# Clustering the Navigation Pattern Using Graph Partitioning

Ei Mon Cho, Khin Thanda Soe
*University of Computer Studies, Yangon*
*eimoncho88@gmail.com, ktdsoe@gmail.com*

## Abstract

*World Wide Web become as information gateway and as a medium for conducting business. People use web pages to retrieve information which may need or demand. Web mining can provide the navigation link to user by using user behavior. This system provides clustering of user navigation patterns based on graph partitioning algorithm. For clustering of user navigation patterns, the system creates an undirected graph based on connectivity between each pair of web pages and assigns weight to edges in such graph. Those weights are then used in clustering algorithm based on graph partitioning. These system results can be used to improve the overall performance of future access and the clusters concerning user navigation pattern can be used for web caching.*

## 1. Introduction

Web users want to have the efficient search to find information easily. Web contains a great amount of information distributed as the common browsing behavior among a group of users. Science many users may have common interests up to a point during their navigation, navigation patterns should capture the overlapping interests or information needs of these users. Navigation patterns should also be capable to distinguish among web pages on their different significant to each pattern.

The processing of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A clustering is a collection of data objects that are similar to one another within the same clustering and are dissimilar to the objects in the other clusters. A clustering based process is adaptable to change and helps single out useful features that distinguish different groups.

Web mining is the extracting of interesting and potentially useful patterns and implicit information from artifact or activity related to World Wide Web. Web mining is a very broad field emerging to solve the issues that arise due to www. By using data mining techniques, there are three directions in the field of web mining; web content mining, web structure mining and web usage mining [1].

Among web mining, web usage mining is the process of identifying browsing patterns by using the user's navigational behavior. This information takes as input the usage data i.e. the data residing in the web server logs. The users, activity when browsing through web sites is registered in these sites' web logs. These data can be identified user navigation patterns concerning the activities in the web site by using data mining techniques [4, 5].

Each access to a web page is recorded in the access log of the web server that hosts it. Web log data produced by web server or proxy server are text files with a row for each HTTP transaction. Web log data can be collected at the server level, client level or proxy level [3, 5]. A web server log records the browsing behavior of site visitors. The data recorded in server logs reflect the concurrent and interleaved access of a web site by multiple users. Typically web log files contain client IP address, request time, request URL, referred, HTTP status code, etc [7].

In this paper, web server log is used to cluster user navigation pattern based on graph partitioning algorithm. As web mining procedure, this system needs to perform some preprocessing tasks: data cleaning, user identification and session identification to web log. After preprocessing steps, undirected graph is created by using connectivity between each pair of web pages. The result of this system is cluster of web pages which can be used in web personalization, recommendation system, web caching and other modifications to the sites [5].

The rest of this paper is organized as follows: section 2 presents background theory. We describe the proposed system of clustering the navigation pattern in section 3 and experimental results in section 4. Section 5 concludes the paper.

## 2. Background Theory

After some preprocessing, there are two processing steps: Navigation Pattern Mining and Clustering base on Graph Partitioning. This system uses the graph partitioning algorithm to mining the user navigation patterns.

## 2.1. Navigation Pattern Mining

To apply graph partitioning algorithm, an undirected graph (M) is created based on connectivity between each pair of web pages. In undirected graph M, degree of connectivity in each pair of pages depends on two main factors: the time position of two pages in a session and occurrence of two pages in a session [2].

Undirected graph M= (V, E)

V= the set of vertices as assigning different pages on web server.

E= the connectivity degree of each two pages in sessions.

For undirected graph M, a weight is measured for approximating the connectivity degree of each two web pages in sessions. To measure weight for graph, there are two concepts related to this measure: "Time Connectivity" and "Frequency Connectivity".
"Time connectivity" measures the degree of visit order for each two pages in a session.

$$TC_{a,b} = \frac{\sum_{i=1}^{N} \frac{T_i}{T_{ab}} \times \frac{f_a(k)}{f_b(k)}}{\sum_{i=1}^{N} \frac{T_i}{T_{ab}}} \qquad (2.1)$$

$TC_{a,b}$=Time connectivity between page a and b

$T_i$=Time duration in ith session that both contain a and b

$T_{ab}$=Difference between request time of a and b

f (k) = K if web page appears in position

N= number of session

"Frequency Connectivity" measures the occurrence of two pages in each session.

$$FC_{a,b} = \frac{N_{ab}}{Max[N_a, N_b]} \qquad (2.2)$$

$FC_{a,b}$=Frequency connectivity between page a and b

$N_{a,b}$=number of session containing both page a and b

$N_a, N_b$= number of session containing only page a and b

Time connectivity and Frequency connectivity are used for weight of each edge in the undirected graph.

$$W_{a,b} = \frac{2 \times TC_{ab} \times FC_{ab}}{TC_{ab} + FC_{ab}} \qquad (3.3)$$

This value is stored in an adjacency list M which contain Wa,b. To limit the number of edge in such graph, weight value is less than a threshold is too little correlated and thus discarded. This threshold is named as MinFreq in this system.

