Chapter 4
Discovery and Analysis of Traversal Path Patterns and Purchase Behavior

4.1 Introduction

In e-commerce, web site design is very important for success as customers want flexibility in accessing the site and locating target contents easily. The navigability of a website is very important. If the users are unable to easily find what they are looking for on the website they might go elsewhere and the website traffic will suffer. This could lead to lower sales and advertisement revenues for the website. Determining the navigability of a website can be quite a subjective task. Accessibility and performance of the web site is heavily depends on the site structure. Thus, it is important to design effective web site structure. Some people like having everything available from a giant menu, while others prefer a system of subcategories that eventually leads them to the desired resource. In addition to website structure and linkage, positioning, formatting and naming of links can greatly affect the navigation of users. After all, if you can't find the link you're not likely to click it. User preferences and web page formatting are unfortunately not something we can discover from web log data. One thing is common for most people though: Very few likes to backtrack more than absolutely necessary when navigating a website. It is much preferred to be able to just click a link to get back to where you were a while ago, instead of hitting the browser back-button many times. Discovering user backtracking behaviors is one of the applications of traversal patterns and can be used to improve the website navigability. Traversal patterns are very useful for finding the frequent patterns in users’ browsing paths, weaknesses in the website structure and determining how the web site is being used. Thus, user traversal path patterns can reveal valuable information for improving web site structure and navigability of web site.
4.2 Web Log Mining and User Traversal Behavior Analysis

When a user navigates the website pages, requests are sent to the web server whenever new pages or resources are accessed by the user. The server receives request and send back the data for the requested resource. The resource requests are stored in log file on the server. Web log files can record a set of transactions in time sequence. If the web-based companies can discover the sequential patterns of the visitors, they can predict users’ visit patterns and target market on a group of users. Server log information could be very precious to the company in the fields of understanding customer behavior, improving customer services and relationship, launching target marketing campaigns, measuring the success of marketing efforts, and so on. Web Usage Mining which is also known as web log mining aims to find out interesting and frequent user access patterns from web browsing data that are stored in web server log. Such discovered knowledge is useful in analyzing how the web pages are accessed or what are seeking for by the users. For the task of applying data mining techniques, information from web data is extracted in order to understand and better serve the needs of users navigating on the web [57]. A user web access sequence is a sequence of web pages ordered by increasing traversal-time. Web usage mining focuses on techniques that could predict user behavior while the user interacts with the web. Usage analysis includes straightforward statistics, such as page access frequency, as well as more sophisticated form of analysis such as finding the common traversal paths through a Website. Within web log analysis the main interests have been user and session identification and sequences of pages being accessed by users. Information from a web log can be used for several purposes. Among them are:

- Caching frequently accessed pages.
- Predicting user navigation.
- Improving sales and advertisement.
- Improving inter-page linkage.
- Improving website structure.
A Web log refers to a collection of records that objectively document visitors’ surfing behaviors on a Web site [56]. The important frequent access patterns are focused and extracted from Web logs rather than directly examining all records in logs. Focusing on visitors’ important access patterns is advantageous in several ways. First, infrequent visiting behaviors recorded in log records can distract or even mask analysis results and therefore should be filtered. By concentrating on frequent patterns, one can uncover prominent surfing behaviors likely to be observed in future visits. Thus, analysis of Web site visiting patterns not only reveals the navigability of the design but also points to areas in which the current structure design could be improved. In addition, the prominent surfing patterns extracted can be applied easily to predict the performance of a Web site and identify desirable enhancements to be implemented.

In an information providing environment where objects are linked together, users can easily travel objects back and forth with links and icons provided. Some node might be revisited because of its location, rather than content. For example, in a WWW environment, to reach a sibling node a user is usually inclined to use backward icon and then a forward selection, instead of opening a new URL. Consequently, to extract meaningful user access patterns from the original log database, the effect of such backward traversals is considered and real access patterns of interest are discovered. It is assumed that a backward reference is mainly made for ease of traveling but not for browsing. Thus the discovery of forward reference patterns is concentrated. Specifically, a backward reference means revisiting previously visited objects by the same user access. In web log sequences, contiguous sequence patterns may appear as subsequences or subsets and mining of these sequences is important since link between web pages is important and to be maintained. Mining of all subsequences would require a lot of time and tedious in sequence mining and hence finding only maximal frequent patterns out of them are potential and less time consuming. The Maximal Forward Path (MFP) as a notion of a maximal forward moving motion in visiting web documents is introduced in order to filter out the redundant pattern from the log source [58]. All the backward traversal actions or Backward References only
occur to users in the process of searching for Web pages that really interest them. Hence it is considered that only the forward browsing motion or Forward Reference contains meaningful information and reflects users’ true browsing patterns. We must give importance to backward references also because it also provides information about users’ navigation patterns and shows whether users are able to navigate easily within a website or not. This demonstrates whether there exists a frequent backward motion which may show that the structure of a web site is not clear.

