Discovery of Significant Usage Patterns from Clusters of Clickstream Data[*]

Lin Lu
(214) 768-1092
llu@engr.smu.edu

Margaret Dunham
(214) 768-3087
mhd@engr.smu.edu

Yu Meng
(214)-768-3080
ymeng@engr.smu.edu

Southern Methodist University
Department of Computer Science and Engineering
Dallas, Texas 75275-0122

ABSTRACT

Discovery of usage patterns from Web data is one of the primary purposes for Web Usage Mining. In this paper, a variation of “user preferred navigational trail” called *Significant Usage Pattern (SUP)* is proposed. SUPs are patterns that are extracted from clustered abstracted clickstream data, with a higher normalized probability of occurrence and may begin/end with specific Web page(s). The novelty of our approach is in the application of clustering to data abstraction based on a new two-phase abstraction technique. In order to generate SUPs, first, the Needleman-Wunsch global alignment algorithm is applied to the sub-abstracted sessionized clickstream data to compute the similarities between each pair of sessions. Based on pair-wise alignment results, a similarity matrix is constructed and then sessions are grouped into clusters according to their similarities. Web sessions are abstracted again using a concept-based abstraction approach and then a first order Markov model is built for each cluster of sessions. The specific navigation paths, i.e. SUPs, with a normalized product of probability along the path above a certain threshold and beginning/ending with specific states are generated from each cluster based on its corresponding Markov model. Experiments conducted using Web log data provided by J.C.Penney show that different clusters of Web sessions may generate very different SUPs.

Keywords

Web Usage Mining, Data Mining, Patterns, Clickstream

1. INTRODUCTION

The detailed records of Web data, such as Web server logs, referrer logs, and so forth, provide enormous amounts of user information. Hidden in these data are the valuable information that implies users’ interests and motivations for visiting a specific website. Research in this area is categorized as *Web Usage Mining (WUM)* [10]. WUM is a branch of Web mining that focuses on applying data mining techniques to discover useful knowledge of user navigation patterns from Web data. It is aimed at improving the Web design and developing corresponding applications to better serve the needs of both users and website owners [22].

There are varieties of usage patterns that have been investigated to examine the Web data from different perspectives and for various purposes. For instance, the maximal frequent forward sequence mines the forward traversal patterns which are maximal and with the backward traversal removed in pattern discovery [8], the maximal frequent sequence examines the sequences that have high frequency of occurrence as well as maximum in length [24], sequential pattern explores the sequence with certain support and is maximal [1], user preferred navigational trail extracts user preferred navigation paths [4] [5], and so forth.

In this paper, a new data mining methodology that involves exploring the *Significant Usage Patterns (SUP)* is introduced. A SUP is defined as a path that is extracted from a Markov model associated with each cluster of user sessions. It may have specific beginning and/or ending states and its corresponding normalized product of probability along the path satisfies a given threshold. SUP is a variation of “user preferred navigational trail” [4] [5]. Comparing to early work, SUP differs in the following four aspects: (1) SUP is extracted from clusters of abstracted user sessions. (2) Users may identify desired beginning and/or ending Web pages for the generated SUPs, for example, the patterns that lead to purchase on a commercial Web site. (3) SUPs are patterns with normalized probability, which makes it easier for users to determine the probability threshold for identifying corresponding patterns. (4) SUP uses a unique two-phase abstraction technique (see sections 3.1 & 3.3).

The focus of this paper is on abstracting and clustering of user sessions as well as extracting SUPs. We assume that the clickstream data has already been sessionized.

The rest of the paper is organized as follows. Section 2 discusses the related work. The methodology related to the alignment, abstraction, and clustering of Web sessions is provided in Section 3. Section 4 gives the analysis of experimental results performed using Web log data provided by J. C. Penney. Finally, some discussion of findings in this study as well as some perspectives for future research in this direction conclude the paper.

WebKDD’05, August 21, Chicago, Illinois, USA.
Copyright 2005 ACM, ISBN: 1-59593-214-3, $5.00.

[*] This work is supported by the National Science Foundation under Grant No. IIS-0208741

2. RELATED WORK

Work relevant to the three main steps involved in mining SUPs: URL abstraction, clustering user sessions of clickstream data, and generating usage patterns, are discussed in detail in the following subsections. We conclude each subsection with a brief examination of how our work fits into the literature.

2.1 URL Abstraction

URL abstraction is the process of generalizing URLs into higher level groups. Page-level aggregation is important for user behavior analysis [22]. In addition, it may lead to much more meaningful clustering results [2]. Since behavior patterns reflected in user sessions often consist of a sequence of low level page views, there is no doubt that pattern discovery using exact URLs will show fewer matches among user sessions, than if abstraction of these pages were performed. Web page abstraction allows the discovery of correlations between user sessions that are frequent enough at an abstract concept level but which may rarely occur at page level. In fact, a lot of pages in a specific web site are usually semantically equivalent which makes web page generalization possible.

In [2], concept-category of page hierarchy was introduced, in which web pages were grouped into categories, called concepts, based on proper analytics and/or metadata information. Since this approach categorized web pages using only the top-most level of the page hierarchy, it could be viewed as a simpler version of generalization-based clustering. A generalization-based page hierarchy was described in [12]. According to this approach, each page was generalized to its higher level. For instance, pages under /school/department/courses would be categorized to "department" pages and pages under /school/department would be classified as "school" pages. While Spiliopoulou et al. employed a content-based taxonomy of web site abstraction, in which taxonomy was defined according to task-based model and each web page was mapped to one of the taxonomy's concepts [21]. In [18], pages were generalized to three categories, namely administrative, informational, and shopping pages, to describe an online nutrition supply store.