## 2.2. Clustering Based on Graph Partitioning

Then we apply graph partitioning algorithm to find groups of strongly correlated page by partitioning graph according to its connected components. Figure1 shows clustering algorithm using graph partitioning. We use the Depth First Search (DFS) which started from a vertex on graph and then search for the connected component reachable from this vertex. We found the component, the algorithm checks if there are any nodes not concerned in the visit. And then DFS is applied by starting from one of the nodes not visited. Then clusters are based on values store in adjacency list M [2, 6, 8].

```
Begin
    L[p]=P;//assign all URL to list

    for each (pi,pj)€L[p] do//for all pair of web pages
        M(i,j)=WeightFormula(Pi,Pj);
        //for computing weight
        Edge (i,j)= M(i,j);
    end for

    for all Edge (u,v)€Graph (E,V) do
        //remove all edge that its weight below than MinFreq

        if Edge  (u,v)< MinFreq then
            remove (Edge());
        end if

    end for

    for all vertices (u)€ do
        Cluster [i]=DFS (u);
        //do DFS algorithm
    end for

    return (Cluster);
end
```

**Figure1. Graph partitioning algorithm for clustering**

## 3. The Proposed System

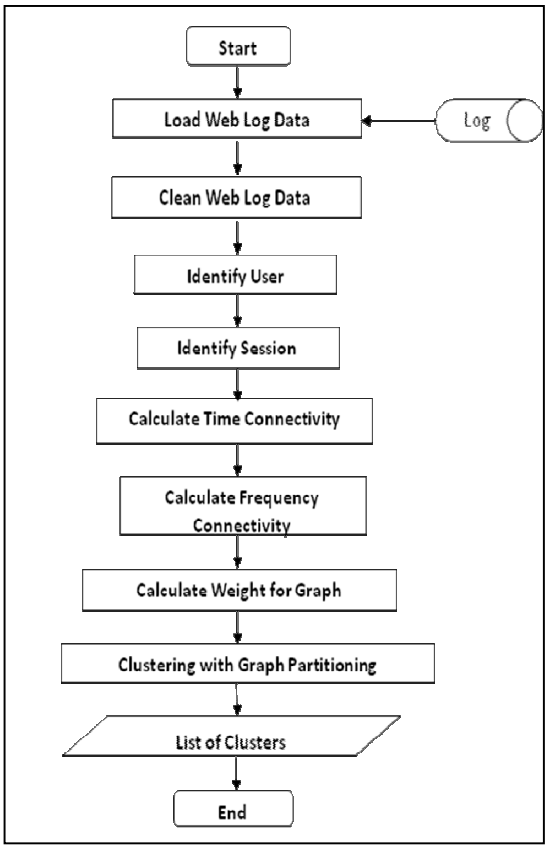


**Figure2. System of clustering the navigation pattern**

Figure2 shows the proposed system that web pages are clustered according user navigation behavior by using graph partitioning. The input of the system is web log files. Some preprocessing steps need to be done on web log entry. These are data cleaning, user identification, session identification.

### 3.1 Data Cleaning

The first step of preprocessing is data cleaning. Depending on the application, web log data may need to be cleaned from entries involving pages. Data cleaning step only performs on this data by removing auxiliary files (e.g. image file) such as .gif, .jpeg, .jpg, etc. After cleaning data, we identified the user according to the IP address.

### 3.2 User Identification

After cleaning log data, the next step is user identification. Many users may be assigned the same IP address and one user may have several different IP addresses. Depending on our system, we assume each IP address as a single user.

### 3.3 Session Identification

The last preprocessing step is to perform session identification by dividing the click stream of each user to session. In this system, minimum time out is set and assumed that consecutive accesses within it belong to the same session .We use minimum time out limitation as 15 minutes.

