4 Analysis of Web Pages of User Interest By Clustering

With the continued growth of Web users, enormous volumes of clickstream (sequential series of pageview requests) data and user data are collected by Web-based organizations in their daily operations. The need to understand the large, complex and information-rich data from which knowledge can be discovered is inevitable in all fields of business, science and engineering. Capturing clickstream data will be helpful to model and analyze the users’ browsing behaviour. This analysis requires the automatic discovery of meaningful patterns and relationships from a large collection of semi-structured data stored in the Web and application server logs.

This chapter focuses on the pattern discovery phase which determines “similar” interests of groups of sessions based on the navigation behaviour. To achieve this,

- The formatted user session file, referred to as an usage model is further subjected to additional post-processing tasks such as transaction normalization.
- Using this data model, “similar” interests of groups of sessions are discovered. This is called as the aggregate usage profile which consists of pages of varying user interest/significance.
- The significance of the pages in each profile is determined. These profiles would later help in various applications of Web usage mining such as Web site improvement, assigning a new user to the appropriate cluster and recommend pages of interest not yet visited by the user to provide personalized Web content.

4.1 Related Survey

A lot of research is being done in the area of Web usage mining. Based on the goals of the analyst and applications, various algorithms can be applied for cluster analysis.
Chapter 4. **Analysis of Web Pages of User Interest By Clustering**

On the whole, clustering is the process of grouping the samples into clusters such that samples within a cluster have high similarity compared to each other but dissimilar to samples in other clusters. As mentioned in [MDLN02], two types of clustering can be performed on usage data: Transaction Clusters and Page Clusters. Each type of clustering is helpful in different applications such as personalization/recommendation, system improvement, Web site structure, business intelligence and user behaviour.

Similarity Measures form the core component for every clustering algorithm. For several years, focus on cluster analysis has been mainly distance-based cluster analysis. There are various distance-based similarity measures such as the Euclidean distance measure, Manhattan distance, Minkowski, Mutual Neighbour Distance (MND), Simple Matching Coefficient, Jaccard Coefficient and Rao’s coefficient. The usage of these similarity measures depends on the features of the samples. However, in [Cha09] a Sequence Alignment Method has been used for measuring similarities between Web pages by considering the URL and the viewing time of the URL. The algorithm for Web Session Clustering Based on Increase of Similarities (WSCBIS) has been implemented along with k-means clustering and ROCK (RObst Clustering using linKs) proving the decrease in time and space complexity.

Research regarding clustering of URLs using Sequence Alignment Method has also been done in [HWV04] in which Web users have been clustered using two different similarity measures: SAM (non-Euclidean distance-based measure) and Association measure (Euclidean distance-based measure). The sequential order of pages is taken into consideration and not the position of the pages. Such sequences are called open sequences. As mentioned in [HWV04], sequences with the same elements occurring in the same order and irrelevant of the positions of the elements are called open sequences. For example, the open sequence (1, 3, 5) occurs in the sequences (4, 1, 2, 3, 6, 5), (1, 2, 3, 4, 2, 5) and (3, 1, 3, 5, 2). Unlike most research work, where users are grouped into clusters with similar pages, it was proved that SAM retrieves sequences
not only with similar pages but the order of pages is also considered compared to the associative measure which is Euclidean-distance based. Hence, users are clustered based on their sequential order of Web navigation. Stochastic methods have been proposed for clustering user transactions for the purpose of user modelling. Since each user may reveal different types of navigation behaviour, the patterns should also capture the overlapping interests of these users.

Mixture Models are able to capture complex, dynamic user behaviour [CMS97]. To determine the user behaviour in Web usage mining systems, [MJJ09] deals with Model-based Clustering Method using EM algorithm. EM algorithm is used for finding the parameter estimates in probabilistic models. The EM algorithm has been compared with the k-means algorithm and the results showed an improvement in the accuracy of the algorithm. A variant of the Model-Based Clustering has been done in [PAV05] in which the interpretation and visualization of model-based clustering schemes using the concept of Correspondence Analysis (CO-AN) has been done. User sessions are clustered using the first-order Markov Model using the EM algorithm. CO-AN is a multi-variate statistical analysis method to interpret and visualize Web users’ navigation patterns. This is helpful for commercial Web sites to understand the customer behaviour and provides scope for site improvement.