In our study, two different abstraction strategies are applied to user sessions before and after clustering process in our model respectively. First, user sessions are sub-abstracted before applying the clustering algorithm to make the sequence alignment approach more meaningful. After clustering user sessions, a concept-based abstraction approach is applied to user sessions in each cluster, which allows us to have more insight into the SUPs associated with each cluster. Both abstraction techniques are based on a concept hierarchy provided about the site.

2.2 Clustering User Sessions of Clickstream Data

In order to mine useful information from user navigation patterns from clickstream data, it is more appropriate to cluster user sessions first. The purpose of clustering is to find groups of users with similar interests and objectives for visiting a specific website. Actually, the knowledge of user groups with similar behavior patterns is extremely valuable for e-commerce applications. With this kind of knowledge, domain experts can infer user demographics in order to perform market segmentations [22].

Various approaches have been introduced in the literature to cluster user sessions [2] [7] [12] [16] [23]. [7] used a mixture of first-order Markov models to partition user sessions with similar navigation patterns into the same cluster. In [12], page accesses in each user session were substituted by a generalization-based page hierarchy scheme. Then, generalized sessions were clustered using a hierarchical clustering algorithm, BIRCH.

Banerjee et al. developed an algorithm that combined both the time spent on a page and *Longest Common Subsequences (LCS)* to cluster user sessions [2]. The LCS algorithm was first applied on all pairs of user sessions. After each LCS path was compacted using a concept-category of page hierarchy, similarities between LCS paths were computed as a function of the time spent on the corresponding pages in the paths weighted by a certain factor. Then, it built an abstract similarity graph for the set of sessions to be clustered. Finally, a graph partition algorithm, called Metis, was used to segment the graph into clusters.

The clustering approach discussed in [16] [23] was based on the sequence alignment method. They took the order of page accesses within the session into consideration when computing the similarities between sessions. More specifically, they used the idea of sequence alignment from bio-informatics to measure the similarity between sessions. Then, sessions were clustered according to their similarities. In [16], Ward's clustering method [15] was used, while [23] applied three clustering algorithms, ROCK [13], CHAMELEON [17], and TURN [11].

The clustering approach used in our work is based on [16] [23], however, a category abstraction of page hierarchy is first considered in measuring the similarities between Web pages. Then the Needleman-Wunsch global alignment algorithm [20] is used to align the sessionized clickstream data based on the page wise similarities and to compute the optimal alignment score between sessions. Finally, the nearest neighbor clustering algorithms is applied to the similarity matrix, resulted from the global alignment of sessions, to cluster the user sessions.

2.3 Generating Usage Patterns

Varieties of browsing patterns have been investigated to examine Web data from different perspectives and for various purposes, such as the maximal frequent forward sequence [8], the maximal frequent sequence [24], the sequential pattern [1], user preferred navigational trail [4] [5], and so forth.

The usage pattern proposed in [4] [5] are the most related to our research. [4] proposed a data-mining model to extract the higher probability trails which represent user preferred navigational paths. In that paper, user sessions were modeled as a *Hypertext Probabilistic Grammar (HPG)*, which can be viewed as an absorbing Markov chain, with two additional states, start (S) and finish (F). The set of strings generated from HPG with higher probability are considered as preferred navigation trails of users. The depth first search algorithm was used to generate the trails given specific support and confidence thresholds. Support and confidence thresholds were used to control the quality and quantity of trails generated by the algorithm. In [5], it proved that the average complexity of the depth first search algorithm used to generate the higher probability trails is linear in the number of web pages accessed.

In our approach, SUPs are higher probability trails that are extracted from clusters of abstracted user sessions, with

Table 1. A comparison

	Clustering	Abstraction	Beginning/ending Web page(s)	Normalized
Sequential Pattern	N	Y*	N	-
Maximal Frequent Sequence	N	N	N	-
Maximal Frequent Forward Sequence	N	N	N	-
User Preferred Navigational Trail	N	N	N	N
Significant Usage Pattern	Y	Y	Y	Y

*Some apply abstraction, such as in [6], while others not, for instance [19].

normalized probabilities of occurrence which make them independent of their length. In addition, SUPs may begin and/or end with specific Web pages of user interests. Table 1 provides the detailed comparison of SUPs with other patterns.

3. METHODOLOGY

To generate SUPs, first, a sequence alignment [16] [23] approach based on the Needleman-Wunsch global alignment algorithm [20] is applied to the sessionized abstracted clickstream data to compute the similarities between each pair of sessions. This approach preserves the sequential relationship between sessions, which reflects the characteristics of chronological sequential order associated with the clickstream data. Based on the pair-wise alignment results, a similarity matrix is constructed and then original un-abstracted sessions are grouped into clusters according to their similarities. By applying clustering on sessions, we are more likely to discover the common and useful usage patterns associated with each cluster. Then, the original Web sessions are abstracted again using a concept-based abstraction approach and then a first order Markov model is built for each cluster of sessions. Finally, the SUPs with normalized product of probability along the path that above a given threshold are extracted from each cluster based on their corresponding Markov model. This process is illustrated in Figure 1. A more detailed description of each step is provided in the following subsections.

