role of the user can be extracted without compromising his/ her privacy. Note that if the same user plays different roles at different times s/he will be clustered into different clusters. This is unlike the approach taken by cookies, which generates static profiles of the users.

3. INSITE: Implementation issues

3.1 Tracking

To make a web site trackable by INSITE, we should embed the INSITE applet agent in each page of the web site. The embedding procedure itself is automated. The agent can be embedded in the dynamic pages by minor modification to the script code and can also be automated. Here, we shall demonstrate tracking of static pages only.
Whenever a visitor views a page, the agent is downloaded in the client machine. Whenever the visitor stops viewing the page (switch to another one or closes the window), the agent sends a single line of text to a java server. Irrespective of where the page is retrieved from (web server, cache or proxy), the agent tracks the page the user is viewing. Agent uses TCP socket to communicate with the server. The agent records the viewing time and the time stamp (of visitation) associated with each page. Each page is identified by a unique combination of its own id and the id of the context to which it belongs. Each user is identified by the host name of the machine s/he is using. For scalability, the server should not be executed in the same machine as the web server. The server buffers the data and periodically stores it in the database as the INSITE log. INSITE tracking is already deployed in three public web sites (digimuse.usc.edu, www.ascusc.org/jcmc/, www.bluesincolor.com). The complete functionality of the INSITE tracking (from user navigation to INSITE log generation and storage) is demonstrated.

3.2 Acquisition

The INSITE log is periodically read into the database. The data is sorted first by host name and then by the timestamp of each entry. Paths are time delimited i.e. two consecutive entries from the same host that are separated by a certain threshold in time indicates the beginning of a new path. Paths are thus readily extractable from the sorted data. Each path goes through a transformation filter [10] that yields the CM-Model of the path. We demonstrate the real time extraction of the CM-Models of the paths from the INSITE log (generated by live navigation of a web site).

3.3 Analysis

Each path is then subjected to a dynamic clustering algorithm that decides the soft membership of the path. Each cluster is represented by a unique CM-Model. We are now testing a number of distance functions to choose the best similarity measure between the CM-Models of a path and a cluster. Experimental validation of the dynamic clustering technique is also underway. However, our preliminary results prove our approach to be very effective in handling the paths and finding their membership to the clusters in real time [10]. Once the membership of a path is decided, the CM-Model of the path updates the CM-Model of the cluster. It also updates the footprint of the user with the id of the cluster with which s/he has been associated. Since a cluster represents certain behaviour, the weighted update of the cluster helps capture the finer changes in a users behaviour (role) within the larger context of the cluster. A persistent, new behaviour from the users gives birth to new clusters. By keeping periodic snapshots of the clusters, we can capture the gradual changes in user behaviour. The database is kept current with the updated CM-Models of the clusters. Due to the very finite number of clusters and the sparse nature of the matrices in the CM-Models, most of the data is kept in memory and does not degrade the performance of the dynamic clustering technique noticeably. We demonstrate the dynamic clustering of the paths and also

the generation of a new cluster in response to new user behaviour.

3.4 Interpretation

The footprint of the user is handed over to the INSITE recommendation engine. The engine queries the database to decide on the salient features of the cluster (most popular pages by hit count, by viewing time, by time stamps etc.) and target the user with customized information accordingly. In our implementation, we have not used the association rules as suggested in [9]. We rather use more straightforward approach of basing our decisions on the salient features of the clusters. During a single session, a visitor can play multiple roles, each captured by the presence of a representative cluster in his/her footprint. We shall demonstrate the feature of extraction of roles by a target advertising application. The footprints are kept in memory. In case of registered users, we can store the footprints in the database for future references.

3.5 Visualization

To facilitate the policy making, INSITE provides the administration and the site owner with a web-based interface for querying the database about the clusters and users (or rather their footprints). We demonstrate this utility through a suite of queries like a) show the clusters that capture the visitors interested in product X and related products, b) show the most (least) recently updated or newborn clusters (also shows the salient features of the clusters), c) show the history of the cluster over a period T, d) show the footprint of all users who have spent over T time on product X, etc.

In our demonstration, the web site is hosted in an apache web server. We use two different browsers (IE and Netscape). The backend database is Oracle. We use a canned version of a public web site for the demonstration.

4. Conclusion

In summary, INSITE tracking captures substantially more information (including accurate temporal information) than the traditional approaches; INSITE acquisition extracts and stores the essence of the captured information in real time by leveraging on the Connectivity matrix model; INSITE analysis possess memory through preserving persistent user behaviors in CM-Models of clusters. Our approach does not compromise the privacy of users; rather it identifies the user through the role s/he is playing. Lastly, our approach is equally applicable to static and dynamic web spaces and works seamlessly with the state of the art load balancing appliances, caching appliances and content mirroring.

5. References

[1] Gomory, S., R. Hoch, J. Lee, M. Podlaseck, and E. Schonberg. Analysis and Visualization of Metrics for Online Merchandizing. In *Proccedings of WEBKDD'99 Workshop on Web Usage Analysis and User Profiling*, San Diego, CA, USA, August,1999.

[2] Cadez, I., D. Heckerman, C. Meek, P. Smyth, and S. White. Visualization of Navigation Patterns on a Web Site Using Model Based Clustering. In *Technical Report MSR-TR-00-18*, Microsoft Research, Microsoft Corporation, Redmond, WA, USA. March 2000.

[3] Mobasher, B., H. Dai, T. Luo, Y. Sun, and J. Zhu. Combining Web Usage and Content Mining for More Effective Personalization. In *Proceedings of the International Conference on E-Commerce and Web Technologies (ECWeb2000*), Greenwich, UK, September 2000.

[4] Shahabi, C., A. Zarkesh, J. Adibi, V. Shah. Knowledge Discovery from Users Web-Page Navigation. In Proceedings of the IEEE RIDE97 Workshop, Apr. 1997.

[5] Raymond, T. Ng. and J. Han. Efficient and effective clustering methods for spatial data mining. In *Proc. of VLDB Conf.*, pages 144-155, September 1994.

[6] Zhang, T., R. Ramakrishnan, and M. Livny. Birch: An efficient data clustering method for very large databases. In *SIGMOD '96*, pages 103-114, Montreal, Canada, June 1996.

[7] Ester, M., H.P. Kriegel, and X. Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. In *Proc. of 4th International Symposium on Large Spatial Databases*, 1995.

[8] Faisal, A., C. Shahabi, M. McLaughlin, F. Betz. INsite: Introduction to a generic paradigm for interpreting user-web space interaction. In *Proceedings of the ACM WIDM*, 1999.

[9] Zarkesh, A., J. Adibi, C. Shahabi, R. Sadri, V. Shah. Analysis and Design of Server Informative WWW-sites. In *Proceedings of the ACM CIKM*, 1997.

[10] Shahabi, C., A. Faisal, F. Banaei Kashani, J. Faruque. A Formal Approach Towards Real-time (and off-line) Analysis of User-Web Space interactions With INSITE. Submitted for publication.