A similar research on Model-Based Clustering has been done in [IHM+03] based on first-order Markov Model and a using a visualization tool, Web Canvas. The results have shown that learning time scales linearly with sample size using model-based clustering compared to agglomerative distance-based methods in which the learning time scales quadratically with sample size. In addition to the discovery of navigation patterns, prediction of future navigation behaviour has been included in [BL08]. Different scoring metrics such as the hit and miss score, the mean absolute error and the ignorance score have been employed to determine the quality of prediction. In [LF08a], two levels of prediction of users’ browsing behaviour have been proposed.
Using Markov Model, browsing behaviour is predicted at the category level and using Bayes Theorem, prediction is done at the Web page level. A combination of Markov model and Bayes theorem results in a two-level prediction of user’s browsing behaviour. The results proved that the hit ratio is effective and accurate in both the levels. An extension of [LF08a] has been dealt with in [LF08b] in which the overlapping or heterogeneous nature of user’s behaviour and improvement in hit ratio has been considered. Fuzzy relational clustering algorithms have been applied for Web usage mining [KJN01, Lab07]. Clustering of relational data using Fuzzy approaches has been implemented using Fuzzy c-Medoids (FCMdd) and Robust Fuzzy c-Medoids (RFCMdd) in [KJN01]. These algorithms have been applied in Web usage mining for discovering user profiles. Similar research has been performed in [Lab07] for discovery of user profiles using Ant clustering algorithm and a linear version of Fuzzy c-Medoids.

Another approach to observing path traversal and clustering has been discussed in [SAZS97] which defines a path similarity measure for a given Web site. The logged data about the user’s paths is clustered using k-means algorithm to aggregate users into groups. A topic that has been gaining momentum is the idea that the behaviour of the users can be learnt by tracking and analyzing their clickstreams.

### 4.2 Methodology

#### 4.2.1 Transaction Normalization

As discussed in Chapter 2, the sessions can be represented as a usage matrix referred to as session-pageview matrix where the $i^{th}$ row represents the pages visited by the user in the session and the $j^{th}$ column represents the sessions in which the page $j$ has been visited. The session-pageview matrix corresponds to the frequency of the pageview $j$ in session $i$. 

Published in Journal of Computer Science 6 (7): 785-793, 2010 ISSN 1549-3636 2010 Science Publications
In this work, the weight of the pageview is further determined by evaluating the importance of a page in terms of the ratio of the frequency of visits to the page with respect to the overall page visits in a session. A numerical weight is assigned to each pageview visited with the purpose of “measuring” its relative importance/interest within the session. If the page has not been visited, the weight of the page is assigned zero. The page visits repeated consecutively have been treated as a single visit to that page. The weights have been normalized to account for variances. After normalization, the cell value \((i, j)\) in the session-pageview matrix now represents the relative frequency of the pageview \(j\) in session \(i\). This matrix is referred to as session-pageweight matrix.

The above session file is represented by using the vector space model. Each session \(s_i\) is modelled as a vector over the \(n\)-dimensional space of pageviews. Each session \(s_i\) is represented as

\[
s_i = \{p_{f_1}, p_{f_2}, p_{f_3}, ..., p_{f_n}\}
\]

where each \(p_{f_j}\) is the relative frequency of pageview \(j\) in session \(i\). This type of weight normalization is referred to as transaction normalization which is beneficial since it captures the relative importance/interest of the pageview in a session.

**Example 1:** Suppose there are 17 pages and the user has visited page 1 twice, page 2 twice and none of the other pages in session 1, then, after normalization, the weights of the pageviews in the session is as follows:

\[
s_1 = 2/4, 2/4, 0, 0, 0, ..., 0
\]

### 4.2.2 Pattern Discovery-Model-Based Session Clustering

Let \(S = \{s_1, s_2, ..., s_m\}\) be a set of \(m\) objects where each object is represented by a vector of pageviews. The goal of this research is to obtain sessions with “similar”
navigational interests. In this context, clustering is usually considered in one of two ways: to cluster transactions or to cluster pageviews. In transaction-based clustering, transactions are grouped together based on the similarity of their browsing behaviour. In pageview-based clustering, pageviews are clustered based on the similarity of the pageviews visited across all transactions. Clustering of transaction records is one of the most commonly used analysis tasks in Web usage mining. Clusters obtained in this way can represent transaction segments based on their common navigational behaviour or interest in various pageviews. The primary motivation behind the use of clustering in Web usage mining is to improve the efficiency and scalability of various mining tasks.