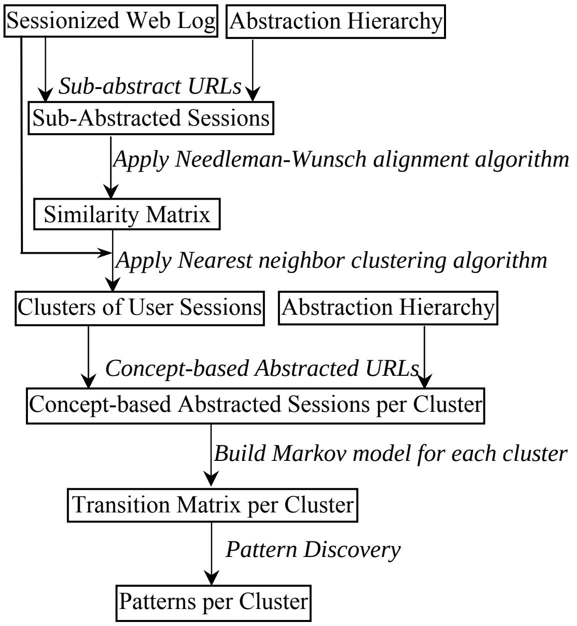

Sessionized Web Log
Abstraction Hierarchy
Sub-abstract URLs
Sub-Abstracted Sessions
Apply Needleman-Wunsch alignment algorithm
Similarity Matrix
Apply Nearest neighbor clustering algorithm
Clusters of User Sessions
Abstraction Hierarchy
Concept-based Abstracted URLs
Concept-based Abstracted Sessions per Cluster
Build Markov model for each cluster
Transition Matrix per Cluster
Pattern Discovery
Patterns per Cluster

Figure 1. Model to generate SUPs

3.1 Create Sub-abstracted Sessions

In this study, we assume that the Web data has already been cleansed and sessionized. Detailed techniques for preprocessing the Web data can be found in [9].

A Web session is a sequence of accessed Web pages by a single user. However, for the sequence alignment result to be more meaningful, we abstract the pages to produce sub-abstracted sessions. We use the term “sub-abstracted” session instead of “abstracted” session, because we do not use a typical abstraction approach, but rather a concept-based abstraction hierarchy, e.g., Department, Category, and Item in e-commerce Web site, plus some specific information, such as Department ID, Category ID in the abstracted session. With this approach, we can preserve certain information to make Web page similarity comparison more meaningful for session alignment described below. A URL in a session is mapped to a *sub-abstracted URL* as follows:

URL -> {<Concept hierarchy keyword> <Unique ID> <|>}

Example 1: Based on the hierarchical structure of J.C. Penney’s Web site, each Web page access in the session sequence is abstracted into three levels of hierarchy, as shown in Figure 2, where D, C, I are the initials for Department, Category, and Item respectively, 1, 2, …, n represent IDs, and vertical bar | is used to separate different levels in the hierarchy.

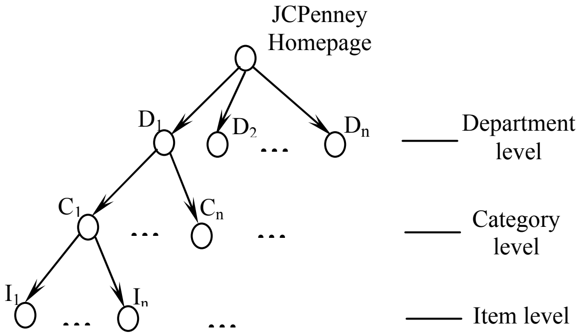


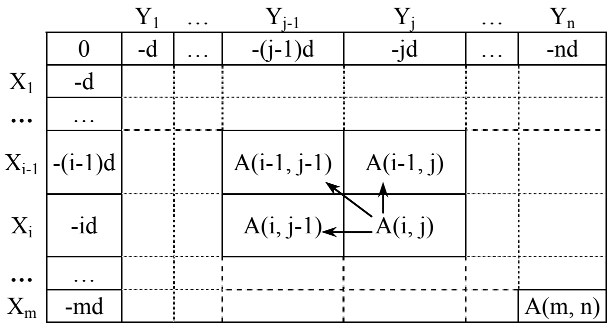
Figure 2. Hierarchy of J.C. Penney Web site

The following is an example of a sub-abstracted session with the last negative number representing the session id (Web pages that do not belong to any department are abstracted as P which stands for general page):

D0|C875|I D0|C875|I P27593 P27592 P28 -507169015

3.2 Session Sequence Alignment

The Needleman-Wunsch alignment algorithm [20] is a dynamic programming algorithm. The basic idea of computing the optimal alignment of two sequences, $X_1…X_m$ and $Y_1…Y_n$, using Needleman-Wunsch alignment algorithm is illustrated in Figure 3. Suppose A(i, j) is the optimal alignment score of aligning $X_1… X_i$ with $Y_1… Y_j$. If we know the alignment scores of A(i-1, j-1), A(i-1,j), and A(i, j-1), then A(i, j) can be computed as A(i, j) = max[A(i-1, j-1)+s(X_i, Y_j); A(i-1, j)+d; A(i, j-1)+d], where s(X_i, Y_j) is the similarity between X_i and Y_j, d is the score of aligning X_i with a gap or aligning Y_j with a gap. That is, an entry A(i, j) depends on three other entries as illustrated in Figure 3. Therefore, we can carry out the computation from upper left corner to lower right corner, A(m,n), which is the optimal alignment score between $X_1…X_m$ and $Y_1…Y_n$. Initially, as shown in Figure 3, set: (1) A(0,0)=0, since it corresponds to align two empty strings of X and Y; (2) A(i, 0)=-d*i, for i = 1…m, which corresponds to align the prefix $X_1... X_i$ with gaps; (3) Similarly, A(0,j)=-d*j, for j=1…n.



