CHAPTER 3

SCOPE OF RESEARCH IN WEB MINING

There are many techniques which have been used for mining data from servers. However, as multiple techniques have been employed in a system, clear boundaries are not known. Some of the previous research where clustering techniques and semantic methods were used hand-in-hand, are discussed below.

3.1 Clustering In Web Usage Mining

One of the widely accepted techniques for data mining is clustering. There are numerous articles published which use clustering method. There exists a variety of algorithms to handle the wide range of data, but which algorithm will suit for a particular scenario depends on nature of the data, how data partitioning is done, partitioning output, etc. [63] provides a proper classification and detailed overview of the algorithms.

Clustering refers to grouping of data based on some common characteristics.

Few research articles considers sequential mining and association mining as part of clustering technique, but these methods have some confidence value and support value which doesn’t provide the bounded classes, rather they provides the result in estimations.

There are different approaches of clustering in data mining, depends on how to cluster and what to cluster. The most common of it is in session clustering. [64] have used the link between the pages in the form of graph as path for each of the sessions and apply the clustering to the paths. [65] describes each session as binary
size of magnitude N, where N represents the pages in the website and it is shown as:

\[ s = \{ p_1, p_2, p_3, p_4, p_5, \ldots, p_N \} \]

\[ P_a = \begin{cases} 
1, & \text{if accessed} \\
0, & \text{else} 
\end{cases} \]

Cosine similarly is defines by using s as a vector form. The topology of the website is represented as synthetic similarity function. Applying CARD [66] to this similarity combination, the clusters are obtained.

A vector is created for each session when the session starts and depending on the distance metric the closest cluster is chosen. Then the frequency of set of pages serves as the variable for choosing it in the cluster. [67] describes similar idea and used k-means type clustering in the place of CARD. The URL’s can be clustered depending on their frequency in the sessions of Association Rule Hypergraph Partitioning, as per [68]. Sliding window was introduced for limiting the ongoing user session so that the previous ones does not affect the clustering process.

Depending on the previous work, [69] suggested latent based method for data clustering. A session per view matrix is created by clustering URL’s and using latent based approach. Single value decomposition algorithm and cosine similarity metric is used for clustering purpose. Weights on pages is assigned depending upon visit time of that page rather than assigning binary weights depending upon if they are accessed or not.

[2] proposed that content mining can be merged with web usage mining. The pageview’s content is merged with usage profiles. During the preprocessing stages feature extraction step is used in all of the web pages and the formal definitions is given below.

\[ p = \langle f_1, w_1 \rangle, \langle f_2, w_2 \rangle, \langle f_3, w_3 \rangle, \langle f_4, w_4 \rangle, \ldots, \langle f_n, w_n \rangle, \rangle \]
Here p is Pageview, fj represents feature, wj represents the corresponding feature weight. The domain expert assigns the weights and this step demands high knowledge and manual type work.

The researchers invert the page view feature to get feature vector list where each feature vector element has associated pageview’s weight. Features are the base for clustering rather than the pageviews. This enables the group enabled features hence more detail is available on feature level rather than page level opposing [70] where the web pages are cluster rather than the features. The content and the usage mining must be collaborated to make recommendations. When both the results are compared then the most scoring pages are recommended.

In this area many unsupervised algorithms also being proposed which have employed Self-Organizing Maps (SOM) algorithm [71]. SOM can generate tight clusters in comparison to k-means.

[72] used another novel approach by using artificial ant colony clustering along with linear genetic programming (LGP). This is inspired from real world clustering of ants and their behavior. Live ants separates dead ants and the ant larvae in separate groups.” The general idea is that isolated items should be picked up and dropped at some other location where more items of that type are present”[72]. This is an ongoing process which must be stopped at some point. The figure below provides a better idea.

![Figure 8 - Ant Clusters at t=1 t=100 and t=900 [72]](image)

### 3.2 Integrating Semantics
The web personalization techniques we have seen or discussed so far for usage data, didn’t considers semantics. The data itself is not explained in the techniques and only a mathematical and statistical approach is being used till now. The click streams are the only base for the association rule mining. Database ID assignment for every item and data abstraction does not affects the clustering methods. The whole data is not used but only a fraction of it is worked upon. Despite of having these advanced techniques, few important questions like the reason for grouping users based on a particular resource are not being answered.

These methods are not able to explain the reason “Why? As they lack the semantic component in their workings and thus cannot penetrate into more complex relations and properties of concepts that reside in the web pages. And so is the aim of semantically enriched web usage mining.” After the rise of the semantic web usage mining there is not much research done for combining the web usage mining with the concepts of ontology. The majority of the research is focused on matching ontologies [49] and ontological extraction from the web pages [73].

Some formal description are presented in [89] and a potential of ontology is shown. The domain level objects information is extracted from user sessions and a user profile is created for each user by combining objects as per their weight and also the merge function.

Concepts in the ontology is represented by a class. It has a set of attributes in the form of \((a_1, a_2, a_3,...,a_n)\). Attributes can be simple literals or they can be complex objects. 2 attributes of same kind are merged in merge function and then a combination of them is represented.

We assume that a domain level ontology is already existing for the website which can be generated by both manual and automated means. The ontological information is extracted and recorded by passing the web pages through an information extraction process. On every object’s attributes, a merger function is assumed to be defined where merging implies that from multiple instances, an aggregate object is created. An example of a sample ontology is given below.
All the oncological objects can be captured which the user visits in his session for creating the user profile. We can take an example where a user visited two different movies in one single session.

Figure 9 - A sample movie ontology

Let’s say the 1\textsuperscript{st} movie is “Spy Game 2002 Action Robert Redford Brad Pitt” and the 2\textsuperscript{nd} one is “Snatch 2000 Comedy Jason Statham Benicio Del Toro” and the user spent 8 seconds viewing the 1\textsuperscript{st} movie and 5 seconds on the 2\textsuperscript{nd}. If we define year merging as timespan, year = [2000-2002]. If we define actor merging as simple actor additions, actor = \{0.8 Robert Redford, 0.8 Brad Pitt, 0.5 Jason Statham, 0.5 Benicio Del Toro\}.

We can take the user session like a domain object set which is extracted from pageview like usage profile. The usage profile is the output of merging of every items of similar kind.

Here, the sessions are clustered and as per the procedure just defined every session’s centroid is calculated hence representing a cluster centroid and user session.

During the recommendation phase we take the session of the current user
and convert that session into a user profile and after that the appropriate cluster is found and items are recommended from that cluster to the user. The in-depth details of the recommendation process is not given by the author.

In another research work [74] apriori algorithm is merged with item to mine frequent patterns. Travel ontology is being used by the author [90]. Here, two level taxonomy is being used. The first one relates the concepts and the second one relates the relations in between the concepts.

The classical properties of apriori algorithm is present in xPMiner algorithm. The level 1 candidates are generated first and then frequent items are selected for generating the level 2 sequences and so the process goes. The apriori algorithm gives the k+1 by using k number of items and then crossing them. The xPMiner used the insertion of property and object along with this.

Inside the recommendation section energetic consultation is checked towards all generated regulations to look if it is a prefix instantiation or not. It means that only prefix of the guidelines need to be matched and a complete instantiation isn't required while using the session. By means of searching all of the objects that supercede the prefix, the objects that provide higher matching to the guideline are decided on and endorsed.