In contrast to the other clustering methods such as partitional clustering and hierarchical clustering wherein the similarity measure is distance-based, model-based clustering employs probability-based approach which is based on finite mixture model. The EM clustering algorithm is an efficient iterative method to determine the MLE when missing or hidden data exists. This algorithm, discussed in Chapter 3, has been deployed for this research study.

Nevertheless, transaction clusters are not an efficient means of capturing the aggregate view of common usage patterns. Each transaction cluster may itself contain a large number of transactions, each consisting of several pageviews. These transaction clusters do not represent an aggregate view. The final goal in clustering transactions is to provide the ability to analyze each cluster and generate an aggregate view so as to use them for tasks such as recommendation. Hence, from each transaction cluster, an aggregate usage profile is derived. A basic approach in deriving an aggregate profile is to calculate the centroid (or the mean vector) of each cluster by finding the ratio of the sum of the pageview weights across transactions to the total number of transactions in the cluster. Thus, the centroid dimension value provides a measure of its significance in the cluster [Mob07b]. The resulting usage patterns are transformed into aggregate usage profiles that provide a comprehensive representation of the similar activities.
or interests of groups of users. This profile can be represented as a vector in the
n-dimensional space which can be used directly in the recommendation phase. The
aggregate usage profile is determined using equation 4.1

\[ Wt(p, up_c) = \sum_{s \in c} \frac{w_p^s}{nc} \]  \hspace{1cm} (4.1)

where \( w_p^s \) represents the weight of the pageview in session \( s \in c \) and \( nc \) represents the
number of sessions in cluster \( c \).

### 4.3 Experimental Design

The preprocessed session file of msnbc.com Web site have been used for this research
work.

**Example 2:**

\[
\begin{align*}
1 & \quad 1 \\
2 & \\
3 & \quad 2 & \quad 2 & \quad 4 & \quad 2 & \quad 2 & \quad 2 \\
\end{align*}
\]

The above example is a sequence of visits in which each record represents a session.
The first row indicates that the user has visited pageview 1 twice. The second row
indicates that a user has visited pageview 2 once. The third row indicates that the user
visited pageview 3 once, pageview 2 visited consecutively twice, then visited pageview
4 once and finally visited pageview 2 consecutively three times.

About 10000 sessions have been selected randomly for this experiment. A portion
of the dataset is as follows:

\[
\begin{align*}
1 & \quad 6 & \quad 6 & \quad 1 \\
1 & \quad 6 & \quad 1 \\
1 & \quad 11 & \quad 1 & \quad 11 & \quad 1 & \quad 14 & \quad 1 & \quad 12 & \quad 1 & \quad 2 & \quad 1 \\
\end{align*}
\]
Chapter 4. **Analysis of Web Pages of User Interest By Clustering**

8 1
1 7 1
4 4

After suppressing the same page visits repeated consecutively in a session using shell script in Linux, the sample dataset is as follows:

1 6 1
1 6 11
1 11 1 11 1 14 1 12 1 2 1
8 1
1 7 1
4

The frequency of the pageviews in each session is shown in table 4.1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

In addition, each page has been given a numerical weight for each session. This indicates the relative importance of each page in the session. If the page has not been visited, its weight is 0 which is represented in table 4.2.

Weka tool has been used for the experimental evaluation for clustering using EM algorithm. A model is estimated from the available samples in the dataset which are generally split into training and testing sets. The model is first designed using
training samples and then it is evaluated based on the performance on the test samples.

In this research, the dataset has been partitioned into 60% of training data and the remaining 40% as test data. The Expectation Maximization clustering algorithm has been applied. The experiment was performed within 10 iterations resulting in 9 clusters with a Maximum Likelihood Estimate of 24.95867. Each cluster represents sessions of “similar” interest in the Web pages or the usage profile which is shown in table 4.3.