		Y_1	…	Y_{j-1}	Y_j	…	Y_n
	0	-d	…	-(j-1)d	-jd	…	-nd
X_1	-d						
…	…						
X_{i-1}	-(i-1)d			A(i-1, j-1)	A(i-1, j)		
X_i	-id			A(i, j-1)	A(i, j)		
…	…						
X_m	-md						A(m, n)

Figure 3. Computing optimal alignment of two sequences using Needleman-Wunsch algorithm

When taking the hierarchical representation of Web pages into consideration, it is reasonable to assume that higher levels in the hierarchy, which have more importance in determining the similarity of two Web pages, should be given more weight. To reflect this in the scoring scheme, first, the longer page representation string in the two Web page representations is determined. Then, a weight is assigned to each level in the hierarchy and its corresponding ID (if any) respectively: the lowest level in longer page representation string is given weight 1 to its ID and weight 2 to its abstract level, the second to the lowest level is given weight 1 to its ID and weight 4 to its abstract level, and so forth. Finally, the two Web page representation strings are compared from the left to the right and stopped at the first pair which they are different. The similarity between two Web pages is determined by the ratio of the sum of the weights of those matching parts to the sum of the total weights. The following is an example of computing the similarities between two Web pages:

Page 1: D0|C875|I	weight=6+1+4+1+2=14
Page 2: D0|C875	weight=6+1+4+1=12
Similarity=12/14=0.857	

Therefore, the similarity value of two Web pages is between 0 and 1, the similarity is 1 when two Web pages are exactly the same, and 0 while two Web pages are totally different.

The scoring scheme used in this study for computing the alignment of two session strings is the same as in [23]. It is defined as follows:

```
if matching     //a pair of Web pages with similarity 1
    score = 20;
else if mis-matching    //a pair of Web pages with similarity 0
    score = –10;
else if gap      //a Web page aligns with a gap
    score = –10;
else  //the pair of Web pages with similarity between 0 and 1
    score = –10 ~ 20;
```

Then, the Needleman-Wunsch global alignment algorithm can be applied to the sub-abstracted Web session data to compute the score corresponding to the optimal alignment of two Web sessions. This is a dynamic programming process which uses the Web page similarity measurement mentioned above as a page matching function. Finally, the optimal alignment score is normalized to represent the similarity between two sessions:

$$\text{Session similarity} = \frac{optimal\ \ alignment\ \ score}{length\ \ of\ \ longer\ \ session}$$

Example 2: Figure 4 provides an example for computing the optimal alignment and the similarity for the following two Web sessions (session ids are ignored in the alignment):

P47104 D0|C0|I D469|C469 D2652|C2652
D469|C16758|I D0|C0|I D469|C469

Thus, the optimal alignment score is 32.1 and the session similarity = 32.1/4 = 8.025

		P47104	D0|C0|I	D469|C469	D2652|C2652
	0	-10	-20	-30	-40
D469|C16758|I	-10	-10	5.7	-4.3	-14.3
D0|C0|I	-20	-20	10	17.1	7.1
D469|C469	-30	-30	0	30	32.1

Figure 4. Computing Web session similarity for Example 2

3.3 Create Concept-based Abstracted Sessions

After the original Web sessions are clustered according to the similarity matrix constructed by the sequence alignment approach, Web sessions are abstracted again using a concept-based abstraction approach.

In this approach, we adopt the same abstraction hierarchy introduced in Section 3.1, which contains Department (D), Category (C), Item (I), and General page (P) in the hierarchy. However, the abstracted page accesses in a session will be represented as a sequence like: $P_1D_1C_1I_1P_2D_2C_2I_2…$, in which each of Pi, Di, Ci, and Ii (i=1, 2…) represents a different page. For example, D_1 (element) and D_2 (element) indicate two different departments. The same applies to Pi, Ci and Ii. In addition, it is also important that for different sessions, the same page may be represented by different elements. For example, shoes department may be represented by D_1 in one session, while be represented by D_2 in another session. The definition of element is based on the sequence of page accesses appeared in a session. In the Markov model, each of these elements will be treated as a state.

A URL in a session is mapped to a *concept based abstracted URL* as follows:

> URL -> <Concept hierarchy keyword> <Unique ID for this concept in this session>

Thus each URL is associated with a lowest level of concept in representing that URL in the concept hierarchy and a unique ID for that specific URL within the session.

By abstracting Web sessions in such a way, it allows us to ignore the irrelevant or detailed information in the dataset while concentrated on more general information. Therefore, it is possible for us to find the general behavior in a group as well as to identify the main user groups.

Example 3: The example given below illustrates the abstraction process in this step (the last negative number representing the session id):

> Original session: D7107|C7121 D7107|C7126|I076bdf3 D7107|C7131|I084fc96 D7107|C7131 P55730 P96 P27 P14 P27592 P28 P33711 -505884861
> Abstracted session: C1 I1 I2 C2 P1 P2 P3 P4 P5 P6 P7 -505884861

3.4 Clustering of Web Sessions

Based on the pair wise session similarity results computed according to above-mentioned techniques, a Web session similarity matrix is constructed. Then, a clustering algorithm can be applied to the matrix to generate clusters. For simplicity, the nearest neighbor clustering algorithm is used in this study. A detailed example of this algorithm can be found in [10].

3.5 Generating Significant Usage Patterns

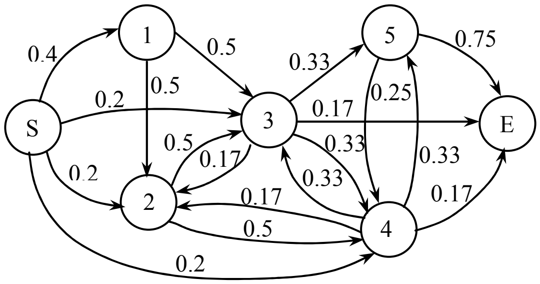
Upon generating clusters of Web sessions, we represent each cluster by a Markov model. The Markov model consists of a set of states and a transition matrix. Each state in the model represents a concept-based abstracted Web page in a cluster of Web sessions, except for two additional states, the "start" state and the "end" state. The transition matrix contains the transition probability between states. Example 4 illustrates this step.