4.3 The Process of User Traversal Pattern Mining

Web spider is a program that can collect and parses all pages of web site by following site hyperlinks. Each parsed web page in web site is further classified as either index page or content page. Index page presents hyperlinks to navigate other web pages. The content page contains the information the visitors are interested and often searched by visitors.

![Diagram of User Traversal Path Pattern Mining]

**Figure 4.1: The User Traversal Path Pattern Mining**
While classifying the web pages, the site structure is also considered as it represents navigation structure. Access patterns are useful to improve the web site structure design. Generally, visitors access set of pages sequentially that is recorded in web server log. The web server log is preprocessed to clean and eliminate information that is not required to find traversal patterns. After preprocessing, user sessions are extracted. A user session is a set of pages visited by the particular user in specified time. From the extracted sessions, traversal patterns are found and analyzed. Figure 4.1 describes the user traversal path pattern mining.

4.4 Session Extraction from Web Server Log

When user navigates the web pages of the website, each request is recorded by the server in server log files and keeps information about each user’s access to the web server. The server log file contains information such as Visitor’s IP address, access date and time, URL of page accessed, the status code returned by the server, bytes transferred to the user etc.

<table>
<thead>
<tr>
<th>URL</th>
<th>Host</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4.5.6.7</td>
<td>12:17:40</td>
</tr>
<tr>
<td>B</td>
<td>4.5.6.7</td>
<td>12:17:56</td>
</tr>
<tr>
<td>C</td>
<td>4.5.6.7</td>
<td>12:18:23</td>
</tr>
<tr>
<td>D</td>
<td>4.5.6.7</td>
<td>12:18:34</td>
</tr>
<tr>
<td>E</td>
<td>4.5.6.7</td>
<td>12:18:52</td>
</tr>
<tr>
<td>F</td>
<td>4.5.6.7</td>
<td>12:19:11</td>
</tr>
<tr>
<td>G</td>
<td>4.5.6.7</td>
<td>12:19:45</td>
</tr>
<tr>
<td>H</td>
<td>4.5.6.7</td>
<td>12:20:06</td>
</tr>
<tr>
<td>I</td>
<td>4.5.6.7</td>
<td>12:20:28</td>
</tr>
<tr>
<td>J</td>
<td>4.5.6.7</td>
<td>12:20:49</td>
</tr>
<tr>
<td>K</td>
<td>4.5.6.7</td>
<td>12:21:14</td>
</tr>
<tr>
<td>L</td>
<td>4.5.6.7</td>
<td>12:21:29</td>
</tr>
<tr>
<td>M</td>
<td>4.5.6.7</td>
<td>12:22:37</td>
</tr>
<tr>
<td>N</td>
<td>4.5.6.7</td>
<td>12:22:42</td>
</tr>
<tr>
<td>O</td>
<td>4.5.6.7</td>
<td>12:22:48</td>
</tr>
<tr>
<td>P</td>
<td>4.5.6.7</td>
<td>12:23:02</td>
</tr>
<tr>
<td>Q</td>
<td>4.5.6.7</td>
<td>12:23:16</td>
</tr>
<tr>
<td>R</td>
<td>4.5.6.7</td>
<td>12:23:51</td>
</tr>
<tr>
<td>S</td>
<td>4.5.6.7</td>
<td>12:24:07</td>
</tr>
<tr>
<td>T</td>
<td>4.5.6.7</td>
<td>12:24:15</td>
</tr>
</tbody>
</table>

Table 4.1: The User Session for Traversal Pattern Analysis
The web log data is usually incomplete and has to be preprocessed to get users sessions for data mining purpose. Web log file is cleaned to remove information which is not needed to find traversal patterns. Also requests entries for which server returned an error code are removed. Cleaned Web log is analyzed to identify visit sessions, each of which constitutes a basic processing unit for the discovery of interesting, prominent access patterns. A user session is a set of pages visited by the same user within the duration of one particular visit to a web site. Each user session consists of only the pages visited by a user in a row. Table 4.1 shows a sample user session generated from web server log for particular user to analyze traversal pattern as an example.