### Table 4.3: Aggregate usage profile

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C0</td>
<td>0.024</td>
<td>0.003</td>
<td>0.002</td>
<td>0.002</td>
<td>0.001</td>
<td>0.005</td>
<td>0.000</td>
<td>0.946</td>
<td>0.001</td>
<td>0.006</td>
<td>0.002</td>
<td>0.000</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>C1</td>
<td>0.146</td>
<td>0.075</td>
<td>0.010</td>
<td>0.008</td>
<td>0.038</td>
<td>0.028</td>
<td>0.003</td>
<td>0.001</td>
<td>0.002</td>
<td>0.238</td>
<td>0.009</td>
<td>0.003</td>
<td>0.002</td>
<td>0.12</td>
<td>0.001</td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>0.000</td>
<td>0.362</td>
<td>0.040</td>
<td>0.028</td>
<td>0.000</td>
<td>0.107</td>
<td>0.047</td>
<td>0.004</td>
<td>0.000</td>
<td>0.011</td>
<td>0.000</td>
<td>0.321</td>
<td>0.061</td>
<td>0.017</td>
<td>0.000</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>0.004</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.000</td>
<td>0.005</td>
<td>0.006</td>
<td>0.472</td>
<td>0.004</td>
<td>0.502</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>0.094</td>
<td>0.015</td>
<td>0.736</td>
<td>0.047</td>
<td>0.003</td>
<td>0.022</td>
<td>0.007</td>
<td>0.006</td>
<td>0.009</td>
<td>0.047</td>
<td>0.005</td>
<td>0.000</td>
<td>0.007</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>0.061</td>
<td>0.008</td>
<td>0.044</td>
<td>0.065</td>
<td>0.035</td>
<td>0.021</td>
<td>0.037</td>
<td>0.035</td>
<td>0.563</td>
<td>0.006</td>
<td>0.018</td>
<td>0.054</td>
<td>0.035</td>
<td>0.009</td>
<td>0.013</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td>0.653</td>
<td>0.089</td>
<td>0.016</td>
<td>0.021</td>
<td>0.000</td>
<td>0.020</td>
<td>0.021</td>
<td>0.000</td>
<td>0.022</td>
<td>0.004</td>
<td>0.030</td>
<td>0.067</td>
<td>0.009</td>
<td>0.049</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>C7</td>
<td>0.028</td>
<td>0.012</td>
<td>0.011</td>
<td>0.000</td>
<td>0.000</td>
<td>0.884</td>
<td>0.000</td>
<td>0.037</td>
<td>0.018</td>
<td>0.002</td>
<td>0.000</td>
<td>0.007</td>
<td>0.000</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>C8</td>
<td>0.045</td>
<td>0.029</td>
<td>0.006</td>
<td>0.635</td>
<td>0.000</td>
<td>0.038</td>
<td>0.080</td>
<td>0.076</td>
<td>0.000</td>
<td>0.012</td>
<td>0.000</td>
<td>0.009</td>
<td>0.016</td>
<td>0.053</td>
<td>0.000</td>
<td>0.003</td>
<td></td>
</tr>
</tbody>
</table>
4.3.1 Evaluation Methods for Session Clustering

The **Precision** \( P(i, j) \), **Recall** \( R(i, j) \) and **Purity** \( Purity(j) \) evaluation measures of each cluster \( j \) for each Web page \( i \) are calculated.

The Precision measure of a cluster with respect to a page determines the ratio of the usage interest of each page in a cluster to the total usage interest of all pages in the corresponding cluster which is given by equation 4.2

\[
P(i, j) = \frac{w_{ij}}{\sum_{i=0}^{k} w_i}
\]  

(4.2)

where \( w_{ij} \) represents the aggregate weight (user interest) of page \( i \) in cluster \( j \). Since the weights have been normalized between \([0,1]\), the weight of the cluster \( \sum w_i \) is always equal to 1. Hence table 4.3 also represents the Precision Measure of page \( i \) in cluster \( j \).

Additionally, to evaluate the extent to which the usage interest of a page is predominant or significant among all the clusters, the Recall measure is used which is defined as the ratio of the usage interest of a page \( i \) to the overall usage interest of the page among all the clusters. The Recall Measure has been modified for this experiment since the relative usage interest is considered and is given by equation 4.3

\[
R(i, j) = \frac{w_{ij}}{\sum_{j=0}^{n} w_j}
\]  

(4.3)

This is represented in table 4.4.