Example 4: Figure 5(a) contains a list of concept based abstracted sessions in a cluster. Assume that each number in the session sequence stands for an abstracted Web page, and it is represented as a state in the Markov model. In addition, a *Start (S)* and an *End (E)* states are introduced in the model and treated as the first and the last states for all sessions in the cluster respectively. Figure 5(b) shows the corresponding Markov model for the sessions listed in Figure 5(a). The weight on each arc is the transition probability from the state where the arc going out to the state where the arc pointing to. The transition probability is computed as the number of corresponding transition occurred divided by the total number of out transitions from the state where the arc leaving.

The Markov model mentioned here is a first-order Markov model. Using the Markov property, the model assumes that the Web page a user visits next is fully dependent on the content of Web page that the user is currently visiting. Research shows that this is reasonable for Web page prediction [14]. The transition matrix for the Markov model records all the user navigation activities within the Web site and it will be used to generate SUPs.

(1) 1, 2, 3, 5, 4
(2) 2, 4, 3, 5
(3) 3, 2, 4, 5
(4) 1, 3, 4, 3
(5) 4, 2, 3, 4, 5

(a)



(b)

Figure 5. Example of building a Markov model for a cluster of abstract sessions

Definition 1: A "*path*" is a totally ordered sequence of states from the Markov Model. The first state in the sequence is identified as the "*beginning state*", while the terminating state is called the "*end state*".

Definition 2: Given a path in the Markov Model, the "*probability of a path*" is:

> *Case 1 (Beginning state identified by user)*: Product of transition probabilities found on all transitions along the path, from beginning to end state.
> *Case 2 (Beginning state not given):* Product of transition probabilities found on all transitions along the path times the transition probability from the Markov Model *Start* state to the beginning state in the path.

Suppose, there exits a path $S_1 \rightarrow S_2 \rightarrow \ldots S_i \rightarrow \ldots \rightarrow S_n$, according to *Definition 2*, the probability of the path, P, is defined as:

$$P = \prod_{i=1}^{n-1} P_{t_i}$$, where Pt_i is the transition probability between two adjacent states.

To illustrate the two cases stated in the definition, we use Example 4, path 1→2→3→4. If state 1 is given by the user, the probability of this path is 0.5×0.5×0.33=0.0825; otherwise, the probability is 0.4×0.5×0.5×0.33=0.033. The purpose of distinguishing between these two scenarios is that: (1) *Case 1*: if a user only gives the end Web page, we assume that the user is more interested in the patterns that lead to that specific end page from the very beginning where Web visitors entering the Web site; (2) *Case 2*: if a user provides both beginning and ending Web pages, we assume that user is more interested in viewing patterns occurring between those two pages.

Considering that the final probability of a path is exponential to the length of the path, in order to set a general rule to specify the probability threshold for generated paths, it is necessary to normalize the probability of a path to eliminate the exponential factor. Therefore, the normalized probability of the path, P_N, is defined as:

$$P_N = \left(\prod_{i=1}^{n-1} P_{t_i} \right)^{\frac{1}{n-1}}$$

, where Pt_i is the transition probability between two adjacent states.

Definition 3: A SUP is a path that may have specific beginning and/or end states, and its corresponding normalized probability greater than a given threshold θ, that is, $P_N > \theta$.

Example 5: To illustrate the concept of SUP, again, we use the above example. Suppose we are interested in patterns with θ > 0.4, ending in state 4, and under two different cases, one is beginning with state 1 and the other one leaves the beginning state undefined. The corresponding SUPs generated under those two circumstances are listed in Table 2. They are generated based on the transition matrix using Depth-first search algorithm.

Table 2. Example of SUPs

θ > 0.4, end state is 4		θ > 0.4, beginning state is 1, end state is 4	
SUP	θ	SUP	θ
S→1→2→3→4	0.45	1→2→3→4	0.46
S→1→2→3→5→4	0.53	1→2→3→5→4	0.56
S→1→2→4	0.46	1→2→4	0.5
S→1→3→4	0.43	1→3→4	0.45
S→1→3→5→4	0.53	1→3→5→4	0.58
S→2→3→5→4	0.45		
S→3→5→4	0.43		

4. EXPERIMENTAL ANALYSIS

4.1 Clickstream Data

The clickstream data used in this study was provided by J. C. Penney. The whole dataset contains one day's Web log data from their online site (jcpenney.com) on 10/5/2003. However, the data itself is not the pure raw log data, instead each recorded click information has already been broken down into several pieces of information, such as category ID, department ID, item ID, session ID, and so forth.

On this specific day, 1,463,180 visit sessions were recorded which consisted of 8,554,665 lines of page view data. However, after removing the sessions generated by robots, we ended up with 593,223 sessions. Then, sessions were divided into two super-clusters: those which contain a purchase and those that do not. The experiments conducted here use the first 2,000 sessions from both purchase and non-purchase clusters, since we assume that sessions from different time frame within a day are equally distributed. Moreover, 2,000 sessions from each cluster is large enough for us to demonstrate the results. Future research will examine techniques to effective sample large Web logs.