4.5 Traversal Pattern Discovery: MFP and MBP

The purpose of pattern discovery is to find the traversal patterns in the user sessions. We detect two types of traversal patterns the Maximal Forward Path (MFP) and Maximal Backward Path (MBP). The Web has a natural graph structure in that pages are linked via hyperlinks. While surfing the web, user may move forward along the graph by selecting a hyperlink in the current page. He may also move backward to any page visited earlier in the same session by selecting a backward icon. A forward reference means user is looking for desired information. A backward reference means the user has found her desired information and is going to looking for something else. A sequence of consecutive forward references may indicate the information for which the user is looking. A maximal forward path is defined as the longest consecutive sequence of forward references before the first backward reference is made to visit some previously visited page in the same session [58]. Thus, the last reference in a maximal forward sequence indicates a content page that is desired by the user. Under such understanding, when a user searches for desired information, her information needs can be modeled by the set of maximal forward paths that occurred during his search process.
MFP defines the traversal data into a more meaningful manner by ignoring the continuous repetition of backward browsing actions. Online browsing movement is not a simple single directional action but it is a dual-directional action. In order to filter out the redundant pattern from the log source, Maximal Forward Reference is a notion of a maximal forward moving motion in visiting Web documents. All the backward traversal actions only occur to users in the process of searching for Web pages that really interest them. Hence, it is assumed that only the forward browsing motion or Forward Reference is reflecting users’ true browsing patterns. When navigating a website, backtracking occurs if it is necessary. It is much preferred to just click a link to get back to where you were ago, instead of hitting the browser back button a zillion times. Discovering user backtracking behaviors is one of the applications of traversal patterns and can be used to improve the website navigability. The Maximal Backward Path demonstrates groups of nodes in the backward sequence combination. This presents a good indication of how well the infrastructure of a site is constructed and arranged. The longer the combination of nodes MBP holds the less organized a site appears to be. This can be interpreted as users having difficulties in finding their desired nodes and hence they are forced to browse each link one after another in order to narrow down the possibilities. If the MBP contains many same combinations then this can inform that this particular reference of linkages is well constructed.

In a maximal forward or backward path a sequence of web log entries originate from the same user in chronological order and no entry appears more than once. The MFP contains entries generated from navigating forward to new unvisited pages and the MBP contains entries generated from navigating backward to pages that have been visited before. The forward and backward references appear in alternate order in the log. A user starts with a forward reference, then may have a backward reference, then may have a new forward reference, and so on. The start of two maximal forward references may be identical, as a user may navigate forward five pages, then backward two pages, then forward again three pages.
Figure 4.2 represents traversal behavior of user according to user session of Table 4.1 and the traversal path is \{A, B, E, F, G, F, E, B, J, K, N, Q, N, K, J, R, S, A, V, W\}. By using traditional analysis methods, for example, nodes B, E and F are showing greater importance than nodes G and Q, but the nodes G and Q are actually the pages containing the information that the user needs. Nodes B, E and F might be pages embedded with all the inter-links in that site and cause an illusion in becoming the most valuable pages. To find MFP, for a set of inter-linked nodes arranged in a hierarchy fashion, the action starts from the highest node or the root node and follows its way down. When the first reverse movement occurs the forward movement is terminated. This results in a collection of nodes which is marked as maximal forward path. Since MFP omits all the reversing directional traveling, it will contain purely the nodes captured during the forward visits. Starting at A, one has forward reference sequences AB, ABE, ABEG and ABEF. Then, at G a backward reference to F is made, meaning that ABEGF is a maximal forward reference or maximal forward path. Now, some caution shall be paid to F. Here, even though a backward reference is made to E, one shall not consider ABGF as a maximal forward reference, because the current backward reference is the second backward reference, the first is G to F. Similarly, ABE shall not be considered as a maximal forward reference because backward reference made from E to B is third backward reference, first is G to F, second is F to E. At B, a forward reference is made to J, thus begins new forward reference sequences ABJ, ABJK, ABJKN, ABJKNQ. ABJKNQ is a maximal forward reference, because the first backward reference occurs at Q after the forward reference.
J. Similarly, one can find the other maximal forward references ABJRS, AVW as presented in Figure 4.3. Where G, Q, S and W are the pages the user desires. The set of the maximal forward references ABEFG, ABJKNQ, ABJRS and AVW describes the user's needs and his actual search behaviors also.

