

Periodicity and Web Browsing Behavior: Same Old Same Old?

Yinghui (Catherine) Yang

Graduate School of Management

University of California, Davis

yyang@ucdavis.edu

<http://faculty.gsm.ucdavis.edu/~yyang>

Balaji Padmanabhan

Department of Information Systems and Decision Sciences

College of Business Administration

University of South Florida

bpadmana@coba.usf.edu

<http://www.coba.usf.edu/isds/faculty/padmanabhan/index.html>

Research questions

In our previous research¹ on user identification from user-centric Web browsing data, we show that the identification accuracies improve with aggregating or pooling user data over sessions. The benefits of aggregation though stop after a certain number of sessions are pooled. One explanation for this is periodicity in browsing behavior. For instance, if a certain prototypical behavior is repeated once every x sessions (i.e. is a *periodic* pattern), then aggregating x sessions can expect to improve user models since such windows will contain this periodic pattern. Figure 1 shows an example set of sessions for a user where {cnn.com, myspace.com} is a periodic pattern with period 2. This example also shows {cnn.com} to be “almost” periodic with period 1 (as we will briefly note below our approach will capture this as well).

S1: cnn.com, yahoo.com, myspace.com

S2: cnn.com

S3: cnn.com amazon.com, gmail.com, myspace.com

S4: amazon.com, gmail.com

S5: gmail.com, cnn.com, yahoo.com, myspace.com

Figure 1. *Example set of sessions illustrating periodicity*

In this research we focus on studying periodicity in Web users’ behavior. In addition to extending our prior work the problem is important for other reasons as well. Periodic patterns – or the lack of a learned

¹ Part of the previous research was presented at the 2007 Winter Conference on Business Intelligence.

periodic pattern – can be a useful indicator for online fraud and can help e-businesses predict and prepare (e.g. content caching) for specific user behavior. In this research we address the following questions:

Q1. How can a given pattern be determined to be “periodic”?

Q2. How can periodic patterns be learned efficiently from Web usage data?

Q3. Do users have periodic patterns when they browse online? and,

Q4. How can periodic patterns be used to better predict user behavior?

Approach

At the core of this work is Q1, which is determining if a given pattern is periodic, and to develop subsequent heuristics that might exploit such a definition to build efficient algorithms for learning periodic patterns. For lack of space here we do not survey existing approaches; instead we present the intuition behind our approach.

We start with user-centric data at a given unit of analysis – i.e. a session, or a day, or a week (or any other user defined unit of analysis). For ease of exposition we will assume this to be a session here. Given a specific pattern that can be tested at each session we record the inter-pattern gaps in the data in terms of how many sessions “separate” subsequent occurrences of this given pattern. The variance of the series of inter-pattern gaps measures how “periodic” this pattern is. For instance for patterns that repeat exactly every x sessions the variance is zero, in the case of {cnn.com} in Figure 1 the variance is very small.

Exactly defining a pattern as periodic then can be done in two ways, both of which we examine in this work. First, a variance cutoff can be used. Second, a null distribution for the variance can be computed in a non-parametric manner using randomization. For instance, if the count of an itemset is 120 and the dataset size is 1000, an expected inter-pattern variance can be computed even if each of the 120 occurrences actually occurs at random (i.e. not periodic) throughout the dataset. If the computed inter-pattern variance is outside the $p\%$ left tail of this distribution the pattern can be defined as periodic using this approach.

Both methods (cutoff and distribution-based) are useful. While the first is simple it may actually be of greater use in cases where a periodic pattern occurs at every single session level – in such cases as can easily be seen the null distribution will be a single point which is just the inter-pattern variance. In cases where the periodicity is higher (> 1) the second approach may be preferred for its objective determination of periodicity.

Domain knowledge is important as well in this research. Periodic Web browsing patterns can take various forms, and domain knowledge can help identify the appropriate representation for patterns for which periodicity is to be evaluated. Naturally, Web users' activities can be grouped by session, day, week or month. The unit of analysis clearly forms a key basis for periodicity since patterns are learned (and tested at) this granularity.

Figure 2 presents the method used to discover periodic patterns.

```

Inputs: Data set  $D = \{ S_1, S_2, \dots, S_N \}$  of  $N$  sessions of a user
Pattern discovery algorithm,  $T$ 
Unit of analysis (e.g. session, day, week or month)
Variance threshold:  $v$ 

Output: A list of periodic patterns.

 $L = \{ \}$ 
Generate a list of candidate patterns  $P = \{ P_1, P_2, \dots, P_M \}$  by applying  $T$  to  $D$ .
Compute  $F_i$  as the number of occurrences of pattern  $P_i$  in  $D$ 

For  $u = 1$  to  $M$  {
  Let  $G$  be the inter-pattern gap for pattern  $P_u$  among  $D$ .
  Calculate variance  $V_G$  for  $G$ .
  Calculate null distribution for variance  $\{V_R\}$  for a list of gaps if  $P_u$  is
  randomly distributed among all transactions in  $D$ .
  If  $V_G < v$  or  $V_G$  lies outside of  $\{V_R\}$ 
     $L = L \cup P_u$ 
}
Print sorted list of patterns in  $L$  according to the variance

```

Figure 2. *Periodic Pattern Discovery*

Our approach may be used to develop intelligent pruning heuristics to integrate periodicity determination into specific pattern discovery algorithms. Given that periodic patterns appear evenly throughout the time period, a frequent pattern that appears a lot initially but that

stops for a large period since may be filtered out. As part of this work we will develop specific such heuristics given specific discovery algorithms.

Periodic patterns can help build better prediction models for user behavior. For example, if we know that a user repeats her activity every day, we can take a day as the unit of analysis when building predictive models for this user. Since this user does not repeat her behavior from session to session, taking a session as the unit of analysis may break any natural rhythm in her activities, complicating any prediction tasks for this user.

Expected contributions

Methodologically, this paper contributes to the existing literature on discovering periodic patterns from time series data by providing a variance-based approach for multivariate mixed type time series data. Traditional pattern strength metrics, such as support, are used in this process to both determine that a pattern is significant (as defined by the underlying discovery algorithm) and to compute null distributions as described. The specific application – learning periodic patterns in Web browsing behavior – is novel as well to our knowledge. Any periodic patterns in this domain can be of significant value for a range of applications as noted previously. Incorporating periodic patterns into the prediction task can also potentially help develop a more accurate prediction model.

Current status of the manuscript

We are currently working on user-centric Web browsing data. Initial analysis of the data indicates that there are indeed periodic browsing patterns for several users. We are currently working on heuristics to incorporate into the discovery procedure in the context of patterns defined as itemsets/rules.