4.2 Result Analysis

The range of scores in the similarity matrix that were generated from applying Needleman-Wunsch global alignment algorithm to the sub-abstracted Web sessions is from -9.4 to 20 for the purchase cluster and –10 to 20 for the non-purchase cluster. The average scores are 3.3 and –0.8 for purchase and non-purchase clusters respectively. These are consistent with the scoring scheme used in this study which defines the similarity score between -10 and 20. After trying different thresholds for the nearest neighbor clustering algorithm, we found that using thresholds 3 and 0 for purchase and non-purchase sessions respectively, both of them result in 3 clusters which give better clustering results. The average session length in the resulting clusters for both purchase and non-purchase clusters are shown in Figure 6. From the figure, it is obvious that purchase sessions are longer than those sessions without purchase on average. It illustrates that users usually request more page views when they are about to purchase something than users who just visit an online store without purchasing. This can be explained by the fact that users normally would like to review the information as well as to compare the price, the quality and etc. for the product(s) of their interest before buying them. In addition, users need to fill out the billing and shipping information as well to commit the purchase. All these factors could lead to longer purchase session.

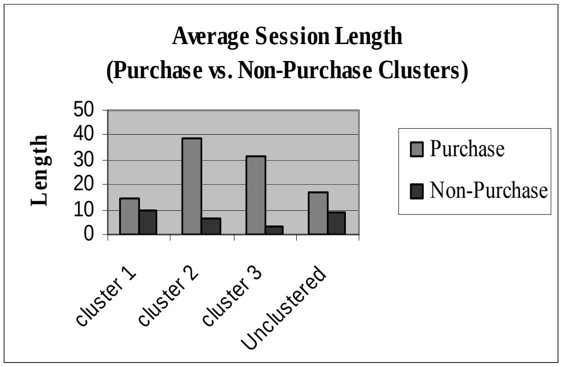


Figure 6. Average session length

Table 3 lists the SUPs generated from the three different clusters in the non-purchase super-cluster. In order to limit the number of SUPs generated from each cluster, we applied different probability threshold to each cluster. From the results in Table 3, it is easy to distinguish patterns among three clusters. In cluster 1, users spend most of their time browsing between different categories. By looking into the sessions in this cluster, we notice that most of the sessions request product pages at some point. However, these kinds of patterns are not dominant when we require a threshold θ>0.3. When we lowered the threshold to θ > 0.25, the generated SUPs also include the following:

S-C1-C1-C2-C3-C4-C5-C5-I1-E
S-C1-C1-I1-C1-C2-C3-C4-C5-E
S-I1-C1-C2-C3-C4-C5-C6-C7-E

Based on the above result, we assume that users in this group are more interested in gathering information of products in different categories. While users in cluster 2 are more interested in reviewing general pages (to gather general information), although some of them may also request some categories and products pages, as shown in the SUPs below (θ > 0.3):

S-P1-P2-P3-C1-I1-E
S-P1-P2-P3-P4-P5-P6-C1-C2-E
S-P1-P2-P3-P4-P5-C4-I6-I7-I8-E

As for cluster 3, since the average session length in this cluster is only 3, we assume users in this group are not serious visitors. This is reflected in their behavior patterns in that after they come to the

Table 3. SUPs in non-purchase cluster

Cluster No.	No. of Sessions	Threshold (θ)	Average Session Length	No. of States	SUPs
1	1746	0.3	9.6	98	1. S-C_1-C_1-C_2-C_3-C_4-C_5-C_6-C_7-E 2. S-C_1-C_1-C_2-C_3-C_4-C_5-E 3. S-C_1-C_1-C_2-C_3-E 4. S-C_1-C_2-C_3-C_3-C_4-C_5-C_6-C_7-E 5. S-C_1-C_2-C_3-C_4-C_4-C_5-C_6-C_7-E 6. S-C_1-C_2-C_3-C_4-C_5-C_5-C_6-C_7-E 7. S-C_1-C_2-C_3-C_4-C_5-C_6-C_6-C_7-E 8. S-C_1-C_2-C_3-C_4-C_5-C_6-C_7-C_7-E 9. S-C_1-C_2-C_3-C_4-C_5-C_6-C_7-C_8-E 10. S-C_1-C_2-C_3-C_4-C_5-C_6-C_7-E 11. S-C_1-C_2-C_3-C_4-C_5-C_6-E 12. S-C_1-C_2-C_3-C_4-C_5-E 13. S-C_1-C_2-C_3-C_4-E 14. S-C_1-C_2-C_3-E
2	241	0.37	6.6	38	1. S-P_1-P_2-P_3-P_3-E 2. S-P_1-P_2-P_3-P_4-P_4-P_5-E 3. S-P_1-P_2-P_3-P_4-P_4-E 4. S-P_1-P_2-P_3-P_4-P_5-P_4-E 5. S-P_1-P_2-P_3-P_4-P_5-P_5-E 6. S-P_1-P_2-P_3-P_4-P_5-P_6-C_1-E 7. S-P_1-P_2-P_3-P_4-P_5-P_6-P_7-E 8. S-P_1-P_2-P_3-P_4-P_5-P_6-E 9. S-P_1-P_2-P_3-P_4-P_5-E 10. S-P_1-P_2-P_3-P_4-C_1-E 11. S-P_1-P_2-P_3-P_4-E 12. S-P_1-P_2-P_3-C_1-E 13. S-P_1-P_2-P_3-E 14. S-P_1-P_2-E
3	13	0.3	3.0	6	1. S-C_1-P_1-P_1-P_2-E 2. S-C_1-P_1-P_1-E 3. S-C_1-P_1-P_2-E 4. S-C_1-P_1-E 5. S-I_1-P_1-P_1-P_2-E 6. S-I_1-P_1-P_1-E 7. S-I_1-P_1-P_2-E 8. S-I_1-P_1-E

Web site for one category or product page and then a couple of general pages, they leave. The corresponding BNF expressions of the SUPs in these three clusters are given in Table 4. In the BNF representation, we ignore the subscript in corresponding P, D, C, and I. BNF notation proves to be a valuable tool to label the significant patterns from each cluster.

