Clustering Web Logs Using Similarity Upper Approximation with Different Similarity Measures

Rajhans Mishra and Pradeep Kumar

Abstract—In this paper we adopted the similarity upper approximation based clustering of web logs using various similarity/distance metrics. The paper shows the viability of our methodology. Web logs capture the information about web sites as well as the sequence of the visit. Sequence of visit provides an important insight about the behavior of the user. Rough set, a soft computing technique, deals with vagueness present in data. It captures the indiscernibility at different levels of granularity. The paper has shown the results on msnbc data set with different similarity measures along with explanation of results.

Index Terms—Clustering, sequential data, similarity upper approximation.

I. INTRODUCTION

Clustering techniques group objects that are similar to each other with respect to different attributes [1]. Clusters can be crisp or overlapping in nature. Overlapping clusters have some objects common and they are also termed as soft clusters [2]. Clusters without any common element to each other are termed as crisp or hard clusters [2]. Clustering is widely used in the field of data mining.

Web logs contain information of web page visit of users. It provides the traversing pattern of users. Traversing pattern of users provide information about their behavior. Users can be grouped on the basis of their web usage pattern.

Web user visit can also be viewed as the sequence of web page visit. Sequence mining deals with finding the useful patterns from the sequence data sets and understanding their behavior [3]. Analysis of sequential data has become increasingly important in various application areas like biological sequences [4], intrusion detection [5] and web access logs.

Web logs contain the information regarding sequence of web pages in which they are visited by the user. All the web pages that are visited by the user in sequence are of different interest for the user. Thus during the clustering of web user sequence of the visited web pages should also be considered besides the web page.

Sequential data captures the order of occurrence of objects. The order of occurrence has an important aspect of the sequential data. Classical clustering algorithms (as K-Means) are not able to perform well in case of sequential data as it is difficult to find pair wise similarity between sequences. [6].

Any user can be a part of more than one cluster as its traversing behavior fluctuates from one area to other depending upon the need of the user.

A student registering himself for a conference will go through the conference web site and will be a member of group called academician. On the same time he will be traversing web sites for booking the flights/hotels and will be put in tourism cluster as he may have to visit places to attend conference/seminar. To address the above issue rough sets can utilized to generate the soft clusters.

Sequence based clustering has been performed to find the groups on the basis of sequential information available in the data. Sungjune et al. [3] have suggested a general sequence-based clustering method by proposing new sequence representation schemes in association with Markov models.

Lingras [7],[8] has focused on rough set clustering for web mining. Unsupervised classification using rough sets along with genetic algorithms has been used to represent clusters as interval sets. Clusters have been formed by the evolution of rough set genomes. A rough set genome has been constructed to contain the objects that have to be classified. A genome has been represented as a string which indicates its membership. Upper and lower approximation of genome has been calculated to find rough clusters.

De & Krishna [9] have considered clustering of web transactions using rough sets and presented a rough approximation based clustering of web transactions from web access logs. Kumar et al. [10] have proposed a hierarchical clustering algorithm using similarity upper approximation.


The current paper deals with finding the outlier (data point which is not grouped in any cluster) and clusters from the web logs preserving the sequential information present in the web logs.

Rest of the paper is organized as follows. Section II describes Rough Sets in brief. Section III discusses the methodology adopted in this paper followed by Experimental Results and Discussion in Section IV. Conclusion has been discussed in Section V.

II. ROUGH SETS

Rough set theory has been introduced by Pawlak [16], deals with uncertainty and vagueness present in data. Rough set is a soft computing technique theory became popular among scientists around the world due to its application in the field of artificial intelligence and cognitive sciences [16]. The
**METHODOLOGY**

This paper is an extension to the paper, Kumar et al. [11] using Jaccard, Dice, Levenshtein and S’M similarity measures on msnbc data set. Kumar et al. [11] have used rough set upper approximation for generating the clusters. The upper approximation gives the group of sequences on the basis of specified value of similarity measure. In this paper we implemented the algorithm proposed by Kumar et al [11] with different similarity measures like Jaccard, Dice, Levenshtein and S’M.

The steps in the algorithm [11] are as follows:

1. The similarity value between the sequential data has been calculated using Jaccard, Dice, and Levenshtein or S’M similarity measures.

2. First upper approximation of sequences has been calculated with a given Similarity Threshold (δ). This step will result in initial clusters.

3. Generated supersets of the clusters formed in the step 2 have been identified and proper subsets are removed. This step results in reduction of number of clusters formed in step 2.