### 3.4 Navigation Pattern Mining

After preprocessing, different pages on web log are assigned as vertices. Then we do two processing steps: Navigation Pattern Mining and Clustering based on Graph Partitioning Algorithm. The values of time connectivity and frequency connectivity are two indicators of degree of connectivity for each pair of web pages. The time connectivity and frequency connectivity are computed using equation 2.1 and equation 2.2 respectively. Then these values are used for computing weight of each edge in the undirected graph from equation 2.3

### 3.5 Clustering Based on Graph Partitioning

Then Depth First Search (DFS) is applied by starting from one vertex of graph for clustering web pages. To limit the number of edge in such graph, weight value is less than a threshold called minimum frequency (MinFreq) are too little correlated and thus discarded. The proposed system produces a list of clusters for user navigation patterns.

## 4. Experimental Results

Web pages to be clustered can be viewed as a set of vertices and edges between the vertices represent the relationship between them. We cluster the web pages rely on Graph partitioning that identifies the clusters by cutting edges from the graph.

The contents of the cluster are increased when the minimum frequency (MinFreq) is increased. Since each edge in the graph represent the similarity between the documents, by removing the edges with the weight algorithm minimizes the similarity between documents in different clusters.

Table 1 shows the number of clusters produced by the system when testing with different minimum frequency on 4 different log files. By analyzing the contents of clusters, we found that MinFreq between 0.3 and 0.8 can give strongly

correlated pages within each cluster. Hence, the resulting cluster contains highly related web pages with these MinFreq values. However, the number of clusters does not depend on MinFreq.

**Table1. Number of clusters based on Minimum Frequency (MF)**

| Data Set | Record | Session | User | Number of clusters | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | **MF 0.1** | **MF 0.2** | **MF 0.3** | **MF 0.4** | **MF 0.5** | **MF 0.6** | **MF 0.7** | **MF 0.8** | **MF 0.9** |
| 1 | 176 | 19 | 13 | 3 | 3 | 3 | 3 | 5 | 8 | 16 | 15 | 14 |
| 2 | 231 | 7 | 5 | 5 | 1 | 2 | 2 | 2 | 3 | 11 | 9 | 12 |
| 3 | 103 | 11 | 6 | 1 | 1 | 1 | 1 | 2 | 4 | 7 | 10 | 10 |
| 4 | 115 | 10 | 6 | 2 | 3 | 3 | 4 | 5 | 5 | 5 | 9 | 9 |

## 5. Conclusion

In this paper, web pages are clustered according to user navigation pattern by using graph partitioning. The web server logs are used as input and the system do web mining steps and produces a list of clusters for user navigation patterns. The advantages of the graph partitioning are its linear time and its quality of clustering. And then these results can be used for predicting user's next request in huge web sites.

## 6. References


**[1]** Albanese, Picariello, Sansone, "A Web Personalization System Based on Web Usage Mining Techniques", www2004, May 17-22, 2004, New York, USA, ACM 1-58113-912-8/04/005, p 288-289

**[2]** Jali, Norwati, Ali, Nasir, "Web User Navigation Pattern Mining Approach Based On Graph Partitioning Algorithm", 2005-2008 JATIT , pp. 1125-1130.

[3] M.Baglioni, U.Ferrara, A.Romei, S.Ruggieri and F.Turini, "Preprocessing and Mining Web Log Data for Web Personalization.", 8th Italian conf.on Artificial Intelligence vol 2829 of LNCS, 2003

[4] Magdalini Eirinaki, "Web Mining: A Roadmap."

[5] Naresh Barsagade, "Web Usage Mining and Pattern Discovery: A survey Paper.", CSE 8331 December 8, 2003

[6] J.Berry, M.Goldberg, "Path Optimization and Near-Greedy Analysis for Graph Partitioning: An Empirical Study", Study proceeding of 1995 symptom or discrete Algorithm, 1995

[7] Wahab, Mahd, Hanafi, Mohsin, "Data Preprocessing on Web server logs for Generalized Association Rule Mining Algorithm."World Academy of Science, Engeenering and Technology 48 2008, pp.190-197

[8] Y.Y.Yao, Hailtonand X.Wang "Page Promper: An Intelligent Agent for web Navigation Creating Using Data Mining Techniques." , Lecture Notes in Computer Science, 2002, Volume 2475/2002, 949, Dol:10, 1007/3-540-45813-67