In a set of hierarchical inter-linked nodes and during a particular session, the MBP starts at a node when a reverse behavior occurs and returns back to the node where a new forward movement invoked. As just discussed, the forward reference sequence ABEFG is the maximal forward path and at G the first backward reference is made to F. Starting at G, one has backward reference sequences GF, GFE, GFE, GFEB. Then at B a forward reference is made to J. It means that GFEB is a maximal backward reference or maximal backward path (MBP). Similarly, other maximal backward paths QNKJ, SA can be found as presented in Figure 4.3. MBP is not necessary the reverse order of a maximal forward path. MBP contains the nodes covered by bidirectional movements that are both forward and backward traveling, where as MFP contains only single directional nodes.

Figure 4.3: Finding MFP and MBP
4.6 Algorithm for Traversal Pattern Detection

The algorithm used to find user traversal path patterns here is based on algorithm presented by Dalei Xing in [59]. It is modified to allow detection of backward references also. The pseudo code of modified algorithm is presented below.

DF is the database to store the mfr patterns in.
DB is the database to store the mbr patterns in.
F is the current mfr pattern.
B is the current mbr pattern.
FWD is a flag indicating forward traversal if true,
backward traversal if false.
S is the referring URI (source) in a log entry.
D is the requested URI (destination) in a log entry.

1. Reset F and B to be empty and set FWD = true.
2. Assign S and D.
3. If S == null :
   // We are beginning a new mfr pattern.
   If F is not empty :
      If FWD == true :
         Store F in DF.
         Reset F to be empty.
      If B is not empty :
         If FWD == false :
            Store B in DB.
            Reset B to be empty.
         Append D to F.
   4. Else if D is already in F :
      If FWD == true :
         // End mfr pattern and start mbr pattern.
         Store F in DF.
Append the last reference in F to B.
Set FWD = false.
// Continue mbr pattern.
Append D to B.
Discard all references after D in F.
5. Else :
If FWD == false :
// End mbr pattern and start mfr pattern.
If B is not empty :
Store B in DB.
Reset B to be empty.
Set FWD = true.
// Continue mfr pattern.
Append D to F.
6. Go to step 2 until no more log entries in session.
7. Go to step 1 until no more sessions.

4.7 Applying Conditional Restrictions With Patterns

The conditional restrictions can be applied with the particular patterns to find the more knowledge about the users’ patterns. The technique can be developed to assist data mining algorithms in order to first refine browsing pattern with relevant conditional restrictions and then support the discovery task. With such a technique pattern identification is significantly faster and more accurate than just using standard discovery methods. The conditional restrictions are taken from standard web log files. The different conditional restrictions can be identified such as restrictions concentrating on factors relating to users’ navigating movements which include MFP and MBP, restrictions concentrating on factors such as time duration of staying on a particular web page, session intervals and periodicity of visits to web pages, restrictions consist of other available information regarding each individual visiting a web site i.e. user. A session is created based on individual temporal attributes, such as
the beginning and end of access time to a site. The periodicity is a time consisting of a series of periodic intervals based on a time cycle unit. It is possible for multi sessions to occur over the same periodic time. A user is determined by a specific IP address, which is assigned to each individual access when connected to the Internet. By applying conditional restrictions with specific patterns, the approach enables data analysts to focus on individual cases with more control while at the same time providing more knowledge about users’ patterns. The value of conditional restrictions can be anything within users’ traversal patterns e.g. the length of the traversal movement, the direction of browsing path, designating nodes inside the browsing pattern, etc. Conditional restrictions are able to refine the data source in order to reduce the processing time with a better return of outcome. Several restrictions can be combined to form a complex restriction type for setting a more detailed filter.

The main entity deals with the type of data mining algorithm to be executed for discovering data patterns from the selected data source. The other entity which is based on the conditional restrictions just mentioned, reduces the selected data source, refines the outcome and passes information back to the main entity. The restrictions can be applied interchangeably and can be modified in accordance to analysts' preferences. Different types of data mining techniques like Associations Rule, Classification, Clustering, Summarization etc are applied for seeking a particular pattern. Each technique has its own merits towards addressing different problems. Different techniques can be adopted flexibly with the customized conditions. As each of the algorithms can be used to seek for potential patterns, it is possible to define variables in order to limit the range of the discovery process. For example, pattern finding using Association Rules to find users’ MFP patterns (A,B,E,F and G) that have time period 4 hours between every Wednesday and Thursday after 18:20 hours and with the range of IP addresses from 192.168.0.20 to 192.168.0.240 can be performed. The data will be reduced according to conditional restrictions values after the execution of mining task and then will be passed to data mining algorithm.
4.8 Traversal Pattern and Purchasing Behavior in E-Commerce

In e-commerce, traversal path patterns represent the navigation behavior of customers. The information about purchasing behavior of customer can be used to find association between purchasing items and this can help in improvement of cross-selling. Considering both traversal pattern and purchasing behavior of customer can add value to association rule finding. Figure 4.4 demonstrates traversal and purchase behavior of customer. The customer traverses and purchases item in following order. First customer starts from A and goes to B where purchases item1. Then, customer sequentially visits E, F and G. On G, customer purchases item2.