We examine SUPs beginning at a specific page, P86806. In the three generated clusters in the non-purchase group, the form of patterns is similar to those which start from "Start" (S) page in the corresponding cluster. Their BNF expressions are given in Table 4.

The SUPs (in BNF notation) generated from the three clusters in purchase group are provided in Table 4 as well. Users in cluster 1 appears to be direct buyers, since the average session length in this cluster is relatively short (14.9) compared to the other two clusters in the purchase group. Customers in this cluster may come to the Web site, pick up the items(s) they want, and then may fill out the required information and leave. The following are some sample SUPs from cluster 1:

S-C_1-I_1-P_1-P_2-P_3-P_4-P_5-P_6-P_7-P_8-P_9-P_{10}-P_{11}-P_{12}-E
S-P_1-P_2-P_3-P_4-P_5-P_6-P_7-P_8-P_9-P_{10}-P_{11}-P_{12}-P_{11}-E
S-I_1-P_1-P_2-P_3-P_4-P_5-P_6-P_7-P_8-P_9-P_{10}-P_{11}-P_{12}-P_{13}-E

SUPs in cluster 2 show that shoppers in this cluster may like to compare the product(s) of their interests or have a long shopping list, since they request many category pages before going to the general pages (possibly for checking out). An example SUP from this cluster is given below:

S-C_1-C_2-C_3-C_4-C_5-C_6-C_7-C_8-C_9-C_{10}-C_{11}-C_{12}-C_{13}-C_{14}-C_{15}-C_{16}-C_{17}-C_{18}-C_{19}-C_{20}-C_{21}-C_{22}-C_{23}-P_4-P_5-P_6-P_7-P_8-P_9-P_{10}-P_{11}-P_{12}-P_{13}-P_{14}-P_{15}-P_{16}-P_{17}-P_{18}-P_{19}-P_{20}-E

Customers in cluster 3 are more like hedonic shoppers, since the patterns show that they first go through several general pages, and then suddenly go to the product pages (probably for purchase) which may stimulated by some information provided in general pages. The following is a sample SUP from this cluster:

S-P_1-P_2-P_3-P_4-P_5-P_6-P_7-P_8-P_9-P_{10}-I_{13}-I_{14}-I_{15}-P_{10}-P_{11}-P_{12}-P_{16}-P_{15}-P_{17}-P_{18}-P_{19}-C_1-C_2-C_3-C_4-C_5-C_6-E

Table 4. Clusters in non-purchase vs. purchase

Cluster	Cluster No.	No. of Sessions	Average Session Length	No. of States	Threshold (θ)	Beginning Web page	SUPs in BNF Notation
Non-Purchase	1	1746	9.6	98	0.3	S	S-{C}-E
					0.25	P86806	P86806-{C}-E
	2	241	6.6	38	0.37	S	S-{P}-[C]-E
					0.34	P86806	P86806-[I]-{P}-E
	3	13	3.0	6	0.3	S	S-<C | I>-{P}-E
					0.2	P86806	P86806-[{P}- [P86806]]-E
Purchase	1	1858	14.9	55	0.47	S	S-[C]-[I]-{P}-E
					0.51	P86806	P86806-[I]-{P}-E
	2	132	39.1	100	0.457	S	S -[{{C}|{I}}]-{P}-E
					0.434	P86806	P86806-[{C }]-{P}-E
	3	10	31.6	47	0.52	S	S-{P}-[{I}]-[{P}]-{C}-E
					0.43	P86806	P86806-[I]-[{P}]-{C}-E

For SUPs starting from page P86806 in the purchase group, a similar pattern is shown to those that start from "Start" (S) page in the corresponding cluster. The BNF expressions of their SUPs are provided in Table 4.

When comparing between SUPs in both purchase and non-purchase clusters, we notice two main differences between them: (1) The average length of SUPs is longer in the purchase cluster than in non-purchase cluster; (2) SUPs in the purchase cluster have higher probability than those in non-purchase cluster. The interpretation for the first difference is as mentioned above, it might due to the fact that users may review the information, compare among products, and fill out the payment and shipping information. The possible explanation for the second phenomenon is that users in the purchase cluster may already have some specific product(s) to purchase in mind when they visit the Web site. Therefore, they may show the similar search patterns, products compare patterns, and purchase patterns, which make the SUPs in purchase cluster having higher probability. While users in the non-purchase cluster may have a more random browsing behavior, since they may not have a specific purpose for visiting the Web site.

From the above result, we can see that SUPs associated with different clusters are different and meaningful. Therefore, by clustering user sessions, it also results in clusters of patterns which may benefit for us to find groups of users with similar interests and motivations for visiting a specific website. In addition, given the flexibility of specifying specific beginning and/or ending Web pages, users can more freely to investigate the patterns of their interests.

5. CONCLUSION AND FUTURE WORK

In this study, we presented a framework to generate Significant Usage Patterns (SUP), which are based on the "user preferred navigational trails" [4] [5]. However, SUPs are generated from clusters of abstracted Web sessions, with normalized probability of occurrence higher than a certain threshold and may have user specified the beginning and/or ending Web page(s). By applying clustering to abstracted user sessions, it is more likely to find groups of users with similar motivations for visiting a specific website. By giving the flexibility for user to specify the beginning and/or ending Web page(s), users can have more control in generating patterns of their interests. While after normalizing probability for SUPs, it makes user much easier to determine the probability threshold for identifying corresponding patterns. The experiments conducted using J.C.Penney's Web data shows that different SUPs associated with different clusters of Web sessions. In addition, to a certain extent, SUPs reflect the interests of users for visiting the J. C. Penney's Web site in the corresponding cluster.