Purity is a simple and transparent evaluation measure. It represents the portion of the cluster corresponding to the largest aggregate weight of the page with respect to the
Table 4.4: Recall measure for each pageview in a cluster

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C0</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.86</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C1</td>
<td>0.14</td>
<td>0.13</td>
<td>0.01</td>
<td>0.01</td>
<td>0.67</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
<td>0.75</td>
<td>0.76</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.09</td>
<td>0.12</td>
<td>0.00</td>
</tr>
<tr>
<td>C2</td>
<td>0.00</td>
<td>0.61</td>
<td>0.05</td>
<td>0.03</td>
<td>0.00</td>
<td>0.10</td>
<td>0.24</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
<td>0.83</td>
<td>0.00</td>
<td>0.00</td>
<td>0.74</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C4</td>
<td>0.09</td>
<td>0.02</td>
<td>0.85</td>
<td>0.06</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.00</td>
<td>0.01</td>
<td>0.15</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C5</td>
<td>0.06</td>
<td>0.01</td>
<td>0.05</td>
<td>0.08</td>
<td>0.28</td>
<td>0.02</td>
<td>0.19</td>
<td>0.03</td>
<td>0.92</td>
<td>0.00</td>
<td>0.06</td>
<td>0.09</td>
<td>0.09</td>
<td>0.01</td>
<td>0.09</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C6</td>
<td>0.62</td>
<td>0.15</td>
<td>0.02</td>
<td>0.03</td>
<td>0.00</td>
<td>0.02</td>
<td>0.11</td>
<td>0.00</td>
<td>0.04</td>
<td>0.01</td>
<td>0.01</td>
<td>0.10</td>
<td>0.11</td>
<td>0.02</td>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C7</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.78</td>
<td>0.00</td>
<td>0.03</td>
<td>0.03</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>C8</td>
<td>0.04</td>
<td>0.05</td>
<td>0.01</td>
<td>0.79</td>
<td>0.00</td>
<td>0.03</td>
<td>0.41</td>
<td>0.07</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
<td>0.01</td>
<td>0.04</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

The cluster which is given by equation 4.4

\[
Purity(j) = \max \frac{w_{ij}}{\sum_{i=0}^{k} w_i}
\]

(4.4)

The average purity of the clustering is the ratio of sum of the purity values of all the clusters to the total number of clusters. It is found to be 61%. The larger the purity result, the better is the performance of the clustering algorithm. This is shown in table 4.5.

### 4.3.2 Pattern Analysis

With reference to table 4.3, each cluster represents the cluster centroid. The centroid represents the mean values of the Web pages contained in each cluster. The aggregate usage profile is represented graphically in Figure 4.1 which shows the user interest on the Web pages.
Table 4.5: Purity measure of each cluster

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 0</td>
<td>0.95</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>0.24</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0.36</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>0.50</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>0.74</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>0.56</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>0.65</td>
</tr>
<tr>
<td>Cluster 7</td>
<td>0.88</td>
</tr>
<tr>
<td>Cluster 8</td>
<td>0.64</td>
</tr>
<tr>
<td><strong>Average Purity</strong></td>
<td><strong>0.61</strong></td>
</tr>
</tbody>
</table>

![Figure 4.1: User Interest on Web pages in clusters](image)

From table 4.3 and table 4.4, using the concept of logical AND, two criteria are taken into consideration. If the usage interest and recall measure are high for Web page $i$ in cluster $j$, then the page $i$ is considered as a significant page in cluster $j$. If one of the measures for the Web page is low, then the overall interest on the Web page becomes low and is not considered as a significant page. This can be observed specifically for Web page 16 in cluster 1, wherein, though the usage measure is low and recall measure is high, it is not considered as a significant page to be taken into consideration which is shown in table 4.6.

From table 4.6, it has been observed that the maximum interest for Web page 1
Table 4.6: Combined effect of Precision and Recall measure for each pageview in a cluster

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C0</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.816</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>C1</td>
<td>0.020</td>
<td>0.009</td>
<td>0.000</td>
<td>0.000</td>
<td>0.056</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.179</td>
<td>0.167</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.083</td>
<td>0.054</td>
</tr>
<tr>
<td>C2</td>
<td>0.000</td>
<td>0.221</td>
<td>0.002</td>
<td>0.001</td>
<td>0.010</td>
<td>0.011</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.267</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.002</td>
<td>0.000</td>
</tr>
<tr>
<td>C3</td>
<td>0.008</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.355</td>
<td>0.371</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>C4</td>
<td>0.004</td>
<td>0.000</td>
<td>0.002</td>
<td>0.005</td>
<td>0.010</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.001</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>C5</td>
<td>0.405</td>
<td>0.013</td>
<td>0.000</td>
<td>0.001</td>
<td>0.000</td>
<td>0.002</td>
<td>0.000</td>
<td>0.001</td>
<td>0.000</td>
<td>0.004</td>
<td>0.000</td>
<td>0.000</td>
<td>0.003</td>
<td>0.000</td>
<td>0.000</td>
<td>0.037</td>
<td>0.000</td>
</tr>
<tr>
<td>C6</td>
<td>0.003</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.692</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>C8</td>
<td>0.002</td>
<td>0.001</td>
<td>0.000</td>
<td>0.500</td>
<td>0.000</td>
<td>0.032</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.004</td>
<td>0.000</td>
</tr>
</tbody>
</table>