![Figure 4.4: Demonstration of Traversal and Purchase Behavior of Customer](image)

In another sequence, Customer visits A, then visits B where customer purchases item1. Next, from B, customer goes to J and then to K where purchases item3. Then, N and Q are visited and on Q, customer purchases item5. Back to N, customer purchases item4. Same way, other traversal sequences are followed. The customer transaction detail with respect to this traversal and purchase behavior is represented in Table 4.2.
Chapter 4: Discovery and Analysis of Traversal Path Patterns and Purchase Behavior

<table>
<thead>
<tr>
<th>Traversal Path</th>
<th>Purchase Detail</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-B-E-F-G</td>
<td>B(item1), G(item2)</td>
</tr>
<tr>
<td>A-B-J-K-N-Q</td>
<td>B(item1), K(item3), N(item4), Q(item5)</td>
</tr>
<tr>
<td>A-B-J-R-S</td>
<td>B(item1), R(item6)</td>
</tr>
</tbody>
</table>

Table 4.2: Customer Transaction Detail

The pattern AVW is not considered as there is no purchase. It is also important to consider backward traversal. Customer purchase item5 on Q then back to N and then purchases item4 at N. Considering both traversal path pattern and purchase behavior is important for finding association between purchasing items. Following statement indicates that considering traversal path ABJKNQ, customer purchases item1 at B then generally also purchases item5 at Q.

\[
\text{ABJKNQ} [B(\text{item1}) \Rightarrow Q(\text{item5})]
\]

Support for this can be defined as \( p/n \) where \( n \) is total number of customers and \( p \) indicates number of customers who traverse ABJKNQ and purchases item1 on B and then purchases item5 on Q. Now support of ABJKN [B(item1)] is \( q/n \) where \( n \) is total number of customers and \( q \) indicates number of customers who traverses ABJKN and purchase item1 on B. The confidence of ABJKNQ [B(item1) => Q(item5)] can be found as under.

\[
\text{Confidence of ABJKNQ} [B(\text{item1}) \Rightarrow Q(\text{item5})]
\]

\[
= \frac{\text{Support of ABJKNQ} [B(\text{item1}), Q(\text{item5})]}{\text{Support of ABJKN} [B(\text{item1})]}
\]

\[
= \frac{p}{q}
\]

The confidence describes that if customer follows traversal path ABJKN and purchases item1 on B then the probability of purchasing item5 on webpage Q is \( p/q \). The support for traversal path pattern indicates number of customers following that
path while traversing. If support for traversal pattern matches with minimum support defined by user then it is called frequent traversal pattern.

4.9 Conclusion

Web mining can be applied to find user traversal path patterns that can be analyzed to improve navigability and structure of web site. It reveals the information that how the web pages are accessed and what the web users are seeking for. The web server log stores browsing behavior of site visitors and it is the source for traversal pattern finding. The web server log is preprocessed to clean and eliminate information that is not required to find traversal patterns. Cleaned Web log is analyzed to identify visit sessions, each of which constitutes a basic processing unit for the discovery of interesting, prominent access patterns. A user session is a set of pages visited by the same user within the duration of one particular visit to a web site. Each user session consists of only the pages visited by a user in a row. We detect two types of traversal patterns the Maximal Forward Path and Maximal Backward Path. A maximal forward path is defined as the longest consecutive sequence of forward references before the first backward reference is made to visit some previously visited page in the same session. Thus, the last reference in a maximal forward sequence indicates a content page that is desired by the user. The Maximal Backward Path demonstrates groups of nodes in the backward sequence combination. This presents a good indication of how well the infrastructure of a site is constructed and arranged. The longer the combination of nodes MBP holds the less organized a site appears to be. The emphasis can also be given to the conditional restrictions which can be applied with specific patterns to find the more knowledge about users’ patterns. In e-commerce, traversal path patterns represent the navigation behavior of customers. The information about purchasing behavior of customer can be used to find association between purchasing items and this can help in improvement of cross-selling. Thus, considering both traversal pattern and purchasing behavior of customer can add value to association rule finding.