To extend this study, different clustering algorithms can also be examined in order to find an optimal one. In addition, patterns in different clusters can be studied in more detail. It will be interesting to see if they can be used to identify different user groups that defined by user applications by developing some techniques. This kind of information will be valuable for Web owners, especially for e-commerce Website, to design different Web pages to target different user groups. Furthermore, by classifying users to predefined clusters based on their current behavior for Web page prediction will be another research direction. Future work can also examine techniques for sampling data after sessionizing to reduce the overhead of the pattern generation process.

6. ACKNOWLEDGEMENTS

Authors would like to thank Dr. Saad Mneimneh for many useful discussions on this study. In addition, we also would like to thank J.C. Penney for providing the dataset for this research.

7. REFERENCES

[1] Agrawal, R. and Srikant, R. *Mining Sequential Patterns*. In Proc. 11 Intl. Conf. On Data Engineering, Taipi, Taiwan, March 1995.

[2] Banerjee, A. and Ghosh, J. *Clickstream Clustering using Weighted Longest Common Subsequences***.** In Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining (Chicago IL, April 2001), 33-40.

[3] Berkhin, P. *Survey of Clustering Data Mining Techniques*. Accrue Software, Technical Report, 2002.

[4] Borges, J. and Levene, M. *Data Mining of User Navigation Patterns*. In Proc. the Workshop on Web Usage Analysis and User Profiling (WEBKDD'99), 31-36, San Diego, August 15, 1999.

[5] Borges, J. and Levene, M. *An average linear time algorithm for web data mining*. International Journal of Information Technology and Decision Making, 3, (2004), 307-320.

[6] Buchner, A. G., Baumgarten, M., Anand, S. S., Mulvenna, M. D. and Hughes, J. G. *Navigation Pattern Discovery from Internet Data*. In Workshop on Web Usage Analysis and User Profiling, August 1999.

[7] Cadez, I. V., Heckerman, D., Meek, C., Smyth, P. and White, S. *Visualization of Navigation Patterns on a Web Site Using Model Based Clustering*. Proc. of 6th ACM SIGKDD Intl' Conf. on Knowledge Discovery and Data Mining, 2000.

[8] Chen, M-S, Park, J. S. and Yu, P. S. *Efficient Data Mining for Path Traversal Patterns*. IEEE Transactions on Knowledge and Data Engineering, 10(2):209-221, March/April, 1998.

[9] Cooley, R., Mobasher, B. and Srivastava, J. *Data preparation for mining world wide web browsing patterns*. Knowledge and Information Systems, 1(1):5-32, 1999.

[10] Dunham, M. H. *Data Mining Introductory and Advanced Topics*. Prentice-Hall, 2003.

[11] Foss, A., Wang, W. and Zaïane, O. R. *A non-parametric approach to web log analysis*. In Proc. of Workshop on Web Mining in First International SIAM Conference on Data Mining, 41-50, Chicago, April 2001.

[12] Fu, Y., Sandhu, K. and Shih, M. *Clustering of web users based on access patterns*. Workshop on Web Usage Analysis and User Profiling (WEBKDD99), August 1999.

[13] Guha, S., Rastogi, R. and Shim, K. *ROCK: a robust clustering algorithm for categorical attributes.* In ICDE, 1999.

[14] Gündüz, Ş. and Özsu, M. T. *A Web page prediction model based on click-stream tree representation of user behavior*. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C, August 24-27, 2003.

[15] Hair, J. F., Andersen, R. E., Tatham, R. L. and Black, W. C. *Multivariate Data Analysis.* Prentice Hall, New Jersey, 1998.

[16] Hay, B., Wets, G. and Vanhoof, K. *Clustering Navigation Patterns on a Website Using a Sequence Alignment Method*. IJCAI's Workshop on Intelligent Techniques for Web Personalization, 2001

[17] Karypis, G., Han, E-H. and Kumar, V. *Chameleon: A hierarchical clustering algorithm using dynamic modeling*. IEEE Computer, 32(8):68-75, August 1999.

[18] Moe, W. W. *Buying, Searching, or Browsing: Differentiating between Online Shoppers Using In-Store Navigational Clickstream*. Journal of Consumer Psychology, 13 (1&2), 29-40, 2003.

[19] Pei, J., Han, J., Mortazavi-Asl, B. and Zhu, H. *Mining Access Patterns Efficiently From Web Logs*. In Proc. of Pacific Asia Conf. on Knowledge Discovery and Data Mining, pp592, Kyoto, Japan, April 2000.

[20] Setubal, C. and Meidanis, J. *Introduction to Computational Molecular Biology*. PWS Publishing Company, 1997.

[21] Spiliopoulou, M., Pohle, C. and Teltzrow, M. *Modelling Web Site Usage with Sequences of Goal-Oriented Tasks*. Multi-Konferenz Wirtschaftsinformatik 2002 vom 9.-11. September 2002 in Nürnberg.

[22] Srivastava, J., Cooley, R., Deshpande, M. and Tan, P. *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data*. SIGKDD Explorations, 1(2):12--23, 2000.

[23] Wang, W. and Zaïane, O. R. *Clustering Web Sessions by Sequence Alignment*. Third International Workshop on Management of Information on the Web in conjunction with 13th International Conference on Database and Expert Systems Applications DEXA'2002, pp 394-398, Aix en Provence, France, September 2-6, 2002.

[24] Xiao, Y-Q and Dunham, M. H. *Efficient mining of traversal patterns*. Data and Knowledge Engineering, 39(2):191-214, November, 2001