(FrontPage) is present in Cluster 6. Similarly, for Web page 2, it occurs in Cluster 2. For Web page 3, the maximum rating occurs in Cluster 4 and for Web page 8, it has been found that the maximum rating appears in Cluster 0. These observations enable us to describe each cluster by assigning it a name [Mal05]. It has been inferred that users/sessions in Cluster 0 are predominantly interested in obtaining information about Weather. Hence Cluster 0 can be labelled as “Weather”.

However, sessions in Cluster 1 indicate that users’ interests are diversified among multiple Web pages, showing almost similar interest in Web page 10 (Living) and Web page 11 (Business) with a lower interest in Web page 15 (Travel), Web page 5 (Opinion) and Web page 17 (msnSummary). Hence, Cluster 1 can be classified as “Living”. Cluster 2 is characterized by interest in Web page 13 (Summary) and Web page 2 (News). Hence Cluster 2 can be classified as “Summary”. Cluster 3 is focused on Web pages 14 and 12. Hence this cluster may be called “Bulletin”. Likewise Cluster 4 is focused on Web page 3 (Technology) and can be labelled as “Technology”. Cluster
Chapter 4. **Analysis of Web Pages of User Interest By Clustering**

5 can be labelled as “Health”. Similarly Cluster 6 is focused on “Frontpage”. Cluster 7 can be labelled as “On-air”. Cluster 8 is characterized by high usage interest in Web page 4 (Local) and a relatively low interest in Web page 7 (Miscellaneous) and, hence can be labelled as “Local”.

A visual representation of the users’ interest on the web pages is shown in Figure 4.2a to Figure 4.2i. From this, one can easily conclude that the user interests are either uni-focused (Cluster 0, Cluster 4, Cluster 5, Cluster 6, Cluster 7) or multi-focused (Cluster 1, Cluster 2, Cluster 3, Cluster 8).

### 4.4 Summary

Data preprocessing results in a user session file which needs to be formatted for mining tasks to be accomplished. Additionally, post-processing of the formatted user session file has been extensively done in this work in order to produce effective clustering of sessions based on user’s interest in the pages of a Web site. Initially, the frequency of page visits has been mapped to the relative user interest within a session and a weighted session-pageview matrix is obtained. In this study, EM clustering algorithm has been employed for grouping Web usage transactions and generating usage profiles. The aggregate usage profiles have been used to analyze user interest in the Web pages.

Experiments are done on real world data set such as msnbc.com to discover the user’s interest in the Web site. The significance of the results will be helpful for organizations in various applications based on their navigational interest and also provide recommendations for page(s) not yet visited by the user. This work deviates from the conventional statistical methods to derive the interpretations and employs the Aggregate Usage Profile, a modified Recall measure and the combination of both results for every pageview to analyze the user interest in various clusters.

Published in Journal of Computer Science 6 (7): 785-793, 2010 ISSN 1549-3636 2010 Science Publications
Chapter 4. Analysis of Web Pages of User Interest By Clustering

(a) Cluster 0

Cluster 0

Web Page

User Interest

(b) Cluster 1

Cluster 1

Web Page

User Interest

(c) Cluster 2

Cluster 2

Web Page

User Interest

(d) Cluster 3

Cluster 3

Web Page

User Interest

(e) Cluster 4

Cluster 4

Web Page

User Interest

(f) Cluster 5

Cluster 5

Web Page

User Interest

(g) Cluster 6

Cluster 6

Web Page

User Interest

(h) Cluster 7

Cluster 7

Web Page

User Interest

(i) Cluster 8

Cluster 8

Web Page

User Interest

Figure 4.2: Interest on Web pages in each cluster

Published in Journal of Computer Science 6 (7): 785-793, 2010 ISSN 1549-3636 2010 Science Publications