4. Cluster centers of other clusters have been removed from the cluster as clusters formed in step 3 can contain cluster centers of other clusters. Clusters formed are soft in nature till step 4.

5. This step deals with constructing hard clusters from the soft clusters found in step 4. The objects, present in more than one clusters have been identified and their similarity values have been compared with respect to each cluster center. The object is assigned to the cluster having highest similarity value between cluster center and object.

The main idea behind adapting various distance/similarity measures in step 1 is to study the impact of various similarity/distance measures while using similarity upper approximation for generating clusters. The Dice and Jaccard measure capture only the content information while Levenshtein measure capture the sequence information. S’M measure is the combination of Jaccard and Levenshtein thus captures the sequential as well as content information.

**IV. EXPERIMENTAL RESULTS & DISCUSSION**

Experiments were performed on a PC having an Intel Core 2 Duo Processor (1.83 GHz) with 2 GB RAM. The proposed algorithm [11] has been implemented in JAVA on the Windows XP platform.

We have used msnbc dataset and used sequences of length six for the purpose of the experiment. The dataset is present (http://archive.ics.uci.edu/m1/datasets/MSNBC.com+Anonymous+Web+Data) and has seventeen categories that are "frontpage", "news", "tech", "local", "opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports". We have converted the categories in to numbers starting from 1 to 17 respectively for the purpose of experiment.

Jaccard, Dice, Levenshtein and S’M similarity measures have been used with the proposed algorithm [1] for finding the clusters and outliers in the msnbc dataset. The combination of number of outliers and number of clusters has been used for validating the results of algorithm with different similarity measures. Too many and too less clusters in the output are extreme solutions and are not desirable. The number of clusters which lies in between the extreme case with less number of outliers is most suitable and appropriate output. The output with too many outliers implies a lot of objects have been left ungrouped and hence not desirable. These outliers can be generated by user while traversing a web site and performing back and forth click.

The algorithm is a hierarchical clustering algorithm which groups the objects in steps to form the clusters. The results are shown in tables Table I-Table II. The experiments have been performed for δ=0.1 on msnbc web navigation dataset. The experiments have been performed on different size of...
data sets which are of sizes 100, 200, 300, 400, 500, 1000, 2000, and 3000 with different similarity measures. In these datasets many user sequences have been repeated. These user sequences have chosen arbitrarily.

The table shows both the number of clusters as well as outliers for different size of data set with various similarity measures. For S3M, experiment were conducted using different value of “p” ranging from p=0.1 to p=0.9. The “p” value in S3M measure capture the weight of sequential information. The higher the value of “p” more sequential aspect is considered, similarly less the value of “p” less sequential aspect is considered. In the table we have quoted the result for p=0.5 (i.e. equal weight is given to sequential aspect as well as content aspect during the clustering) and p=0.8 (more weight has been given to sequential aspect than content aspect). Results suggest that S3M perform better than other similarity measures The number of outliers discovered and number of clusters formed using S3M is reasonable. The results also suggest that while clustering the web user, both aspects i.e. sequential aspect and content aspect, should be considered.

<table>
<thead>
<tr>
<th>TABLE I: NUMBER OF OUTLIERS AND CLUSTERS WITH DIFFERENT SIMILARITY MEASURES IN DATASET OF DIFFERENT SIZES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Set-100</td>
</tr>
<tr>
<td>Jaccard</td>
</tr>
<tr>
<td>Outliers</td>
</tr>
<tr>
<td>44</td>
</tr>
<tr>
<td>44</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>23</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

Note: For S3M bold values are considered as they indicate better result.

<table>
<thead>
<tr>
<th>TABLE II: NUMBER OF OUTLIERS AND CLUSTERS WITH DIFFERENT SIMILARITY MEASURES IN DATASET OF DIFFERENT SIZES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Set-500</td>
</tr>
<tr>
<td>Jaccard</td>
</tr>
<tr>
<td>Outliers</td>
</tr>
<tr>
<td>465</td>
</tr>
<tr>
<td>465</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>249</td>
</tr>
<tr>
<td>48</td>
</tr>
</tbody>
</table>

V. CONCLUSION

The paper has implemented the algorithm proposed by Kumar et al. [11] with four similarity measures that are Jaccard, Dice, Levenshtein and S3M. The algorithm [11] uses rough set upper approximation to find out the clusters of users and outliers. The results are better with S3M measure as it captures the sequence and content both. Jaccard and Dice only measure the similarity based on content and Levenshtein measures the similarity based on sequence. S3M measure the similarity based on content and sequence hence produces better result.

REFERENCES